Eerste prioriteit: in kaart brengen van het datalandschap

De uitdaging voor databeheerders

Malcolm Chisholm

De afgelopen jaren zijn voor IT niet gemakkelijk geweest. Het barsten van de internet-zeepbel had een aanzienlijke afname van IT-medewerkers tot gevolg. Outsourcing, en speciaal offshoring, dragen ook een flinke steen bij aan deze trend. Feitelijk is er sprake van een periode van IT-moeheid, met niet alleen dalende investeringen, maar ook tanende belangstelling voor IT.

Databeheer blijkt hier nog meer onder te hebben geleden dan IT in zijn algemeenheid. Het waarom staat ter discussie. In het ene kamp wordt vermoed dat databeheer te weinig toegevoegde waarde aan de business heeft geleverd; het andere kamp denkt dat databeheer alleen op de lange termijn bijdraagt aan de resultaten en dus niet past in de huidige trend van korte-termijn opbrengsten. Wat de reden ook mag zijn, op databeheer zou nooit zijn bezuinigd als de bestuurders ervan overtuigd waren dat het een essentiële functie had. Klaarblijkelijk is deze perceptie nooit ontstaan in de hoofden van de beslissers.

Op dit moment zijn er onmiskenbare tekenen van hernieuwde belangstelling bij het management van de informatie-assets van bedrijven. SOA's, Master Data Management en de noodzaak van Data Governance zijn allemaal 'hot topics'. Senior Executives zijn weer bereid om geld in IT-projecten te steken die voor opgeschoonde data en geïntegreerde databases moeten gaan zorgen. We zijn dus op het punt aangekomen waar databeheer een tweede kans krijgt. Het is maar de vraag of databeheer er deze keer wel in slaagt om haar waarde aan te tonen, of dat het opnieuw op een mislukking uitdraait. Ik moet toegeven, dat als de huidige fase waarin databeheer zich bevindt alleen maar een voortzetting is van de voorafgaande fase, het zeer zeker een risico is.

Scheiding tussen logisch en fysiek

In het verleden hebben de meeste beheerafdelingen hun werk uiteen zien vallen in twee categorieën. De eerste is die van datamodellering en dan speciaal de logische datamodellering. Het begrijpen van de abstracte concepten achter datamodellering en het op de juiste wijze kunnen toepassen van de normalisatieregels worden bij dit soort werk beschouwd als de belangrijkste vaardigheden voor databeheer-medewerkers.

Er zijn afdelingen die dit ook uitbreiden naar dimensionele modellering voor datawarehouses en datamarts. In mijn ervaring is het echter helemaal niet gebruikelijk om datamodellering te doen binnen het databeheer. Het blijkt dat datawarehouses en datamarts dikwijls geïmplementeerd worden als stand-alone projecten met hun eigen modelleerders. De tweede categorie databeheerwerkzaamheden betreft het optreden als 'bibliothecaris' van een data dictionary van data-elementen. Soms zijn ook business termen in de data dictionary opgenomen.

> Voor de meeste medewerkers lijkt het antwoord op alle vragen te liggen in datamodellering

Naast deze twee werkzaamheden moeten ter ondersteuning talloze activiteiten worden uitgevoerd, zoals de ontwikkeling van standaards, en de opbouw van analytische kennis van de databeheermedewerkers. Wat het meest opvalt is dat de werkzaamheden van databeheerafdelingen volledig gericht zijn op het logische niveau. Het werk richt zich zelden op het fysieke niveau door het specificeren van fysieke datamodellen of het verbinden van logische datamodellen aan reverse-engineered fysieke datamodellen.
Zo zien we dat er een duidelijke kink zit in de verbinding tussen de werelden van de logische en de fysieke data. Inderdaad, voor veel databeheerafdelingen wordt het feit dat de data daadwerkelijk zijn geïmplementeerd op een fysiek niveau beschouwd als een hinderlijke bijkomstigheid die het best gewoon genegeerd kan worden. De medewerkers op dit soort afdelingen

hebben een hekel aan fysieke data en doen hun best ze te vermijden. Zelfs als ze op een project worden gezet dat duidelijk óók fysieke data omvat, zoals Master Data Management, proberen ze vooral hun werkzaamheden terug te brengen tot het logische niveau en alleen maar datamodellen te produceren.

De problemen van modellen

Logische datamodellen zijn prachtig en erg nuttig, maar ze hebben hun beperkingen. De eerste is dat ze geen erg rijk spectrum aan metadata bevatten. Wat we meestal zien is:

- Naam entiteit;
- Definitie entiteit;
- Naam attribuut;
- Definitie attribuut;
- Datatype attribuut (soms);
- relatie met parent- en child-entiteiten;
- relatie met Optionality;
- relatie met Cardinality;
- relatie Verb Phrase (soms, en meestal niet erg handig).

Bovendien kunnen de datamodelleer-tools niet worden ingesteld op het vastleggen van andere soorten metadata. In theorie kan een data-analist altijd een plekje vinden om aanvullende metadata op te slaan, bijvoorbeeld in een notes field, of een user-defined property. Zelfs als dat kan, zijn de tools er niet voor ontworpen om zulke metadata te managen. Dus informatie over eigenaar, beheerder, entity life cycles, productiestatus, entity categorisatie voor beheeracties en dergelijke kan niet op de juiste manier worden behandeld.
De business redenen die zorgen voor een hernieuwde interesse in databeheer hebben te maken met het op brede schaal delen van data binnen een onderneming. Om dat te bewerkstelligen is het nodig te weten waar de data zich bevinden, wat ze betekenen, welke kwaliteitsrisico's er in het spel zijn. Het betekent ook dat er implementatie moet plaatsvinden van governance-programma's, masterdata management, datakwaliteits-certificering, enzovoort. De metadata die voor dit alles nodig zijn gaan de mogelijkheden van een datamodel ver te boven; ook de vereiste functionaliteit verschilt nogal van die van het gebruikelijke datamodellerings-tool.

Toch lijkt voor de meeste medewerkers van databeheerafdelingen het antwoord op alle vragen te liggen in datamodellering. Deze medewerkers zullen zichzelf eerder zien als datamodelleerders dan als data-analisten. Dat is een groot probleem. Modellering is een activiteit met een grote ontwerpcomponent. De business kan een team datamodelleerders vertellen dat ze te maken krijgen met klanten, leveranciers, banken, onderaannemers enzovoort. De modelleerders gaan dan waarschijnlijk de discussie aan hoe dit kan worden geabstraheerd naar een Party entity, gemanaged door verschillende rollen en relaties, waarvan geen enkele ook maar enige betekenis heeft voor de business gebruikers (tenzij het kan worden terugvertaald naar business termen). Dan wordt een logisch datamodel gebouwd en er wordt een softwarepakket gekocht en geïmplementeerd om zo aan de behoeftes van de business te voldoen. Wat heeft in dit scenario een logisch datamodel voor zin? Er zal gezegd worden dat het gebruikt kan worden bij de aanschaf van de software, om te zien of die beantwoordt aan de eisen van de business. Maar als er in het modelontwerp gekozen is voor zaken zoals Party entity's, vergelijken we eigenlijk de oplossing van de datamodelleerder met die van het softwarepakket. Een logisch model kan men terugmappen naar business-niveau, maar het is nog steeds een ontwerp op een ander abstractieniveau en beschrijft het niet specifiek.

> De meest urgente behoefte ligt bij de koppeling met de business

Ontwikkeling op maat is momenteel tamelijk zeldzaam, maar het gebeurt nog steeds. In die gevallen wordt de IT-afdeling die verantwoordelijk is voor de implementatie van de fysieke database voorzien van de logische datamodellen. Vaak worden deze modellen slechts behandeld als uitgangsposities of suggesties. Er is meestal geen enkel proces dat ervoor zorgt dat het logische ontwerp wordt gemapped naar het fysieke ontwerp en dat de twee in de pas gehouden worden als in de loop der tijd veranderingen plaatsvinden.
Het eindresultaat van dit alles is dat het enige waarover we beschikken logische datamodellen zijn en dat deze niet overeenkomen met fysiek geïmplementeerde databases.

De Black Box repository

De andere categorie werkzaamheden die traditioneel door databeheerders wordt uitgevoerd, is het bouwen van een data dictionary van data-elementdefinities. Dikwijls worden deze artefacten opgeslagen in een repository die niet door de hele onderneming kan worden benaderd. Ik ben implementaties tegengekomen waarbij het nodig was om een formeel verzoek in te dienen bij de databeheerafdeling voor het verkrijgen van rapporten over wat in de repository zat. Dit lijkt misschien een uitzonderlijk geval, maar repository's die niet breed toegankelijk zijn komen zeer veel voor.
Een stuk van de weerzin van databeheerders om repository's met data-elementdefinities open te stellen, blijkt te maken te hebben met de kwaliteit van de inhoud. Hier zijn wat voorbeelden:

- Data Element: accountnummer;
- Definitie: een uniek nummer voor een account;
- Entiteit: Employee Project Assignment;
- Definitie: deze entiteit geeft een many-to-many relatie weer tussen de medewerker-entiteit en de project-entiteit.

Seminar Master Data Management

Malcolm Chisholm verzorgt op 21 november 2007 een seminar Master Data Management in de Holiday Inn te Leiden.
Meer informatie op www.arrayseminars.nl.

De eerste definitie voert alleen maar terug naar de attribuutnaam en zegt helemaal niets over de manier waarop een accountnummer wordt gegenereerd, of het een intelligente sleutel is (hetgeen ze vaak zijn), hoe het wordt beheerd, enzovoort. De tweede definitie is geschreven voor datamodelleerders en heeft geen relevantie voor de business. Naar mijn ervaring is de kwaliteit van de definities in data dictionary's tamelijk laag en daarmee is hun bruikbaarheid gelimiteerd.
Een tweede reden om een repository binnen ondernemingen af te schermen is dat als gebruikers kunnen zien dat een data-element bestaat, ze ook willen weten waar het is geïmplementeerd, en zoals we net bespraken bestaat er zelden een verbinding tussen de logische en de fysieke niveaus. In plaats van uit het pluche van het logische niveau te komen, geven de meeste databeheerafdelingen er de voorkeur aan om hun metadata niet toegankelijk te maken voor pottenkijkers uit de rest van de onderneming.

Een ander probleem is dat als de data-elementdefinities moeten worden verzameld, de databeheerder dikwijls probeert iemand anders die opdracht te laten uitvoeren. Dat gebeurt dan door het definiëren van 'standaards' die bedoeld zijn om mensen buiten de databeheerafdeling het werk toe te wijzen. Zo kan er bijvoorbeeld een 'standaard' zijn dat men bij elk systeemontwikkelingsproject de data-elementen waarmee het project werkt, in moet voeren in de data dictionary. Andere mensen vertellen wat ze moeten doen ten behoeve van een hoger doel, werkt maar tot bepaalde hoogte. Het werkt in elk geval niet als het projecten vertraagt of tot geen enkel aantoonbaar resultaat voor de onderneming leidt.

De fysieke uitdaging

De focus van databeheer op het logische niveau wordt versterkt door de activiteiten op het gebied van datamodellering, alsmede door de functionaliteit van de gebruikelijke datamodelleringstools. Het wordt nog verder versterkt door het besef dat een doorsnee business gebruiker alleen maar van de data wil weten wat de definities zijn. Als deze focus niet verandert, dan zal databeheer niet het hoofd kunnen bieden aan de zware uitdagingen waar vele ondernemingen mee te maken hebben. Wat moet de databeheerder dan wél doen? Om deze vraag helemaal te beantwoorden valt buiten het kader van dit artikel, maar we concentreren ons op één competentie die alle moderne databeheerafdelingen moeten ontwikkelen: fysiek databeheer. Dit betekent niet dat men zich moet gaan bemoeien met

productie-ondersteuning of database-beheer. Het betekent veel meer dat ervoor gezorgd moet worden dat de onderneming haar informatie-assets volledig kan benutten.

In kaart brengen

Het is voor de meeste bedrijven vrijwel onmogelijk om te begrijpen welke fysieke data ze hebben en in welke databases ze worden bewaard. Enkele individuele medewerkers weten gedeeltelijk welke data waar zijn, maar meestal is hun blikveld beperkt tot hun naaste omgeving. Daarom snijdt het hout om te spreken van het fysieke-datalandschap. Het bestaat echt en moet in kaart gebracht worden. Zonder dat dat is gebeurd, is het onmogelijk om te weten welke informatie de onderneming bezit en waar die informatie zich bevindt. Er bestaat geen logische-datalandschap en vragen over welke data het bedrijf bezit en op welke locaties, hebben op het logische niveau geen enkele betekenis.

Het is een nalatigheid van de databeheerafdeling dat kennis over het datalandschap bij zoveel bedrijven ontbreekt. De databeheerafdeling voelt zich gelukkig met allerhande standaards, maar heeft meestal verzuimd om iets op te nemen over het verzamelen van data ten aanzien van fysiek geïmplementeerde databases. Dat betekent dat ondernemingen allemaal stuklopen, tenzij ze hun datalandschap in kaart brengen. Helaas rijst hierbij een zeer groot probleem, omdat dit een enorme hoeveelheid analistentijd vergt. Bovendien gaat het waarschijnlijk zo lang duren, dat tegen de tijd dat de analisten één gebied in kaart hebben gebracht en bezig zijn met het volgende, de eerste intussen alweer is veranderd.

De enige realistische manier om het datalandschap in kaart te brengen is om geautomatiseerde tools te gebruiken. Gelukkig zijn er inmiddels dergelijke tools op de markt. Daarbij moet de databeheerafdeling een repository installeren waarin deze metadata kunnen worden opgeslagen, toegankelijk voor het grootst mogelijke publiek. Maar belangrijker is nog dat de databeheerafdeling moet begrijpen wat ze met die informatie moeten doen, waarbij de meest urgente behoefte ligt bij de koppeling met de business.

Bouw en onderhoud van de logische-fysieke koppeling

Het in kaart brengen van het datalandschap heeft wel wat weg van het in kaart brengen van het menselijk genoom. In het geval van het menselijk genoom kunnen we vaststellen waar een gen zich bevindt en wat zijn structuur is, maar dat is alles. We kunnen uit de kaart niet opmaken wat het gen doet of hoe belangrijk het is. Dit geldt ook voor het datalandschap. Het in kaart brengen levert alleen maar een lijst op met databases, tabellen en kolommen. Het is de verantwoordelijkheid van databeheer om voor de processen te zorgen die deze metadata met het logische niveau verbinden.

Dat betekent dat de metadata in de logische-datamodellen, die traditioneel door de databeheerder werden gemaakt, geïmporteerd moeten worden in de repository waar het fysieke landschap is opgeslagen en daaraan worden gekoppeld. Misschien dat dit de eerste keer op een hoger niveau kan geschieden – zeg, subject area naar database. Het duurt erg lang eer we op het punt zijn aangeland waar alle entiteiten en attributen aan tabellen en kolommen zijn gemapped. Het is ook waarschijnlijk dat er zich tijdens het proces problemen openbaren waarmee eerst moet worden afgerekend.

> Ontwikkeling op maat is tamelijk zeldzaam maar gebeurt nog steeds

We mogen niet vergeten dat het logische niveau wordt gebruikt in plaats van het business niveau, omdat meestal conceptuele datamodellen op business niveau en de bijbehorende glossaria met business termen niet bestaan. Logische datamodellen zijn het beste alternatief en meestal al aanwezig. Zoals gesteld, bevatten logische datamodellen hoe dan ook grote elementen van ontwerp die nauwelijks meer zijn dan een rechtstreekse beschrijving van de business. Het is uiteindelijk de verantwoordelijkheid van databeheer om het fysieke-datalandschap te koppelen aan het business niveau.

Databeheer zonder beperkingen

Met begrip van het datalandschap en met aan elkaar gekoppelde business-, logische- en fysieke-dataniveaus, kan de databeheerafdeling eindelijk haar naam waarmaken. Ze kunnen zaken als governance implementeren op alle mogelijke manieren. Veranderingen in het fysieke landschap bijvoorbeeld, kunnen worden opgespoord en aangepast aan change control requests die helemaal terugvoeren naar het niveau van business informatie. Data kunnen een tag krijgen met een versie of een kwaliteitscertificatie. Ownership en stewardship kunnen proactief worden beheerd. Dat zijn de activiteiten die een moderne databeheerafdeling niet alleen maar relevant, maar essentieel maken voor de onderneming. Het vraagt om een andere instelling, maar als een paar bedrijven er succes mee behalen zal de druk op databeheerafdelingen die nog op de oude manier werken geweldig groot worden. Medewerkers die hun instelling niet kunnen veranderen, zijn enorm in het nadeel. Bij deze zee aan veranderingen kun je beter de golf voorblijven dan meegesleept worden.

Dit is een bewerkte en vertaalde versie van de originele Engelse tekst, die u kunt vinden op onze website www.dbm.nl onder 'specials', 'extra materiaal'. In geval van discussies is de originele Engelse tekst doorslaggevend.

Malcolm Chisholm is directeur van Askget.com Inc te New Jersey.