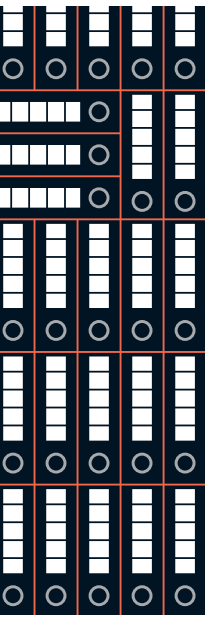


Taak van de databeheerder is een stuk complexer geworden

Correcte definities zijn noodzakelijk

Jan Henderyckx

Geen sector zo onderhevig aan trends als de IT. Is Data Governance niet meer dan de nieuwste voorbijgaande hype in IT-land of gaat er meer achter schuil?

De implementatie van business processen heeft een vloed van ingrijpende veranderingen achter de rug met steeds vaker crossfunctioneel gebruik van informatie en almaar omvangrijkere informatiesystemen, die bijgevolg nog moeilijk te overzien zijn. En ook de wetgeving laat zich niet onbetuigd, met striktere regelgeving omtrent de omgang met gegevens.
De nieuwe uitdagingen, in combinatie met de striktere wetgeving, leiden als vanzelf tot de behoefte aan een data governance-project binnen organisaties. Wat dit nu precies behelst en met welke aspecten rekening te houden, diepen we graag voor u uit in dit artikel, dat concreet ingaat op de informatiedefinitie, het databeheer en de databeveiliging. In de artikelenreeks rond CIM, gepubliceerd in DB/M 1, 2, 3 en 4 van dit jaar, is het aspect datakwaliteit al uitvoerig aan bod gekomen.

Informatiedefinitie

Een doelmatige Data Governance staat of valt met de correcte definitie van data items. Zonder correcte definitie hebben de overige governance regels weinig zin. Zoals vastgesteld is de behoefte aan governance vooral een feit in complexe, omvangrijke omgevingen. Het is dan vooral zaak eventuele conflicten en divergerende definities snel vast te stellen.

Business glossary

Deze bewaking van de definities is nooit louter een Data Governance probleem, want het gaat hier om business concepten die worden toegepast in de beschrijving van business concepten, business objecten en die uiteindelijk hun weerslag vinden in de informatie-persistentielaag.
Ondubbelzinnige definities vormen dus de brug tussen deze drie architecturale lagen, zoals te zien in afbeelding 1. De 'governance' van de definities kan daarom ook geen Data Governance probleem zijn, maar is een corporate governance probleem dat moet worden beschouwd als een wezenlijk onderdeel van de bedrijfs-GRC (Governance, Risk and Compliance) strategie.
Afhankelijk van de invalshoek kan deze definitie zich manifesteren als een glossary, een datamodel, een objectbeschrijving of een velddefinitie. Net als met vele applicatie-architecturen trapt men hier wel eens in de val van een silobenadering. Iedereen heeft wel een eigen tool of SLDC-document (Software Development Life Cycle) om deze informatie op te slaan en te onderhouden. Silo's zijn echter verre van efficiënt: het ultieme doel bestaat namelijk in het vinden van een synergie in de definities en dit kan alleen maar door uit te gaan van het geheel, veeleer dan van de afzonderlijke delen.

> Data lineage biedt een inzicht in wat bij een wijziging de gerelateerde objecten zijn

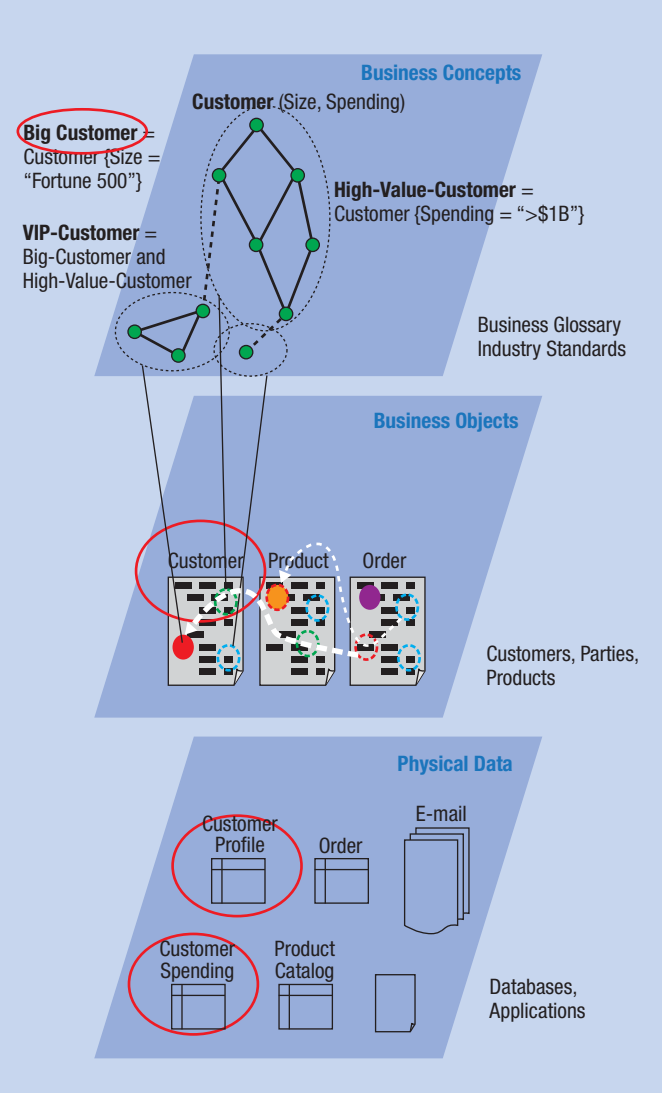
Het synchroniseren van de verschillende bronnen en het aanmaken van lange consolidatierapporten leidt slechts tot een voortdurende divergentie en tot bergen ongelezen papier. Eén enkele repository, wellicht met diverse perspectieven, is de enige werkbare oplossing. Wikipedia is een schoolvoorbeeld van hoe het openen en annoteerbaar maken van informatie uiteindelijk leidt tot een betere definitie.
Het concrete voorbeeld van Wikipedia is trouwens een geslaagde analogie, want ook bij Wikipedia is de rol van steward of bewaker van informatie een sleutelelement. Binnen de organisatie kan ieder naar believen definities aanvullen en verbeteren,

maar zonder een hoofdverantwoordelijke met ultiem beslissingsrecht voor elk concept zou één en ander hopeloos in het honderd draaien.
Het verplicht maken van het definiëren van concepten is derhalve een eerste stap in het doen toenemen van de maturiteit van de organisatie. Bij het uitvaardigen van regels geldt door de bank genomen dat regels met intrinsieke toegevoegde waarde beter worden gevolgd. Het hergebruiken van de definitie binnen de levenscyclus van het concept vergroot de kans op een correcte invulling van deze definitie. Dit brengt ons meteen bij de problematiek van de naamgeving van objecten en data elementen.

Controlled vocabulary

De ISO 11179-5 standaard (zie http://metadata-standards.org) beschrijft de regels waaraan een ondubbelzinnige naamgeving moet voldoen. Het is onontbeerlijk dat objecten worden benoemd via een eindige set van concepten. Dit leidt als vanzelf tot een convergentie van de definities. Een conflict tussen twee gebruikers van dezelfde term geeft aldus aanleiding tot een verduidelijking van de omschrijving of tot het ontstaan van een nieuw business concept met een duidelijker afgebakende betekenis, meteen de link met de business glossary. Data elementen kunnen alleen maar worden benoemd aan de hand van concepten opgenomen in de glossary.
Deze aanpak van naamgeving gaat heel wat verder dan wat de meeste bedrijven toepassen. Een regel die alleen maar de maximale lengte van een naam vaststelt, en de tekens die kunnen worden gebruikt, geeft technisch gesproken wel een eindige set van mogelijke namen maar leidt niet tot automatische convergentie van de definities.
Met andere woorden, lengte (max 18 tekens) en tekendeel van ('a'-'z','A'-'Z','0'-'9','_','-') kunnen niet verhinderen dat een data item 'klant_lengte' en 'klant_hoogte' worden gedefinieerd. Zijn lengte en hoogte aan elkaar gelijk? Dat zal afhangen van de respectievelijke definitie. Zonder strikter afgebakende naamgevingsregels wordt deze vraag zelfs niet gesteld.

Een striktere definitie van data item naam toetst deze voorwaarde wel af. Een bijkomende regel is: naamdeel komt voor in (lijst van gedefinieerde naamdelen, de controlled vocabulary). De keuze tussen lengte en hoogte ligt nu wel vast omdat de concepten verplicht moeten voorkomen in de lijst van gekende naamdelen. De uniekheid van definitie is een rechtstreeks gevolg van de naamgevingsregel. Merk op dat de lijst niet werkt met data item namen maar met naamdelen. Indien de controlled vocabulary bestaat uit volledige data item namen kan er weer divergentie ontstaan. Naast het verplicht gebruik van gedefinieerde concepten is het ook noodzakelijk om het soort van concept vast te leggen en vaste patronen te definiëren.
ISO 11179-5 kent vier soorten naamdelen. Object Class Term (het object dat wordt beschreven), Property Term (de eigenschap van het data item), Representation Term (de voorstelling van de



Afbeelding 1: Drie architecturale lagen.

eigenschap) en Qualifier Term (een bijkomende differentiatie van de Class of Property term). Om zeker tot een semantische convergentie te komen, leggen we ook nog de sequentie van de gebruikte termen op: C - n(Q) - P - R.

Indien de representation term overlapt met de property term wordt de term slechts één keer gebruikt. Aan de hand van een voorbeeld wordt alles veel duidelijker. In het concept Cost Budget Period Total Amount onderkennen we de Class term 'Cost', de Property term Amount, de Representation term Amount en de Qualifier terms Budget Period en Total. Volgens de naamgevingssequentie geeft dit een data item met de naam Cost Budget Period Total Amount (Amount). De volgorde van Budget Period en Total kunnen geen aanleiding geven tot andere definities. De qualifiers zijn daardoor sequentieneutraal. Door het toepassen van deze regels komen we erg dicht in de buurt van ons beoogde doel van unieke definitie. Elk van de individuele delen wordt nu opgenomen in de controlled vocabulary zodat ze op een éénduidige manier herbruikbaar zijn.

Het toepassen van deze methodologie op een bestaande omgeving zal niet eenvoudig zijn, omdat er wordt ingegrepen in tal van processen. Conflicten zullen hoe dan ook ontstaan en al snel zal blijken dat sommige items eenzelfde naam dragen, maar een andere betekenis hebben en dat andere items hetzelfde betekenen maar toch een andere naam dragen. Het is de taak van de data administrator om dergelijke knopen door te hakken en met de business data steward overleg te plegen over de correcte definities.

> Het waardeverval bepaalt hoe de informatie mag worden gebruikt in de verschillende omgevingen

Naast een definitie krijgt elk concept tevens een afgekorte naam als brug tussen een conceptuele naamgeving en de fysieke implementatie. Er is nu een één-op-één relatie tussen een fysieke data item naam en de daaraan gekoppelde concepten uit de controlled vocabulary. Deze link vormt de basis voor de data lineage, onontbeerlijk voor het databeheer.

Databeheer

Veeleer dan een statisch gegeven evolueren data samen met de wijzigende procesbehoeften. Daarbij doelen we niet zozeer op de inhoud dan wel op de metadata van de data items. Steeds als een tabel een nieuwe structuur meekrijgt of een data item van definitie wisselt, moet de impact hiervan worden vastgesteld. Daarnaast moet de wijziging op een veilige manier worden doorgevoerd. De vraag naar de impact kan worden beantwoord via een data lineage oplossing waarbij voor de aanpassing van het object gebruik kan worden gemaakt van change management software.

Data lineage

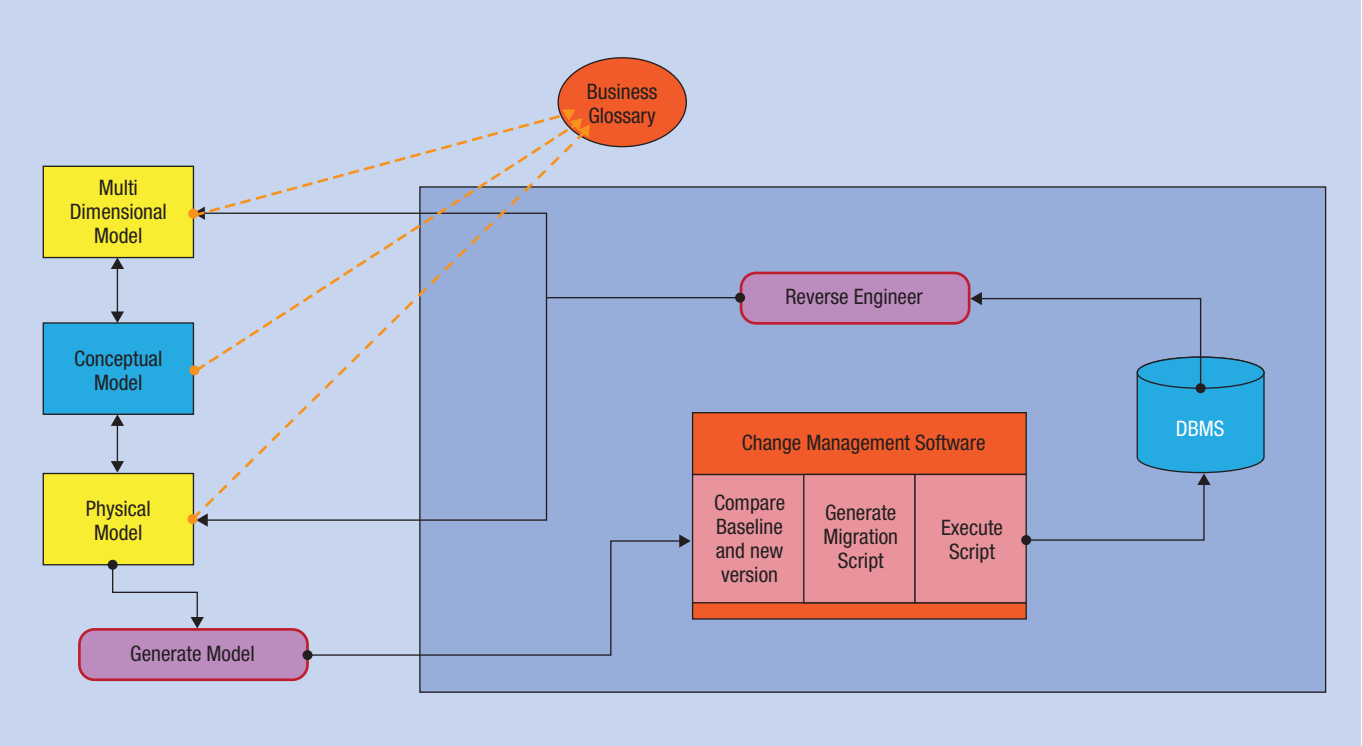
Data lineage biedt een inzicht in wat bij een wijziging de gerelateerde objecten zijn, zoals programma's, andere tabellen, data flows, transformatieprocessen enzovoort. Ondersteund door een coherente datadefinitie is hiermee alvast een eerste hindernis adequaat genomen; namelijk wat houdt verband met wat. Zonder correcte definities en een duidelijke naamgeving daarentegen is het haast onbegonnen werk alle mogelijke correlaties tussen objecten in kaart te brengen. In de praktijk zien we dan ook dat het aanpassen van data items vaak *plug and pray* is. Het aanvullen van de verbanden tussen de data items en processen is een niet-triviale taak die meestal niet volledig onder controle is. Er zijn wel repository's zoals de informatie uit de databanken, maar de link met het proces is daar niet altijd terug te vinden door de manier waarop de databanken worden benaderd. Bij het gebruik van dynamische SQL statements bevat de databank geen informatie over de gebruikte statements tot op het moment van effectieve uitvoering. Het is noodzakelijk om de broncode van de processen te verwerken zodat de statements worden gevonden voor de eigenlijke uitvoering. Dit lijkt een taak waarbij je liever een beroep zou doen op een software-leverancier, gezien de complexiteit en het dynamische karakter van de te verwerken componenten. Jammer genoeg vinden we een lineage oplossing nog al te vaak in een louter BI-omgeving en ontbreekt de link met de operationele gegevens.

Schema-evolutie

Aanpassingen aan een database schema zijn vaak complexe operaties waarbij meerdere Gigabytes aan gegevens worden ontladen en herladen in nieuw gedefinieerde objecten. Anno 2008 zou je verwachten dat alle schema-aanpassingen online en zonder complexe operaties kunnen verlopen, maar de praktijk is weerbarstiger.
De aanpassingen aan het database schema hangen af van welke versie van de applicatieve logica wordt geïnstalleerd op de machines. Het is dan ook best mogelijk dat de implementatie van een nieuw schema in de verschillende fasen van het in productie brengen telkens anders is. Dit betekent dat het voor de DBA niet volstaat om het script, dat de databank aanpassing doorvoert, te kopiëren naar de volgende omgeving. Elke omgeving vraagt een specifiek script. In de volledig geautomatiseerde keten van software management is dit vaak nog de enige stap die uit de pre-geïndustrialiseerde wereld komt. Met name de stap: neem een DBA met ervaring in de hoop dat die geen fouten maakt. Aangezien we nog steeds met mensen werken is het evident dat we op het vlak van risk management met deze oplossing niet echt hoog scoren. Het hoeft echter niet op deze manier te verlopen. In afbeelding 2 zien we hoe het anders kan. Maak gebruik van change management software en koppel deze als het even kan aan data modeling software. Het aanmaken van het script gebeurt daarbij door software op basis van het database schema, dat via de CMDB (configuration management database) wordt aangeleverd, en de actuele toestand van het DBMS schema. De DBA dient nu alleen nog toe te zien dat er geen fouten ontstaan. Hiermee zet ook de DBA de stap in de 21ste eeuw en neemt het risico bij het doorvoeren van een change flink af.

Databeveiliging

Het beveiligen van data behelst uiteraard meer dan louter het beheer van de toegangsrechten. De problematiek van het beheer van omgevingen en transfer van gegevens tussen omgevingen mag daarbij zeker niet uit het oog worden verloren.
Bij de ontwikkeling van toepassingen start men in een development-omgeving met een erg beperkte set van gegevens. Het gaat daarbij alleen om data die nodig zijn voor functionele testen. De inhoud moet voldoen aan basisvalidatieregels, maar er is geen behoefte aan grote volumes, die de testen alleen maar



Afbeelding 2: Change management software gekoppeld aan data modeling software.

kunnen bemoeilijken. Deze functionele testgegevens zijn dan ook niet van kritieke aard. Bij de volgende stappen, unit- en systeem-testing, is er wel behoefte aan grotere volumes gegevens, soms zelfs in dezelfde hoeveelheden als in de productie. Vraag is wat daarbij de aangewezen manier is om deze gegevens aan te maken en te beheren. Een belangrijke parameter hierbij is de tijdsgevoeligheid van de gegevens: hun waardeverval.

Het waardeverval van informatie

Stel dat we kunnen beschikken over de lottocijfers van komende zaterdag. De waarde van deze informatie is erg afhankelijkheid van het tijdstip waarop we er over beschikken. Op vrijdagochtend nog van onschatbare waarde, op zondag al een stuk minder. Of; de naam van een politie-informant is informatie die wellicht pas in waarde afneemt na het overlijden van de informant, en zelfs dan kan het vrijgeven van deze informatie minder prettige gevolgen hebben.

Het waardeverval van de informatie bepaalt dus hoe de informatie mag worden gebruikt in de verschillende omgevingen. Voor de lottocijfers volstaat het te wachten tot na de trekking, om ze dan te kopiëren naar minder beveiligde omgevingen. In het geval van de informant is het niet toegestaan de informatie te kopiëren zonder deze eerst te versleutelen. Gelukkig bestaat er masking-software om de inhoud van velden op dergelijke manier om te vormen zodat ze niet terug te leiden zijn naar de oorspronkelijke waarde, maar waarbij de gegevens toch in overeenstemming blijven met de databank en de applicatieve integriteitsregels. Het is vooral de gegevensintegriteit die het moeilijk maakt om de productie-informatie eventjes op een veilige manier zelf te gaan versleutelen. De eenvoudigste en goedkoopste oplossing is dus het laden van productiegegevens met de nodige vertraging, bijvoorbeeld de backup van vorige maand. Hebben de gegevens echter een waardevervalperiode die te groot is, dan is het verstandig te investeren in software die versleutelen en intelligent genereren mogelijk maakt.

Conclusie

Het correct beheren van data behelst duidelijk meer dan het louter waken over operationele aspecten zoals beschikbaarheid, schaalbaarheid en performance. Dit zijn aspecten die in de meeste organisaties inmiddels onder controle zijn, maar zoals is duidelijk gemaakt, komt er heel wat meer kijken bij een afdoende Data Governance. Het onderkennen van de extra uitdagingen en er een pertinente invulling aan geven kan het verschil maken tussen compliance of overtreding van de regels.

De databeheerder krijgt er door deze extra regels onmiskenbaar een pak taken bij met een mogelijk enorme impact op de totale organisatie. Het uitrollen van een glossary en aanpassen van een nieuwe naamgevingsstandaard zijn intensieve taken die niet even in een maand afgehandeld zijn, maar vaak processen van lange adem zijn, over meerdere jaren gespreid.

De taak van databeheerder is dus een stuk complexer geworden, maar biedt tegelijkertijd een mooie uitdaging. Het gaat daarbij al lang niet meer alleen om de infrastructuur, maar komt de algehele bedrijfsvoering ten goede.

Jan Henderyckx is onafhankelijk consultant, spreker en auteur op het gebied van informatie-architectuur en databases.