Informatyzacja przedsiębiorstw

Izabela Szczęch

Politechnika Poznańska

Plan wykładu

Elementy Business Intelligence

- Przetwarzanie OLTP vs OLAP
- Hurtownie danych – podstawowe pojęcia
- Proces ETL

Cele informatyzacji przedsiębiorstw

- Dwa podstawowe cele informatyzacji przedsiębiorstw:
  - **usprawnienie pracy pojedynczego pracownika operacyjnego** (sprzedawcy, magazyniera, księgowego, urzędnika) poprzez automatyzację realizowanych przez niego działań, charakteryzujących się ściśle określoną procedurą postępowania i cykliczną powtarzalnością
  - **racjonalizacja działania całego przedsiębiorstwa w wyniku wspomagania decyzji kadry zarządzającej** przez dostarczenie danych analitycznych opisujących bieżący stan i historię działania danej firmy
    - Narzędzia analityczne udostępniające informacje statystyczne o bieżącym stanie firmy, występujących trendach i korelacjach między różnymi czynnikami

Rodzaje aplikacji systemu informatycznego

- W kontekście takich celów informatyzacji można wyróżnić dwa rodzaje aplikacji systemu informatycznego
  - **aplikacje operacyjne**
  - **aplikacje analityczne**

Aplikacje operacyjne

- ich celem jest wspomaganie pracy pojedynczych pracowników operacyjnych
- charakteryzują się prostym przetwarzaniem (odczyt, wstawianie, modyfikacja i usuwanie danych)
- przetwarzanie zazwyczaj ograniczone do niewielkiego zbioru danych szczegółowych

Aplikacja operacyjne

Modelem przetwarzania właściwym dla aplikacji operacyjnych jest tzw. **przetwarzanie transakcyjne** (**OLTP** – **O**n-**L**ine **T**ransaction **P**rocessing) oparte na pojęciu elementarnej jednostki przetwarzania - transakcji

- Właściwości transakcji – ACID

  (atomowość, spójność, izolacja, trwałość)

- **Główne zadania OLTP**: efektywne przetwarzanie dużej liczby współbieżnych transakcji, zapewnienie spójności danych

- Podstawowe modele danych w zastosowaniach OLTP:

  - hierarchiczny
  - sieciowy
  - relacyjny
  - post-relacyjny

Aplikacja analityczne

Aplikacje analityczne

- ich celem jest wspomaganie pracy kadry zarządzającej
- charakteryzują się dużo większą złożonością przetwarzania niż aplikacje operacyjne
- przetwarzanie zorientowane na wspieranie procesów decyzyjnych, czyli przetwarzanie danych historycznych, zagregowanych, często skonsolidowanych z różnych źródeł

Aplikacja analityczne

Modelem przetwarzania właściwym dla aplikacji analitycznych jest tzw. **przetwarzanie analityczne** (**OLAP** – **O**n-**L**ine **A**nalytical **P**rocessing)

- **Główne zadania OLAP**: efektywne wielowymiarowe przetwarzanie dużych wolumenów danych

Przykładowe zapytania analityczne:

- *Ile sprzedano samochodów w Wielkopolsce w zeszłym roku*?
- *Ile sprzedano samochodów osobowych w Poznaniu w ostatnich 10 latach*?

OLTP vs. OLAP

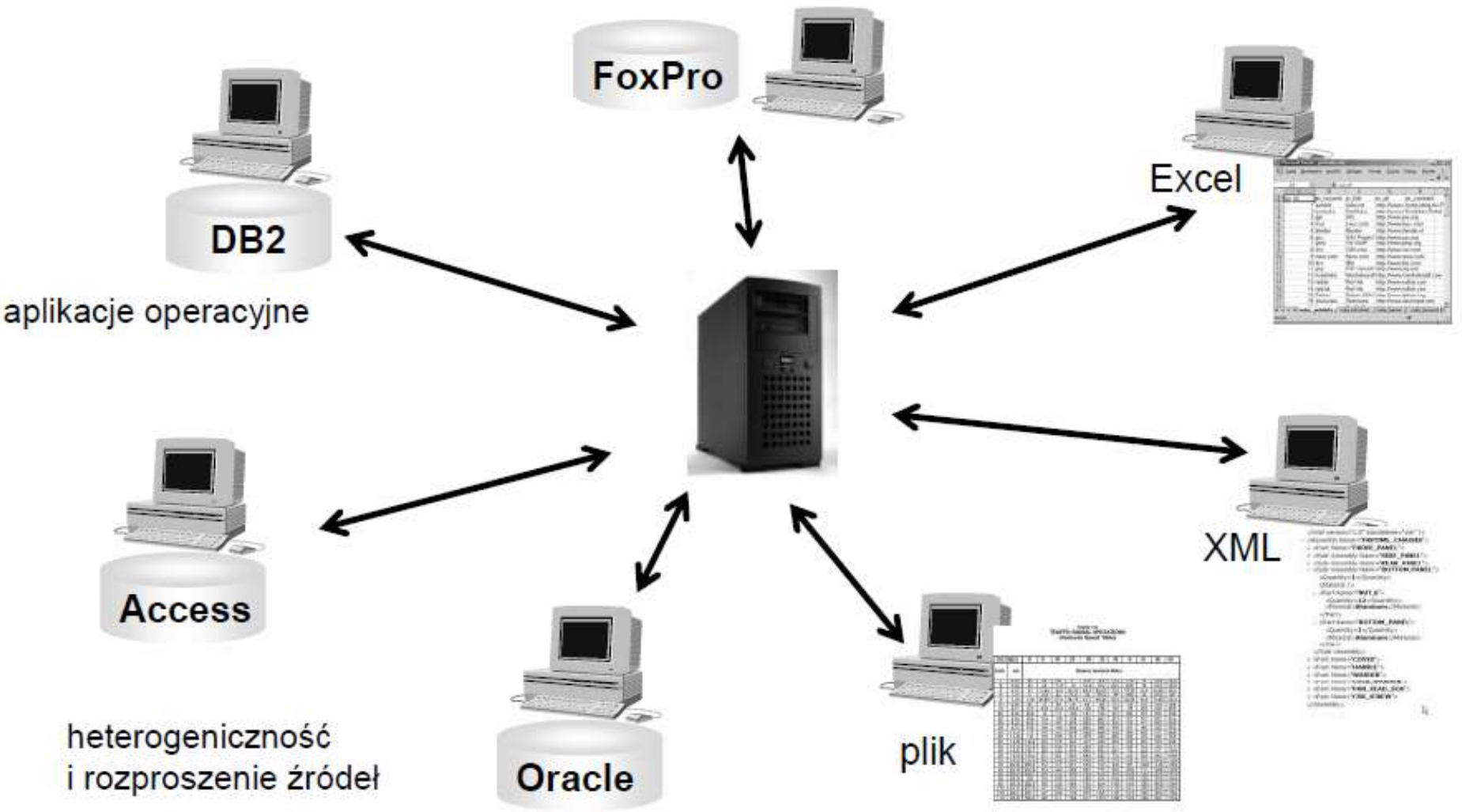
Porównanie cech przetwarzania OLTP i OLAP

Cecha	**OLTP**	**OLAP**
Profil użytkownika	pracownik operacyjny	analityk, decydent
Zapytania	ograniczony, dobrze znany zbiór	częściowo nieprzewidywalne
Czas odpowiedzi	sekundy	minuty/godziny
Rozmiar przetwarzanych danych	kilka rekordów	miliony rekordów
Perspektywa czasowa	dane bieżące	dane historyczne
Tryb dostępu do danych	odczyt/zapis	głównie odczyt
Model danych	dwuwymiarowy	wielowymiarowy

OLTP vs. OLAP

- Z uwagi na odmienność charakterystyki przetwarzania transakcyjnego i analitycznego, rozwiązania stosowane w konwencjonalnych SZBD są nieprzydatne do eksploatacji aplikacji analitycznych
- Główne problemy:
  - heterogeniczność i rozproszenie eksploatowanych systemów OLTP
    - wielość struktur
    - różna funkcjonalność
    - różne modele danych
    - dane są rozmieszczone w geograficznie różnych lokalizacjach
  - równoczesna eksploatacja aplikacji operacyjnych i analitycznych może być przyczyną niskiej efektywności działania systemu

Problematyka integracji danych



OLTP vs. OLAP

- Implementacja systemów wspomagających OLAP wymaga:
  - zintegrowania heterogenicznych, często rozproszonych źródłowych systemów obsługi bieżącej przedsiębiorstwa (OLTP)
  - efektywnego przetwarzania analitycznego (nowa architektura, technologia właściwa dla OLAP)

Hurtownia danych

- Jedną z najczęściej stosowanych technik integracji danych jest ich transformacja do wspólnego modelu i składowanie w centralnym systemie, zwanym hurtownią (magazynem) danych (ang. data warehouse)

- **Hurtownia danych** to „tematycznie zorientowana, zintegrowana, zmienna w czasie, nieulotna kolekcja danych, wykorzystywana w przedsiębiorstwach głównie do wspomagania podejmowania decyzji”

- Ważne nazwiska: Bill Inmon, Ralph Kimball

Hurtownia danych

- **Tematycznie zorientowana**
  - ukierunkowana na dobrze zdefiniowany cel biznesowy przedsiębiorstwa
  - ukierunkowanie inne niż operacyjnych baz danych
- **Zintegrowana**
  - integracja różnych (heterogeniczne) źródeł danych
  - konwersja i integracja przenoszonych danych
  - usunięcie niespójności w zbieranych danych (konwencje nazewnictwa, kodowania pomiędzy różnymi źródłami danych)

Hurtownia danych

- **Zmienna z czasie**
  - horyzont czasowy jest znacząco większy niż w przypadku operacyjnej bazy danych
  - bazy operacyjne przechowują aktualne wartości danych i nie zawsze zawierają element czasu, a hurtownia przechowuje pełną historię i zawsze zawiera elementy związane z czasem
- **Nieulotna**
  - dane operacyjne są regularnie uaktualniane, podczas gdy trans-akcje w systemach OLTP modyfikują, usuwają, wstawiają rekordy
  - w hurtowniach danych dane są doładowywane
  - w hurtowniach danych nie ma uaktualniania danych w tradycyjnym znaczeniu, dane podlegają tylko operacji odczytu

Hurtownia danych

- Hurtownie danych - krótka charakterystyka
  - implementowane jako ogromne bazy danych
  - niezależne od operacyjnych baz danych, na których działają aplikacje operacyjne
  - izolacja przetwarzania operacyjnego i analitycznego
  - wybrane dane z baz danych systemów obsługi bieżącej (OLTP) są replikowane i magazynowane w hurtowni w celu późniejszego przetwarzania analitycznego

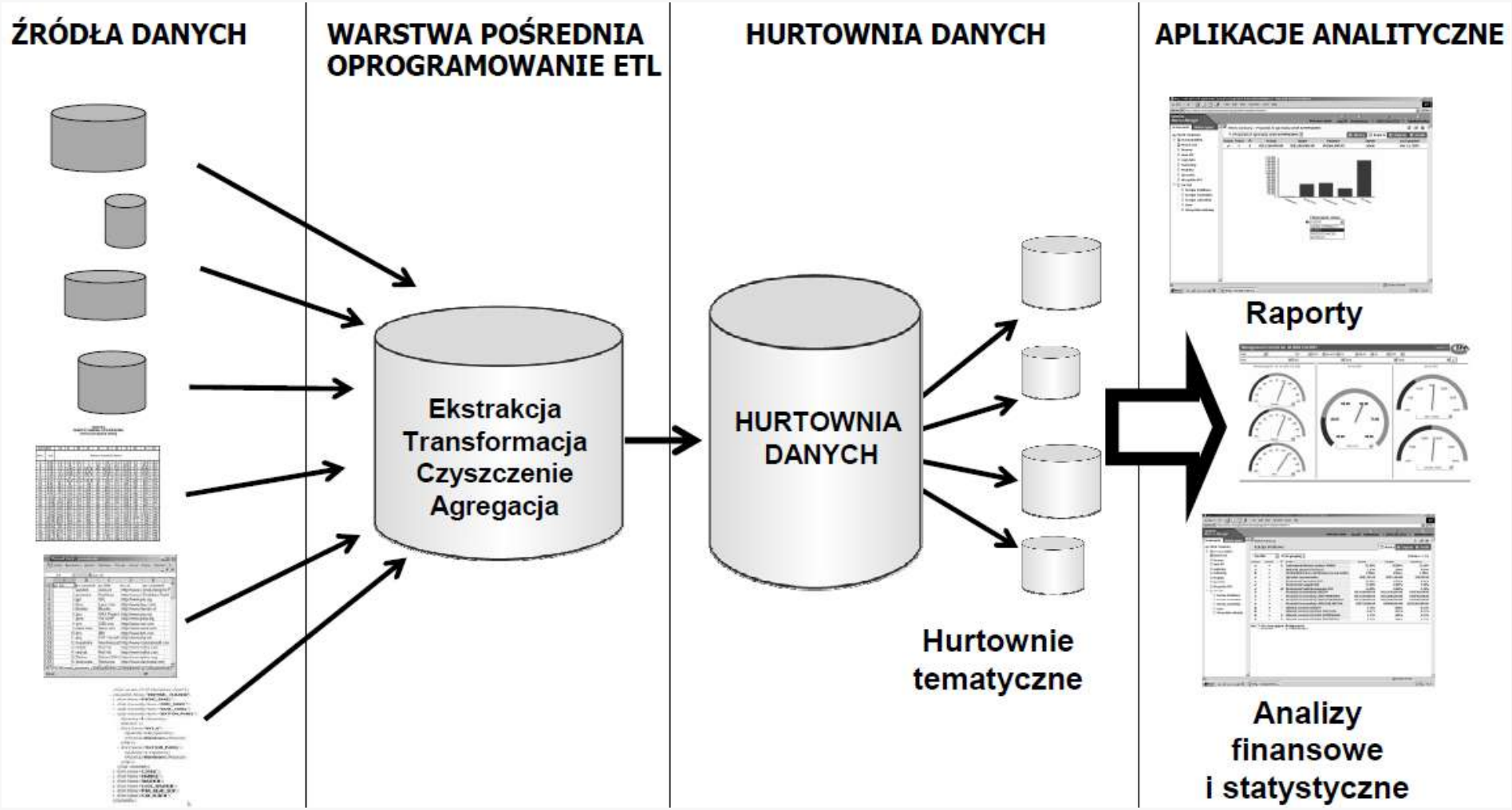
Podstawowe kategorie danych

- W hurtowni danych przechowywane są następujące cztery podstawowe kategorie danych:
  - elementarne pozyskane bezpośrednio ze źródłowych baz danych
  - historyczne tworzone w momencie pojawienia się nowych wartości danych już przechowywanych
  - zagregowane i sumaryczne o różnym stopniu przetworzenia
  - metadane (dane opisujące semantykę, pochodzenie, algorytmy wyznaczania pozostałych kategorii danych)

Zalety systemów hurtowni danych

- Wysoka wydajność zapytań
- Zapytania są niewidoczne poza hurtownią
- Brak ingerencji w dane operacyjne
- Możliwość pracy w przypadku braku dostępu do źródła danych
- Wspieranie specjalnych rodzajów zapytań
- Dodatkowe informacje udostępniane przez hurtownie danych

Podstawowa architektura systemu z hurtownią danych



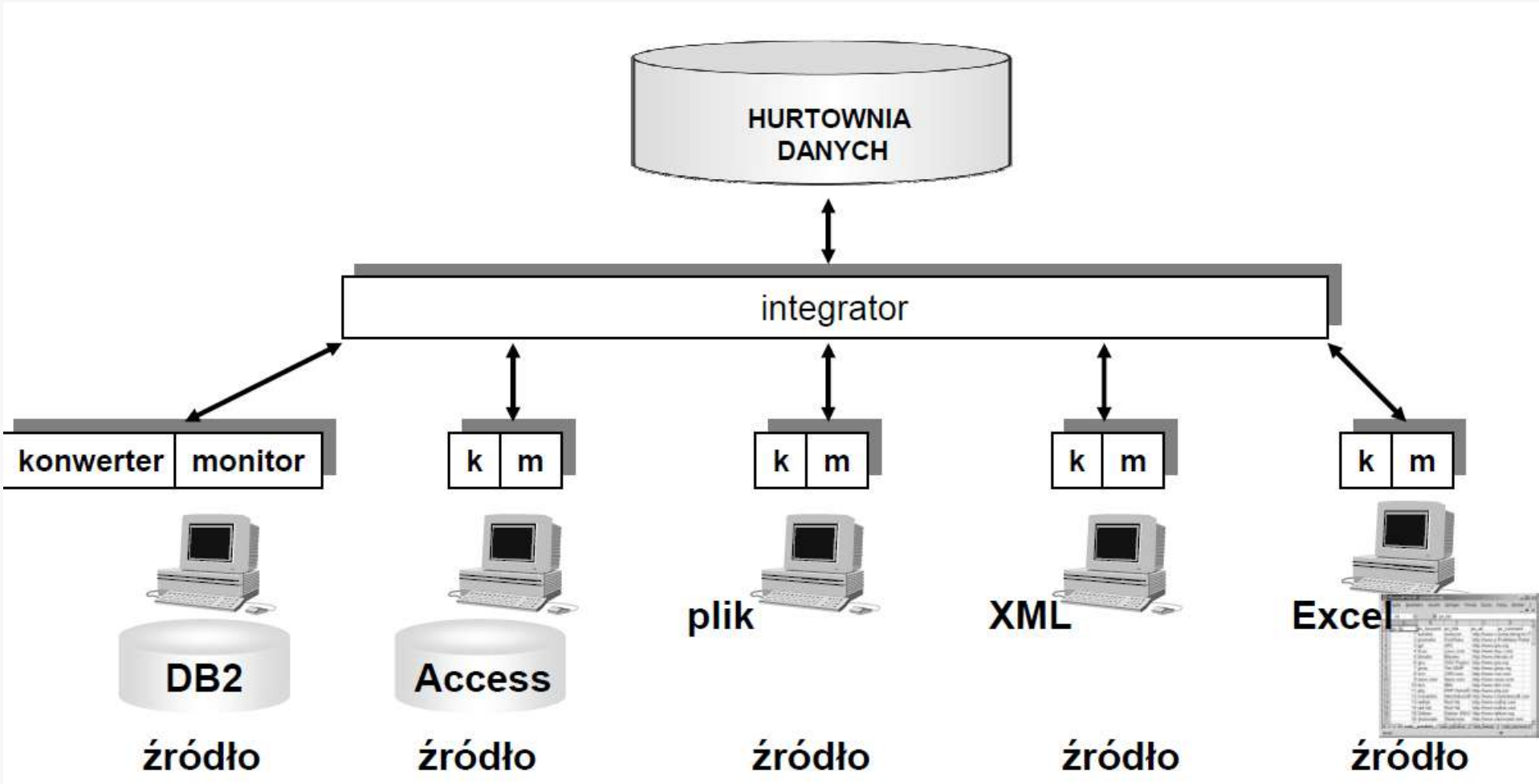
Źródła danych

- **Źródła danych** to systemy zewnętrzne w stosunku do hurtowni danych, często heterogeniczne, o różnej funkcjonalności i stopniu technologicznego zaawansowania
- Autonomia źródła przejawia się w niezależnych od jakichkolwiek systemów zewnętrznych zmianach schematu źródła oraz modyfikacjach jego danych przez lokalnych użytkowników
- Autonomicznie administrowane źródło może narzucać ograniczenia w dostępie do danych (np. dla utrzymania swej wydajności pozwalać na pobieranie swoich danych przez systemy zewnętrzne tylko w określonych porach dnia)

Proces ETL

- Heterogeniczne i rozproszone źródła danych zasilają hurtownię danymi za pośrednictwem warstwy oprogramowania **ETL** (**E**xtraction **T**ransformation **L**oading)
- Oprogramowanie ETL realizuje tzw. **procesy ETL**, składające się z trzech następujących faz:
  - odczytu danych ze źródeł (Extraction)
  - transformacji ich do wspólnego modelu wykorzystywanego w hurtowni wraz z usunięciem wszelkich niespójności (Transformation)
  - wczytanie danych do magazynu (Loading)

Oprogramowanie ETL



Oprogramowanie ETL

- Z każdym z heterogenicznych źródeł (aplikacje operacyjne) jest związana dedykowana dla niego warstwa oprogramowania o nazwie **konwerter/monitor**
- Zadaniem modułu **konwertera** jest transformowanie danych z formatu wykorzystywanego w źródle, do formatu wykorzystywanego w hurtowni
- Przykład:
  Źródło przechowuje dane w dokumentach tekstowych, a hurtownia została zaprojektowana z wykorzystaniem modelu relacyjnego. Konwerter musi zapewnić poprawne odwzorowanie danych z plików w struktury modelu relacyjnego

Oprogramowanie ETL

- Zadaniem modułu **monitora** jest wykrywanie zmian w danych źródłowych i ich przekazywanie do warstwy oprogramowania integratora (po uprzedniej konwersji do modelu danych hurtowni)
- Sposób wykrywania zmian w danych źródłowych zależy od własności samych źródeł. W tym kontekście źródła można podzielić na:
  - źródła aktywne
  - źródła utrzymujące dzienniki operacji
  - źródła przepytywalne
  - źródła wspierające migawki

Podział źródeł danych ze względu na sposób wykrywania zmian

- **Źródła aktywne** (ang. active sources) posiadają zaimplementowane mechanizmy wyzwalaczy, które informują monitor o zmianach zachodzących w danych źródłowych
- **Źródła utrzymujące dzienniki operacji** (ang. logged sources) pozwalają na wykrywane zmian poprzez analizę zawartości dziennika w module monitora
- **Źródła przepytywalne** (ang. queryable sources) umożliwiają wydawanie do nich zapytań i w celu wykrycia zmian w danych źródłowych monitor okresowo wydaje zapytania do takich źródeł
- **Źródła wspierające migawki** (ang. snapshot sources) umożliwiają tworzenie migawek, czyli obrazów stanu źródła z określonego momentu, a moduł monitora wykrywa zmiany poprzez porównanie migawek z kolejnych momentów

Proces ETL - ekstrakcja danych

- Konwersja danych ze źródeł do hurtowni rozpoczyna się od procesu **ekstrakcji danych** (ang. data extraction)
- Dane są odczytywane ze struktur źródłowych (plików tekstowych, arkuszy kalkulacyjnych, stron WWW, baz danych) za pomocą bramek (ang. gateways) i standardowych interfejsów (ODBC, JDBC, dostarczane przez producentów oprogramowania ze strukturami źródłowymi)
- Niejednokrotnie trzeba implementować wyspecjalizowane procedury do ekstrakcji danych ze źródeł niestandardowych

Proces ETL – ekstrakcja danych

Problemy - przykład:

Wyekstrahuj ze źródeł dane o sprzedaży.

- Co oznacza „sprzedaż”?

  otrzymanie zamówienia od klienta?

  realizacja zamówienia dla klienta?

  wystawienie faktury za realizowane zamówienie?

- Częstym problemem jest brak w źródle tabeli `SPRZEDAŻ`
  (a istnienie np. tabeli `ZAMÓWIENIE` z polem `STATUS_ZAMÓWIENIA`)

Proces ETL – transformacja danych

- Kolejnym etapem jest **transformacja i czyszczenie danych** w celu zapewnienia odpowiedniej jakości i poprawności danych
- Przykłady konfliktów i „zabrudzeń” danych:
  - różne formaty danych dla tego samego atrybutu (np. płeć zapisana jako: M/K, kobieta/mężczyzna, M/F, 0/1)
  - różne formaty daty (np. dd-mm-rrrr, rr-mm-dd)
  - różne długości pól (np. adres przechowywany raz w polu o 20 znakach, raz w polu o 50 znakach)
  - niespójne wartości tych samych danych spowodowane błędami przy wprowadzaniu

Proces ETL – transformacja danych

- Przykłady konfliktów i „zabrudzeń” danych:
  - różne standardy nazewnictwa: homonimy i synonimy
  - niezgodność między wartością atrybutu a jego nazwą (np. pole `NAZWA` może zawierać nazwę firmy lub nazwisko indywidualnego klienta)
  - brakujące wartości, które zgodnie ze schematem hurtowni powinny być wypełnione
  - redundantne informacje na temat jakiegoś obiektu świata rzeczywistego
  - …

Proces ETL – transformacja danych

- Podstawowe metody czyszczenia danych:
  - **konwersja i normalizacja** (transformacja i standaryzacja heterogenicznych formatów danych,
    np. ustalenie formatu daty na dd-mm-rrr)
  - **czyszczenie specjalne** (uspójnianie wartości pola na podstawie słownika synonimów, np. słowniki imion, nazw geograficznych, nazw farmaceutycznych)
  - **czyszczenie oparte na regułach** (np. zastąp „magister” przez „mgr”)

Proces ETL – transformacja danych

- Podstawowe metody czyszczenia danych:
  - „**deduplication**” - technika mająca na celu zapewnienie, że w hurtowni występuje jeden dokładny wpis dla każdego obiektu świata rzeczywistego
  - „**householding**” – technika mająca na celu grupowanie klientów indywidualnych na podstawie np. gospodarstwa domowego, organizacji, do której należą (dodatkowe aspekty marketingowe, np. „direct advertising”)

Jan Kowalski	123	ul. Kosowska
J. Kowalski	123	ulica Kosowska
Janek Kowalski	321	ul. Kossowska
Kowalski, Jan	123	al. Kossowska

Czy to ta sama osoba?

Proces ETL – ładowanie danych

- Po wyczyszczeniu danych następuje etap **ładowania danych** do hurtowni danych, którym zarządza **moduł integratora**
- Proces ładowania danych realizuje także
  - agregację danych
  - wzbogacenie danych o wymiar czasowy
  - łączenie danych z różnych źródeł
  - sprawdzanie więzów integralnościowych
  - budowanie struktur indeksowych

Proces ETL – ładowanie danych

- Ładowanie danych zwykle odbywa się w trybie wsadowym, jest to proces bardzo czasochłonny
- Aplikacja związana z modułem integratora musi pozwalać administratorowi na monitorowanie procesu ładowania, zawieszanie i wznawianie ładowania, restartowanie po awarii
- Proces ładowania danych może być traktowany jako jedna transakcja tworząca nową bazę danych

Odświeżanie danych

- Źródła danych nieprzerwanie zmieniają swoją zawartość, co wymusza uaktualnianie zawartości hurtowni danych
- Dostępność danych aktualnych wpływa na:
  - jakość wyników analiz
  - decyzje biznesowe
- Po załadowaniu danych zmiany następujące w źródłach są monitorowane poprzez moduł monitora i propagowane do hurtowni podczas procesu **odświeżania hurtowni**
- Proces odświeżania jest realizowany przez proces ETL

Odświeżanie danych

Istotne problemy przy procesie odświeżania danych:

- **Jak** odświeżać (sposób odświeżania)
  - w pełni
  - przyrostowo
- **Kiedy** odświeżać (moment odświeżania)
  - okresowo
  - automatycznie
  - na żądanie
- **Co przesyłać** (rodzaj przesyłanych obiektów)
  - dane (data shipping)
  - polecenia (transaction shipping)

Repozytorium metadanych

- **Repozytorium metadanych** to składnik hurtowni danych, przechowujący informacje wspomagające zarządzanie hurtownią
- Repozytorium zazwyczaj zawiera następujące informacje:
  - listę źródłowych baz danych wraz z opisem ich zawartości
  - opisy i charakterystyki bramek między bazami źródłowymi a hurtownią
  - schemat hurtowni
  - definicję perspektyw i danych wyliczalnych przechowywanych w hurtowni
  - opisy wymiarów i hierarchii
  - predefiniowane zapytania i raporty

Repozytorium metadanych

- Repozytorium zazwyczaj zawiera następujące informacje:
  - lokalizację tematycznych hurtowni danych
  - zasady czyszczenia, transformacji danych źródłowych
  - zasady odświeżania danych
  - profile użytkowników i grup użytkowników
  - dane dotyczące bezpieczeństwa hurtowni (autoryzacja, prawa dostępu itp.)

Literatura

- R. Kimball, M. Ross,
  The Data Warehouse Toolkit: The Complete Guide to Dimensional Modeling,
  John Wiley & Sons 2002
- M. Jarke, M. Lenzerini, Y. Vassiliou, P. Vassiliadis,
  Hurtownie danych. Podstawy organizacji i funkcjonowania,
  Wydawnictwa Szkolne i Pedagogiczne 2003
- Z. Królikowski,
  Hurtownie danych: logiczne i fizyczne struktury danych,
  Wydawnictwo Politechniki Poznańskiej 2007
- Materiały e-learningowe z przedmiotu „**Zaawansowane systemy baz danych**” (http://wazniak.mimuw.edu.pl/)