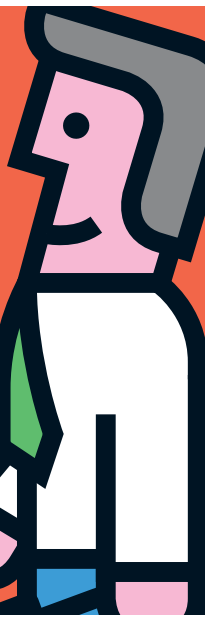


Verscheidenheid aan oplossingen is beschikbaar

Open Source Data Quality

Jos van Dongen

In juni 2008 gaf ik een presentatie over Open Source BI en het 'data quality' vakje in het componentenoverzicht was nog opvallend rood gekleurd. Rood met als betekenis: nog geen producten voor dit onderdeel beschikbaar. De ontwikkelingen op dit vlak zijn de laatste tijd echter in een stroomversnelling geraakt: hoogste tijd dus om hier aandacht aan te besteden. En wat biedt dan een beter kader dan het thema Data Governance?

Problemen met datakwaliteit kosten het bedrijfsleven volgens een schatting van The Data Warehousing Institute jaarlijks honderden miljoenen euro's. De grootste ellende ontstaat al tijdens het invoeren van gegevens, maar ook de verspreiding van informatie over verschillende systemen, verkeerd uitgevoerde data-integratieprocessen en gefragmenteerde validatieregels dragen bij aan de vernietiging van bedrijfskapitaal (want dat is het). Datakwaliteit is dus vooral een organisatorisch probleem; tools leveren hier slechts een bescheiden doch onmisbare bijdrage. Tegen de tijd dat we tools in kunnen zetten is het kwaad al geschied en gaat het meestal om het verbeteren van gegevens-kwaliteit, in plaats van het voorkomen van problemen. Data quality tools zijn er in verschillende soorten en maten, en als we naar Open Source oplossingen kijken blijkt ook daar inmiddels een verscheidenheid aan oplossingen beschikbaar.

Functies van DQ tools

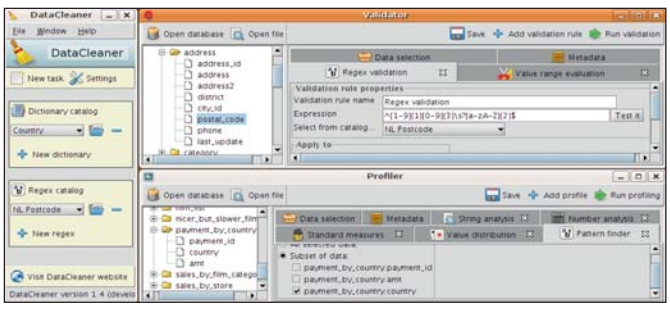
Elders in dit vakblad heeft u al kunnen lezen wat er precies wordt verstaan onder Data Governance en data quality. Als we naar de voorhanden zijnde tools kijken kunnen we ruwweg de volgende soorten toepassingen onderscheiden:

- Data profiling, inzicht krijgen in de opbouw en samenstelling van gegevens;
- Data standaardisatie, afdwingen van regels voor gegevens-kwaliteit;
- Geocoding, valideren en corrigeren van naam/adresgegevens;
- Matching/linking, koppelen van dezelfde soorten gegevens uit verschillende bronnen.

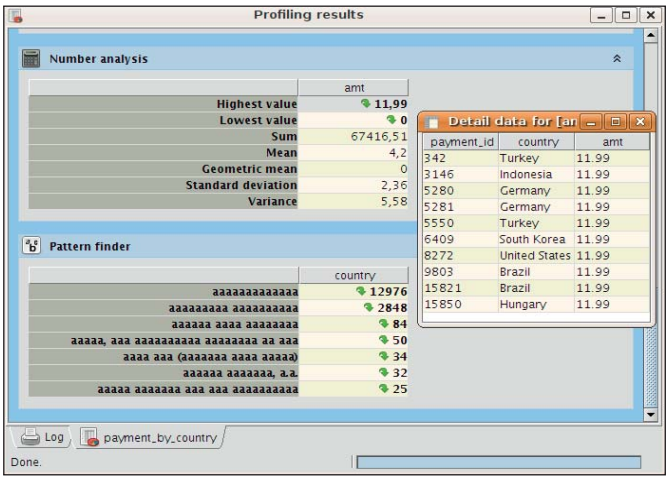
Voor de ondersteuning van deze functies zijn verschillende commerciële oplossingen beschikbaar. Data profiling maakt vaak al standaard onderdeel uit van ETL-tools (bijvoorbeeld Data Integrator of Oracle Warehouse Builder) en ook binnen de Open Source gemeenschap zijn hiervoor verschillende tools beschikbaar, waarover later meer. Kijken we naar meer geavanceerde functies als adresvalidatie, gegevensstandaardisatie, rapportage en het ontdubbelen en matchen/linken van gegevens dan ziet het er wat minder fraai uit, maar ook hier zijn enkele veelbelovende initiatieven gaande.

Data profiling

Het kunstje van het maken van een gegevensprofiel wordt het best beheerst en is ook relatief eenvoudig te implementeren, althans voor zover het de basis betreft. De basis bestaat uit het bepalen van kolomprofielen, waarbij per database-kolom informatie als aantal unieke waarden, aantal NULL waarden, aantal lege velden en minimum/maximum/gemiddelde veldlengte (bij strings) of waarde (bij numerieke gegevens) wordt berekend. Inderdaad, dit zou ook met SQL in een query editor kunnen. Lastiger wordt het al als er een top/bottom-10 berekend dient te worden, of statistische uitkomsten als mediaan, standaard deviatie, variantie en laagste/hoogste kwartiel. In de meeste gevallen is het ook mogelijk een frequentietabel te laten bepalen, zodat de verdeling van de waarden binnen de kolom op een grafische wijze zichtbaar wordt gemaakt.



Afbeelding 1: Datacleaner.



Afbeelding 2: Datacleaner output.

Soms is het inzichtelijk maken van relaties tussen twee tabellen ook een onderdeel van data profiling, bijvoorbeeld om het aantal klanten te bepalen dat nog nooit een order heeft geplaatst (waardoor het eigenlijk geen klanten zijn). Een volgende stap ligt meer op het terrein van datakwaliteit en hier komen zaken als patroonherkenning op basis van regular expressions en domeinvalidatie aan bod.

Eobjects Datacleaner

Zoals gezegd zijn de meeste ontwikkelingen op data quality-gebied erg recent. Het Deense Datacleaner van Eobjects heeft pas in maart van dit jaar een alfa release opgeleverd, maar is inmiddels al bij een stabiele 1.4 versie aangeland. Het pakket bestaat uit de onderdelen profile, validate en compare die alle drie vanuit een centrale console kunnen worden gestart. De naam Datacleaner is overigens 'op de groei' gekozen omdat er nog niet zo veel te cleanen valt. Wat al wel kan is het uitvoeren van een grote hoeveelheid analysetaken op elke gewenste database. Allereerst dienen er connecties te worden gedefinieerd en bestaat de mogelijkheid om catalogs aan te maken. Een catalog is een lijst met beschikbare domeinvalidatie-bestanden (database of tekst) of regular expressions. Met name de werking van deze laatste optie is goed doordacht. Regular expressions zijn krachtige hulpmiddelen voor data analyse maar meestal niet voor de 'faint at heart'. Eobjects maakt de drempel wat lager door enerzijds een aantal veel gebruikte controles mee te leveren die vanuit een regex properties-bestand kunnen worden geopend, anderzijds door een validatie op de expressie uit te voeren en de mogelijkheid te bieden om een en ander te testen en uit te proberen. Vooral deze laatste optie is erg handig om direct te kunnen zien of een expressie het gewenste resultaat oplevert. In afbeelding 1 is goed te zien hoe het pakket in elkaar zit. Links staat het hoofdscherm waarin ook de catalogs kunnen worden gedefinieerd en van waaruit nieuwe taken worden gestart. Rechts staat bovenaan een validatietaak met daarin een regular expressievalidatie voor de controle van Nederlanse postcodes, en onderaan is de definitie van een profile te zien.

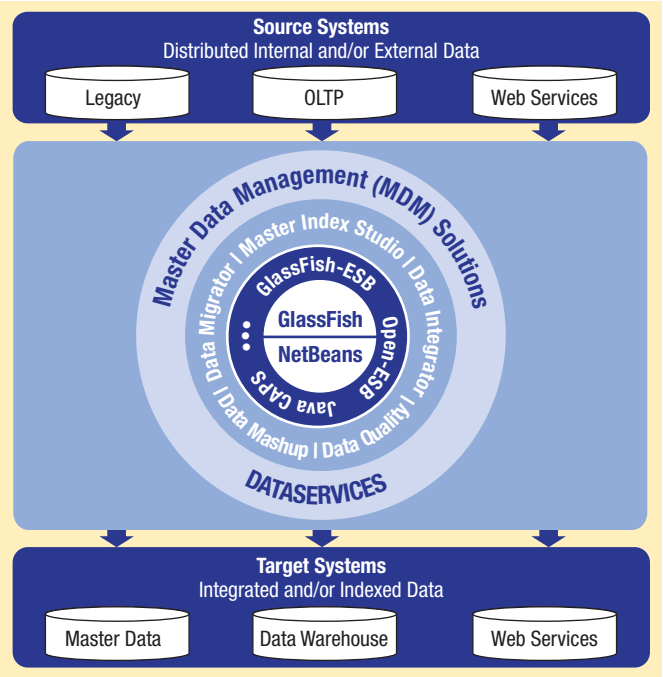
Elke taak kan bestaan uit verschillende profile-opties. Zodra een taak wordt uitgevoerd komt er een uitvoerscherm in beeld met daarin de gevraagde resultaten. Afbeelding 2 laat een subset van deze resultaten zien, namelijk die van de numerieke analyse en de patroonherkenner. Zoals in dit soort tools gebruikelijk is, kan direct ingezoomd worden op de resultaten. Zo is te zien dat de hoogste waarde (11,99) maar een paar keer voorkomt.
De toekomst van Datacleaner ziet er beloftevol uit: versie 2.0 zal web based worden en Birt gebruiken voor rapportage, waardoor ook een flink arsenaal aan grafische mogelijkheden beschikbaar komt. Tevens zullen opties voor scheduling en monitoring van data cleaning jobs toegevoegd worden, waardoor het pakket een geduchte concurrent kan zijn voor de commerciële tegenhangers.

Talend Data Quality

Op 23 juni jongstleden kondigde Talend met de voor hun gebruikelijke bombarie de "eerste Open Source Data Profiling oplossing" aan, gevolgd door een aankondiging op 20 augustus van een complete Data Quality suite die geïntegreerd gaat worden in Talend's Open Studio data-integratieproduct. Uiteraard zijn ze hierin ook weer de eerste, nou ja, laten we zeggen: de eerste die het op deze manier van de daken schreeuwt en er een hoop persaandacht mee krijgt. Dat dan weer wel.
De complete DQ suite is nog niet beschikbaar dus we moeten het in dit artikel doen met versie 1.1 van de profiler. Globaal is de werking hetzelfde als van vergelijkbare pakketten, dus we geven eerst aan wat er geanalyseerd dient te worden, dan welke analyses op de data losgelaten moeten worden en tot slot doen we het uitvoeren en bekijken van de resultaten. In tegenstelling tot Eobjects beschikt Talend wel over grafische uitvoer, maar daar houden de voordelen ook wel op. Het pakket bevindt zich nog duidelijk in een bèta-fase, wat vooral blijkt uit de krakkemikkige schermopbouw en -afhandeling, het nog niet werken van de doorklikmogelijkheid op resultaten zodat de detailrecords getoond worden en het domweg niet berekenen van waarden die vervolgens als 'not available' worden weergegeven.
Een andere onvolkomenheid betreft de ontbrekende synchronisatie met de metadata. Bij het aanmaken van een connectie worden de metadata van de te analyseren database ingelezen. Wordt er daarna een tabel of view toegevoegd of gewijzigd, dan is deze informatie alleen beschikbaar te krijgen door een nieuwe connectie te maken. Tja, dat alles opgeteld bij de haperende graphics, vaste scherm layout en vastlopen bij grote tabellen geeft als resultaat dat deze tool op dit moment niet de eerste keuze zal zijn voor data profiling.

Glassfish Mural

Voordat we ingaan op de data quality opties eerst even een (korte) Glassfish introductie. Dit is een door Sun gestart initiatief met als doel om een volgende generatie Java applicatieserver te ontwikkelen. Hier blijft het echter niet bij: het project omvat een groot aantal subprojecten, waarvan het destijds embryonale Data Integrator al eens onder de loep is genomen. Mural is het



Afbeelding 3: Mural architectuur.

subproject dat zich bezighoudt met Master Data Management in de volle breedte. Data profiling hoort hierbij, maar ook cleansing, deduplication, merging en matching van gegevens uit verschillende bronnen. In afbeelding 3 is een schematische weergave van het platform te zien. Alles start met een Master Index applicatie die wordt gebouwd met de Master Index (MI) Studio. In tegenstelling tot de eerder besproken producten gaat het hier niet bepaald om een 'tooltje', maar om een volledige ontwikkelomgeving, waarin alles wat met master data te maken heeft gerealiseerd kan worden.

Mural Data Quality bevat verschillende engines die hiervoor ingezet kunnen worden: match, standaardisatie, profiling en cleansing. Deze kunnen afzonderlijk in een applicatie worden gebruikt maar ook als (web)service beschikbaar worden gesteld aan andere componenten, zoals de Data Integrator. Hierdoor kunnen in één proces onderdelen voor extractie, matching en cleansing worden geïntegreerd.

Mural kent twee omgevingen voor MI applicaties, een ontwikkel- en een runtime-omgeving. De ontwikkelomgeving is de MI Studio die als plugin voor Netbeans, de standaard Java (en meer) ontwikkelomgeving van Sun, beschikbaar is. Met behulp van Netbeans wordt een MI applicatie gecreëerd en geconfigureerd. Deze laatste stap is de belangrijkste in het proces, maar ook de meest bewerkelijke. Configuratie betekent onder andere het aangeven van welke validaties op welke database-velden dienen te worden uitgevoerd, naast het definiëren van connecties, parameters, etcetera. Alle configuratie-informatie wordt opgeslagen in XML-bestanden die voor een deel bewerkt kunnen worden met de Configuration Editor. Inderdaad, voor een deel. Het andere deel dient handmatig te worden bewerkt met behulp van de XML editor. Hoewel de ontwikkelomgeving alles wat wordt ingevoerd valideert, waardoor de kans op (syntactische) fouten klein is, werpt dit toch een drempel op. Even snel een data-analyse uitvoeren is er dan ook niet bij, gezien het feit dat na het ontwikkelen de oplossing nog uitgerold moet worden naar de runtime-omgeving, de al eerder genoemde Glassfish applicatieserver. Pas hierna kan er daadwerkelijk gewerkt worden met de MI applicatie, en even wat aanpassen betekent weer de ontwikkel/configuratie/uitrol-cyclus opnieuw doorlopen.

Conclusie

De eerste conclusie is uiteraard dat er in een paar maanden tijd een hoop kan veranderen, en dat de open source wereld inmiddels ook oog heeft voor de echt belangrijke zaken rondom data management. Wat niet bij de afzonderlijke oplossingen is aangegeven (maar niet onbelangrijk om te vermelden) is het feit dat ze allemaal op Java zijn gebaseerd. Het maakt dus niet uit welk Windows-, Linux- of Unix-platform u heeft, deze spullen werken altijd. Tenminste: áls ze werken. Er is namelijk nóg een project (sourceforge.net/projects/dataquality) dat er veelbelovend uitziet, maar dat niet voorbij configuratie- en connectie-errors wilde komen.

Wat node mist bij de eenvoudiger pakketten van Eobjects en Talend is een repository waarin de analyseresultaten worden opgeslagen. Dit zou enerzijds de werking aanzienlijk versnellen, omdat nu elke keer alles moet worden doorgerekend. Anderzijds zou dit meer mogelijkheden voor rapportage bieden, ook buiten het pakket om. Mural werkt wel op basis van een database die tijdens het configuratieproces dient te worden aangemaakt en waarin alle resultaten worden opgeslagen.

Het eindoordeel is natuurlijk niet verrassend: de mensen van Talend hebben nog wel wat werk te verzetten voordat hun Profiler zinnig ingezet kan worden. Eobjects Datacleaner is een prima te gebruiken oplossing waar helaas de repository en grafische mogelijkheden (nog) ontbreken, maar gezien de plannen voor versie 2.0 is dit slechts een kwestie van tijd. De zwaargewicht in dit gezelschap is uiteraard Mural, waarmee niet alleen profiling maar ook ontdubbeling en standaardisatie gerealiseerd kunnen worden. Gezien de (soms bizar) hoge kosten die vergelijkbare commerciële tools met zich meebrengen, zou dit wel eens een heel interessant product kunnen zijn. Bent u echter alleen geïnteresseerd in een stuk analyse/profiling en werkt u al met bijvoorbeeld Data Integrator of OWB, blijf dat dan vooral lekker doen.

Referenties

The Data Warehousing Institute (www.tdwi.org)
Wikipedia (http://en.wikipedia.org/wiki/Data_quality)
Eobjects (www.eobjects.dk/sourgeforge.net/projects/datacleaner)
Talend (www.talend.com)
Mural (mural.dev.java.net)

Jos van Dongen

Jos van Dongen (jvdongen@tholis.com) is Senior Consultant bij Tholis Consulting.