Oracle Warehouse Builder (2)

Analyse van een bèta versie

Oracle Warehouse Builder is Oracle's product voor ETL – Extract, Transform en Load van data, uit diverse bronnen via een transformatie naar een doel-database, meestal een datawarehouse. Oracle Warehouse Builder 10g Release 2 – ook wel aangeduid als de Paris Release – speelde de hoofdrol in een artikel in Optimize nummer 3. In dit artikel kijken we met name naar de Data Profiler in OWB, een volledig nieuw stuk functionaliteit.

In het vorige artikel hebben we gezien hoe je in OWB diverse databronnen – flat file, Oracle-database: tabel, view, queue, PL/SQL API, third party-database of ERP-applicatie – kunt registreren. Ook de doelen voor ETL kunnen worden geregistreerd of ontworpen met en gegenereerd vanuit OWB. Warehouse Builder bevat visuele ontwerp-tools voor het ontwerp van tabellen en views maar ook een Star- of Snowflake-schema voor een datawarehouse, gevormd door Dimensions, Measures en Hierarchy's. Bronnen en doelen worden in zogenaamde Mappings op visuele wijze met elkaar verbonden. Binnen een Mapping kunnen allerlei operatoren en transformaties worden opgenomen, van Filter, Deduplicate en Join tot Match Name and Address, Pivot en Custom (PL/SQL-gebaseerde) Operators. OWB kan vervolgens de voor Oracle 9iR2 of 10g geoptimaliseerde PL/SQL-code genereren om de transformatie uit te voeren. Deze code kan de ETL-processen Set-based of Record-based uitvoeren. Alle nieuwe database-features, van Bulk Collect to Merge, worden toegepast in deze code. Ook is de code geïntegreerd in het run-time framework van OWB voor auditing en logging, foutafhandeling en scheduling. Tenslotte ging het artikel kort in op de mogelijkheden om met OWB op basis van de data-targets complete Business Intelligence-modules te ontwerpen en genereren – een ingerichte Oracle Discoverer End User Layer of een BI Beans-applicatie.
Hoewel dit eerste artikel in belangrijke mate het ontwerp en de generatie van het datawarehouse beschreef, bleek hier al dat Warehouse Builder een vlag is die de lading lang niet dekt: OWB kan veel meer dan alleen een datawarehouse construeren. Gewoon relationeel database design, ontwerp en generatie van ETL processen en van BI-applicaties zijn aanpalende activiteiten, maar gaan toch al een stukje verder. In dit artikel kijken we met name naar de Data Profiler in OWB, een volledig nieuw stuk functionaliteit en wel heel ver af van Warehouse Building.

Nieuwe ontwikkelingen

Voor we in de Data Profiler duiken, noem ik kort een aantal recente ontwikkelingen rondom de Oracle-producten en features voor business intelligence en datawarehousing. Allereerst OWB 10gR2 zelf: in het vorige artikel gaf ik – op voorspraak van Jean-Pierre Dijcks, Oracle Product Manager voor OWB – augustus 2005 als release datum. Helaas, maar de 10gR2 Paris-release is inmiddels al verschillende keren uitgesteld. Op dit moment is de verwachting dat we de productierelease – de bèta-test is inmiddels al wel aan zijn derde release toe – in de laatste maanden van 2005 kunnen verwachten, met heel misschien kort na Oracle Open World een Developer Preview release op OTN.
Bijna iedere organisatie die Oracle-producten gebruikt zal kunnen beschikken over OWB. Hetzij via licenties op de 10gDS Developer Suite – met ondermeer ook Oracle Designer, Oracle Forms, Discoverer en JDeveloper-, hetzij via een Oracle 9iR2 of 10g Enterprise Edition van de database. Bij beide is OWB inbegrepen. De marktpositie van Oracle Warehouse Builder wordt door Gartner zeer positief ingeschat. In het Magic Quadrant voor ETL producten, gedateerd 11 mei 2005, bevindt OWB zich in het Leaders-kwadrant, met alleen Ascential Software (IBM) en Informatica voor zich. Op geruime afstand vormen Business Objects, SAS en Microsoft het achtervolgende peloton. Deze inschatting van Gartner is overigens nog gebaseerd op de huidige 10gR1 release van Warehouse Builder.
Het is interessant om te zien in ondermeer de 10gR2 release van de database en ook in de aankondigingen op Oracle Open World 2005 – volop aan de gang op het moment dat ik dit artikel schrijf – dat Oracle zeer actief is op het gebied van busi-

Advertentie

ness intelligence en datawarehousing. Dit jaar hebben we al de Discoverer 10g Drake release gezien met ondermeer de geleidelijke afstoting van Discover Desktop (Client/Server), de toenemende benutting van standaard database functionaliteit als Materialized Views en Analytische Functies, de losmaking van Oracle Portal en de functionele uitbreidingen op basis van het BI Beans framework dat ook binnen Oracle Reports, JDeveloper, OLAP Excel Plugin en Enterprise Planning & Budgetting wordt gebruikt. Daarnaast Discoverer OLAP dat de OLAP Kubussen en Analytical Workspaces in de database toegankelijk maakt voor eindgebruikers.

Data mining

Binnen de E-Business Suite (release 11.10.5) is een nieuw mechanisme ontwikkeld voor de rapportage die tot nu toe volledig op basis van Oracle Reports werd gedaan. Dit mechanisme is XMLPublisher gedoopt. Tijdens Open World krijgt dit framework voor flexibeler rapportages vrij veel aandacht. Het ligt voor de hand dat het op afzienbare termijn ook los van Oracle Applications beschikbaar komt, wellicht als complementair product voor Oracle Reports. Een recente aankondiging daarover suggereert eind 2005 als release moment.
De belangrijkste uitbreiding overigens van Oracle Reports 10gR2 lijkt te zijn het publiceren van rapporten in Excel formaat. Overigens was er tijdens Oracle Open World ook een brainstorm-achtige presentatie over Report Center – meer een concept dan een product lijkt het. Report Center is een dashboard-achtige interface voor eindgebruikers om zelf Discoverer rapporten samen te stellen.
De meeste focus vanuit Oracle binnen het domein van Business Intelligence lijkt te liggen op data mining. Sinds in Oracle RDBMS 9iR2 het vroegere Oracle Darwin, een standalone tool voor datamining, werd opgenomen in de kern van de database is in kleine stappen een verdere integratie tussen de datamining-engine en andere componenten tot stand gebracht. In 10gR1 bijvoorbeeld via de PL/SQL Package DBMS_FREQUENT_ITEMSET en DBMS_DATAMINING die een PL/SQL interface voor ODM (Oracle Data Mining) boden. Data Mining heeft ook een Java API. Deze is in 10gR2 volledig herzien, teneinde de internationale standaard JSR-173 te implementeren. Daarnaast is er Oracle Data Miner, een stand-alone GUI waarin modellen kunnen worden ontwikkeld en getest. In 10gR2 zijn nieuw algoritmes beschikbaar binnen Data Mining: Decision Tree en Anomaly Detection. Ook is in 10gR2 een aantal Data Mining operatoren toegevoegd aan SQL, zoals PREDICTION, PREDICTION_PROBABILITY en PREDICTION_DETAILS. Dat betekent dat in gewone SQL statements gebruik gemaakt kan worden van resultaten van Data Mining modellen. Bijvoorbeeld: selecteer alle klanten die volgens onze Campagne model een bepaalde aankoop zouden gaan doen (data mining model) en join met de actuele ordergegevens (gewoon relationeel). De term die Oracle veelal gaat hanteren voor het toepassen van Data Mining is 'Predictive Analysis'. Data Mining wordt vaak ingezet om op basis van historische gegevens patronen te ontdekken die worden gebruikt om toekomstige gebeurtenissen te voorspellen. Oracle Spreadsheet Add-in for Predictive Analytics is de pakkende naam van een met 10gR2 nieuw gelanceerde plugin voor Excel voor het doen van 'voorspellende analyse op basis van Data Mining modellen'. Tenslotte prijst Oracle het nieuwe 10gR2 supplied package DBMS_PREDICTIVE_ANALYTICS aan als de toegang tot Data Mining voor een zeer breed publiek.

Database

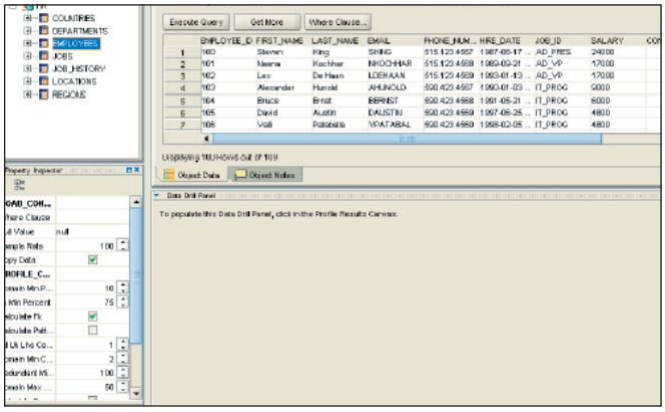
Tenslotte kort nog even over de Oracle 10gR2 database en enkele verbeteringen daarin die relevant zijn met het oog op business intelligence en datawarehouses. De functionaliteit van de MODEL-clause in SQL statements is uitgebreid: UPSERT kan eenvoudiger toegepast worden en ook kunnen we gebruik maken van Analytical Expressions in de Model Clause. De database gebruikt nieuwe algoritmes voor het uitvoeren van SORT en AGGREGATIE operaties. Deze leiden tot flinke performance-verbeteringen, bijvoorbeeld tot een factor 5 voor ORDER BY clauses in query's die veel records opleveren of bij het aanmaken van Indexen. Ook aggregaties zoals SUM en AVG kunnen aanzienlijk sneller wordt uitgevoerd, tot twee á drie keer zo snel.

In 10gR2 kan een Query Rewrite gebruikmaken van meer dan één Materialized View – opnieuw potentieel een performance verbeteraar en in elk geval leidend tot een vereenvoudiging van het ontwerp van Materialized Views. Change Data Capture – een op Oracle Streams gebaseerd mechanisme om asynchroon wijzigingen tussen databases te communiceren, bijvoorbeeld een operationele database die near-realtime een Data Warehouse bijwerkt – is flexibeler geworden. CDC kan tussen verschillende database-versies op verschillende O/S platformen worden gebruikt. Het beheer van Partitioned Tables is sterk vereenvoudigd. Bij gebruik van Dimension Hierarchies kan met de clausule SKIP WHEN NULL worden aangegeven dat een bepaalde DIMENSION op sommige niveaus een NULL-'waarde' kan bevatten, zonder dat dat de hiërarchie breekt. Hiermee kunnen flexibeler hiërarchieën worden gedefinieerd en kunnen eenvoudiger aggregaties op verschillende niveaus van de hiërarchie worden uitgevoerd.
Een zeer interessante optie die in 10gR2 is toegevoegd heeft betrekking op INSERT, UPDATE, MERGE en DELETE-operaties. Zeker de eerste drie spelen een grote rol bij het laden van data in het datawarehouse. Vaak is de laadoperatie een zware klus die grote hoeveelheden data omvat. Het kan dan ook erg zuur zijn als na succesvolle verwerking van honderdduizenden of miljoenen records er een paar records zijn die een cons-

traint overtreden waardoor het gehele statement faalt en wordt teruggedraaid. De nieuwe DML Error Logging functionaliteit biedt de mogelijkheid om bij een DML statement aan te geven dat in geval van fouten bij een record het statement als geheel moet worden doorgezet. Het record dat de fout oplevert wordt in een aparte tabel gelogd, onder vermelding van de Error Code. Fouten die op deze manier kunnen worden afgehandeld voor individuele records zijn ondermeer Mismatch met Kolom-definitie (waarde te groot, verkeerde data type), Constraint overtreding (verplichte waarde ontbreekt, waarde voldoet niet aan Check Constraint, waarde is niet uniek of bevat geen geldige referentie naar een master) of exception in een row level trigger. Dankzij DML Error Logging kunnen Triggers en Constraints gewoon enabled blijven tijdens ETL-operaties, zonder dat het risico bestaat dat grote batches in het zicht van de haven stranden!

Data Profiler

Veel van de functionaliteit in Oracle Warehouse Builder 10gR2 bestond in minder uitgebreide, minder geavanceerde vorm ook al in vroegere versies van het product. Dat geldt niet voor de Data Profiler. Deze component is nieuw in deze Paris release. Data Profiler is de eerste stap op weg naar betere datakwaliteit, een van de thema's voor OWB 10gR2.
Stel je het volgende scenario voor: je krijgt van je manager een database voorgezet. Er is geen data model, geen technische documentatie en bij een eerste inspectie van de Data Dictionary blijken de tabel- en kolomnamen onbegrijpelijk, zijn er geen declaratieve constraints – geen primary keys, unique keys laat staan foreign keys – en zijn alle kolommen van het type VARCHAR2(4000). De opdracht: migreer de data uit deze database naar jullie eigen applicatietabellen. Toegegeven, een wat vergezocht scenario. Maar onderdelen van die situatie zijn niet zo uitzonderlijk: geen of incompleet of achterhaald datamodel, ontbrekende constraints, onduidelijk kolomdefinities, mogelijk redundante, gedenormaliseerde kolommen of additionele bedrijfsregels waar bestaande records zich niet aan houden. In al dat soort situaties kan de Data Profiler uitkomst bieden.
Data Profiling is een twee-stap proces. In de eerste stap wordt een batch proces in gang gezet. Dit proces scant een sets database-tabellen, neemt steekproeven van de data, gebruikt statistische analyses en datamining-algoritmes om kennis op te bouwen van de onderzochte tabellen en gegevens. De resultaten van dit profiling-proces worden vastgelegd in de OWB-repository. Nota bene: deze eerste noodzakelijke stap voor het doen van Data Profiling kan zeer langdurig zijn. Voor tabellen met tienduizenden of meer records is in elk geval sampling – steekproefgewijs onderzoeken – aan te bevelen, maar zelfs dan kan profiling uren tot dagen duren. Gezien het feit dat dit een strikt eenmalige operatie is, hoeft dat verder geen problemen op te leveren. Echter, voor de kennismaking met Data Profiling raad



Afbeelding 1. Tabblad Profile Object

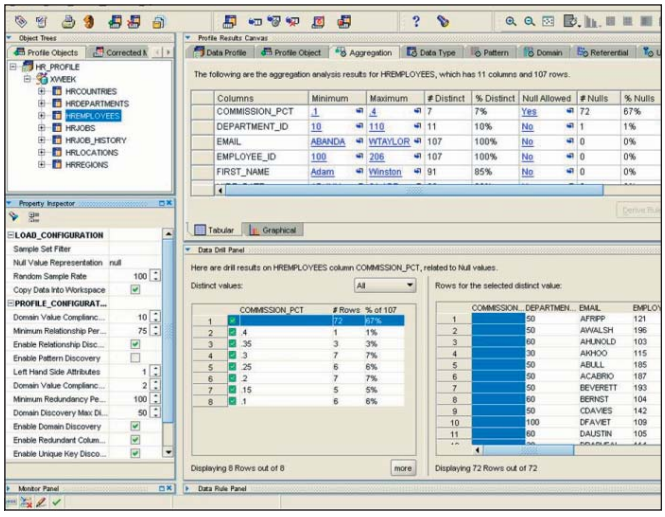
ik aan een overzichtelijke tabel- en gegevens-collectie te laten profilen. In dit artikel kijken we naar de resultaten van dataprofiling van de Employees-tabel (met kleine aanpassingen) in het HR demo-schema van de Oracle database.

Categorieën

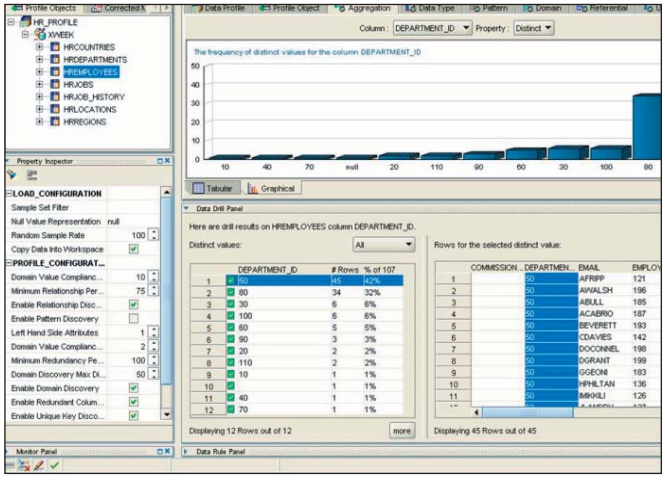
De tweede stap van Data Profiling is de interactieve verwerking van de Data Profiling resultaten. Deze resultaten zijn in de volgende categorieën te verdelen:

Profile Object
Dit tabblad geeft een overzicht van de data-set waarover het profile is bepaald. Dit kunnen alle data uit de tabellen zijn, maar ook een steekproef. Linksonder in de Property Inspector zijn de Profile Configuration Settings zichtbaar. Deze geven aan naar welke soort informatie tijdens data-profiling is gezocht, hoe groot de steekproef was of onder welke omstandigheden een waarde als Domain-waarde wordt herkend.

Aggregation
Op dit tabblad worden per kolom statistische aggregatiegegevens getoond, afgeleid uit alle records in de tabel. Ondermeer de minimum en maximum kolomwaarde, de mediaan, het aantal verschillende waarden en het percentage verschillende waarden, het aantal NULL's en de zogenaamde Six-Sigma waarde die de kwaliteit van de data in de kolom beschrijft. Per kolom kan een histogram worden opgevraagd dat grafisch de voorkomens van verschillende waarden in de kolom toont. Per waarde kunnen de rijen uit de tabel worden opgevraagd die een dergelijke waarde bevatten. Deze meeste van deze resultaten kan je trouwens ook in PL/SQL verkrijgen met het supplied package dbms_stat_funcs, de summary procedure.
Uit de figuur kunnen we afleiden dat er 91 verschillende waarden voor FIRST_NAME voorkomen en dat dit betekent dat 85% van de FIRST_NAME waarden uniek is. Voor een kolom die niet zo'n duidelijke betekenis heeft als FIRST_NAME zou zo'n percentage aanleiding kunnen zijn te onderzoeken of er



Afbeelding 2. Op het tabblad 'Aggregation' worden per kolom statistische aggregatie gegevens getoond.

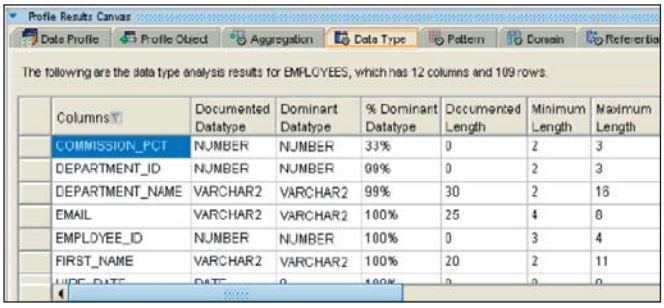


Afbeelding 3. Per kolom kan een histogram worden opgevraagd.

sprake is van een gemankeerde Uniqueness constraint. De kolom Null Allowed geeft de suggestie voor een Custom Data Rule. Voor kolom DEPARTMENT_ID stelt de Data Profiler voor om een Custom Data Rule 'Nulls not allowed' te introduceren aangezien bij slechts 1% van de records de waarde ontbreekt.

Data Type

Dit tabblad geeft informatie over de data types van de waarden in onze kolom. Het kan best zijn dat kolommen als VARCHAR2 zijn gedefinieerd maar in de praktijk eigenlijk vrijwel alleen maar numerieke of datum-waarden bevatten. Als dat het geval is geeft dat aanleiding om het database-ontwerp te heroverwegen. Hoe beter de Data Types gedefinieerd zijn, hoe hoger de kwaliteit van de data en ook hoe beter de Query Performance. Dit tabblad geeft ook informatie over de daadwerkelijke lengte van de waarden in de kolom, afgezet tegen de gedefinieerde



Columns	Documented Datatype	Dominant Datatype	% Dominant Datatype	Documented Length	Minimum Length	Maximum Length
COMMISSION_PCT	NUMBER	NUMBER	33%	0	2	3
DEPARTMENT_ID	NUMBER	NUMBER	99%	0	2	3
DEPARTMENT_NAME	VARCHAR2	VARCHAR2	99%	30	2	16
EMAIL	VARCHAR2	VARCHAR2	100%	25	4	8
EMPLOYEE_ID	NUMBER	NUMBER	100%	0	3	4
FIRST_NAME	VARCHAR2	VARCHAR2	100%	20	2	11

Afbeelding 4. Dit tabblad geeft informatie over de data types van de waarden in onze kolom.

kolombreedte. Ook hiervoor geldt: je kunt beter de kolom zo goed mogelijk op maat ontwerpen dan hem voor het gemak maar zo breed mogelijk te maken.
In dit voorbeeld zien we dat DEPARTMENT_NAME met een breedte van 30 is gedefinieerd maar in de praktijk niet langer is dan 16 karakters. We zien hier geen VARCHAR2 kolom die hoofdzakelijk NUMBER waarden bevat. Licht onhandig is wellicht het feit dat het percentage voor Dominant Datatype ook NULL's meerekent. Het Dominant Datatype voor de kolom COMMISSION_PCT is NUMBER, maar dat komt toch maar in 33% van de gevallen voor. Echter, in de overige 67% van de gevallen bevat de kolom geen waarde: 100% van de feitelijke waarden voor COMMISSION_PCT is NUMBER.

Het Data Type-tabblad geeft ook weer een aantal six-sigma kwaliteitsindicatoren, bijvoorbeeld het percentage rijen waarin de waarden afwijken van de gespecificeerde Kolom Breedte, Scale of Precisie.

Pattern

Het Pattern Tabblad geeft per kolom aan of er een patroon – Regular Expression – ontdekt is dat het formaat voor de meerderheid van de rijen in een kolom beschrijft. Voorbeelden van Patterns zijn bijvoorbeeld 9999AA voor postcode of formaten voor Datum, Tijd, Email, Url of Sofi-nummer. Per patroon geeft het tabblad weer welk percentage van de rijen er met zijn waarde voor de kolom aan voldeed. De gebruiker kan het gevonden patroon bevestigen of eventueel verfijnen. Dit kan later bij het corrigeren van data via Custom Data Rules worden toegepast. Het opsporen van de patronen kost veel tijd tijdens de eerste stap van Profiling en kan desgewenst overgeslagen worden.

Domain

Op het Domain Tabblad toont de Data Profiler of er voor bepaalde kolommen Domeinen van toegestane waarden van toepassing lijken. Je kunt een drempelwaarde instellen, die het minimale percentage van de rijen aangeeft dat een bepaalde waarde moet bevatten alvorens die waarde als domeinwaarde

De Caesar Groep is een open en eerlijke ICT-onderneming die oplossingen biedt met aantoonbaar rendement voor de klant. Wij nemen daarbij de doelstellingen van de klant als uitgangspunt en geven garantie op het behalen van rendement. Caesar beschikt over uitgebreide technologie-expertise (.NET, Java, Oracle, Microsoft, Progress en Infrastructuur) en kennis van de markten waarin wij ons begeven. Onze diensten omvatten advies, projecten, detachering, implementatie en beheer. Caesar heeft ruim 300 medewerkers en is gevestigd in Utrecht; ons bedrijf is door Management Team verkozen tot beste IT-adviseur van 2003.

BEN JIJ TOE AAN VERDERE PROFESSIONALISERING EN RESULTAATGERICHT, DAN ZIJN WIJ OP ZOEK NAAR JOU!

RESULTAATGERICHTE ICT-ERS DIE WERKEN AAN RENDEMENT

ORACLE-SPECIALISTEN MET ÉÉN OF MEER VAN DE VOLGENDE EXPERTISES:

DBA-SPECIALIST
Je hebt kennis van de database en de applicatieserver (9ias), met betrekking tot installatie, inrichting, back-up & recovery, performance en security.

JAVA/J2EE ARCHITECT
Je hebt aantoonbare kennis van Java/J2EE. Verder ben je bekend met JDeveloper, ADF, JSP, Struts, servlets, EJB, RUP methodiek én OO analyse en design (UML).

BI-SPECIALIST
Je hebt kennis van Oracle Warehouse Builder en Discoverer (en Oracle Reports als pre). Je hebt ervaring met ontwerpen en realiseren van ETL processen.

DESIGNER/DEVELOPER SPECIALIST
Je hebt aantoonbare ervaring met minimaal één van de volgende zaken: Designer 6i, 9i of 10g en/of Developer (forms en reports) 6i, 9i, of 10g.

Je hebt 1 tot 4 jaar aantoonbare ervaring in je vakgebied, hebt een opleiding op HBO-niveau, bent ondernemingsgericht en werkt graag mee in projecten. Je salaris is afhankelijk van je ervaring en kennisniveau. Ons flexibel en individueel in te vullen arbeidsvoorwaardenpakket is zonder meer uitstekend te noemen. Bovendien besteden wij veel aandacht aan opleidingen.

INTERESSE?
Mail je reactie met CV en motivatie naar Karin van Oostrom, k.oostrom@caesar.nl

www.caesar.nl

CAESAR GROEP

ICT OPTIMA FORMA

Sogeti is één van de top-5 ICT-bedrijven van Nederland. Wij staan garant voor interessant werk, met Oracle-opdrachten bij interessante klanten. We profileren ons daarbij met kwaliteit en vakmanschap. Klanten zien dit terug in onze dienstverlening en in de door ons ontwikkelde en tot standaard uitgegroeide methoden zoals DYA®, Regatta®, TMap® en TPI®.

Innovatieve engineering met Oracle
Binnen de divisie Distributed Software Engineering (DSE) werken engineers die gebruik maken van 'leading technology', waaronder Oracle. Onze Oracle-community bestaat uit ruim 70 enthousiaste professionals die resultaatgericht werken en met heel hun hart voor het ontwikkelen van systemen gaan. We maken klantgerichte oplossingen, waarbij we werken met onder andere J2EE, Oracle 8i, 9i, 9iAS, 10g, Oracle DBA-kennis, Oracle Warehousebuilder en Discoverer, JDeveloper en JHeadstart. Ook mobiele oplossingen op basis van het Oracle-platform ontbreken niet.

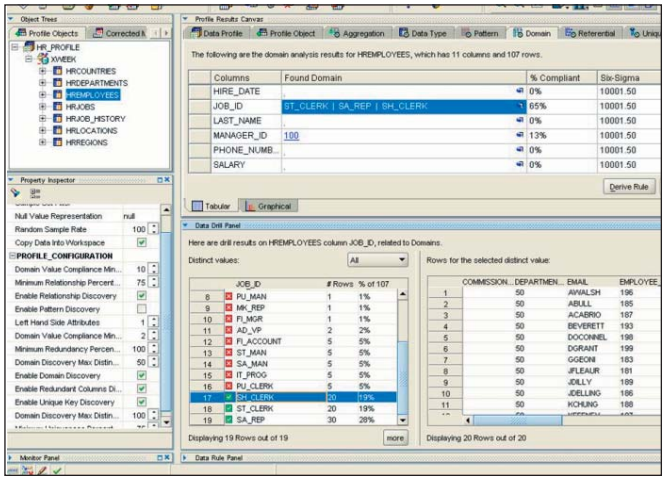
Gezien de sterke stijging van het aantal opdrachten op het Oracle-vakgebied zoeken wij:
- Ervaren applicatieontwikkelaars
- Ervaren DBA's
- Young Professionals

We zoeken voor deze functies vakmensen met een HBO-werk- en denkniveau en aantoonbare ervaring die klinkende resultaten kunnen behalen door onze klanten te ondersteunen en te adviseren omtrent het gebruik en de mogelijkheden van Oracle. Wij bieden je goede arbeidsvoorwaarden en uitstekende opleidings- en certificerings-mogelijkheden.

Ben je geïnteresseerd? Stuur dan je cv met motivatie naar dse-hrm@sogeti.nl. Heb je nog vragen, dan kun je bellen met onze HR manager Merie Sigmond op 020 660 66 00 + nakiesnummer 7614.

Sogeti Nederland B.V. Postbus 76, 4130 EB Vianen
Tel (020) 660 66 00 Fax (020) 660 67 21 **www.sogeti.nl**

SOGETI
Realisme in ICT



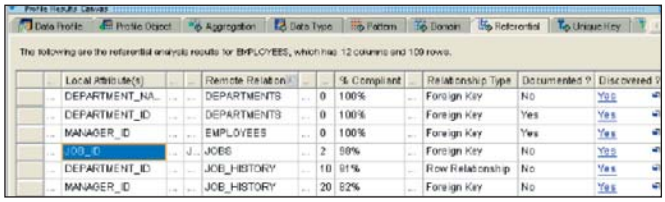
Afbeelding 5. Data Profiler toont of er voor bepaalde kolommen Domeinen van toegestane waarden van toepassing lijken.

wordt beschouwd. In het voorbeeld van kolom JOB_ID in onderstaande figuur zien we dat er 19 verschillende waarden zijn aangetroffen waarvan er drie in meer dan 10% van de rijen voorkomen. Als we de drempel verlagen tot 5% vinden we een domein met acht waarden waarmee 88% van de rijen is afgedekt. In dit tabblad kunnen we stap voor stap het domein samenstellen. Het uiteindelijke domein kan worden vastgelegd als Custom Data Rule. Die rule kunnen we toepassen bij Data Cleansing of Correctie en bij het Controleren van nieuw ontvangen data.

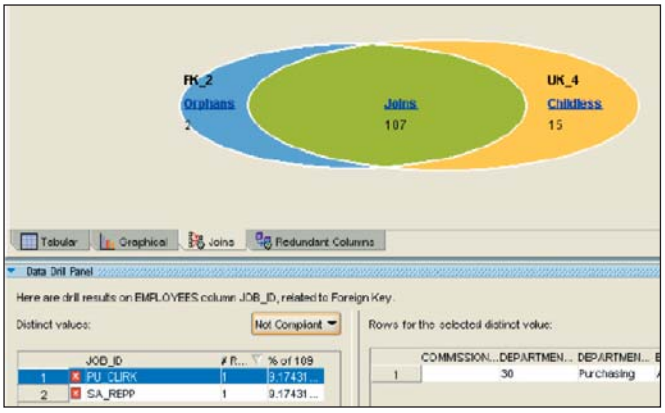
Referential

Het tabblad Referential komt de Data Profiler met een analyse van mogelijke Referentiële relaties tussen tabellen. Let wel: hiervoor worden niet de Foreign Key definities uit de Data Dictionary gebruikt. De resultaten op dit tabblad zijn afgeleid uit statistische analyse van de gesamplede data. Met de uitkomsten kunnen we wel kandidaat Foreign Keys benoemen. Afbeelding 6 laat zien dat er wat Data Profiler betreft een relatie lijkt te zijn tussen de JOB_ID kolom in tabel EMPLOYEES en de gelijknamige kolom in tabel JOBS. De relatie blijkt te gelden voor 98% van de records in tabel Employees.

Kijken we naar een grafische representatie (afbeelding 7), dan zien we al snel dat er twee Orphan-records zijn: records die



Afbeelding 6. Er lijkt een relatie te zijn tussen de JOB_ID kolom in tabel EMPLOYEES en de gelijknamige kolom in tabel JOBS.

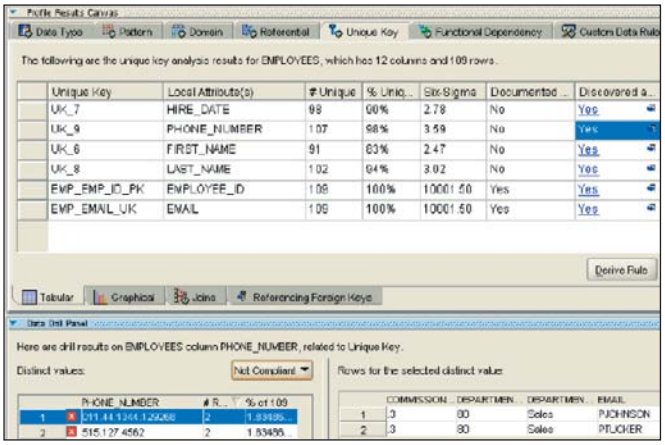


Afbeelding 7. In de grafische representatie zien we dat er twee Orphan-records zijn.

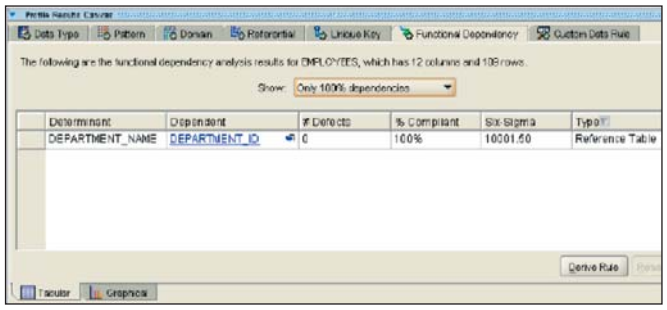
geen master hebben volgens deze kandidaat Foreign Key en dat er 15 Master-Records zijn waarvoor er geen kind bestaat – een toegestane Job-waarde die nog niet gebruikt wordt. Als we van mening zijn dat een door Data Profiler gevonden relatie daadwerkelijk een echte Foreign Key beschrijft, kunnen we dat vastleggen door middel van een Custom Data Rule.

Unique Key

Kolommen kunnen met een Unique Constraint zijn gedefinieerd als uniek, iedere waarde is verschillend. Data Profiler gaat op zoek naar kolommen waarvoor dat niet is vastgelegd met een Unique Key constraint, maar waarvoor dat wel lijkt te gelden. Kolommen waarvoor een groot percentage van de rijen daadwerkelijk een unieke waarde heeft worden gepresenteerd als kandidaten voor een echte regel die dat afdwingt. In afbeelding 8 blijkt dat PHONE_NUMBER maar liefst 98% verschillende waarden bevat. Dit zou erop kunnen duiden dat uit de bedrijfslogica volgt dat deze kolom als uniek kan of moet worden gedefinieerd.



Afbeelding 8. PHONE_NUMBER bevat maar liefst 98% verschillende waarden.

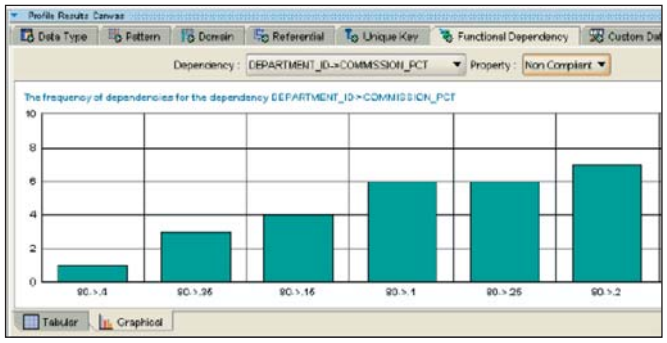


Afbeelding 9. De kolom DEPARTMENT_NAME in tabel EMPLOYEES volgt in alle gevallen uit DEPARTMENT_ID.

Functional Dependency

Het is mogelijk dat een bepaalde kolom functioneel afhankelijk is van een andere kolom. Dat wil zeggen: als de waarde van de andere kolom bekend is, dan weten we of kunnen we afleiden wat de waarde van afhankelijke kolom is. Dat duidt op redundantie of gedenormaliseerde data. In afbeelding 9 zien we dat de kolom DEPARTMENT_NAME in tabel EMPLOYEES voor 100% dus in alle gevallen volgt uit DEPARTMENT_ID. Dat suggereert dat deze kolom wellicht kan worden verwijderd uit de tabel EMPLOYEES en kan worden ondergebracht in een look-up tabel. Uiteraard kan OWB dat voor ons verzorgen.

In het grafische tabblad bij de Functional Dependency's krijgen we nog een ander inzicht: hier zien we voor bepaalde kolom-combinaties, in afbeelding 10 voor de combinatie DEPARTMENT_ID => COMMISSION_PCT, het aantal records waar niet aan een vaste combinatie wordt voldaan. Bijvoorbeeld DEPARTMENT_ID 80 (het Sales Department) heeft in op één na alle gevallen een COMMISSION_PCT van 40%. Combinaties die geen toeval zijn, maar die volgen uit bedrijfsregels – bijvoorbeeld voor JOB_ID=>COMMISSION_PCT zou kunnen gelden SA_REP=> NOT NULL en alle overige waarden=> NULL – kunnen we als Custom Data Rule vastleggen en vervolgens gebruiken voor correctie en toekomstige controle.



Afbeelding 10. Hier zien we voor bepaalde kolom-combinaties het aantal records waar niet aan een vaste combinatie wordt voldaan.

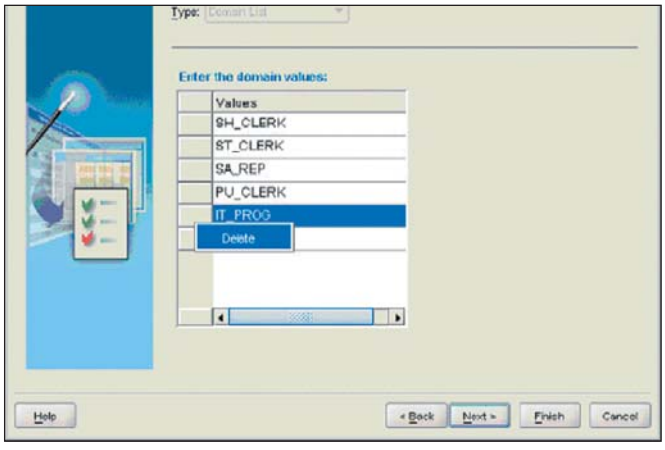
Custom Data Rule

In het Custom Data Rule tabblad kunnen we eigen regels definiëren, van verschillende types, en vervolgens Data Profiler laten rapporteren in welke gevallen aan die regel voldaan wordt en welke rijen dat niet doen. Een eenvoudige regel zou kunnen zijn de waarde van (de redundante) kolom INCOME wordt berekend als de som van Salary en Bonus. Dit tabblad laat ons deze regel definiëren en vervolgens inspecteren of er records zijn waar niet aan die voorwaarde wordt voldaan.

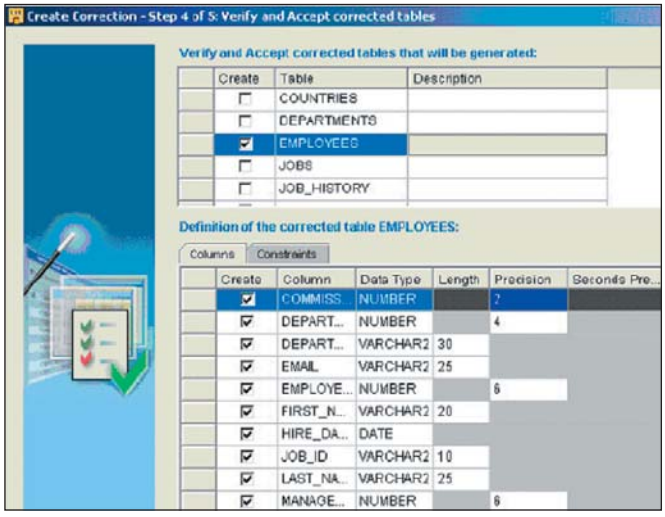
Schema en Data Correction

De resultaten die zijn gepresenteerd door Data Profiler geven niet alleen inzicht in de structuur en de kwaliteit van de data in de onderzochte tabellen, ze geven in veel gevallen ook aanleiding voor verdere actie. Vaak legt Data Profiler tekortkomingen bloot, hetzij in de definities van de kolommen – denk aan verkeerde datatypes of groottes, ontbrekende uniqueness constraints of redundante kolommen – hetzij in de data zelf. Beide soorten tekortkomingen kunnen met Oracle Warehouse Builder opgelost worden.
De eerste stap in dat proces zijn we al tegengekomen: het vastleggen van de Custom Data Rules. Oracle Warehouse Builder toont ons in de verschillende tabbladen suggesties voor de data-rules, maar het is aan ons om die suggesties te accepteren en om te zetten in daadwerkelijke data-rules, die de basis vormen voor Data Correctie en Data Auditing. We volgen het voorbeeld van het voorgestelde Domein voor de kolom JOB_ID. We geven aan dat we dit Domein willen gaan vastleggen als Custom Data Rule. De Derive Data Rule wizard wordt gestart, en bevat de door OWB afgeleide regel. Nu hebben we de kans om de Data Rule verder aan te scherpen. De waarde IT_PROG die door OWB in het domein is opgenomen blijkt niet correct; we kunnen deze waarde uit het domein verwijderen.

De tweede stap is het definiëren van zogenaamde Corrections. In Corrections leggen we vast welke Data Rules we willen gaan



Afbeelding 11. We kunnen de waarde IT_PROG, die door OWB in het domein is opgenomen, uit het domein verwijderen.

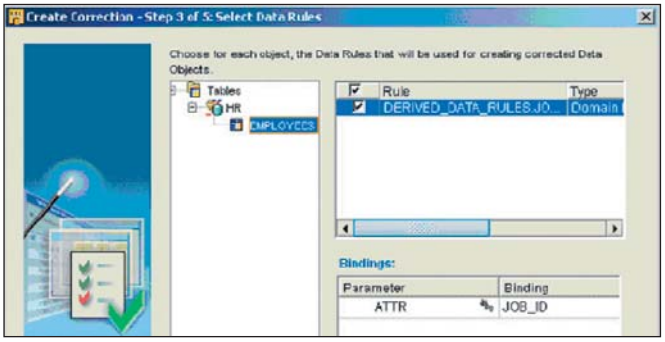


Afbeelding 12. In Corrections leggen we vast welke Data Rules we willen gaan afdwingen.

afdwingen en ook wat OWB moet doen met records die niet voldoen aan deze Data Rules. We kiezen ervoor om een Correction te creëren voor tabel EMPLOYEES. OWB presenteert de tabeldefinitie die het op grond van de profiling resultaten en de daarvan afgeleide Data Rules voor ogen heeft. Zowel kolomdefinities als constraints worden door OWB aangescherpt naar aanleiding van de Data Rules.

Vervolgens kunnen we de Data Rules selecteren die we door OWB afgedwongen willen hebben. In dit geval kiezen we de Domain Rule die voor kolom JOB_ID is vastgelegd.

Oracle Warehouse Builder kan nu Target Database Objecten en een Mapping creëren en genereren. Dat betekent dat OWB ons de gecorrigeerde tabel- en constraint-definities oplevert. En ook de PL/SQL- en SQL-code om de data uit de huidige tabel(len) te migreren naar de doel-tabellen. Nu komt echter nog een heel belangrijk punt: de schoning van data. We weten dat een klein percentage van de records niet past in de gecorrigeerde tabelstructuur of afwijkt van de Data Rules die zijn gespecificeerd. We kunnen dus niet zonder meer alle bestaande gegevens naar de gecorrigeerde structuur gaan migreren. OWB laat ons per Data Rule vastleggen wat er moet gebeuren met afwijkende records: door de vingers zien, rapporteren, verwerpen en proberen op te schonen. Voor het opschonen zijn verschillende strategieën beschikbaar: plaats in een aparte tabel waarna handmatige correctie kan volgen, roep een custom PL/SQL-functie die een gecorrigeerde waarde afleidt, maak de waarde leeg (NULL), gebruik OWB-library's om een waarde af te leiden. Voor domeinen kan OWB bijvoorbeeld de waarde vervangen door de geldige domeinwaarde die het meest lijkt op de kolomwaarde. Die laatste strategie kan volautomatisch typefouten als SH_CLIRK corrigeren tot de geldige domeinwaarde SH_CLERK.



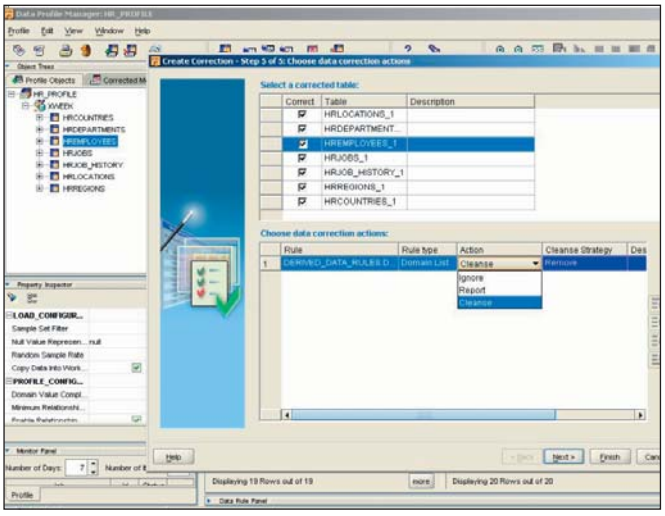
Afbeelding 13. We kunnen de Data Rules selecteren die we door OWB afgedwongen willen hebben.

Data Auditor

Data Profiling en Correction zoals tot nu toe beschreven lijkt een strikt eenmalig proces. Echter, in een typische datawarehouse-omgeving worden frequent nieuwe data aangeleverd uit externe systemen. Ook de kwaliteit van die nieuw ontvangen data is een punt van aandacht. De regels die door middel van Data Profiling zijn gevonden en na eventuele handmatige verfijning zijn vastgelegd kunnen worden gebruikt door OWB om van nieuw ontvangen data de kwaliteit te bepalen. Hiertoe kunnen zogenaamde Data Auditors worden gedefinieerd. Dit zijn achtergrondprocessen die de data doorlichten en statistische gegevens verzamelen die de kwaliteit van de data op grond van de data-rules in kaart brengen. Data Auditors kunnen rapportages opleveren en ook afwijkende gegevens automatisch naar audit- en error-tabellen schrijven. In geval van overschrijding van kwaliteitsnormen kunnen ook alerts verzonden worden.

Overige Functionaliteit

Oracle Warehouse Builder 10gR2 bevat een aantal additionele faciliteiten die ik kort wil aanstippen. Zo kunnen de Mappings



Afbeelding 14. Voor domeinen kan OWB de waarde vervangen door de geldige domeinwaarde die het meest lijkt op de kolomwaarde.

Advertentie

of ETL-processen die in OWB ontworpen en gegenereerd zijn ook als Processen of Jobs gescheduled worden. OWB kan Oracle Workflow aansturing genereren, maar integreert ook met DBMS_SCHEDULER. De workflows of Process Flows zoals ze binnen OWB worden aangeduid worden net als de Mappings grafisch ontwikkeld, en bevatten eventueel logica die bijvoorbeeld een volgende processtap pas start als drie voorafgaande stappen zijn afgerond of een stap een bepaald resultaat heeft opgeleverd. Nota bene: de Mappings zijn de processtappen binnen de Process Flow; de Mappings of ETL-processen worden binnen een Process Flow gecoördineerd en op elkaar afgestemd, zodat bijvoorbeeld data in de juiste volgorde – eerst dimensies, dan feiten – worden geladen.
OWB ondersteunt near-realtime data capture. Dat betekent dat OWB code kan genereren voor Real Time Mappings die bouwen op Oracle Streams om CDC (Changed Data Capture) te realiseren. In geval van Real Time Mapping is er niet een batch job die periodiek wordt uitgevoerd om data uit bronnen te lezen en naar een datawarehouse over te hevelen. In plaats daarvan logt de bron-database relevante wijzigingen asynchroon op basis van de REDO log files – zodat de transacties in deze database geen performanceverlies ondervinden als gevolg van de CDC – naar een Queue. Oracle Streams distribueert de gegevens naar geïnteresseerde target Queues die de log-gegevens ontvangen, interpreteren en verder verwerken. Met minimale vertraging kunnen zo wijzigingen uit operationele databases binnengehaald worden in het datawarehouse.

Eigenschappen

OWB heeft een Command Line Interface en een Java- en PL/SQL-API waarlangs de OWB Repository kan worden benaderd op programmatische wijze. Daarnaast kunnen binnen OWB zogenaamde 'Experts' worden ontwikkeld; dit zijn wizards die door normale OWB-gebruikers worden toegepast om bepaalde taken vereenvoudigd uit te voeren. De Experts zijn meestal op maat gemaakt voor een organisatie door een OWB Administrator. OWB biedt tenslotte een HTML Web interface, die read-only toegang biedt tot zowel de ontwerp-informatie als de audit- en logging-gegevens van de processen die vanuit OWB's Control Center zijn uitgevoerd. Denk daarbij aan generatie en deployment van database-objecten en het uitvoeren van Mappings, oftewel het doen van ETL-operaties.
OWB biedt een grote scala aan User Definable eigenschappen. Dit loopt van eigen icons in de GUI van OWB tot het toevoegen van eigenschappen aan bestaande elementen in de OWB Repository of zelfs het uitbreiden van die Repository met eigen element-types en associaties. Dit doet heel sterk denken aan vergelijkbare functionaliteit in de Oracle Designer repository.
OWB biedt Impact Analyse en Lineage, ook voor deze Custom Types. Dit houdt in dat van een object in OWB een nauwkeurig overzicht (Lineage) kan worden verkregen door welke andere objecten een object wordt beïnvloed – bijvoorbeeld uit welke bronnen een kubus data ontvangt – of welke objecten worden geraakt als een object wordt aangepast (Impact Analyse).
Daarnaast heeft OWB min of meer standaard multi-user repository-faciliteiten als User en Role Management, Version Control met een beperkte mate van Compare/Diffing & Merging en Auditing.

Conclusie

Alles bij elkaar moeten we er nu al meer dan twee jaar op wachten, maar Oracle Warehouse Builder 10g Release 2 is er nu toch echt bijna. De functionaliteit is zeer de moeite waard. Zeker voor wie complete datawarehouses en business intelligence-applicaties ontwikkelt, maar eigenlijk ook voor iedere organisatie met veel verschillende database-applicaties en datastromen tussen databases.
De OWB Data Profiler geeft ons de gelegenheid om van een wat onduidelijke data-set op eenvoudige wijze een prima inzicht te krijgen in de onderlinge samenhang van records en tabellen, van waarschijnlijke formaten en andere business rules, van denormalisaties en afwijkende data. Data Auditing en Data Cleansing met eventueel Match-Merge is een krachtig mechanisme om de kwaliteit van data te monitoren en eventueel bij te sturen. Zodra we de regels waaraan onze data zich zouden moeten houden in kaart hebben gebracht, eventueel op basis van de Data Profiler-resultaten aangevuld met handmatige definities, kunnen we bestaande en binnenkomende data aan de hand van die regels valideren en zelfs laten corrigeren. Op grond van simpele, declaratief instelbare correctieregels of eventueel volledig handmatig ontwikkelde PL/SQL-code kan OWB het datacorrectieproces volledig genereren en periodiek laten uitvoeren. OWB creëert geen tovercode. OWB bevat een eigen bibliotheek van waardevolle functies maar genereert verder gewone PL/SQL- en SQL-code, die we ook zelf kunnen schrijven. Alleen OWB-code wordt tientallen malen zo snel tot stand gebracht en hoeft niet door ons getest te worden. Productiviteit en kwaliteit als ook meer inzicht in onze data. Dat is wat OWB met Data Profiler en aanvullende componenten biedt: leuk en vooral zeer waardevol. Een aanrader dus om Oracle Warehouse Builder eens nader te bekijken.

Lucas Jellema (jellema@amis.nl) is sinds 2002 werkzaam bij AMIS Service in Nieuwegein, als Expertise Manager Technologie en Technisch Consultant. Daarvoor werkte hij ruim acht jaar bij Oracle, ondermeer binnen het iDevelopment Center of Excellence. Hij houdt zich onder meer bezig met Java, XML/XSLT en andere webtechnologie als ook de Oracle database en tools voor applicatie ontwikkeling.