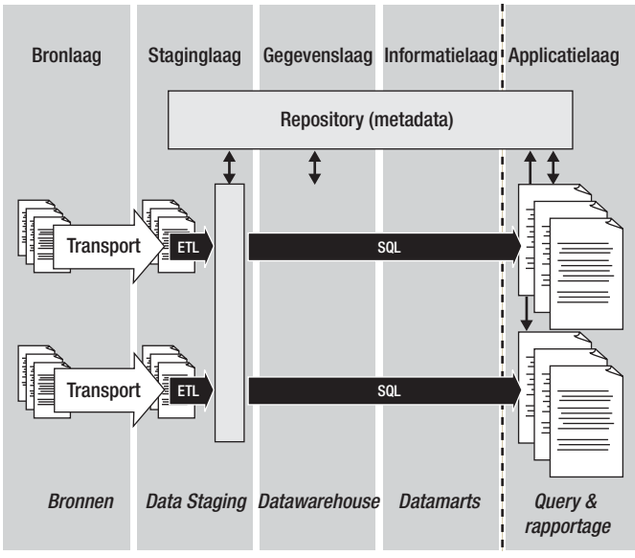
Bouw uw eigen toolbox

Doe-het-zelf Data Profiling

Martin Misseyer e.a.

In een datawarehouse-project is de verhouding tussen aandacht voor datakwantiteit en datakwaliteit tamelijk scheef. Dit wil zeggen dat er doorgaans veel meer aandacht is voor datakwantiteit dan voor datakwaliteit. Immers, datakwantiteit is direct zichtbaar en eenvoudig meetbaar; de verwerking van grote(re) hoeveelheden data krijgt bovendien vaak expliciet aandacht.

Datakwaliteit is veel minder zichtbaar en lastiger meetbaar. In veel gevallen wordt er niet direct naar de datakwaliteit gekeken in de zin van 'wat is de bron van' maar meer 'hoe krijgen we het er in'; de aandacht voor datakwaliteit is hiermee dus vaak impliciet. Hoewel er, vanzelfsprekend, in datawarehouse-literatuur wel over datakwaliteit wordt gesproken, veelal in termen als 'Data Profiling' en 'data cleansing', zouden datawarehouse professionals op de hoogte moeten zijn van de problematiek rondom data.
We hebben geconstateerd dat er in de praktijk nog steeds weinig expliciet tijd besteed wordt aan de kwaliteit van de geleverde data aan het datawarehouse.
Afgezien van het feit dat datakwaliteit direct is gerelateerd aan data management in een organisatie en, alszodanig, een meer centrale plaats verdient in de organisatie, is datakwaliteit een telkens terugkerend probleem in een datawarehouse-project. Met name inzicht, signalering en analyse, in de kwaliteit van de aangeleverde data bepaalt in sterke mate de vervolgacties (inzicht in datawaliteit is essentieel om zinvolle uitspraken te kunnen doen voor bijvoorbeeld de projectplanning, maar beperkt ook het aantal noodzakelijke iteraties tijdens het ontwerp en de realisatie) en tevens het succes van het op te leveren datawarehouse.
In relatie tot het totaalconcept Xtreme Data Warehousing (zie DB/M 7), is bij Ordina VisionWorks onderzocht hoe de bron- en data analyse-activiteiten vergemakkelijkt kunnen worden, door hiervoor een ondersteunend hulpmiddel in te zetten. Al snel ontstond een alternatief dat tussen de 'persoonlijke tool' en de 'markt tools' in zit. In dit artikel wordt het ontwerp en het gebruik van de 'Data Profiling toolbox' beschreven.



Afbeelding 1: Architectuur Data Profiling toolbox.

Verbazingwekkend

Wanneer een organisatie begint met een project dat de ontwikkeling van een datawarehouse behelst, weet men dat het flink geld kan gaan kosten. Natuurlijk kan men hierbij besparen door de juiste keuzen te maken wat betreft platform, hulpmiddelen en infrastructuur. Desalniettemin zullen de kosten van een datawarehouse-omgeving van enig omvang al gauw enkele tonnen in euro's bedragen, en soms zelfs een orde van grootte meer. Het is verbazingwekkend te noemen dat er doorgaans zeer veel geld gemoeid is met het aankunnen van datakwantiteit, dat wil zeggen dat er veel wordt uitgegeven in een datawarehouse-project om een grote hoeveelheid data in korte tijd te kunnen verwerken, maar nauwelijks geld wordt besteed aan het inzichtelijk maken, laat staan verbeteren, van de kwaliteit van de data. Datakwaliteit, uit te drukken als onvolledige/ontbrekende data, onscherpe data, foutieve data, enzovoort, kan vervelend tot en met gevaarlijk zijn. Vervelend omdat bijvoorbeeld de 'overig' of 'onbekend' categorie heel groot kan zijn, waardoor een rapport meer vragen oproept dan inzichten verschaft. Gevaarlijk omdat er pertinent onwaarheden c.q. tegenstrijdigheden worden gemeld, waarop gerapporteerd, gestuurd en dergelijke zou worden.

Data Profiling

Data Profiling is slechts een beperkte activiteit in het traject dat moet leiden tot het verbeteren en/of vasthouden van de datakwaliteit in een datawarehouse. Data Profiling houdt in het

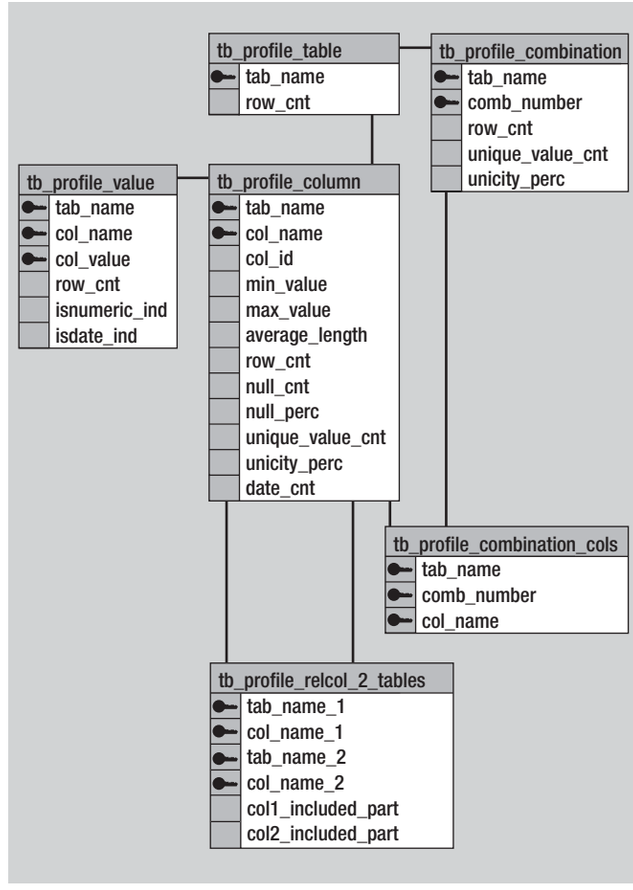
signaleren/analyseren van data om de kwaliteit ervan te bepalen, terwijl het feitelijke poetsen, 'data cleansing' genoemd, om de kwaliteit van de data te verhogen, pas kan worden gedaan als men weet wat er te poetsen valt (nog even afgezien van de locatie waar zou moeten worden gepoetst).

Om brondata te analyseren dienen er extracties uit bronnen te worden verkregen, die vervolgens kunnen worden onderzocht op kwaliteit. De signalerings- en analysefunctie van Data Profiling houdt in dat we data feitelijk onderzoeken op vier soorten ('niveaus') constraints, te weten: a. *kolom* (waarden), b. *rij* (relaties tussen kolommen), c. *tabel* en d. *inter-tabel*. Enerzijds dient te worden onderzocht of de door de bron aangegeven constraints stroken met de geleverde datasets, anderzijds dienen we vast te stellen of er constraints zijn af te leiden uit de geleverde datasets. Meer concreet houdt profiling op kolomniveau in: het verkrijgen van inzicht in inhoud van een individuele kolom (zoals bijvoorbeeld waardenverdeling, gebruikte datatype, null/not null-waarde, minimale, maximale, en gemiddelde lengte van de waarde, uniciteit van de kolomwaarden) in vergelijking tot dat wat vooraf opgegeven is door de bron. Profiling op rijniveau houdt in dat we inzicht willen krijgen in mogelijke sleutels (meerdere kolommen). Profiling op inter-tabelniveau houdt in dat we inzicht willen krijgen in betrouwbaarheid van integriteit tussen tabellen (datasets). Op kolom-, rij- en tabel niveau willen we inzicht in belangrijke statistieken (aantallen). Syntactisch willen we controleren of bijvoorbeeld datatypes kloppen en of de inhoud geldig is (31-02-2004 is geen geldige datum en postcode 100 AB bestaat niet). Tenslotte willen we ook semantisch een en ander controleren, zoals bijvoorbeeld een relationele vergelijking (bijvoorbeeld parent/child relaties, einddatum voor begindatum, enzovoorts), sleutel identificatie, en bijvoorbeeld een toetsing aan business rules (bijvoorbeeld: bedrag X kan 'business wise' nooit nul zijn). Naast het verzamelen van allerlei statistieken/eigenschappen zoals hiervoor genoemd, is het vaak ook wenselijk om inzicht te krijgen in bijvoorbeeld de mate van overlap tussen twee kolommen binnen een dataset, of bijvoorbeeld tussen twee kolommen van twee verschillende datasets. Uiteraard kan dit laatste worden doorgevoerd tot en met het vergelijken van (alternatieve) sleutels van verschillende datasets.

Alternatieven

Wanneer men aan Data Profiling wil doen, rekening houdend met het feit dat we gewenste query-, analyse- en rapportagefunctionaliteit willen hebben, dan zijn er de volgende alternatieven:

Persoonlijk Data Profiling hulpmiddel.
De professionals onderscheiden zich door met behulp van 'het betere hobbywerk' in bijvoorbeeld Excel gegevenssets te verzamelen in aparte werkbladen, en werkbladen te definiëren voor de hierboven genoemde drie niveaus van Data Profiling. Voordelen zijn voornamelijk te vinden in het gebruik van de eenvoudige infrastructuur, eenvoudige programmering, en dergelijke.



Afbeelding 2: Datamodel Data Profiling repository.

Nadelen liggen eveneens op het vlak van de infrastructuur, namelijk de geldende beperkingen (omvang datasets en programmering in Excel), de persoonlijke inspanning om het hulpmiddel te maken (en hiermee de inefficiëntie omdat 'iedereen het wiel uitvindt'), en de beperkte overdraagbaarheid.

Generiek hulpmiddel met specifieke Data Profiling-functionaliteit.
In de grotere datawarehouse-projecten wordt er al gauw gekozen voor serieuze tooling voor de inrichting van de verschillende architectuurcomponenten (data-opslag, dataverwerking en informatie-ontsluiting). Zo worden ETL-tools steeds meer uitgebreid met ondersteunende functionaliteit, zoals bijvoorbeeld die voor Data Profiling (zie bijvoorbeeld het artikel over UDS en Informatica in DB/M nummer 5 van dit jaar). Dit alternatief behelst tevens een kleine Data Profiling-specifieke repository (datamodel) waarin de eerder genoemde Data Profiling-informatie wordt weggeschreven. Hier bovenop zouden (dynamisch) met een hulpmiddel (Excel) rapportages kunnen worden gedefinieerd. Voordelen van dit alternatief zijn het (her)gebruik van de betreffende tooling (met name ETL en DBMS), immers meer gebruik van dezelfde tooling is kostenbesparend. Daarnaast zijn deze tools bedoeld voor performance en kunnen grote hoeveelheden data aan (het datakwantiteit deel). Nadelen van dit alternatief

zijn de (tot nu toe) beperkte Data Profiling-functionaliteit en de benodigde inspanning om de beoogde Data Profiling toolbox te realiseren.

Specifieke Data Profiling-hulpmiddelen.
Vanzelfsprekend zijn er diverse krachtige Data Profiling tools op de markt. De bovenkant van deze markt wordt gedomineerd door een handjevol leveranciers die prachtige en rijke producten leveren. Voordelen liggen met name op het gebied van functionaliteit, en integratie met andere producten van dezelfde en/of andere leveranciers. Nadelen liggen eveneens op het gebied van de functionaliteit – er zijn vaak zoveel analyses en parametriseringen mogelijk, dat het gebruik van deze tools veel tijd kan vergen, op zijn minst een aanzienlijke inleertijd. Natuurlijk zijn de kosten van deze 'toptools' een groot nadeel.
Elk van de bovenstaande alternatieven heeft zijn waarde in het kader van Data Profiling. Echter, de genoemde voor- en nadelen worden beschouwd, en gecombineerd met de behoefte aan een krachtig, compact en eenvoudig Data Profiling-hulpmiddel, dat ook onderdeel uitmaakt van een totaalconcept, dan luidt de conclusie dat er een vierde alternatief te beschrijven valt. De genoemde voor- en nadelen, gecombineerd met de behoefte aan een krachtig, compact en eenvoudig Data Profiling hulpmiddel, billijken een vierde alternatief.

De Data Profiling Toolbox

Om op een eenvoudige en snelle manier datasets te kunnen verwerken met behulp van de Data Profiling toolbox, is er een aantal uitgangspunten gedefinieerd waaraan dient te zijn voldaan. De belangrijkste uitgangspunten worden hierna beschreven.

Datasets (door een bron geleverde extractiebestanden):
- worden toegelicht met brondocumentatie (kolom/attribuut, dataset structuur, formattering en inhoud);
- worden alleen voor dit doel in tijdelijke profiling staging-tabellen geladen (onderscheiden met 'DWH_DP_' prefix);
- worden in een vast CSV-formaat ontvangen;
- bevatten alleen numerieke velden met een decimale PUNT geleverd.

Staging-componenten (staging-tabellen ten behoeve van Data Profiling-onderzoek):
- zijnde Data Profiling staging-tabellen worden alleen voorzien van een technische unieke sleutel (record-nummer);
- datum formaat voor extractie: YYYYMMDD;
- alle attributen in de doeltabellen worden gedefinieerd als varchar(4000);
- een NULL-waarde wordt omgezet in een geschreven string 'NULL';
- geen constraints aanleggen, na laden wel indexen aanmaken op alle kolommen.

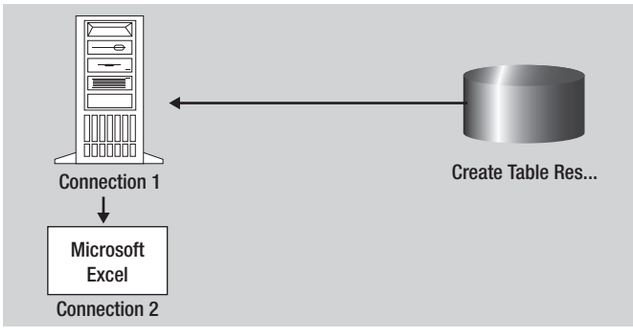
Procescomponenten (ETL-processen ten behoeve van Data Profiling-onderzoek):
- gebruik maken van standaard SQL;
- programmering in DBMS/Stored Procedures.

Rapportages (informatie over de geprofileerde datasets):
- met betrekking tot eigenschappen en statistieken worden gevoed vanuit de Data Profiling repository;
- met betrekking tot dataset-data worden rechtstreeks op de Data Profiling staging-tabellen gedraaid.

Afbeelding 1 toont de architectuur van de Data Profiling toolbox. De onderscheiden architectuurcomponenten zijn achtereenvolgens: a. een repository, b. de staging omgeving, c. Data Profiling scripts (ETL) en d. rapportages.
De globale werking van de Data Profiling toolbox is als volgt: de van bronsystemen ontvangen datasets (extractiebestanden) worden 1:1 in staging profiling-tabellen verwerkt, waarbij de

Nr.	Naam	Stap	Component	Beschrijving
Verplichte stappen				
1a	CREATE_PROFILING_TABLES	Definitie	Laden (technisch)	Definiëren (DDL) van data profiling staging-tabellen
1b	INSERT_INTO_STAGING	Vullen		Dataset toevoegen (DML INSERT) aan data profiling staging-tabellen
Verplichte stappen				
2a	CLEAN_NULLS	NULL-waarden	Transformeren (technisch)	Vertaal alle NULL-waarden in "NULL" string
2b	TRIM_TAB_COLUMNS	Spaties		Verwijder alle voorloop- en sluitspaties
2c	CREATE_INDEXES	Indexen		Creëer indexen op de attributen aan (performance)
Door analist te bepalen analyses				
3a	TAB_COL	Standaard	Analyseren (functioneel)	Bereken en verzamel eigenschappen en statistieken op waarde-, kolom- en datasetniveau
3b	COMBINATION	1 tabel		Voer een combinatie-onderzoek uit tussen attributen binnen een tabel
3c	RELCOL_2_TABLES	2 tabellen		Voer een combinatie-onderzoek uit tussen twee attributen van twee verschillende tabellen
Door beheerder te bepalen stappen (in overleg met de analist)				
4a	CLEAN_PROFILING	Verwijder data	Opruimen (technisch)	Verwijder data en metadata data profiling-resultaten
4b	DROP_PROFILING	Verwijder alles		Verwijder alle profiling-componenten

Afbeelding 3: Overzicht van de ETL-procescomponenten ten behoeve van Data Profiling.



Afbeelding 4: ETL-proces Excel-extractie.

specifieke Data Profiling scripts over de eerder genoemde drie niveaus informatie verzamelen en opslaan in de Data Profiling repository. Voor het uitvoeren van de Data Profiling scripts dient een ETL-hoofdproces te worden gedefinieerd.
Afbeelding 2 toont het eenvoudige datamodel van de Data Profiling repository. Dit datamodel wordt, evenals de Data Profiling scripts, gecreeerd door het DDL-script aan te roepen.
Uit het datamodel kan worden afgeleid dat er statistieken worden verzameld op dataset- ('tb_profile_table'), op kolom- ('tb_profile_column') en op inhoud- ('tb_profile_value') niveau. Tevens kan de uniciteit worden bepaald tussen twee kolommen binnen een dataset ('tb_profile_combination'/ 'tb_profile_combination_cols') en kan de mate van overlap van twee kolommen in verschillende datasets ('tb_profiler_relcol_2_tables') worden onderzocht.
Mocht het wenselijk zijn om profiling-resultaten historisch te bewaren, dan zal de primaire sleutel van het repository-model moeten worden uitgebreid met een datumattribuut. Tenslotte merken we op dat niet alle attributen van het Data Profiling repository-datamodel worden getoond.
Uitgaande van hergebruik van de (bestaande) datawarehouse-architectuur, is de dataverwerking gebaseerd op ETL-processen, die gebruik maken van verschillende procescomponenten. Deze procescomponenten zijn onder te verdelen in laad-, transformatie- en analyseprocessen. De tabel in afbeelding 3 geeft een overzicht van de verschillende procescomponenten.

De eerste categorie procescomponenten heeft te maken met het laden van de datasets in de Data Profiling staging-tabellen. Hiervoor wordt bijvoorbeeld een 'standaard' DTS-package gemaakt, een Transact-SQL (SQL Server) of PL/SQL (Oracle) procedure. Vanzelfsprekend dienen de tabellen eerst te worden gedefinieerd. Dit vindt plaats door een DDL-script met CREATE TABLE statements uit te voeren (het voert te ver om voor te schrijven dit via een datamodelleerhulpmiddel te doen, hoewel dit uit oogpunt van formele(re) documentatie kan).
De tweede categorie procescomponenten betreft de technische stappen die dienen te worden uitgevoerd om een bepaalde NULL-waardedefinitie te hanteren (2a), om spaties te verwijderen (2b), en om uit oogpunt van performance op alle gedefinieerde attributen een index te leggen (2c). Nota bene: activiteiten 2a en 2b behelzen het 'standaardiseren' van data. De derde categorie proces-componenten wordt gebruikt op basis van de Data Profiling-analysebehoefte. Concreet houdt dit in dat altijd de standaardanalyse wordt uitgevoerd, en dat afhankelijk van de situatie er combinatie-onderzoek gewenst is (1-tabel of 2-tabellen). De vierde categorie procescomponenten is bedoeld om de Data Profiling-omgeving te schonen (4a) dan wel om te verwijderen (4b).

De procedure voor het verwijderen van voorloop- en sluitspaties ziet er als volgt uit:

```
procedure TRIM_TAB_COLUMNS (p_tab_name IN VARCHAR2)
is
cursor   C1 is
select   c.column_name
from     ALL_TAB_COLUMNS c
where    upper(c.table_name) = upper(p_tab_name);

h_stmt       VARCHAR2(16000);
h_num_cols   INTEGER;
v_procedure  VARCHAR2(28) := 'TRIM_TAB_COLUMNS';
v_err_msg    VARCHAR2(256);

BEGIN
h_num_cols  := 0;
h_stmt      := 'update ' || p_tab_name || 'set';
For R1 in C1
Loop
h_num_cols  := h_num_cols + 1;
h_stmt      := h_stmt || R1.column_name || ' =
                  trim(' || R1.column_name || '),';
End loop;
IF h_num_cols > 0 THEN
h_stmt      := substr(h_stmt, 1, length(h_stmt)-1);
execute immediate h_stmt;
COMMIT;
END IF;

exception
when OTHERS
then
rollback;
v_err_msg := SUBSTR(SQLERRM, 1, 256);
dwh_pck_profiling.log_error(v_procedure, v_err_msg);
End TRIM_TAB_COLUMNS;
```

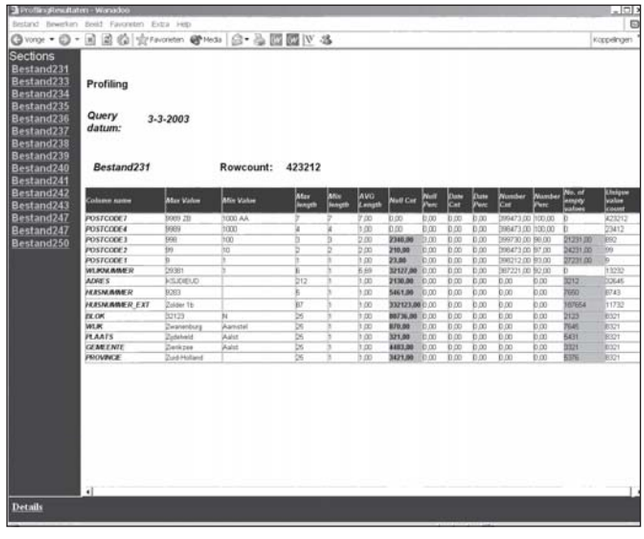
De analist die zich met Data Profiling bezighoudt zal zich met name bezig moeten houden met het bepalen van welke analyses moeten worden uitgevoerd. Daarnaast zal hij/zij een keuze moeten maken om handmatig data te profileren of om hiervoor een geïntegreerd 'overall' ETL-proces voor te definiëren.
Het eerste werkt gemakkelijk wanneer het niet te veel data(sets) betreft en er dus interactief kan worden gewerkt. Is de hoeveel-

heid data(sets) groot, dan is het verstandig om te streven naar een 'overall' ETL-proces, eventueel per bron.

Rapportage

Voor de presentatie van de Data Profiling-resultaten kan gebruik worden gemaakt van de (beoogde) BI-tool, en als deze (nog) niet aanwezig is, van een query-tool of zelfs Excel. Wanneer er onderliggend ook SQL Server wordt gebruikt en DTS voorhanden is, dan is het tamelijk simpel om eenvoudige 'rapportages' te genereren in Excel. Afbeelding 4 geeft een voorbeeld van een basaal DTS-proces aan. Het te genereren Excel-werkblad kan meer specifiek worden gemaakt door de onderliggende SQL-query te parametriseren (restricties op te nemen). Doet men dit niet dan is het alsnog via de Excel autofilter-functie mogelijk om te drillen.
Vanzelfsprekend kunnen rapportages ook fraaier gepresenteerd worden. Zo is het handig om gebruik te maken van de Data Profiling repository. Wanneer er een BI-tool wordt gebruikt is het mogelijk om snel inzicht in de Data Profiling-resultaten te krijgen. Afbeelding 5 geeft een voorbeeld gemaakt met BusinessObjects. In dit geval is er een universe gedefinieerd op de Data Profiling repository. Deze universe kan worden uitgebreid met de staging-tabellen, zodat er via de BI-tool direct in de staging-tabellen kan worden gekeken naar de individuele data. Tenslotte kan de lay-out van de rapportages verder worden verfraaid.

De besproken Data Profiling toolbox is, zoals inmiddels wel duidelijk is geworden, beschikbaar in de smaken SQL Server en Oracle, en is gebruikt bij verschillende klanten, zowel in project-vorm als op individuele basis. In de praktijk is ervaring opgedaan met het op verschillende manieren implementeren van proces-componenten (DTS, Transact-SQL en PL/SQL), hetzelfde geldt voor de modellering van staging-tabellen (modelleer-tool versus DDL-script). Uiteraard wordt de keuze uit deze alternatieven bepaald door de situatie en beïnvloed door persoonlijke smaak en/



Afbeelding 5: Voorbeeld van profiling-rapportage via een BI-tool.

of werkwijze. Wanneer eenmaal data(sets) zijn verkregen, spreekt het voor zich dat wanneer binnen één of enkele uren de eerste Data Profiling-resultaten toonbaar zijn, het nut van de Data Profiling toolbox is aangetoond. Daarnaast is het een compliment om te constateren dat ook na je vertrek de Data Profiling toolbox blijft worden gebruikt.

Conclusie

Data Profiling zal, indien gedegen en volledig uitgevoerd, er toe leiden dat:

- 'bron-pijnpunten' worden blootgelegd. Deze zouden dienen te worden opgelost in de bron zelf. Hoewel dit waarschijnlijk meer oplostijd zal vergen, verdient dit zich dubbel en dwars terug tijdens rapportage-validaties later in het datawarehouse-project;
- de bewustwording in de organisatie van datakwaliteit als een 'ongoing' proces wellicht ook voor overige, buiten de scope van het datawarehouse-traject gelegen bronnen aangewend kan worden;
- Data Profiling-acties (indirect) ook zullen leiden tot een hervalidatie van betreffende onderliggende business rules (onder meer correctheid, geldigheid, actualiteit).

De conclusie is dat met een zeer beperkte inspanning een eenvoudig doch krachtig product is te fabriceren dat het gat opvult tussen enerzijds hobbywerk en anderzijds structurele, maar kostbare oplossingen op het gebied van Data Profiling. Dit gat moet worden opgevuld omdat we hebben geconstateerd dat er een sterke behoefte bestaat om snel en effectief onderzoek te kunnen doen naar de kwaliteit van door bronsystemen aan het (in oprichting c.q. in onderhoud zijnde) datawarehouse geleverde data.
Het spreekt voor zich dat wanneer hiertoe de noodzaak bestaat om op een meer tactisch/strategisch niveau om te gaan met Data Profiling, de besproken Data Profiling toolbox niet volstaat, en dat de keuze voor een specifieke Data Profiling tool een logische is. Dezelfde redenering kan worden gehanteerd wanneer er in (grote) datawarehouse-projecten veel en intensief zal moeten worden geprofileerd, en wellicht data moeten worden geschoond ('gecleansed').
Dan is het zeker aan te bevelen om al direct een keuze te maken voor een geïntegreerde oplossing waar Data Profiling onderdeel van uitmaakt.

Martin Misseyer, René Nobel, Ben Meester en **Edwin Weber**
Dr. Martin P. Misseyer (martin.misseyer@ordina.nl) is Profession Leader bij Ordina VisionWorks.
René Nobel, Ben Meester en Edwin Weber zijn consultants bij Ordina VisionWorks.