Handvatten voor auditors en BI-ontwikkelaars

Betrouwbaarheid datawarehouses naar hoger niveau

30 januari 2018 Alfred Colenbrander

Datawarehouses lijken door de vele ontwikkelingen op datagebied zoals big data en internet of things, tegenwoordig vergane glorie en rijp voor de sloop. Toegegeven: vernieuwend zijn datawarehouses al lang niet meer. Maar is business intelligence, waar datawarehouses een belangrijke rol in spelen, dan helemaal geen spannend onderwerp meer? Hebben datawarehouses geen waarde meer voor organisaties? En een belangrijke vraag voor financial auditors en IT-auditors: hoe bekend zijn zij eigenlijk met de wereld van datawarehouses?

In dit artikel staan het concept van datawarehousing en de bijbehorende risico's centraal. Over datawarehouses is uitgebreid geschreven en over de betrouwbaarheid van bedrijfsinformatie ook. Maar over de specifieke combinatie ervan, de betrouwbaarheid van bedrijfsinformatie in datawarehouses, is vrij weinig geschreven. Daarom moet ik me in dit artikel baseren op meer algemene literatuur over datawarehouses en op mijn eigen ervaringen. De opbouw van het artikel is als volgt. Eerst beschrijf ik het concept van datawarehouses (DWH) en de bijbehorende risico's. Daarna sta ik stil bij de aandacht die het DWH in het algemeen krijgt van financial auditors en IT-auditors. Hierbij ga ik ook in op controles van het DWH door de beherende organisaties zelf, aangezien dat een opstap is naar een beter stelsel van controlemaatregelen. Vervolgens behandel ik het stelsel van controlemaatregelen bij Wageningen University & Research. Op basis hiervan adviseer ik welke controlemaatregelen in het algemeen zijn toe te voegen aan de controlemix in en rondom datawarehouses. Vervolgens schets ik de impact van enkele actuele ontwikkelingen op het terrein van DWH's. Tot slot beantwoord ik de drie eerder genoemde vragen.

Het concept datawarehousing

In de jaren tachtig van de vorige eeuw werden rapportages vaak 'op order' geproduceerd. [ABBA04, ABBA05] Daarbij werden query's rechtstreeks gedraaid op de transactionele bronsystemen. In de ideale wereld is dat ook wenselijk, want de informatie is dan altijd actueel. Maar bij het draaien van de query's op de transactionele brondatabases deden zich verschillende problemen voor. Zo werd het transactionele systeem zwaar belast, duurde het lang voordat de query's output opleverden en was er daarna nog veel tijd

nodig om de output vanuit de verschillende transactionele databases te combineren en bewerken. Als oplossing voor deze problemen is in de jaren negentig het produceren van rapportages 'op voorraad' geïntroduceerd. [ABBA04, ABBA05] De relevante gegevens in de transactionele brondatabases werden daarbij overgebracht naar een speciale database, een datawarehouse, waarin deze gegevens met elkaar werden gecombineerd, bewerkt en klaargezet voor rapportages. Deze manier van rapporteren was primair bedoeld om managementinformatie over meerdere informatiedomeinen heen te genereren en niet voor operationele informatie[1]. [ABBA05] Eén keer per dag de gegevens in het DWH actualiseren was dan ook frequent genoeg. Dan de betrouwbaarheid van de informatie: het DWH werd een middel om besluitvorming te ondersteunen en de betrouwbaarheidseisen lagen dus hoog. 100 procent betrouwbaarheid was niet haalbaar, bijvoorbeeld door data-inconsistenties tussen meerdere bronsystemen[2]. [SPOO05] De eis werd daarom dat de informatie in het DWH zo betrouwbaar moest zijn dat er weinig risico was op het nemen van verkeerde beslissingen. [VELD14]

Datawarehouses leveren tegenwoordig niet alleen managementinformatie, maar worden ook veel gebruikt voor andere toepassingen, zoals verschillende vormen van *analytics*.[3] Toch is er in principe weinig veranderd aan het concept van datawarehousing. De nieuwe generatie BI-tools, waaronder bekende namen als Tableau en Qlik, profileert zich niet als datawarehouse-product, maar past ondanks bepaalde vereenvoudigingen in feite hetzelfde concept toe. En dat concept is voor business intelligence (BI) specialisten bekend terrein. Dit geldt echter niet voor eindgebruikers en ook niet voor de meeste auditors. Gebruikers zien alleen het eindresultaat, de rapportages, zonder zich te realiseren wat er allemaal 'onder de motorkap' gebeurt. Het is natuurlijk ook niet vreemd dat gebruikers daar weinig interesse voor hebben[4]. Voor auditors zou dat anders moeten liggen omdat datawarehouses een belangrijk middel voor informatievoorziening vormen en de betrouwbaarheidseisen hoog zijn. Voor een goed begrip van de betrouwbaarheid en controleerbaarheid[5] van BI-toepassingen is enig inzicht nodig in de technische opzet van datawarehouses. Hierover gaat de volgende paragraaf.

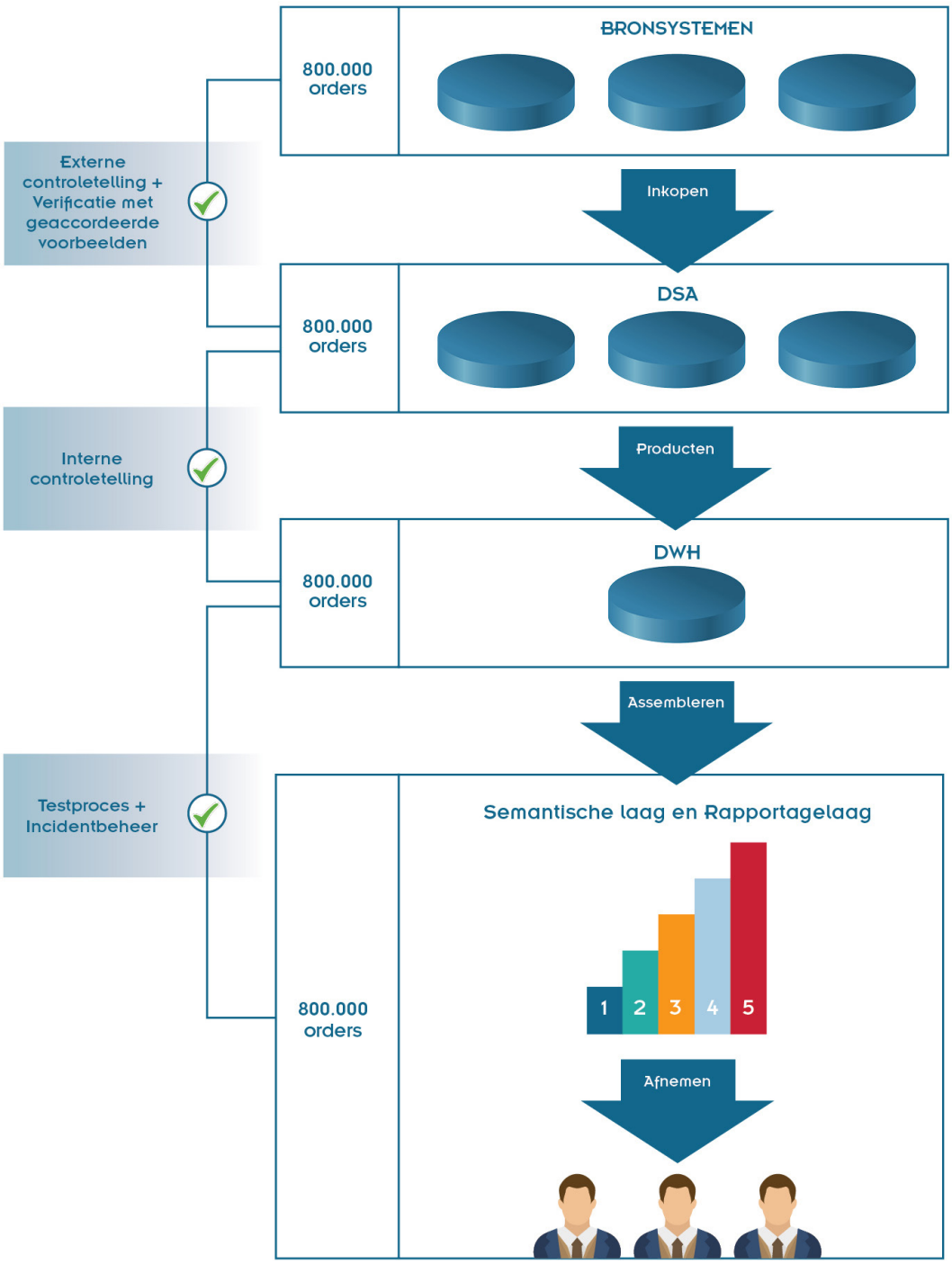
Datawarehouses: technische opzet en risico's

Belangrijke kenmerken van de technische opzet van datawarehouses zijn dat de gegevensverwerking via een keten plaatsvindt en dat er via *Extract-Transform-Load* (ETL) een bepaald datamodel wordt gevuld. Daarnaast worden specifieke DWH-mechanismen toegepast.

Datalogistiek

Om bronsystemen niet te belasten, worden de relevante brondata een-op-een overgehaald naar een *Data Staging Area* (DSA-laag). Dan volgen transformaties die de

aanbodgestuurde data, volgens de datastructuur van de bronsystemen, omzetten naar vraaggestuurde data, optimaal gestructureerd voor bevraging, in de *Datawarehouse Area* (DWH-laag). Om automatisch rapportage-query's te kunnen laten genereren, worden de in de rapportages gewenste velden beschikbaar gemaakt voor gebruikers via een semantische laag, de laag die de vertaling verzorgt van de technische veldnamen naar leesbare bedrijfstermen. Binnen deze laag worden ook elementen toegevoegd die niet in een eerdere schakel konden worden toegevoegd. Een voorbeeld is een formule die afhankelijk is van de selectie van een gebruiker. Ten slotte is er nog de rapportagelaag, die de beschikbare velden via geautomatiseerde rapport-query's met elkaar combineert tot een rapport. Een belangrijke doelstelling is dat de rapportages eenvoudig in elkaar zitten. Zoveel mogelijk logica zou moeten zitten in de transformaties van DSA naar DWH, zodat in de rapportages zelf weinig logica hoeft te worden ingesteld en de rapportages allemaal gebaseerd zijn op dezelfde logica. Vanuit deze benadering bestaat de BI-keten als productieketen uit vier schakels, naar analogie van een productiebedrijf aan te duiden als inkopen, produceren, assembleren en afnemen (zie figuur 1)[6].

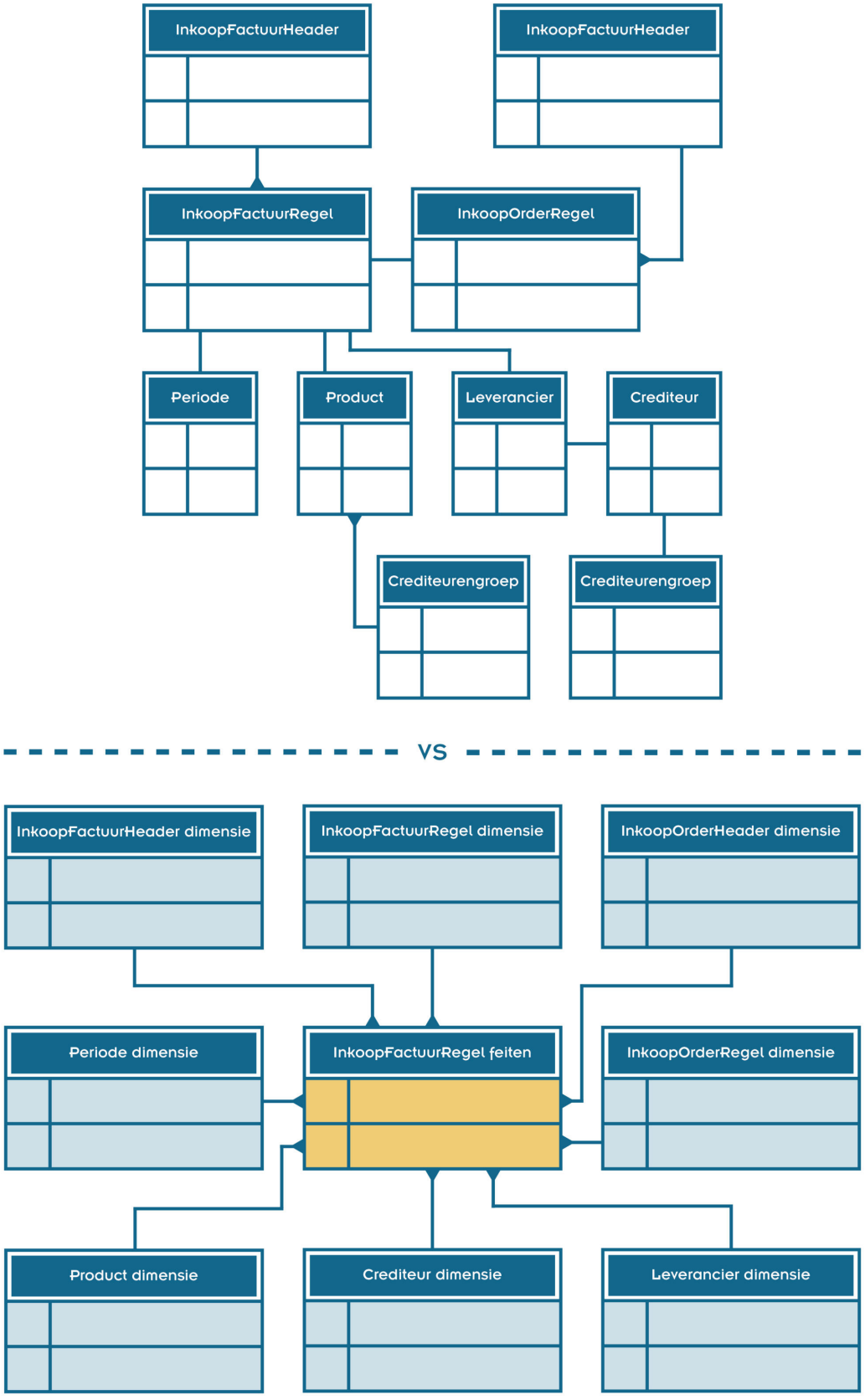


Figuur 1: De BI-keten en haar schakels

Dimensioneel datamodel

Bij het transformeren van aanbodgestuurde data naar vraaggestuurde data worden de data meestal 'dimensioneel' gemodelleerd voor hoge performance van rapportage-query's en voor gemak bij het samenstellen van rapporten. [KIMB02; KIMB08] Dit houdt in dat de genormaliseerde tabellenstructuur van de transactionele systemen, die gericht is op efficiënte opslag, wordt 'gedenormaliseerd' tot een 'dimensioneel model' dat ook wel 'stermodel' wordt genoemd[7]. Elke ster bevat één zogenaamde feitentabel, dat wil zeggen een centrale tabel met 'meetwaarden'. Deze meetwaarden zijn getallen, zoals

inkoopfactuurbedragen. Om de feitentabel heen zitten zogenaamde dimensietabellen. Met behulp van de dimensies worden de meetwaarden weergegeven in de context van de bedrijfsvoering. Voorbeelden van dimensies zijn leverancier, product en periode (zie figuur 2).



Figuur 2: Voorbeeld van een genormaliseerd datamodel (boven) versus een dimensioneel datamodel (onder)

Relaties binnen het stermodel

In tegenstelling tot het genormaliseerde datamodel gaan de relaties binnen een ster slechts één niveau diep. [KIMB02] De dimensietabellen hebben allemaal één lijntje met de centrale feitentabel en verder zijn er geen lijntjes. Immers, hoe meer tabellen moeten worden gecombineerd, hoe langer het duurt om deze als één dataset in een rapport te krijgen. De relaties binnen de ster gaan via *inner joins*, aangezien *outer joins* normaliter *performance killers* blijken[8]. Het gebruik van de inner join betekent wel dat regels kunnen wegvallen, bijvoorbeeld als een inkooporder wel voorkomt in de centrale feitentabel maar er geen corresponderende waarde is in de project dimensie. Om dat onterecht wegvallen te voorkomen, wordt in de regel een algemene 'dummy'-waarde toegevoegd in de dimensies. De inkooporder en het orderbedrag worden dan wel weergegeven in de output, maar met een 'dummy'-project.

Surrogaatsleutels

De relaties tussen de centrale feitentabel en de dimensietabellen worden gelegd door surrogaatsleutels. [KIMB02] Dit zijn betekenisloze getallen die worden gebruikt om eventuele problemen met natuurlijke sleutels, zoals projectnummers, te voorkomen[9]. Surrogaatsleutels worden ook gebruikt om de noodzaak van samengestelde sleutels te vermijden. Bij samengestelde sleutels moeten relaties worden gelegd op een combinatie van meerdere velden en dit is ten eerste lastig in gebruik en ten tweede een potentiële performance killer. Het gebruik van surrogaatsleutels brengt wel een grotere kans op fouten met zich mee. Als bijvoorbeeld een dimensietabel wél is geactualiseerd maar de feitentabel niet, dan zijn de surrogaatsleutelwaarden niet meer consistent. Concreet betekent dit dat bijvoorbeeld een inkooporder in de output is gekoppeld aan een onjuist project of dat salariskosten in de output is gekoppeld aan een onjuiste medewerker. Het spreekt voor zich dat dit niet mag gebeuren.

Opbouw van historie

Voor bepaalde BI-toepassingen is het gewenst om historie op te bouwen. Dit vindt dan plaats binnen het DWH. Met opbouw van historie is het mogelijk om informatie zoals die op een bepaald moment was, precies te reproduceren en te tonen. [KIMB02]; [INMO96] Dit kan worden gezien als 'tijdreizen'. Door de bril van vandaag is een bepaalde inkooporder bijvoorbeeld gesloten, maar door de bril van drie maanden geleden is die betreffende inkooporder nog open.

Incrementeel laden

Om te voorkomen dat historische gegevens onnodig opnieuw worden ingelezen, wordt er in een DWH regelmatig gewerkt met mechanismen om data incrementeel te laden. [LANS12] Er wordt dan eerst gecontroleerd of een aangeboden record al in het DWH aanwezig is. In dat geval wordt het record niet opnieuw ingeladen. Dit mechanisme luistert echter heel nauw. Soms wordt de controle gebaseerd op een *timestamp* dat al aanwezig is in het aangeboden record. Meestal wordt de controle gebaseerd op de

waarden van alle of enkele velden in het record. Maar mocht de basis voor deze controle niet geheel juist zijn, dan kan dit leiden tot onjuiste records of zelfs dubbele records. Als de aangeleverde timestamps niet altijd kloppen of er niet op alle (relevante) velden wordt gecontroleerd, dan kunnen de hiervoor genoemde problemen bij het laden van records optreden. Het instellen van incrementeel laden vergt dus veel aandacht.

Uit bovenstaande beschrijving van de technische opzet van datawarehouses blijkt dat er binnen de BI-keten veel bewerkingen plaatsvinden. Hierdoor is er in principe een reëel risico van verlies van volledigheid en juistheid van informatie. [LANS12; LINT15] Later in dit artikel (paragraaf 'Controle betrouwbaarheid DWH bij WUR') behandel ik de uitkomst van een risicoanalyse om vervolgens tot een set controlemaatregelen te komen.

Beperkte aandacht auditors voor het DWH

Mijn eigen ervaring leert dat auditors datawarehouses maar zeer beperkt controleren. Dit was al zo in de tijd dat ik IT-audits uitvoerde voor accountants en ik ervaar het nog steeds als BI-coördinator bij Wageningen University & Research (WUR). Regelmatig heb ik met externe auditors te maken, maar het datawarehouse voor onze centrale informatievoorziening hebben ze nog nauwelijks onder de loep genomen. Dat is toch opmerkelijk. Voor de uitvoering van controlewerkzaamheden door auditors zijn er 'Nadere voorschriften controle- en overige standaarden' (NV COS) en deze zijn meestal gericht op financiële informatie of financiële overzichten. [NBA16] Een DWH levert onder andere die financiële informatie. Toch is het DWH zelden object van onderzoek[10] of wordt er om het DWH heen gecontroleerd. Was de IT-auditor niet in het leven geroepen, juist om een black box open te kunnen maken voor accountants? De informatie in een DWH is wel relevant, want de onderliggende bronsystemen zijn weer wel object van onderzoek voor auditors. De reden is dat die bronsystemen direct of indirect de financiële gegevens bevatten die de basis zijn voor de jaarrekening. Kort door de bocht gezegd is het DWH voor onder andere financiële informatie een relevant verlengstuk van de bronsystemen, terwijl toch meestal alleen die onderliggende bronsystemen object van onderzoek zijn.

DWH wordt slecht begrepen

Natuurlijk is het niet zo dat een DWH nooit object van onderzoek is. Het is dan ook zinvol om aan te geven hoe auditors een DWH controleren als zij het DWH wél in het onderzoek betrekken. Hierbij baseer ik me niet op onderzoek, maar mijn ervaringen en indrukken zijn als volgt:

- Een auditor vraagt me tijdens een interview hoe de verwerking van bronsysteem naar rapportage plaatsvindt. Ik beschrijf dan stap voor stap en met schema's hoe de BI-keten in elkaar zit. Daarbij beantwoord ik wat vragen en vervolgens krijg ik het verzoek van de

auditor om op basis van het DWH een overzicht op te leveren, bijvoorbeeld een verrijkt grootboekoverzicht of een besteloverzicht, inclusief de aansluiting op totalen met een lijst uit het bronsysteem. De opzet is dan getoetst en de output ook. Maar de betrouwbaarheid van de gegevensverwerking binnen de BI-keten is op deze manier niet getoetst! Het toetsen van de werking is achterwege gebleven. Wellicht doordat de BI-keten te complex lijkt of doordat de auditor zich niet realiseert wat er mis kan gaan binnen de BI-keten.

- Wat ik ook heb ervaren is dat auditors graag willen steunen op de output van bepaalde rapporten en dan de rapportage-query's opvragen. Die zijn vrij eenvoudig op te leveren, maar de vraag is of een auditor die query goed kan beoordelen. En belangrijker nog: die rapportage-query's betreffen slechts een deel van de eerder beschreven BI-keten. Die rapportage-query's betreffen de rapportagelaag en die laag is niet het meest riskante deel van de keten doordat in deze laag maar een beperkte hoeveelheid bewerkingen plaatsvindt. Zoals eerder beschreven worden de meeste bewerkingen toegepast in de DWH-laag en hier liggen dus de grootste risico's.

Wel of niet steunen op het DWH?

Bij de vraag of auditors in de praktijk wel of niet steunen op een DWH komen er volgens mij drie situaties voor:

1. Een auditor doet geen poging om te steunen op een DWH. Dit zal vaak voorkomen, en het maakt de controleopdracht lastig als de output van een bepaald DWH-rapport relevant is voor die controleopdracht. Er moet dan om het DWH heen worden gecontroleerd.
2. Een auditor besluit te steunen op een DWH, maar doet dit zonder grondig onderzoek. Dit betekent een detectierisico, het risico dat een auditor een fout die door het interne controlestelsel van de organisatie is 'geglipt' niet detecteert. En een verhoogd detectierisico betekent een verhoogd risico dat de auditor tot een fout oordeel komt. [VEEN14]
3. Een auditor die wel goed thuis is in datawarehouses weegt na grondig onderzoek af of er wel of niet kan worden gesteund op het DWH van de klant. Er zijn echter niet veel auditors met voldoende kennis van datawarehouses en als die kennis er wel is, zal de conclusie veelal zijn dat er niet kan worden gesteund op het DWH. Standaard bieden BI-oplossingen namelijk weinig toereikende controles op de volledigheid en juistheid van de gegevensverwerking. [LANS12; LINT15]

Terwijl organisaties die gebruikmaken van een DWH zelf wel steunen op haar DWH-rapportages, laten auditors de DWH dus meestal links liggen of besteden ze er onvoldoende aandacht aan. En als auditors het DWH wel in het vizier hebben en erop zouden willen steunen, dan zullen zij ervaren dat er zonder aanvullende maatregelen te weinig waarborgen zijn om op het DWH te kunnen steunen. Dit komt doordat DWH-programmatuur standaard niet is uitgerust met toereikende controlemaatregelen.

Controle DHW door de organisatie zelf

Over de controles die een IT-afdeling uitvoert kan ik, hoewel ik hierbij moet generaliseren, vrij kort zijn. Het laden van het DWH is een batchproces dat normaliter tijdens de nacht draait. 's Ochtends controleren de IT-medewerkers of het *batchproces* zonder foutmeldingen is voltooid[11]. Zo niet, dan grijpen ze in en laden ze het DWH eventueel opnieuw. Zijn er geen foutmeldingen, dan is de veronderstelling dat de vulling

van het DWH volledig en juist is. Op zich is dat een logische gedachte, maar de controle op foutmeldingen dekt niet alle risico's op onbetrouwbare gegevensverwerking af. Een batchproces kan technisch best zonder foutmeldingen draaien, terwijl er toch een inconsistente dataset wordt gegenereerd.

Onbetrouwbare informatie komt vaak aan het licht vanuit de gebruikers. Zeker de gebruikers die goed thuis zijn in een of meer bronsystemen zien vrij snel afwijkingen tussen DWH-rapportages en de informatie in een bronsysteem. Bijvoorbeeld als er uren vanuit de tijdregistratie ontbreken of als er te veel inkooporders worden gerapporteerd onder de eerdergenoemde 'dummy'-waarde. Op zich is het mooi als fouten tijdig aan het licht komen vanuit de gebruikers. De IT-afdeling kan hierdoor snel ingrijpen. Maar de onjuiste informatie kan zelfs dan al hebben geleid tot verkeerde beslissingen. Als een projectleider bijvoorbeeld niet doorhad dat er uren ontbraken op een project, dan zou die tijdens een projectoverleg kunnen zijn uitgegaan van een te hoog resterend projectbudget. Daarnaast dragen fouten zeker niet bij aan het vertrouwen van gebruikers in het DWH. Een toereikende en proactieve controle is onmisbaar.

Controle betrouwbaarheid DWH bij WUR

Het centrale datawarehouse van Wageningen University & Research wordt elke nacht bijgewerkt met gegevens vanuit de applicaties voor Projecten, Onderwijs, HR, Inkoop en Financiën. Uiteraard vinden er controles op foutmeldingen plaats, maar die vallen buiten het bestek van dit artikel. Hier richt ik me op de controles op de processen in de verschillende lagen van het DWH.

De controles die op het DWH van WUR worden uitgevoerd, zijn voor een groot deel gebaseerd op risico's die inherent zijn aan de eerder beschreven technische opzet van datawarehouses. Uiteraard bestaat op hoog niveau het risico dat de gegevensverwerking onvolledig of onjuist plaatsvindt. Maar om risicobeperkende maatregelen te implementeren moeten we inzoomen op de BI-keten om te kunnen vaststellen op welke punten er iets fout kan gaan. Er is geen kant-en-klaar overzicht met DWH-risico's en maatregelen, en daarom hebben we die risico's zelf geïnventariseerd voor onze organisatie. Voor een groot deel konden we risico's vooraf al voorzien, maar sommige risico's bleken pas uit de praktijk, tijdens het testen en helaas ook na de ingebruikname van het DWH. Het DWH is live gegaan in het voorjaar van 2015. Tijdens de eerste maanden waren er regelmatig nachten waarin zich problemen voordeden tijdens het laden of transformeren. Gebruikers attendeerden ons er dan de volgende ochtend op dat de informatie in het DWH onbetrouwbaar was[12]. De focus lag de eerste maanden daarom vooral op *bugfixing*. In 2016 hebben we vervolgens het DWH stabieler gemaakt door risicobeperkende maatregelen in te bouwen.

Toen we eenmaal in beeld hadden waar in de DWH-keten de risico's lagen, konden we door analyse van die risico's een set maatregelen ontwikkelen. In tabel 1 staan onze inschattingen van de kans van optreden van risico's en hun impact bij optreden. De risico's vloeien vooral voort uit de eerder beschreven technische opzet van datawarehouses.

Risico	Object	Betreft Volledigheid of Juistheid ?	Risico-inschatting in productionele fase	Toelichting	Mogelijke aanvullende maatregel(en)
Onbetrouwbare gegevens-verwerking	DWH-laag: Surrogaat-sleutels	Juistheid	Hoog	Onjuiste combinaties van tabelwaarden kunnen optreden als niet alle tabellen succesvol zijn bijgewerkt. Deze situatie kan op een aantal dagen in het jaar voorkomen.	- Controle van tabellen op de bijwerkdatum. - Verificatie van output met geaccordeerde voorbeelden. - Tabellen alleen vrijgeven als ze allemaal zijn bijgewerkt. - Vervangen surrogaatsleutels door natuurlijke sleutels.
	DWH-laag: Consistentie van dummy's	Volledigheid	Hoog	Het wegvallen van feiten met dummy-verwijzingen kan optreden als dummy-waarden niet succesvol zijn verwerkt in de dimensietabellen. Deze situatie kan op een aantal dagen in het jaar voorkomen.	- Het inbouwen van een extra controle op het bestaan van dummy's voorafgaande aan het vrijgeven van bijgewerkte tabellen. - Bij kritieke dimensies kunnen Inner Joins tussen feitentabel en dimensietabel worden vervangen door Outer Joins.
	DWH-laag: Incrementeel laden	Volledigheid en Juistheid	Hoog	Records kunnen dubbel worden geladen of records kunnen ten onrechte niet worden bijgewerkt als foutgevoeligheden voor incrementeel laden niet voldoende zijn getest.	- Nagaan van de foutgevoeligheid van timestamps en velden waarop het mechanisme voor incrementeel laden controleert. In geval van foutgevoeligheid niet meer werken met dit mechanisme. - Nagaan of het mogelijk is om bij het actualiseren (delen van) de data integraal in plaats van incrementeel te laden.
	DWH-laag: Complexe transformaties	Juistheid	Midden	Een deel van de output kan onjuist zijn als de toegepaste transformaties complexiteit bevatten en niet alle relevante scenario's getest zijn.	- Interne en externe controletellingen op feitentabellen en kritieke dimensietabellen. - Verificatie van output met geaccordeerde voorbeelden. - Dagelijkse controle op logische groei van feitentotalen.
	DWH-laag: Onzuivere filters	Volledigheid	Midden	Het wegvallen van records kan optreden als er in de transformaties filtering wordt toegepast die niet volledig zuiver is.	- Interne en externe controletellingen op feitentabellen en kritieke dimensietabellen. - Verificatie van output met geaccordeerde voorbeelden. - Dagelijkse controle op logische groei van feitentotalen."
	DSA-laag	Volledigheid en Juistheid	Laag	Normaliter heeft deze laag alleen betrekking heeft op een kopieerslag zonder bewerkingen.	Geen, want testen en incidentbe-heer zouden het risico al voldoende moeten afdekken.
	Semantische laag	Volledigheid en Juistheid	Laag	De kans dat gegevens onbetrouwbaar raken in deze laag is initieel hoog, maar netto laag doordat fouten in tabelrelaties, formules en filters zo structureel van aard zijn dat ze nooit door de testen zouden kunnen zijn gekomen.	Geen, want testen en incidentbeheer zouden het risico al voldoende moeten afdekken.
	Rap-portagelaag	Volledigheid en Juistheid	Laag	De kans dat gegevens onbetrouwbaar raken in deze laag is initieel hoog, maar netto laag doordat fouten in rapportformules en -filters zo structureel van aard zijn dat ze nooit door de testen zouden kunnen zijn gekomen.	Geen, want testen en incidentbeheer zouden het risico al voldoende moeten afdekken.

Tabel 1: Risk-Control matrix voor datawarehouse WUR

Zoals in tabel 1 is weergegeven, zijn er diverse controlemaatregelen mogelijk om het risico op onbetrouwbare gegevensverwerking te beperken. Voor het DWH van WUR hebben we gekozen voor:

1. het DWH niet incrementeel, maar vooralsnog elke dag integraal laden. Zolang integraal laden haalbaar is, worden hiermee fouten die inherent zijn aan het incrementeel laden, vermeden;
2. tabellen waarop de rapportages worden gebaseerd alleen vrijgeven als ze allemaal zijn bijgewerkt. Tijdens het vullen van het DWH worden daarom eerst concept-tabellen geladen. De rapportagetabellen bevatten tijdens het laden nog de gecontroleerde vulling van een dag eerder. Pas als de diverse controles op de concept-tabellen allemaal een positief resultaat geven, worden de rapportagetabellen gevuld met de inhoud van de concept-tabellen. Indien de controles een fout opleveren, dan behouden de rapportagetabellen de gecontroleerde vulling van een dag eerder. Dit mechanisme noemen we het 'automatische fallback mechanisme'. De controles die op de hiervoor genoemde concept-tabellen worden uitgevoerd, zijn:
   - verificatie van output met geaccordeerde voorbeelden. Er wordt dan bijvoorbeeld bij een afgesloten project gecontroleerd of de begrotings- en realisatiecijfers en medewerkers, die op het project hebben gewerkt, overeenkomen met een geaccordeerd voorbeeld;
   - controle of er een algemene 'dummy'-waarde, zoals beschreven in de paragraaf over de relaties binnen het stermodel, is toegevoegd aan de dimensietabellen. Dit voorkomt dat feiten, zoals inkooporders waarbij geen project is ingevuld, wegvallen tijdens het genereren van rapportages;
   - interne controletellingen op feitentabellen en kritieke dimensietabellen. Hiermee wordt bedoeld dat totaalaantallen en totaalbedragen van de DWH-tabellen worden vergeleken met die in het DSA;
   - externe controletellingen op feitentabellen en kritieke dimensietabellen. Hiermee wordt bedoeld dat totaalaantallen en totaalbedragen van de DWH-tabellen worden vergeleken met de totalen van de bronsystemen. Een eenvoudig voorbeeld van deze controletellingen, de controle of in elke schakel van de BI-keten het totaal van 800.000 orders aanwezig is, is opgenomen in figuur 1.

3. dagelijks een logbestand automatisch laten vullen met de resultaten van het laadproces. Deze logging geeft niet alleen per dag aan of het laden succesvol was, maar ook in welke stap eventuele fouten zijn opgetreden en wat de resultaten waren van de automatische vergelijking van controletellingen. Deze maatregel stond niet in de oorspronkelijke risk control matrix, maar geeft wel zeer nuttige mogelijkheden om fouten gedurende een periode te analyseren en verantwoording af te leggen over de betrouwbaarheid van het DWH gedurende een boekjaar.

Bovenstaande beschrijving biedt een goede basis om controlemaatregelen toe te voegen die standaard in datawarehouses ontbreken. Maar ook de beschreven controlemaatregelen bieden natuurlijk geen 100 procent garantie voor betrouwbare informatie. Binnen WUR hanteren wij dan ook nog de volgende uitgangspunten om een betrouwbare informatievoorziening zo goed mogelijk te waarborgen:

- liever informatie die een aantal dagen niet is bijgewerkt maar wel betrouwbaar is dan onbetrouwbare informatie. Of nog extremer: Liever een dag geen informatie dan onbetrouwbare informatie.

- indien het fallback-mechanisme in werking is gegaan en informatie dus minder actueel is, dan dienen gebruikers hierover automatisch te worden geïnformeerd met behulp van een 'verkeerslicht' in het DWH.
- indien er ondanks de controles en het fallback-mechanisme toch nog fouten door het vangnet zijn gekomen, dan krijgen gebruikers dit zo spoedig mogelijk te horen en zorgt IT zo snel mogelijk voor correctie.
- problemen met bronsystemen, zoals slechte datakwaliteit, hebben wij niet meegenomen in de risicoanalyse. Althans, wij hebben er geen specifieke controlemaatregelen voor gedefinieerd binnen het DWH-object13. Onze redenering is dat de juiste plek om problemen in een bronsysteem te corrigeren het bronsysteem zelf is en niet het DWH. [LANS12, P198] Wel moet de gebruikersorganisatie zich bewust zijn van dit uitgangspunt. Als hulpmiddel hebben wij voor de gebruikersorganisatie enkele controlerapporten beschikbaar gemaakt die inzicht geven in de datakwaliteit van de bronsystemen.

De beschreven controlemaatregelen en uitgangspunten zorgen ervoor dat het DWH voor WUR een goede basis is voor betrouwbare informatievoorziening. Uiteraard is de implementatie van automatische controles geen reden om achterover te leunen. De werking dient per slot van rekening dagelijks te worden bewaakt door de organisatie zelf. En de auditor zou jaarlijks die werking moeten toetsen.

De risico's bij BI- en DWH-oplossingen en de mogelijke maatregelen en handvatten om die risico's te beperken zijn in het voorafgaande beschreven. Vervolgens is het nuttig om kort stil te staan bij de impact van enkele ontwikkelingen op het vlak van datawarehousing.

Een blik op DWH-ontwikkelingen

Zoals eerder aangegeven is er, ondanks de verschillende toepassingen, aan het concept van datawarehousing niet zoveel veranderd. Maar natuurlijk zijn er wel ontwikkelingen. Zo zien we sinds enkele jaren steeds meer ETL-tools om een 'Virtueel Datawarehouse' (of 'Logisch Datawarehouse') te creëren. De naam zegt het eigenlijk al: in plaats van een fysiek DWH is er dan sprake van een virtueel DWH. Rapportagetools maken daar verbinding mee en merken geen verschil met een fysiek datawarehouse. Voordelen van het virtuele datawarehouse zijn [LANS12, p.48, 88-89, 137, 149]:

- Gegevens hoeven niet te worden gekopieerd en op meerdere plekken te worden opgeslagen. In plaats daarvan worden gegevens real-time opgevraagd uit het bronsysteem en near real-time aangeboden aan rapportages.
- Het niet hoeven creëren van fysieke tabellen in een DSA en DWH betekent dat ontbrekende gegevens sneller zijn toe te voegen aan het datawarehouse en de rapportages. Wijzigingsverzoeken van gebruikers zijn dan te realiseren binnen een veel kortere *time-to-market*; dit is de marktintroductietijd ofwel de tijd tussen het idee en het beschikbaar zijn van het product.

- De software vertaalt de door de ontwikkelaar samengestelde ETL naar één efficiënte *pushdown query* per bronsysteem. Dit scheelt in de hoeveelheid ontwikkelwerk en in de omvang van op te halen data.
- De ETL-software bevat veelal standaardconnectoren met 'big data'-platformen. Het is daardoor mogelijk om diverse bronsystemen, waaronder een eventueel bestaand fysiek DWH, te combineren met 'big data'-bronnen in één virtueel datawarehouse.

Kortom: de informatie is dan actueler en het DWH kan vanwege de kortere marktintroductietijd bij gewenste uitbreidingen en mogelijkheden voor 'big data' verschuiven van een *system of record* in de richting van een *system of innovation*. [GART12]

Impact van het virtuele DWH en big data voor gebruikers en auditors

Nieuwe technieken brengen in de regel ook andere risico's met zich mee. Bij het virtuele DWH kan ik daar alleen theoretisch iets over zeggen, maar het concept zit zo in elkaar dat de risico's op onbetrouwbare informatie zouden moeten afnemen. De kopieer- en bewerkingsketen wordt korter, waardoor er minder plekken in de keten zijn waar informatie onvolledig of onjuist kan worden. Daarnaast maakt de kortere keten de controleerbaarheid van de gegevensverwerking eenvoudiger en overzichtelijker: per bronsysteem is er in principe nog maar één grote pushdown query in plaats van een heel stelsel van query's. Voor auditors zie ik de omvang van de benodigde controlewerkzaamheden dan ook afnemen in vergelijking met de benodigde controlewerkzaamheden in het geval van een fysiek DWH.

Bij het integreren van big data in het (virtuele) DWH zie ik echter wel een groter betrouwbaarheidsrisico, aangezien er bij big data minder houvast is voor bijvoorbeeld het vergelijken van controletellingen tussen DWH en bronsystemen. Vanuit een business-perspectief staat tegenover dat risico wel een groot voordeel bij het integreren van big data in een DWH. Opname in het DWH betekent namelijk dat de big data goed worden geprepareerd voordat ze worden gebruikt in rapportages en analyses. De laatste jaren is weleens geroepen dat big data een einde zouden maken aan de wereld van datawarehousing, maar hand-in-hand bieden ze juist veel mogelijkheden. [KOBI15] Geprepareerde data maakt data-analyse beter mogelijk en combinatie van die geprepareerde big data met reeds in het DWH aanwezige data biedt dwarsverbanden tussen big data en 'traditionele' data.

Tot slot

Na de behandeling van het concept van datawarehousing, de risico's op onbetrouwbare informatie, de mogelijke beheersmaatregelen en een blik op ontwikkelingen, is het tijd

om terug te komen op de drie vragen uit het begin van dit artikel: is BI geen spannend onderwerp meer? Hebben datawarehouses geen waarde meer voor organisaties? Hoe bekend zijn financial auditors en IT-auditors eigenlijk met datawarehouses?

Voor BI-specialisten is het wellicht niet meer zo spannend om een traditioneel datawarehouse op te tuigen. Toch kan het waarborgen van de blijvende betrouwbaarheid ook voor hen een uitdaging zijn. Toereikende betrouwbaarheidscontroles zijn namelijk niet aanwezig in de standaardoplossingen. Daarnaast zijn de klantvragen steeds uitdagender: zij wensen real-time inzicht en de mogelijkheid om snel databronnen, waaronder 'big data', toe te kunnen voegen aan hun informatieplatform. Er zijn dus nog genoeg uitdagingen op BI-gebied.

Binnen de BI is de recente ontwikkeling *analytics* sterk in opkomst. Door de eigen bedrijfsgegevens en eventuele (externe) big data goed te combineren en te analyseren, ontstaan inzichten om de bedrijfsprocessen te optimaliseren en meerwaarde te bieden aan klanten. Hierdoor heeft de BI een nieuwe boost gekregen en is duidelijk geworden dat de onderliggende datawarehouses juist veel waarde bieden – wellicht meer dan ooit.

Voor de gemiddelde auditor zijn datawarehouses zeker geen vertrouwd terrein. Voor een deugdelijk onderzoek naar BI-omgevingen en de output daarvan zullen zij zich moeten verdiepen in de architectuur en de kenmerken van de BI-keten. Zoals in dit artikel is aangegeven, zijn er diverse plekken in de keten waar informatie onvolledig en/onjuist kan worden.

Afrondend, waar moeten gebruikers en auditors op letten? Besef vooral dat naarmate er meer bronsystemen en kopieer-/bewerkingsslagen zijn, de kans op data-inconsistenties en dus de kans op onbetrouwbare informatie toeneemt. Er zijn dan aanvullende organisatorische en technische beheersmaatregelen nodig. In dit artikel is een set maatregelen beschreven. Wat er in de regel ontbreekt is een mechanisme dat geladen gegevens pas vrijgeeft als alle benodigde controles, zoals een vergelijking van totaalbedragen in het DWH met die van de bronsystemen, succesvol zijn en de controleuitkomsten zijn vastgelegd in logging. Uiteraard zijn er ook organisatorische afspraken nodig over het gebruik van informatie: hoe kunnen gebruikers nagaan of bepaalde informatie op een bepaalde dag kan worden gebruikt? Wanneer en waarvoor mag bepaalde informatie niet worden gebruikt? Hoe moet de gebruiker handelen als bepaalde beschikbare informatie toch niet betrouwbaar blijkt te zijn?

Ik hoop dat dit artikel handvatten levert om de betrouwbaarheid van informatie in datawarehouses op een hoger niveau te krijgen.

Noten

1 Voor de meer operationele informatie, zoals een openstaande debiteurenlijst, was de gebruiker aangewezen op het transactionele systeem. [ABBA04]

2 Naast technische uitdagingen zijn er uiteraard ook organisatorische uitdagingen, zoals het hanteren van uniforme definities binnen verschillende afdelingen en hun systemen. [DIJK01]

3 Hiertoe zijn onder andere 'Process Mining', 'Descriptive Analytics' en 'Predictive Analytics' te rekenen.

4 Overigens zou het in het kader van verwachtingenmanagement wel goed zijn als gebruikers zich wat meer zouden realiseren dat er eerst veel onder de motorkap moet plaatsvinden voordat managementinformatie kan worden opgeleverd.

5 Uiteraard zijn er meer kwaliteitsaspecten bij datawarehouses, zoals 'exclusiviteit' en 'onderhoudbaarheid'. In dit artikel richt ik me echter op de betrouwbaarheid van de gegevensverwerking en de controleerbaarheid daarvan.

6 Soms worden delen van het datawarehouse nog gekopieerd en eventueel geaggregeerd naar 'Data Marts' voor specifieke gebruikersgroepen. [ABBA04] In dit artikel wordt deze schakel echter buiten beschouwing gelaten.

7 Er zijn verschillende methoden om tot een vraaggestuurd datamodel te komen. Bekende methoden zijn de Kimball benadering, de Inmon benadering en 'Data Vault'. De laatste twee maken gebruik van een extra laag waarin gegevens van meerdere bronnen eerst in één genormaliseerd model worden gebracht voordat deze worden getransformeerd naar een dimensioneel model. Dit artikel gaat echter uit van de Kimball benadering: transformeren naar een dimensioneel model zonder extra tussenlaag.

8 Bij 'Inner Joins' worden alleen de overeenkomstige rijen van twee tabellen weergegeven. Bij 'Outer Joins' wordt een van de twee tabellen aangemerkt als basistabel en wordt altijd de rij van de basistabel weergegeven, ook als er geen overeenkomstige rij is in de tweede tabel.

9 Natuurlijke sleutels, zoals een projectnummer, worden wel eens aangepast binnen een bronapplicatie.

10 Het datawarehouse is uiteraard wel object van onderzoek als het specifiek wordt gebruikt voor 'Continuous Controls Monitoring'. [HOFF07]

11 Zo'n foutmelding kan bijvoorbeeld betrekking hebben op het tijdelijk niet beschikbaar zijn van een benodigde brontabel.

12 Dit betekende niet dat het DWH maandenlang onbetrouwbaar was. Fouten hadden veelal te maken met problemen tijdens het laden en/of transformeren. Die problemen deden zich soms wel voor en vaak ook niet.

13 Uiteraard zijn er wel controlemaatregelen mogelijk, zoals 'cleansing'. [KIMB02] Hierbij worden bijvoorbeeld verschillende veldwaarden waarmee hetzelfde wordt bedoeld, zoals 'Male' en 'M', vertaald naar de waarde 'Mannelijk'.

Literatuur

[ABBA04] Abbas, S., *Kracht van de vernieuwing; Visies op ICT*. 2004.

[ABBA05] Abbas, S., *A new era without Data Warehouses; An architectural vision*. 2005.

[DIJK01] Dijk, C. van, Saher, E. von, Kalk, P., *Managementinformatie II; De lijn van strategie naar stuurinformatie*. Wolters Kluwer, 1e druk, 2001.

[GART12] Gartner, *Gartner Says Adopting a Pace-Layered Application Strategy Can Accelerate Innovation*, persbericht 14 februari 2012. https://www.gartner.com/newsroom/id/1923014, geraadpleegd op 16 november 2017.

[HOFF07] Hoffer, R.M., The value of continuous auditing, EDPACS, 35 (2007) 6, p.1-19.

[INMO96] Inmon, W.H., *Building the Data Warehouse.* John Wiley & Sons Inc., second edition, 1996.

[KIMB02] Kimball, R., Ross, M., *The Data Warehouse Toolkit; The Complete Guide to Dimensional Modeling* (2nd edition), John Wiley & Sons, Inc.; 2002.

[KIMB08] Kimball, R., et al. *The Data Warehouse Lifecycle Toolkit*, 2nd ed. John Wiley and Sons, Inc.; 2008.

[KOBI15] Kobielus, J., Nee, het datawarehouse is nog niet dood, Hadoop maakt geen einde aan de noodzaak van een datawarehouse, *cio.nl*, 1 mei 2015.

[LANS12] Lans, R. van der, *Data Virtualization for Business Intelligence Systems*; *Revolutionizing Data Integration for Data Warehouses.* 2012.

[LINT15] Linthicum, D., Gebruik de cloud niet als datawarehouse, *cio.nl*, 3 juli 2015.

[NBA16] *Nadere voorschriften controle- en overige standaarden (NV COS)*. Vastgesteld bij bestuursbesluit van 6 december 2016. https://www.nba.nl/tools/hra-2017/?document=429, geraadpleegd op 16 november 2017.

[SPOO05] Spoor, L.L., Roozen, F.A., Betrouwbaarheid van periodieke bestuurlijke informatie, *Handboek Management Accounting*, april 2005.

[VEEN14] Veenhof, B-J., Werkhoven, M. van, Automatisering in het kader van de audit – lastige tijden voor auditpraktijk, FM.NL, 7 maart 2014. http://financieel-management.nl/artikel/automatisering-in-het-kader-van-de-audit-lastige-tijden-voor-auditpraktijk, geraadpleegd op 16 november 2017.

[VELD14] Veld, C. in 't, *Sturen op managementinformatie.* Focus op verbeteren, position paper, 7 november 2014. https://www.focusopverbeteren.nl/wp-content/uploads/pdf/Managementinformatie.pdf, geraadpleegd op 16 november 2017.

Drs. A.B. Colenbrander RE | Consultant bij *Wageningen University & Research, IT Informatiesystemen*

Alfred Colenbrander is sinds vier jaar werkzaam bij Wageningen University & Research. Als consultant op het gebied van Business Intelligence (BI) besteedt hij veel aandacht aan financiële informatie. Eerder was Alfred een kleine tien jaar werkzaam in de accountancy. Daar vervulde hij vooral de rol van externe IT-auditor. Later deed hij ook veel consultancywerk op het gebied van administratieve organisatie, ERP en BI. Bestuurlijke Informatievoorziening is de rode draad in Alfreds loopbaan.

De auteur heeft dit artikel op persoonlijke titel geschreven.