

# Zapewnianie jakości danych ładowanych do systemów analitycznych – omówienie możliwości narzędzi wbudowanych w Oracle Warehouse Builder 11g i Oracle Data Integrator 10g

Mariusz Masewicz  
Politechnika Poznańska

*Mariusz.Masewicz@cs.put.poznan.pl*

**Abstrakt.** Systemem analitycznym nazywamy każdy system, który gromadzi w sobie dane z wielu różnych źródeł umożliwiając swoim użytkownikom wykonywanie zadań (analizy danych, szukanie powiązań, wnioskowanie), które byłyby trudne do realizacji, bez takiego właśnie centralnego repozytorium informacji. Takimi sztanarowymi przykładami rozwiązań analitycznych są hurtownie danych, czy też systemy CRM.

Przez wiele lat najważniejszym zadaniem twórców takich systemów było uzyskanie dostępu do systemów źródłowych (przeważnie są to systemy „operacyjne” z punktu widzenia prowadzonego biznesu), pobranie z nich danych i po ewentualnej transformacji załadowanie tych danych do systemu analitycznego. Powstało wiele narzędzi wspierających twórców systemów analitycznych na etapie projektowania procesów Ekstrakcji danych ze źródeł, ich ewentualnej Transformacji i wreszcie Ładowania do docelowych struktur systemu analitycznego. Przykładami takich narzędzi są wymienione w tytule niniejszego referatu Oracle Warehouse Builder 11g i Oracle Data Integrator 10g

Obecnie jednak coraz większy nacisk kładzie się na to, aby dane prezentowane w systemie analitycznym jak najlepiej odzwierciedlały prezentowaną rzeczywistość. Stąd konieczność stosowania narzędzi wspomagających uzyskiwanie największej jakości danych trafiających do systemu analitycznego. Najczęściej występujące problemy to: „gdzie znajdują się aktualne dane klienta – w systemie wystawiającym faktury, czy w dziale windykacji?”, „czy Jan Kowalski z Poznania i Jan Kowalski z Warszawy to dwóch różnych klientów czy też jeden, ale często wysyłany przez pracodawcę w delegację do stolicy? I co zrobić z Janem KowOlskim?” i wreszcie „kto jest aktualnym patronem ulicy J. Dąbrowskiego w Poznaniu?”

Na szczęście coraz więcej narzędzi wspierających projektowanie i zarządzanie procesów ETL (w tym także tytułowe narzędzia) posiada moduły wspierające zapewnianie jakości ładowanych danych. Na rynku dostępne są też słowniki oraz wzorce poprawności danych, które można dołączyć do omawianych narzędzi jako kolejne rozszerzenia ich funkcjonalności w zakresie zapewnienia jakości danych.

Niniejszy referat prezentuje możliwości narzędzi Oracle Warehouse Builder 11g i Oracle Data Integrator 10g jako filtrów wychwytyjących nieprawidłowości podczas ładowania danych i próbujących naprawić przekłamanne dane.

## 1. Wstęp

Większość współczesnych przedsiębiorstw (począwszy od „jednoosobowych działalności gospodarczych”, a na potężnych międzynarodowych korporacjach skończywszy) działa obecnie w rzeczywistości przesiąkniętej ogromną ilością danych. Dane te – gromadzone w różnych systemach informatycznych – potrzebne są im zarówno do codziennej mozolnej pracy, jak i do podejmowania decyzji, które mogą zdecydować o dalszych losach danego przedsiębiorstwa. Stąd bardzo istotne jest to, aby na każdym kroku zapewniać najwyższą możliwą jakość przetwarzanych danych. Szczególnym miejscem systemu informatycznego, w którym jakość zaczyna mieć największe znaczenie jest system analityczny (hurtownia danych) służący do wspomaganie pracy (podejmowania decyzji) zarządu przedsiębiorstwa. To właśnie użytkownicy hurtowni danych są tymi, którzy na własnej skórze odczuwają wszelki – najmniejsze nawet – niedociągnięcia w pozyskiwaniu i składowaniu danych źródłowych oraz ich późniejszym udostępnianiu do celów analitycznych.

W kolejnych rozdziałach niniejszej publikacji omówione zostaną zagrożenia dla jakości danych wynikające z błędnego ich pozyskiwania, integrowania i udostępniania w hurtowniach danych, następnie przedstawione zostaną rozwiązania wbudowane w narzędzia firmy Oracle służące do projektowania i wdrażania hurtowni danych. Omówione zostaną tu dwa narzędzia: Oracle Warehouse Builder 11g i Oracle Data Integrator 10g

## 2. Jakość danych

Czym jest jakość danych? Czy można ją jakoś zmierzyć? Zanim odpowiemy sobie na to pytanie przyjrzyjmy się stosowanym w świecie definicjom jakości. Większość z nas słyszała zapewne o normach z rodziny ISO 9000. W dużym uproszczeniu można przyjąć, że twórcy tych norm wyszli z założenia, że nic nie stoi na przeszkodzie, aby „produkować gwoździe z plasteliny”. Taka produkcja będzie zgodna z tym szacownym „systemem zapewniania jakości”, ale muszą być spełnione trzy podstawowe warunki:

- proces będzie doskonale udokumentowany
- proces będzie powtarzalny
- klient będzie zadowolony

Bazując na powyższym podejściu można stworzyć następującą **definicję jakości danych w systemach informatycznych**:

- Dane zgromadzone w systemie będziemy nazywali danymi o wysokiej jakości jeżeli informacja w nich zawarta odzwierciedla reprezentowaną rzeczywistość w sposób założony przez projektanta systemu
- Jakość [danych|produktów] – stopień w jakim [dane|produkty] są w stanie realizować cele do których są przeznaczone

Krótkie przeszukanie Internetu prowadzi do szeregu ciekawych informacji na temat wpływu danych o „złej jakości” na systemy informatyczne:

- Raporty amerykańskiego Data Warehousing Institute mówią, że problemy z jakością danych w USA kosztują przedsiębiorstwa 600 mld USD rocznie
- Koszty „czyszczenia” danych mogą stanowić nawet 80% budżetu przeznaczonego na wdrożenie hurtowni danych
- Ponad 50% projektów CRM zakończyło się niepowodzeniem z powodu złej jakości danych

- W raporcie Gartnera z 2003 r. pt. „Strategiczne podejście do poprawy jakości danych” stwierdza się, że 50% firm wdrażających strategię CRM nie zdaje sobie sprawy ze znaczącego problemu, jakim jest jakość danych.
- Według administracji rządowej USA 15-20% danych w typowej organizacji (włączając w to organizacje rządowe) jest błędnych lub nieużytecznych, co przekłada się na miliardowe straty każdego roku

Coraz częściej mamy do czynienia z sytuacją w której znaczna część danych pochodzi spoza przedsiębiorstwa. Takie dane są pozyskiwane i ładowane do hurtowni danych ze źródeł o różnej wiarygodności. W takiej sytuacji istotne jest to, czy istnieją mechanizmy zapewniania kompatybilności pomiędzy źródłem danych a docelowym miejscem ich przetwarzania. Z najnowszych raportów firmy AMR Research wynika, że o ile wśród użytkowników systemów analitycznych istnieje świadomość potrzeby zapewniania jakości danych odpowiednie mechanizmy są wdrożone w niewielkiej liczbie systemów tego typu. Szczególnie źle wygląda sytuacja właśnie w przypadku pozyskiwania danych z zewnątrz (od partnerów biznesowych). A koszty posiadania w systemie nieprawidłowych danych są różne – od podjęcia błędnych decyzji przez zarząd przedsiębiorstwa do przestojów w produkcji.

Kolejnym kosztem wynikającym z przetwarzania błędnych danych w systemach analitycznych jest narastająca niechęć ich użytkowników do korzystania z takiego systemu, co może wprost doprowadzić do fiaska takiego systemu.

Jak więc z problemem zapewniania jakości danych radzą sobie inni? Oto założenia opracowane przez rząd Kanady dla dostawców oprogramowania tzw. data quality framework:

- Dokładność (Accuracy)
- Porównywalność (Comparability)
- Aktualność (Timeliness)
- Użyteczność (Usability)
- Przydatność (Relevance)

Dodatkowe cechy danych o wysokiej jakości

- Spójność
- Kompletność
- Poprawność
- Dostępność
- Unikalność
- Wiarygodność

Tak więc projektant systemu informatycznego, a już w szczególności hurtowni danych musi pamiętać o następujących aspektach zapewnienia jakości danych:

- Jakość definicji danych
- Jakość architektury danych
- Jakość zawartości danych
- Jakość prezentacji danych

### 3. Zagrożenia dla jakości danych

Analizując problemy na jakie napotykają twórcy procesów ETL (ekstrakcja, transformacja i ładowanie danych do hurtowni danych) można wymienić główne płaszczyzny zagrożenia dla jakości danych przetwarzanych w projektowanym systemie analitycznym

- Błędy w projekcie systemu
- Błędne wykonanie systemu
- Błędne działanie użytkowników systemu
- Błędne działanie urządzeń wspierających działanie systemu
- Błędy na etapie analizowania danych

Do najczęstszych błędów w projekcie systemu można zaliczyć:

- Nerozpoznane reguły walidacji danych (ile informacji jest zakodowanych w numerze PESEL? i czy na ich podstawie walidować wprowadzane dane?)
- Złe struktury danych
- Nerozpoznane zależności pomiędzy danymi
  - Dane o wartościach „wyspanych z palca”
  - „Osierocone” rekordy

Z kolei błędne wykonanie systemu często objawia się następującymi problemami:

- Złe zapisywanie danych w bazie
  - Przekłamania na drodze klient <-> baza danych (np. brak konwersji kodowania ISO-8859-2 <-> WIN1250)
  - Błędne przypisywanie pól formularza do kolumn w bazie danych
- Brak sprawdzania poprawności danych
  - Nadużywanie pól typu „memo” bez najmniejszej nawet walidacji danych
  - Pola o różnej semantyce – zależnej od aktualnego kontekstu
- Brak sprawdzania zależności między danymi

Jak już wspomniano – kolejnym źródłem błędnych danych w systemie są jego użytkownicy. W tym wypadku zagrożeniem dla jakości danych mogą być:

- Błędny system nie pozwalający na wprowadzenie poprawnych/jakichkolwiek danych (np.: brak możliwości zapamiętania numeru PESEL, który zostanie przez system uznany za błędny)
- Integracja systemów wymuszająca wprowadzanie/zachowanie błędnych danych
- Celowy atak na integralność danych
- Zmęczenie pracownika
- Kłopoty z pozyskaniem danych (nie przygotowany petent)
- Niskie kwalifikacje pracownika

Oprócz użytkowników to także urządzenia ułatwiające/automatyzujące proces wprowadzania danych mogą przyczyniać się do powstawiania zagrożeń dla jakości tych danych. Urządzeniami tego typu mogą być:

- skanery
- czytniki kodów
- czujniki

Problemy wynikające z błędnej konfiguracji ww. urządzeń to na przykład:

- czytnik kodu poprawnie odczytuje kod produktu, ale nie jest on „wprowadzony” do bazy, więc kasjer ręcznie wprowadza kod „zbliżony” (w pewnej sieci supermarketów autor tego artykułu dość często kupując jogurt ma później na paragonie piwo)

Wreszcie – pomimo dołożenia wszelkich starań – błędy powstają także na etapie analizowania danych i interpretowania wyników analiz. W tej grupie najczęstsze błędy to:

- Użycie niewłaściwych funkcji
- Użycie niewłaściwych metod zbierania/analizowania danych
- Błędne wnioski (np.: „80% kobiet nie widzi różnicy”)

### 3. Zagrożenia dla jakości danych

#### 3.1. Błędne dane – brak słowników danych

Przykład danych wprowadzanych „z ręki” przez użytkowników pewnego systemu kadrowego

...	<b>Adres_miasto</b>	...
...	Poznań	...
...	poznan	...
...	Poznan	...

Przykładowe zapytanie: Ilu naszych pracowników mieszka w mieście o nazwie „Poznań”?

#### 3.2. Błędne dane – duplikaty rekordów

W pewnej sieci hipermarketów pojawili się klienci o prawie takich samych danych osobowych

Imie	Nazwisko	Kod	Miasto
Jan	Kowalski	00-001	Warszawa
Jan	Kowalski	00-002	Kraków

Przykładowe zapytanie: Czy to jest ta sama osoba, tylko po przeprowadzce drugi raz wprowadzona do bazy, czy też dwie różne osoby? Co jeżeli w jednej z lokalizacji klient ów zachowuje się jak stateczny mąż i ojciec – czyli typowy koszyk zawiera jedno piwo i paczkę pieluszek, a w drugiej lokalizacji typowy koszyk to ogromne ilości napojów wysokokowych i do tego chip-sy?

#### 3.3. Błędne dane – brak weryfikacji zależności pomiędzy danymi

Kolejny przykład danych pochodzących z systemu kadrowego:

ID_pracownika	Stanowisko	Pensja
101	Prezes	10000
102	Sprzedawca	800
103	Sprzedawca	800
104	Sprzedawca	80000
105	Sprzedawca	800

Przykładowe zapytania: Znajdź pracowników, którzy zarabiają więcej niż stawka przewidziana na ich stanowisku. Porównaj średnie zarobki sprzedawców ze średnimi zarobkami pracowników zarządu,

### 3.4. Błędne dane – brak standardu modelowania danych

Dwa różne systemy, dwa różne sposoby zapisywania danych teleadresowych

Imie	Nazwisko	Kod	Miasto
Jan	Kowalski	00-001	Warszawa

Imie i Nazwisko	Adres
Jan Kowalski	00-001 Warszawa

Problem: Jak załadować te dane do hurtowni nie tracąc przy tym informacji?

### 3.5. Błędne dane – brak możliwości porównania danych

Dwa różne magazyny/oddziały. Dwa różne modele danych. Dwa różne standardy raportowania...

#### Oddział I

Nazwa_towaru	ID_towaru	Ilość_sprzedanych	Przychód_netto
Zszywacz	1001	10 szt.	100
Klej	1002	0,5 litra	50
Spinacz	1003	3 opakowania	20

#### Oddział II

Grupa towarów	Przychód_brutto
Materiały biurowe	230

Problem: Który oddział jest lepszy?

### 3.6. Semantyka danych zależna od kontekstu

Fragment rzeczywistego systemu. Autor tego systemu do dzisiaj jest z niego dumny, gdyż udało mu się zbudować system oparty o jedną tabelę w bazie danych. Co prawda po pewnym czasie zupełnie stracił rozeznanie w tym, co tak naprawdę w danym kontekście oznaczają poszczególne pola każdego z rekordów...

...	Typ_klienta	Status_klienta	...
...	indywidualny	1 *)	...
...	korporacyjny	1 **)	...

\*) nieaktywny

\*\*\*) właśnie pozyskany

Problem: Odpowiedz bez dokumentacji do tego systemu: Co oznacza status 1 dla klienta indywidualnego i czy jest to wartość poprawna?

### 3.7. Błędna analiza danych

Dane w hurtowni danych agregowane na różnych poziomach

ID_klienta	ID_towaru	Data	Cena
101	1001	2007-01-01	100
101	1002	2007-01-01	100
101	1003	2007-01-01	100
101	1004	2007-01-01	100
101	1005	2007-01-02	1000
<b>średnia</b>			<b>280</b>

ID_klienta	Data	Srednia_cena
101	2007-01-01	100
101	2007-01-02	1000
<b>średnia</b>		<b>550</b>

Przykładowe zapytanie: Ile wynosi średnia cena towarów kupowanych przez klienta o identyfikatorze 101?

### 3.8. Niedostępne dane

Przykład rozmowy z call-center operatora telekomunikacyjnego (T: telefonistka, K: potencjalny klient)

T: Mamy dla Pana świetną ofertę: produkt X...	
	K: Ale ja już od pół roku używam produktu X
T: Hmm... Być może. Nie mam dostępu do systemu transakcyjnego – mi się tylko wyświetla, że mam zadzwonić do Pana W takim razie przepraszam że niepokoiłam	

Problemy:

- Jaki jest koszt takiej rozmowy:
  - cena „impulsu”,
  - koszt pracy telefonistki,
  - irytacja klienta
- Ile takich rozmów odbyło się dzisiaj?

## 4. Jak poprawić jakość danych

Istnieje szereg metod wspomagających projektanta, a później administratora hurtowni danych w utrzymywaniu najwyższego poziomu jakości danych składowanych i przetwarzanych w systemie analitycznym. Do najczęściej stosowanych należą:

- Profilowanie/analiza – podsumowanie stanu danych na dzień dzisiejszy; identyfikacja źródeł danych, krótkie ich opisanie, wyłowienie oczywistych błędów i niespójności, porównanie danych z wzorcami (np. statystykami rozkładu danych)
- Standaryzacja jakości – to określenie standardów składowania i formatowania danych w bazach źródłowych, wzorcowa baza metadanych oraz identyfikacja źródeł danych, które ze standardem nie są zgodne, stopień zgodności danych ze standardem

- Integracja – ujednoczenie danych w systemach, normalizacja systemów relacyjnych, automatyczne lub ręczne dostosowanie danych do wypracowanego standardu
- Wzbogacanie – dodawanie wartości składowanym danym, np. normalizacja danych teleadresowych
- Monitoring – proces, w ramach którego organizacja będzie zapewniać jakość danych od razu (a priori), a nie w wyniku kosztownych i długotrwałych procedur „czyszczących” (a posteriori)

W związku z tym, iż proces zapewniania wysokiej jakości danych może być kosztowny (w wymiarze finansowym, ale także nakładu pracy, czy też zwiększenia czasu i skomplikowania procesów przetwarzających dane) bardzo często przeprowadza się go w wielu etapach wprowadzając następujące reguły postępowania:

- Ustalenie priorytetów – warto wskazać dane o największym znaczeniu. części danych nie opłaca się poprawiać, jeśli ich wartość biznesowa jest mała w stosunku do wydatków koniecznych na podniesienie ich jakości
- Zaangażowanie właścicieli danych – w celu ustalenia przyczyn powstawania błędów i przekłamań w danych, a także stwierdzenia, jaki stopień czystości danych będzie optymalny
- Dbanie o wysoką jakość nowych danych
- Trwała współpraca pracowników działu IT z pracownikami merytorycznymi przy bieżącej poprawie danych i właściwym oznakowaniu danych, których jakość budzi wątpliwości, a których z tych czy innych przyczyn nie można poprawić

Z dotychczasowych rozważań można już wnioskować, że proces zapewniania jakości danych jest bardzo odpowiedzialnym, ale i niezwykle trudnym zadaniem. Na pewno nie podoła mu pojedynczy projektant, a później administrator hurtowni danych. Dlatego też często sięga się po dodatkowe mechanizmy wspomagające utrzymywanie właściwej jakości danych w systemach informatycznych. Najczęściej stosuje się następujące podejścia:

- Czyszczenie danych siłami własnymi – proces żmudny (często ręczny), pozwalający chwilo (jeżeli nie wyeliminuje się przyczyn) osiągnąć satysfakcjonujące rezultaty
- Zakup słowników danych (coraz więcej firm oferuje ciekawe słowniki)
- Zakup oprogramowania wspierającego proces czyszczenia danych
- Zakup usług czyszczenia danych

## 5. Narzędzia firmy Oracle

Rynek produktów wspierających projektowanie i utrzymywanie hurtowni danych jest bardzo bogaty. Firma Oracle posiada w swojej ofercie kilka takich produktów z których najpopularniejsze to: Oracle Warehouse Builder 11g i Oracle Data Integrator 10g. Oba te produkty pozwalają zaprojektować wygląd docelowej hurtowni danych oraz zaprojektować procesy ETL odpowiedzialne za pozyskanie danych z systemów źródłowych oraz ich transformację i ładowanie do projektowanej hurtowni danych. Także oba te narzędzia mogą służyć do uruchamiania i nadzorowania wcześniej zaprojektowanych procesów ETL. W obu tych narzędziach znajdują się też moduły pozwalające zadbać o to, aby dane przez nie przepływają miały odpowiednią jakość. W najnowszych wersjach tych produktów mamy do czynienia z tymi samymi narzędziami do profilowania i czyszczenia danych – różnią się one jedynie interfejsem użytkownika i co już jest być może mniej istotne dla użytkownika – sposobem i miejscem działania, co wprost wynika z różnic pomiędzy tymi narzędziami.

Istotnym czynnikiem różniącym oba te produkty jest ich pochodzenie, technologia w której zostały wykonane i filozofia działania. Oracle Warehouse Builder to okręt flagowy w marynarce wojennej z banderą Oracle. Od początku był on tworzony z tą myślą, że przy jego pomocy będzie się projektować hurtownie danych działające w różnych bazach danych firmy Oracle, czerpiące swe dane także z baz danych Oracle, a jeżeli zajdzie potrzeba połączenia się do innego systemu – to poprzez usługi heterogeniczne baz danych Oracle.

Profile Results Canvas

Unique Key Functional Dependency Referential Data Rule  
Data Profile Profile Object Aggregation Data Type Pattern Domain

Here are the aggregation analysis results for EMPLOYEES, which has 12 columns and 109 rows.

Columns	Minimum	Maximum	# Distinct	% Distinct	NOT NULL	Discard
LAST_NAME	Abel	Zlotkey	102	93.6%	Yes	Yes
MANAGER_ID	100	205	18	16.5%	No	Yes
PHONE_NUMB...	011.44....	650.50...	107	98.2%	No	Yes
SALARY	2100	24000	57	52.3%	No	Yes

Derive Data Rule Remove Data Rule

Tabular Graphical

Data Drill Panel

Here are drill results on EMPLOYEES column SALARY related to Maximum value.

Distinct values: All

	SALARY	# Rows	% of 109
1	24000	1	9%
2	17000	2	1.8%
3	14000	1	9%

Displaying 57 Rows out of 57 more

Rows for the selected distinct value:

	PHONE_NUMB...	SALARY
1	515.123.4567	24000

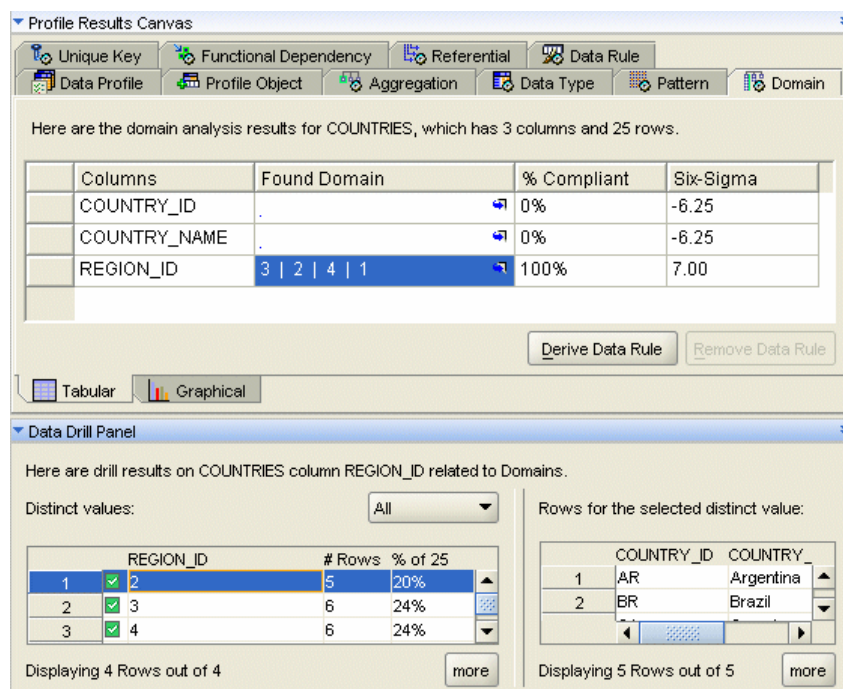
Displaying 1 Rows out of 1 more

Rys. 1. Data Profiler w WHB

Jeżeli chcielibyśmy pozostać przy terminologii związanej z marynarką wojenną – to Oracle Data Integrator należałoby porównać do krążownika, który jest w stanie dotrzeć do każdej przeszkody (bazy danych) i wytaczając swoje działa (sterownik JDBC) połączyć się z dowolnym systemem w celu pobrania z niego danych lub zbudowania w nim struktur hurtowni danych.

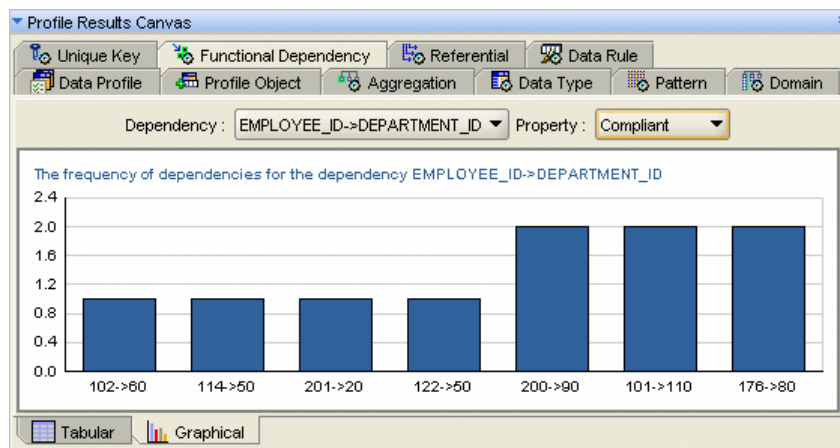
Podstawowym mechanizmem zapewniania jakości danych przepływających przez procesy ETL projektowane w tych narzędziach jest Profiler. Jest to narzędzie, które na podstawie przedstawionej mu próbki danych buduje profil statystyczny tych danych (wartości maksymalne i minimalne w danym polu, rozkład statystyczny np.: 90% studentów politechniki ma atrybut „płeć” ustawiony na wartość „M”, liczba i procent niepowtarzających się wartości, liczba i procent pól z wartością pustą, dla danych liczbowych: wartość średnią, medianę i odchylenie standardowe, ...). Taki profil może później posłużyć do walidowania procesu ładowania rzeczywistych danych (próba załadowania danych o studentach, gdzie w grupie istnieje przewaga kobiet powinna zakończyć się wygenerowaniem alarmu). Profil ten można wykorzystać do zbudowania automatycznych reguł korekcji danych.

Profiler jest też w stanie wykryć w przedstawionej próbce danych pewne zależności funkcyjne, unikalność atrybutów, czy też fakt, że zawartość pewnego atrybutu tworzy domenę o zamkniętym zbiorze wartości.



Rys. 2. Ekran prezentujący odkryte domeny wartości

Informacje o wykrytych zależnościach funkcyjnych mogą być prezentowane w formie wykresów (histogramów) przedstawiających częstości występowania danej zależności



Rys. 3. Ekran prezentujący informacje o wykrytych zależnościach funkcyjnych

Jeżeli profilowaniu podlega zbiór tabel źródłowych Profiler potrafi także wykryć zależności referencyjne pomiędzy tabelami.

Here are the referential analysis results for EMPLOYEES, which has 12 columns and 109 rows.

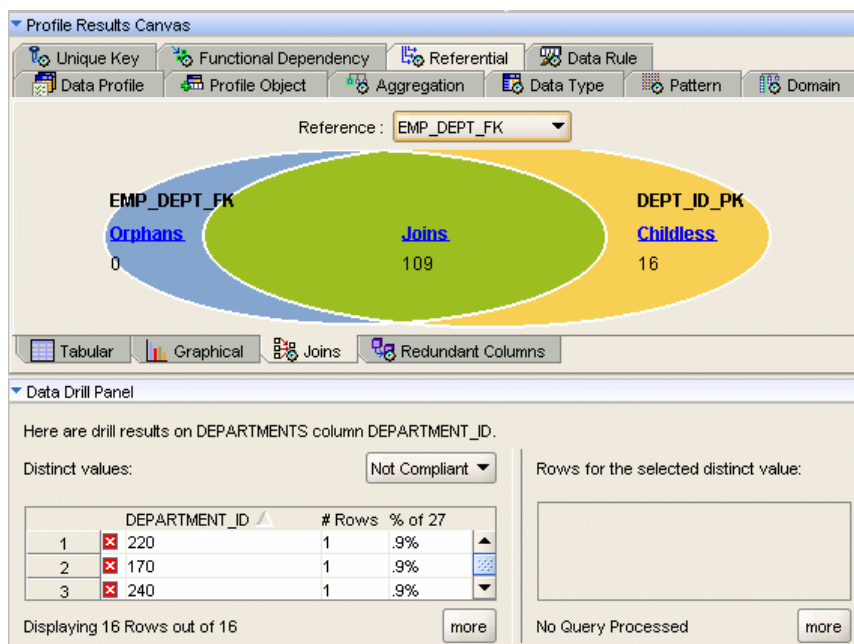
Relationship	Type	Documen...	Discovered ?	Local Attribute(s)	Remote
EMP_DEPT_FK	Foreign Key	Yes	Yes	DEPARTMENT_ID	DEPT_ID
EMP_MANAGE...	Foreign Key	Yes	Yes	MANAGER_ID	EMP_EM
RR_5	Row Relationship	No	Yes	DEPARTMENT_ID	RR_6

Buttons: Derive Data Rule, Remove Data Rule

Bottom tabs: Tabular, Graphical, Joins, Redundant Columns

Rys. 4. Ekran prezentujący wykryte zależności referencyjne

W przypadku zależności referencyjnych Profiler pozwala na graficzną prezentację wyników łączenia tabel prezentując jakie części danych z łączonych tabel udaje się połączyć, a ile danych nie ma odpowiedników w drugiej relacji i do ewentualnego ich wykorzystania/ujawnienia należałoby wykonać połączenie zewnętrzne (OUTER JOIN)



Rys. 5. Graficzna prezentacja wyniku łączenia dwóch tabel

Kolejną właściwością Profilera jest możliwość definiowania reguł dla testowanych danych i późniejszego badania zgodności danych ze zdefiniowaną regułą. Reguły mogą być definiowane jako: domeny z listami, zakresami lub wzorcami poprawnych wartości, wyrażenia regularne reprezentujące format danych, zależności funkcyjne, referencyjne i unikalne, wzorce nazw i adresów, czy wreszcie dowolne wzorce walidowane przy pomocy odpowiedniego zapytania w języku SQL.

Reguły zdefiniowane w WHB lub DI mogą następnie posłużyć do budowania mechanizmów monitorowania i automatycznej korekty ładowanych danych. Na podstawie reguł WHB, oraz DI same potrafią przygotować tak zwane mapowanie korygujące dane, w ramach którego ładowane dane mogą być wzbogacane o informacje potrzebne do spełniania ograniczeń integralnościowych (np. unikalności), zmieniony może zostać typ danych, dane teled adresowe mogą zostać poprawione i zapisane w odpowiednim formacie – system posiada wbudowane wzorce dla kilkudziesięciu sposobów zapisywania danych teled adresowych, stosowanych w różnych krajach. Zdefiniowano

także wzorce dla zapisu informacji o nazwach (przedsiębiorstw lub danych osobowych ludzi), numerach identyfikacyjnych (numery SSN, PESEL, telefony, ...) i wielu innych typowych danych gromadzonych w systemach informatycznych i prezentowanych tam na wiele sposobów.

Mechanizmy automatycznej korekty mogą także dopasowywać kontrolowane wartości (o ile nie spełniają one reguły) do jakiejś wartości zadanej. W zależności od zdefiniowanej metody mogą te dane dopasowywać do wartości maksymalnej, minimalnej, podobnej do aktualnej, ale spełniającej wybraną regułę (znajdującą się w domenie).

System potrafi także usuwać z ładowanego zbioru danych rekordy o powtarzających się wartościach – o ile oczywiście zostanie zdefiniowana korekta bazująca na regule mówiącej, że dane powinny być unikalne.

Zdefiniowane reguły i oparte o nich mechanizmy korekty danych mogą zostać zaimplementowane jako integralna część procesów ETL i wówczas są one wykonywane podczas ładowania danych. Mogą one także być wykorzystywane niezależnie do walidowania jakości danych w źródłach lub hurtowni danych.

## 6. Podsumowanie

Oprócz niezawodności dostępu do danych – drugim aspektem świadczącym o sukcesie przedsięwzięcia informatycznego jest jakość tych danych.

Zła jakość danych – ich niekompletność, niepoprawność może być przyczyną niechęci użytkowników do korzystania z nich, jako źródła niewiarygodnych informacji. Informacja powstała na podstawie złych danych prowadzi do wyciągania błędnych wniosków i podejmowania złych decyzji biznesowych, prowadzących do strat w przedsiębiorstwie oraz osłabienia i obniżenia pozycji rynkowej przedsiębiorstwa.

W celu podnoszenia jakości i użyteczności swoich danych, przedsiębiorstwa powinny przede wszystkim przeprowadzać analizę pochodzenia danych złej jakości i przyczyn występowania błędów w danych. Dopiero później, na podstawie wyników analizy przedsiębiorstwo powinno opracować skuteczny proces poprawy jakości danych oraz wybrać odpowiednie narzędzie, które będzie wspomagać proces.

Po wykonaniu czyszczenia danych należy dbać o to, żeby dane nie uległy już więcej „zanieczyszczeniu”.

Narzędzia firmy Oracle: Oracle Warehouse Builder 11g i Oracle Data Integrator 10g pozwalają na zbudowanie hurtowni danych i procesów odpowiedzialnych za jej utrzymanie z uwzględnieniem najwyższych standardów zapewniania jakości danych.

## Bibliografia

Warehouse Builder Concepts – [http://www.oracle.com/pls/db112/to\\_toc?pathname=owb.112/e10581/toc.htm](http://www.oracle.com/pls/db112/to_toc?pathname=owb.112/e10581/toc.htm)

Warehouse Builder Data Modeling, ETL, and Data Quality Guide – [http://www.oracle.com/pls/db112/to\\_toc?pathname=owb.112/e10935/toc.htm](http://www.oracle.com/pls/db112/to_toc?pathname=owb.112/e10935/toc.htm)

Oracle Data Integrator – <http://www.oracle.com/technology/products/oracle-data-integrator/index.html>

Oracle Data Integrator 10.1.3.5.0 Information – [http://www.oracle.com/technology/products/oracle-data-integrator/10.1.3/htdocs/1013\\_support.html#docs](http://www.oracle.com/technology/products/oracle-data-integrator/10.1.3/htdocs/1013_support.html#docs)

Oracle Data Profiling and Data Quality for Data Integrator – <http://www.oracle.com/technology/products/oracle-data-quality/index.html>

Oracle Data Profiling Datasheet – [http://www.oracle.com/technology/products/oracle-data-quality/pdf/oracledp\\_datasheet.pdf](http://www.oracle.com/technology/products/oracle-data-quality/pdf/oracledp_datasheet.pdf)

Oracle Data Quality for Data Integrator Datasheet – [http://www.oracle.com/technology/products/oracle-data-quality/pdf/oracledq\\_datasheet.pdf](http://www.oracle.com/technology/products/oracle-data-quality/pdf/oracledq_datasheet.pdf)

Getting Started with Oracle Data Quality for Data Integrator guide – [http://www.oracle.com/technology/products/oracle-data-quality/pdf/oracledq\\_gs\\_guide.pdf](http://www.oracle.com/technology/products/oracle-data-quality/pdf/oracledq_gs_guide.pdf)

Oracle Data Quality Tutorial – [http://www.oracle.com/technology/products/oracle-data-quality/pdf/oracledq\\_tutorial.pdf](http://www.oracle.com/technology/products/oracle-data-quality/pdf/oracledq_tutorial.pdf)

Comprehensive Data Quality with Oracle Data Integrator – [http://www.oracle.com/technology/products/oracle-data-integrator/pdf/oracledi\\_comprehensive\\_quality.pdf](http://www.oracle.com/technology/products/oracle-data-integrator/pdf/oracledi_comprehensive_quality.pdf)

Oracle Data Integrator User's Guide – [http://www.oracle.com/technology/products/oracle-data-integrator/10.1.3/htdocs/documentation/oracledi\\_users.pdf](http://www.oracle.com/technology/products/oracle-data-integrator/10.1.3/htdocs/documentation/oracledi_users.pdf)