Procesy ETL

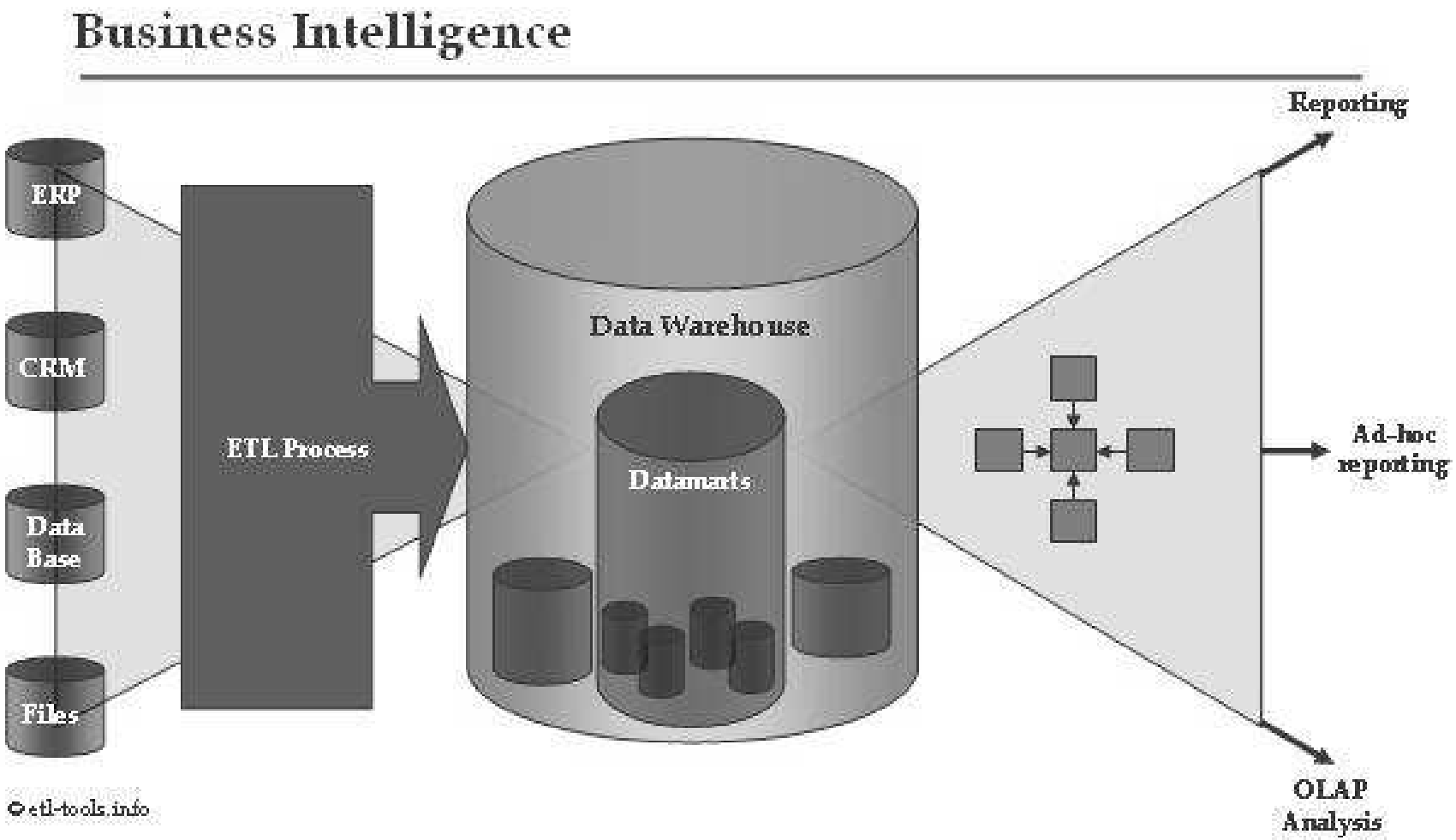
Paweł Szołtysek

10 maja 2009

Agenda

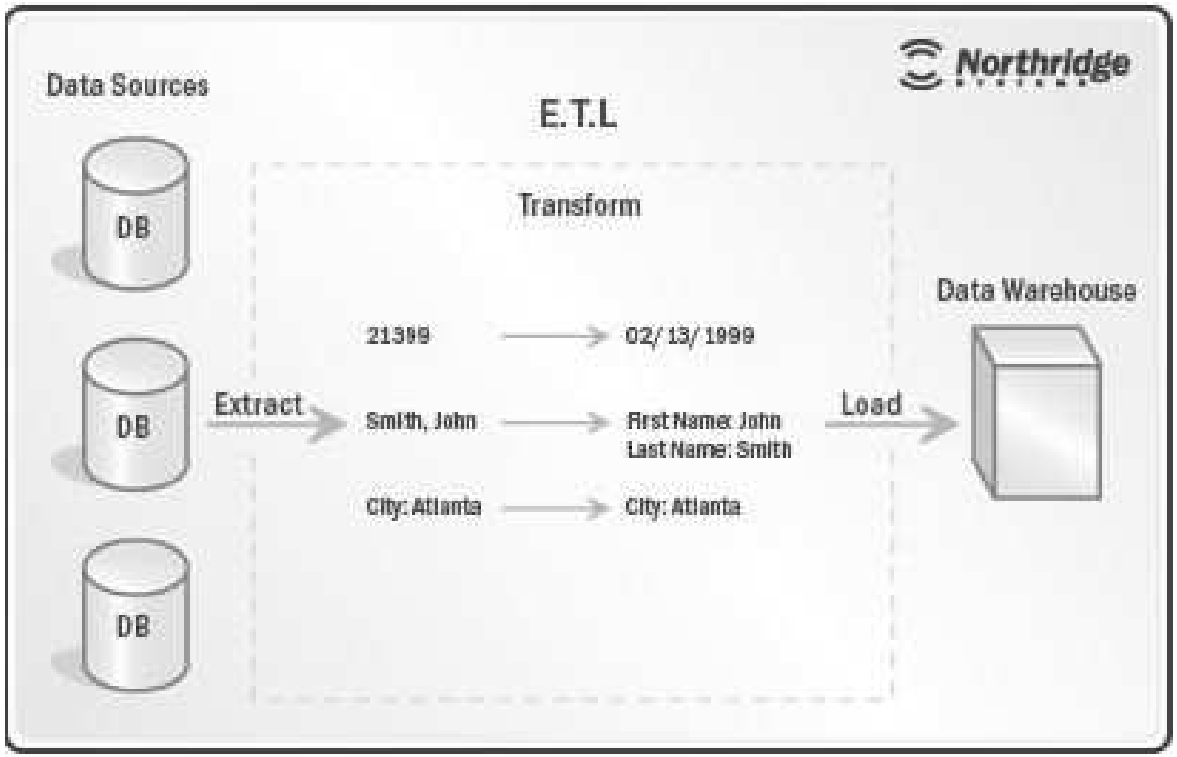
Agenda
Wstęp
ETL
ETL w praktyce
Błędy w ETL

Zagadnienie Business Inteligence



Czym jest proces ETL?

Dane: dowolny zbiór danych ze źródeł zewnętrznych.
Szukane: hurtownia danych bazująca na wskazanych danych.
Rozwiązanie: przekształcenie danych.



ETL = **E**xtract, **T**ransform, **L**oad

ETL: transformacja danych

Transformacja danych to zbiór funkcji które przekształcają przygotowane dane źródłowe do formy akceptowalnej przez hurtownię danych.
Zależnie od formy w jakiej dane znajdują się na wejściu tej części procesu, mogą zostać poddane różnym funkcjom które dostosują je do wymaganego formatu. W szczególności można wyróżnić np:

- ✔ Tłumaczenie kodowanych wartości
- ✔ Obliczanie wartości
- ✔ Filtrowanie, sortowanie, agregacje
- ✔ Selekcja, łączenie

W tym przypadku także przewiduje się sprawdzenie poprawności przetworzonych danych.

ETL: transformacja danych

Transformacja danych to zbiór funkcji które przekształcają przygotowane dane źródłowe do formy akceptowalnej przez hurtownię danych.
Zależnie od formy w jakiej dane znajdują się na wejściu tej części procesu, mogą zostać poddane różnym funkcjom które dostosują je do wymaganego formatu. W szczególności można wyróżnić np:

- ✔ Tłumaczenie kodowanych wartości
- ✔ Obliczanie wartości
- ✔ Filtrowanie, sortowanie, agregacje
- ✔ Selekcja, łączenie

W tym przypadku także przewiduje się sprawdzenie poprawności przetworzonych danych.

ETL: transformacja danych

Transformacja danych to zbiór funkcji które przekształcają przygotowane dane źródłowe do formy akceptowalnej przez hurtownię danych.
Zależnie od formy w jakiej dane znajdują się na wejściu tej części procesu, mogą zostać poddane różnym funkcjom które dostosują je do wymaganego formatu. W szczególności można wyróżnić np:

- ✔ Tłumaczenie kodowanych wartości
- ✔ Obliczanie wartości
- ✔ Filtrowanie, sortowanie, agregacje
- ✔ Selekcja, łączenie

W tym przypadku także przewiduje się sprawdzenie poprawności przetworzonych danych.

ETL: transformacja danych

Transformacja danych to zbiór funkcji które przekształcają przygotowane dane źródłowe do formy akceptowalnej przez hurtownię danych.
Zależnie od formy w jakiej dane znajdują się na wejściu tej części procesu, mogą zostać poddane różnym funkcjom które dostosują je do wymaganego formatu. W szczególności można wyróżnić np:

- ✔ Tłumaczenie kodowanych wartości
- ✔ Obliczanie wartości
- ✔ Filtrowanie, sortowanie, agregacje
- ✔ Selekcja, łączenie

W tym przypadku także przewiduje się sprawdzenie poprawności przetworzonych danych.

ETL: ekstrakcja danych

Problemy związane z ekstrakcją danych:

- ✔ Lokalizacja.
- ✔ Dostępność.
- ✔ Problem konsolidacji.
- ✔ Różnorodność organizacji zbiorów danych.
- ✔ Różnorodność formatów.

Proces ekstrakcji kończy się przygotowanym dla etapu transformacji zbiorem danych, w dobrze określonym formacie. Przewiduje się sprawdzenie poprawności formatu przygotowanych danych.

ETL: ekstrakcja danych

Problemy związane z ekstrakcją danych:

- ✔ Lokalizacja.
- ✔ Dostępność.
- ✔ Problem konsolidacji.
- ✔ Różnorodność organizacji zbiorów danych.
- ✔ Różnorodność formatów.

Proces ekstrakcji kończy się przygotowanym dla etapu transformacji zbiorem danych, w dobrze określonym formacie. Przewiduje się sprawdzenie poprawności formatu przygotowanych danych.

ETL: ekstrakcja danych

Problemy związane z ekstrakcją danych:

- ✔ Lokalizacja.
- ✔ Dostępność.
- ✔ Problem konsolidacji.
- ✔ Różnorodność organizacji zbiorów danych.
- ✔ Różnorodność formatów.

Proces ekstrakcji kończy się przygotowanym dla etapu transformacji zbiorem danych, w dobrze określonym formacie. Przewiduje się sprawdzenie poprawności formatu przygotowanych danych.

ETL: ekstrakcja danych

Problemy związane z ekstrakcją danych:

- ✔ Lokalizacja.
- ✔ Dostępność.
- ✔ Problem konsolidacji.
- ✔ Różnorodność organizacji zbiorów danych.
- ✔ Różnorodność formatów.

Proces ekstrakcji kończy się przygotowanym dla etapu transformacji zbiorem danych, w dobrze określonym formacie. Przewiduje się sprawdzenie poprawności formatu przygotowanych danych.

ETL: ekstrakcja danych

Problemy związane z ekstrakcją danych:

- ✔ Lokalizacja.
- ✔ Dostępność.
- ✔ Problem konsolidacji.
- ✔ Różnorodność organizacji zbiorów danych.
- ✔ Różnorodność formatów.

Proces ekstrakcji kończy się przygotowanym dla etapu transformacji zbiorem danych, w dobrze określonym formacie. Przewiduje się sprawdzenie poprawności formatu przygotowanych danych.

ETL: ekstrakcja danych

Problemy związane z ekstrakcją danych:

- ✔ Lokalizacja.
- ✔ Dostępność.
- ✔ Problem konsolidacji.
- ✔ Różnorodność organizacji zbiorów danych.
- ✔ Różnorodność formatów.

Proces ekstrakcji kończy się przygotowanym dla etapu transformacji zbiorem danych, w dobrze określonym formacie. Przewiduje się sprawdzenie poprawności formatu przygotowanych danych.

ETL: ładowanie danych

- ✔ Przetworzone dane trafiają do hurtowni danych.
- ✔ Zależnie od potrzeb, mogą one napisywać aktualnie znajdujące się dane, lub też dodawać je w celach udostępniania danych historycznych.
- ✔ Etap ten ma powiązanie z implementacją hurtowni danych, zarówno jeśli chodzi o typy użytych silników, jak i wygląd logiczny poszczególnych tabel (unikalność, integracja, wymagane atrybuty itp).

ETL: ładowanie danych

- ✔ Przetworzone dane trafiają do hurtowni danych.
- ✔ Zależnie od potrzeb, mogą one napisywać aktualnie znajdujące się dane, lub też dodawać je w celach udostępniania danych historycznych.
- ✔ Etap ten ma powiązanie z implementacją hurtowni danych, zarówno jeśli chodzi o typy użytych silników, jak i wygląd logiczny poszczególnych tabel (unikalność, integracja, wymagane atrybuty itp).

ETL: ładowanie danych

- ✔ Przetworzone dane trafiają do hurtowni danych.
- ✔ Zależnie od potrzeb, mogą one napisywać aktualnie znajdujące się dane, lub też dodawać je w celach udostępniania danych historycznych.
- ✔ Etap ten ma powiązanie z implementacją hurtowni danych, zarówno jeśli chodzi o typy użytych silników, jak i wygląd logiczny poszczególnych tabel (unikalność, integracja, wymagane atrybuty itp).

Wykorzystanie procesów ETL

Procesy ETL mogą powstawać w różnych, niekiedy mało intuicyjnych celach.

Przykład: Badanie jakości danych za pomocą procesu ETL.
Cel: Implementacja procesu ETL przeprowadzającego testy jakości i spójności danych, generującego raporty z rekordami które nie spełniają reguł walidacji.
Sytuacja: Hurtownia danych jest zasilana niskiej jakości danymi. Przeprowadza się testy znakowe (wzorce znakowe, niedozwolonych liter, wystąpienia myślników, kropek itp.) i referencyjne (spójność z architekturą i modelem danych).

Wykorzystanie procesów ETL

Procesy ETL mogą powstawać w różnych, niekiedy mało intuicyjnych celach.

Przykład: Badanie jakości danych za pomocą procesu ETL.

Cel: Implementacja procesu ETL przeprowadzającego testy jakości i spójności danych, generującego raporty z rekordami które nie spełniają reguł walidacji.

Sytuacja: Hurtownia danych jest zasilana niskiej jakości danymi. Przeprowadza się testy znakowe (wzorce znakowe, niedozwolonych liter, wystąpienia myślników, kropek itp.) i referencyjne (spójność z architekturą i modelem danych).

Wykorzystanie procesów ETL

Procesy ETL mogą powstawać w różnych, niekiedy mało intuicyjnych celach.
Przykład: Badanie jakości danych za pomocą procesu ETL.
Cel: Implementacja procesu ETL przeprowadzającego testy jakości i spójności danych, generującego raporty z rekordami które nie spełniają reguł walidacji.
Sytuacja: Hurtownia danych jest zasilana niskiej jakości danymi. Przeprowadza się testy znakowe (wzorce znakowe, niedozwolonych liter, wystąpienia myślników, kropek itp.) i referencyjne (spójność z architekturą i modelem danych).

Wykorzystanie procesów ETL

Przykładowe problemy jakości danych:

- ✔ Data realizacji zamówienia wcześniejsza od daty wprowadzania zamówienia
- ✔ Niepoprawny numer faktury, np. zawierający niedozwolone znaki lub niepoprawną ilość znaków
- ✔ Zły adres, np. kombinacja kod pocztowy i nazwa ulicy niezgodny ze słownikiem adresowym
- ✔ Numer telefonu niezgodny ze wzorcem – gdy przykładowo zawiera spacje lub myślniki a w hurtowni danych zdefiniowano go jako tylko liczby

Błędy w ETL

To nie było zamierzone: na 7 slajdów dwa traktują o możliwych błędach w procesach ETL.

Błędy te można podzielić na kilka różnych kategorii (np. ze względu na ich umiejscowienie czy na pojawienie się nieprawidłowych danych) w zależności od których można je wykrywać łatwiej lub trudniej.

Z wiedzy na temat testowania wiadomo, że nigdy nie możemy być pewni wyniku działania aplikacji.

Trzeba więc nie tylko zapobiegać, ale i przeciwdziałać...

Błędy w ETL

To nie było zamierzone: na 7 slajdów dwa traktują o możliwych błędach w procesach ETL.

Błędy te można podzielić na kilka różnych kategorii (np. ze względu na ich umiejscowienie czy na pojawienie się nieprawidłowych danych) w zależności od których można je wykrywać łatwiej lub trudniej.

Z wiedzy na temat testowania wiadomo, że nigdy nie możemy być pewni wyniku działania aplikacji.

Trzeba więc nie tylko zapobiegać, ale i przeciwdziałać...

Błędy w ETL

To nie było zamierzone: na 7 slajdów dwa traktują o możliwych błędach w procesach ETL.
Błędy te można podzielić na kilka różnych kategorii (np. ze względu na ich umiejscowienie czy na pojawienie się nieprawidłowych danych) w zależności od których można je wykrywać łatwiej lub trudniej.
Z wiedzy na temat testowania wiadomo, że nigdy nie możemy być pewni wyniku działania aplikacji.
Trzeba więc nie tylko zapobiegać, ale i przeciwdziałać...

Błędy w ETL

To nie było zamierzone: na 7 slajdów dwa traktują o możliwych błędach w procesach ETL.
Błędy te można podzielić na kilka różnych kategorii (np. ze względu na ich umiejscowienie czy na pojawienie się nieprawidłowych danych) w zależności od których można je wykrywać łatwiej lub trudniej.
Z wiedzy na temat testowania wiadomo, że nigdy nie możemy być pewni wyniku działania aplikacji.
Trzeba więc nie tylko zapobiegać, ale i przeciwdziałać...

Błędy w ETL

To nie było zamierzone: na 7 slajdów dwa traktują o możliwych błędach w procesach ETL.
Błędy te można podzielić na kilka różnych kategorii (np. ze względu na ich umiejscowienie czy na pojawienie się nieprawidłowych danych) w zależności od których można je wykrywać łatwiej lub trudniej.
Z wiedzy na temat testowania wiadomo, że nigdy nie możemy być pewni wyniku działania aplikacji.
Trzeba więc nie tylko zapobiegać, ale i przeciwdziałać... a aby przeciwdziałać, należy wykrywać.

O pracy

Temat pracy: *Automatyczne wykrywanie błędów w procesach ETL.*
Plan pracy:

- ✔ Opis przebiegu procesu ETL wraz z występującymi w poszczególnych etapach błędami.
- ✔ Klasyfikacja opisanych błędów w procesach ETL.
- ✔ Opis dostępnych metod detekcji błędów każdej klasy.
- ✔ Implementacja wybranej metody detekcji błędów.
- ✔ Weryfikacja skuteczności zaimplementowanej metody na podstawie badań.
- ✔ Przedstawienie wniosków.

O pracy

Przewidywany termin realizacji pracy: (prawdopodobnie) wakacje 2009.
Problem: o ile każde narzędzie ETL posiada algorytmy wykrywania błędów, o tyle firmy je produkujące nie chcą dzielić się autorskimi sposobami radzenia sobie z nimi.