

Hurtownie danych - przegląd technologii

Robert Wrembel

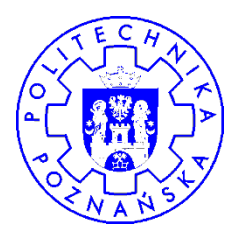
Politechnika Poznańska

Instytut Informatyki

`Robert.Wrembel@cs.put.poznan.pl`

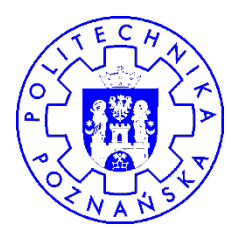
`www.cs.put.poznan.pl/rwrembel`





Plan wykładów

- ⇒ **Wprowadzenie - integracja danych**
- ⇒ **Architektury hurtowni danych**
- ⇒ **Modelowanie (ROLAP, MOLAP)**
- ⇒ **Zasilanie i odświeżanie hurtowni**
- ⇒ **Indeksowanie danych**
- ⇒ **Optymalizacja zapytań gwiazdzistych**
- ⇒ **Perspektywy zmaterializowane**
- ⇒ **Partycjonowanie danych i indeksów**
- ⇒ **Kompresja danych**
- ⇒ **Przetwarzanie równoległe**
- ⇒ **Wsparcie SQL dla analiz biznesowych**
- ⇒ **Metadane**
- ⇒ **Kierunki badawczo-rozwojowe**

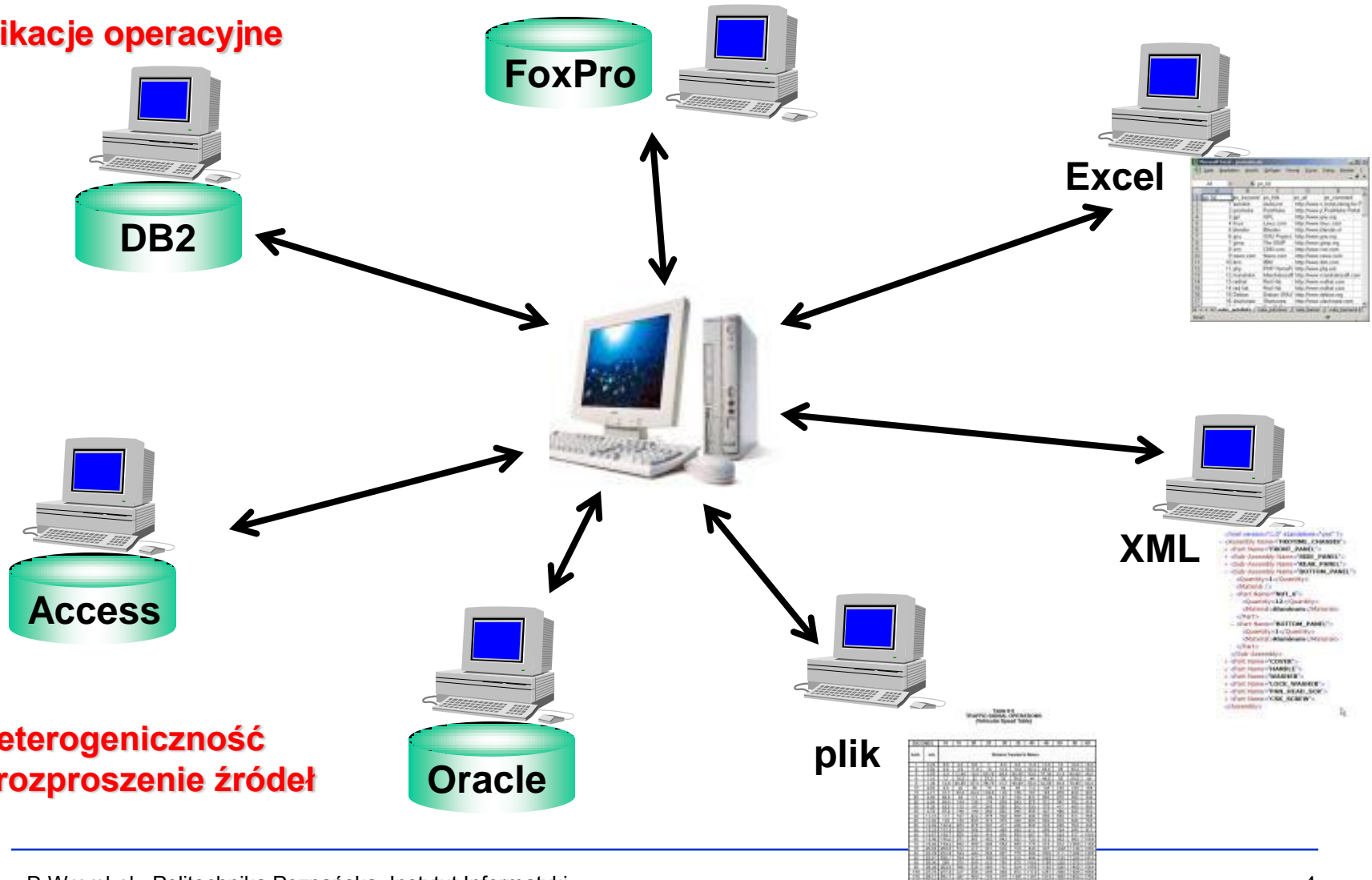


Architektury integracji danych

- ⇒ **Wprowadzenie do problematyki integracji danych**
- ⇒ **Architektury integracyjne**
 - **systemy mediacyjne**
 - **systemy hurtowni danych**
- ⇒ **HD i OLAP**

Problematyka integracji danych

aplikacje operacyjne

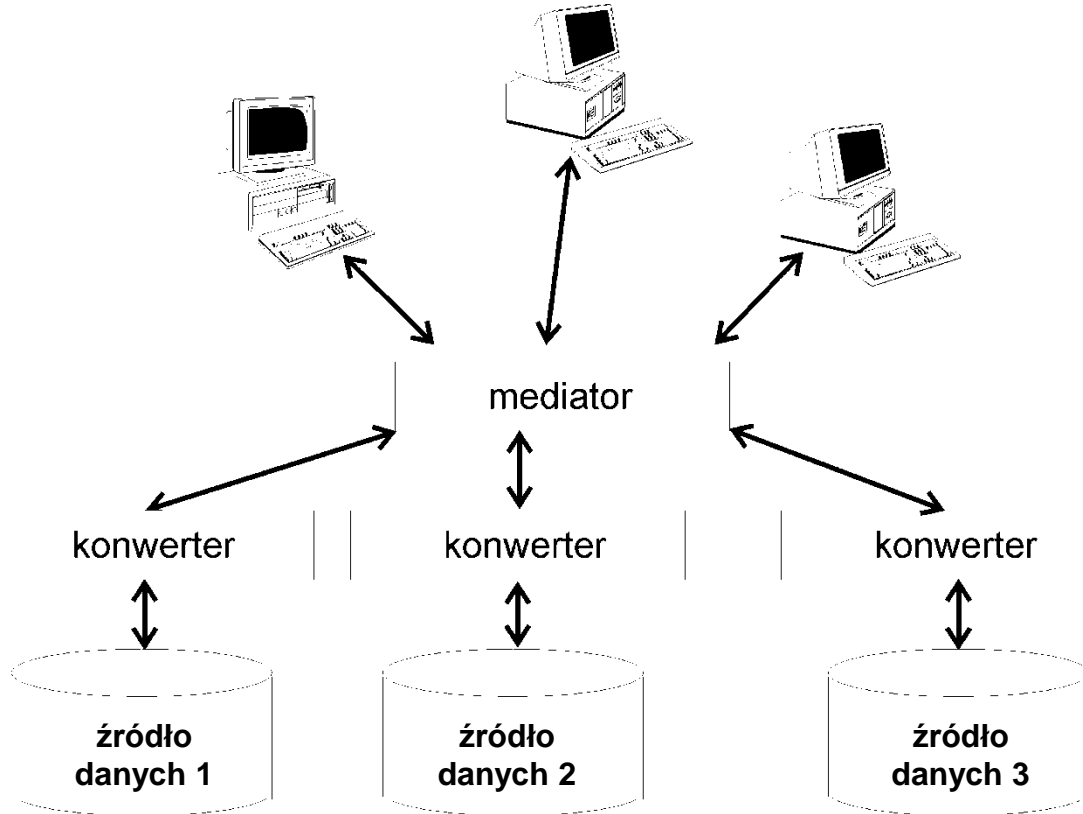




Heterogeniczność źródeł

- **Różni producenci/technologie**
- **Różna funkcjonalność**
 - bazy danych / nie bazy danych
 - dialekty SQL
 - sposoby dostępu i przetwarzania danych
- **Różne modele danych**
 - hierarchiczne, sieciowe
 - relacyjne
 - obiektowe
 - obiektowo-relacyjne
 - wielowymiarowe
 - semistrukturalne
- **Architektury integracyjne**
 - system mediacyjny
 - hurtownia (magazyn) danych

System mediacyjny



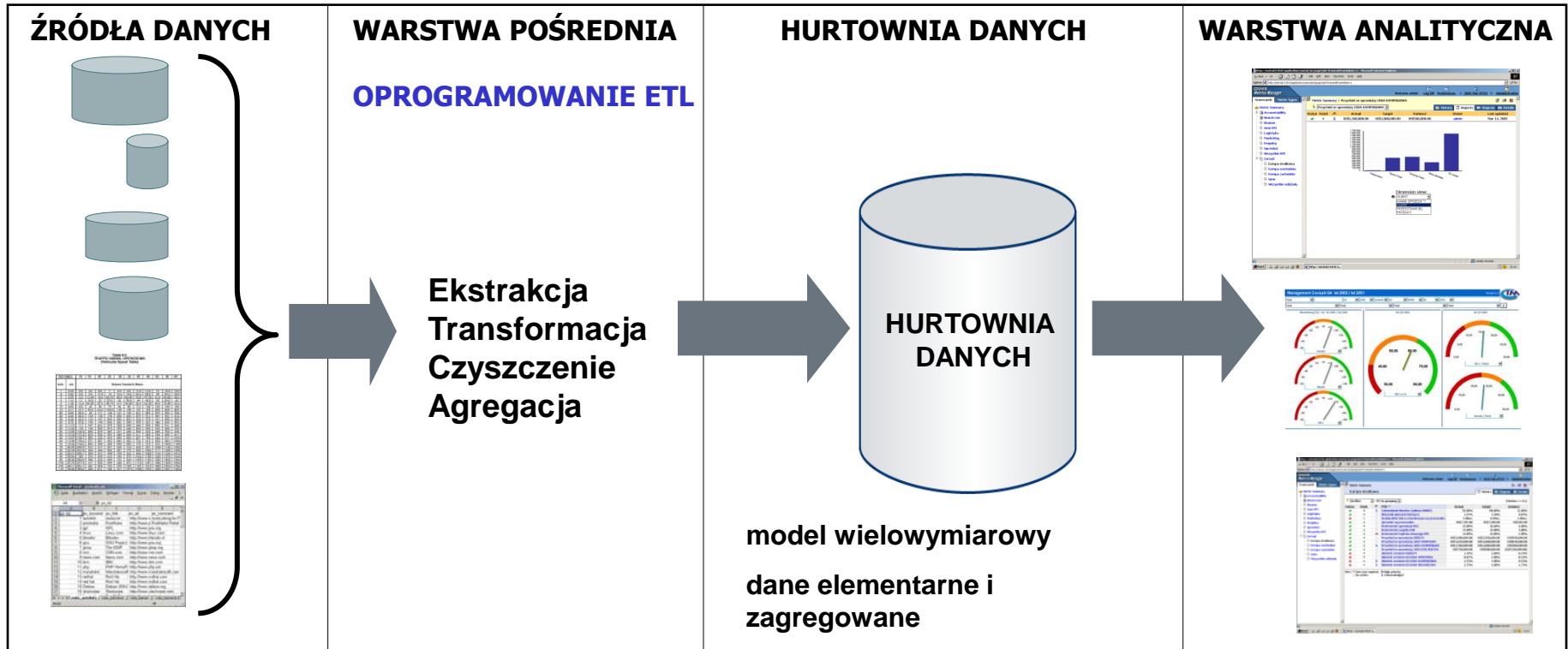
Wady

- czas dostępu do danych
- niedostępność źródeł
- konwersja zapytań i danych

➤ Zalety

- brak redundancji danych
- dostęp do danych aktualnych

Architektura 1 (podstawowa)



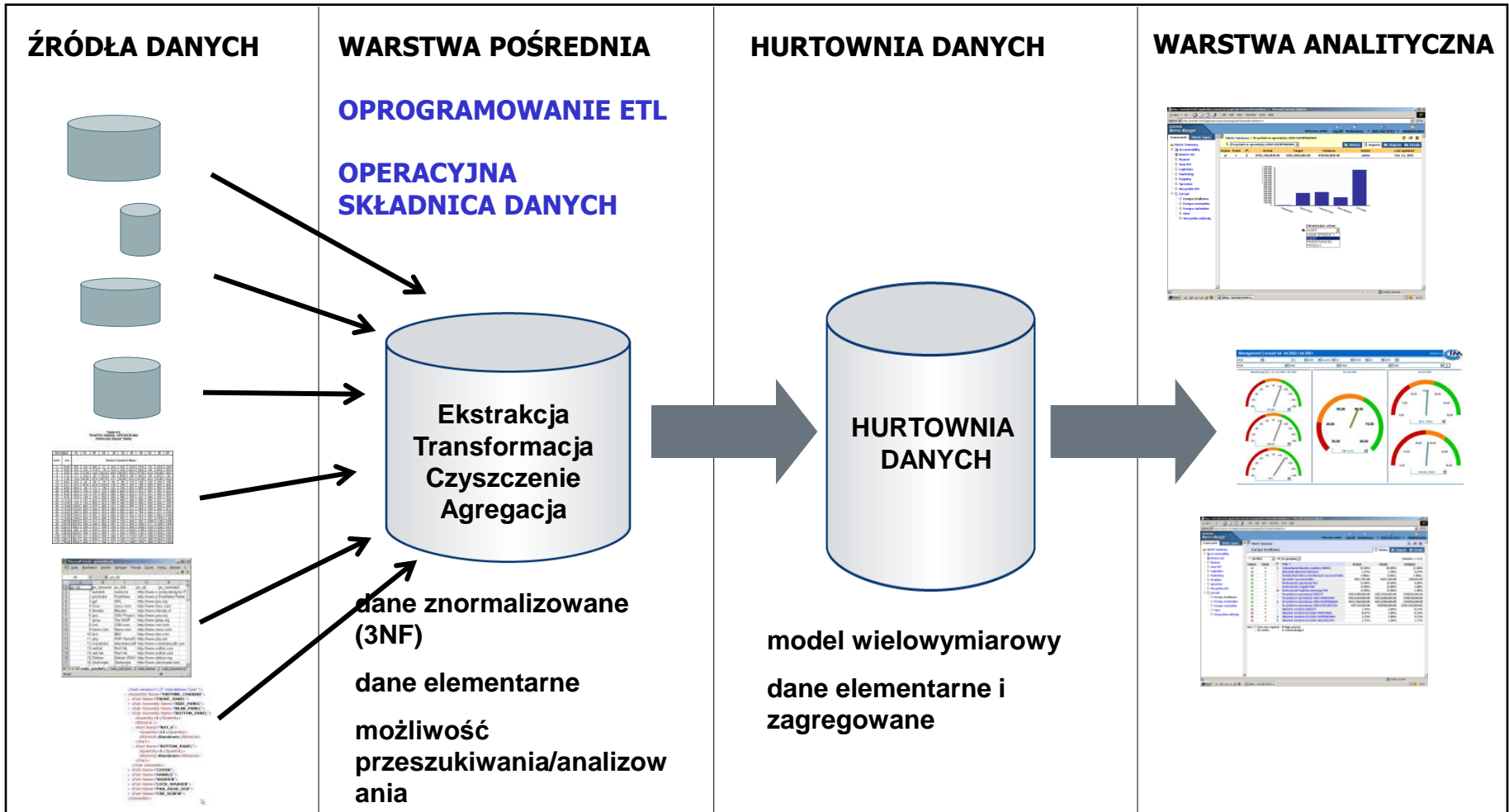
➤ Zalety

- dane zintegrowane (spójna struktura i wartości)
- szybkość dostępu do danych
- niezależność od awarii źródeł

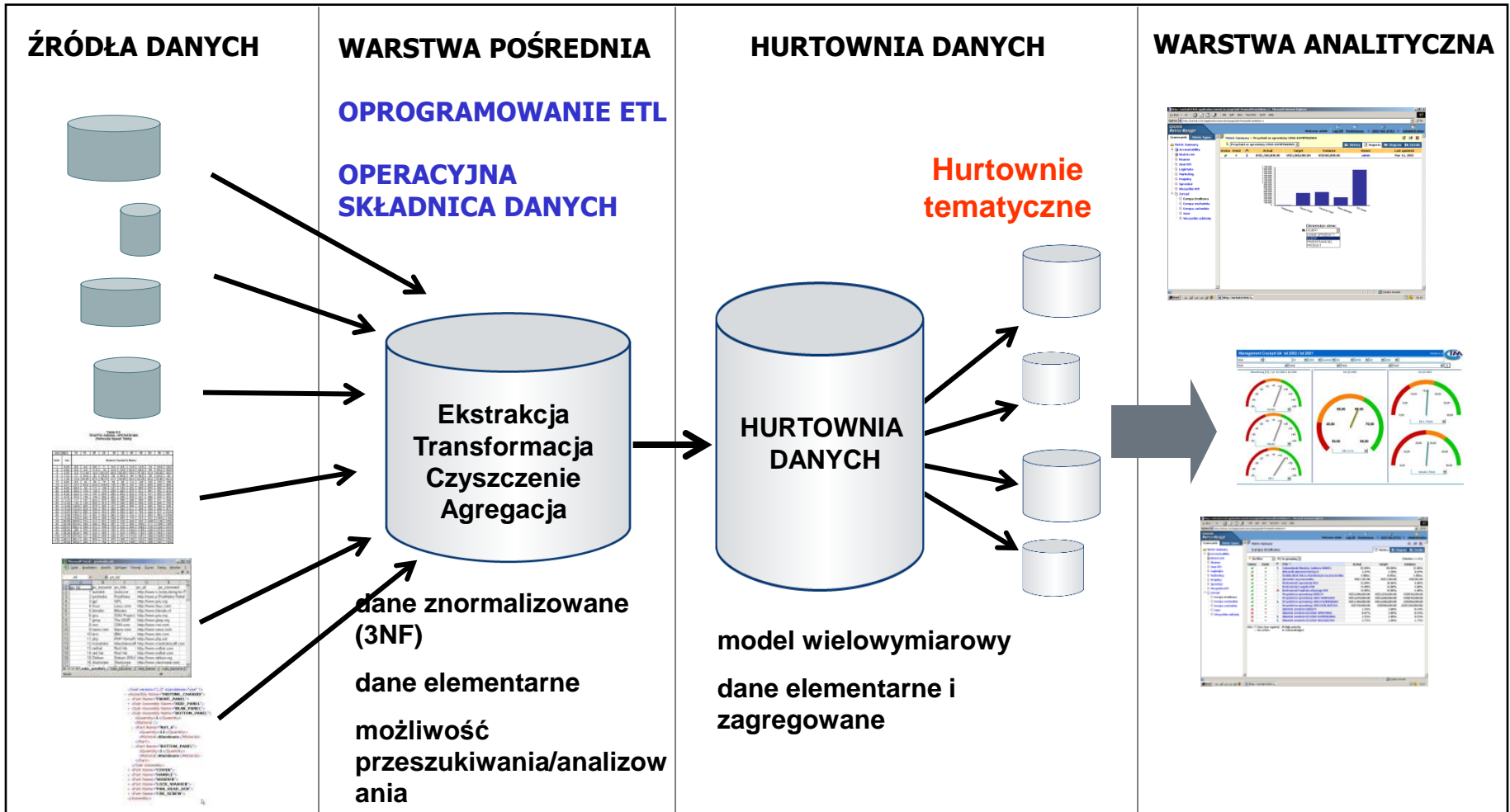
➤ Wady

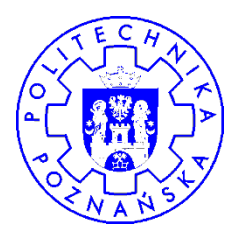
- redundancja danych
- odświeżanie danych

Architektura 2



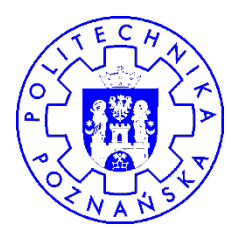
Architektura 3



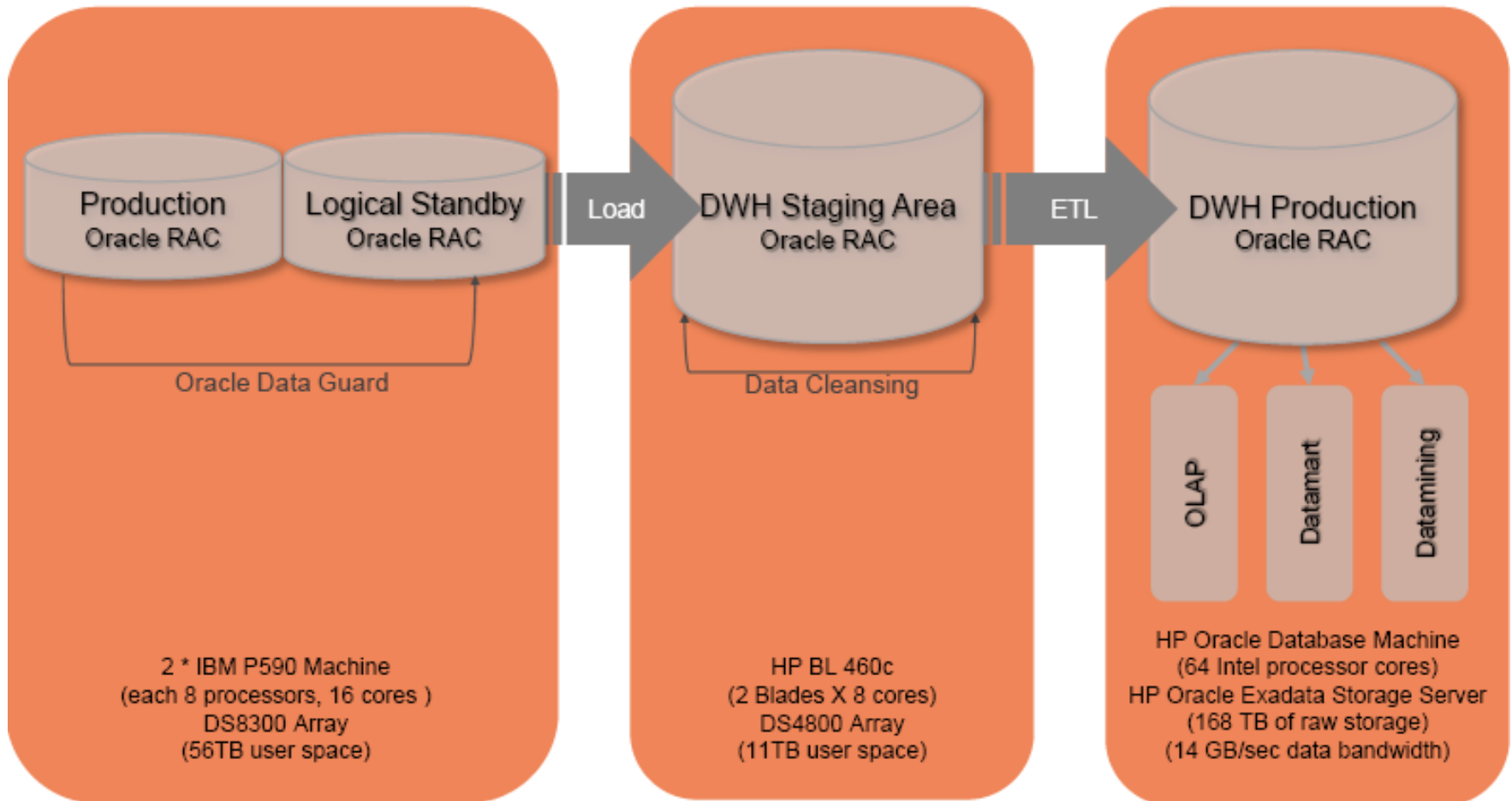


Systemy komercyjne

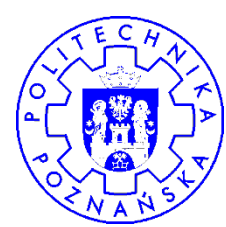
- ⇒ **Oracle8i, Oracle9i, Oracle10g/11g – Oracle Corporation,**
- ⇒ **DB2 UDB – IBM,**
- ⇒ **Sybase IQ, Sybase Adaptive Server Enterprise – Sybase, Inc.,**
- ⇒ **MS SQL Server – Microsoft,**
- ⇒ **SAP Business Warehouse – SAP,**
- ⇒ **Adabas C i Adabas D – Software AG,**
- ⇒ **Teradata – NCR Corporation,**
- ⇒ **Hyperion Essbase OLAP Server – Hyperion Solutions Corporation**
- ⇒ **Red Brick Warehouse – Red Brick Systems**



DWS Allegro



C. Maar, R. Kudliński: Allegro on the way from XLS based controlling to a modern BI environment. Konferencja HD i BI, Warszawa, 2008



Cele stosowania MD

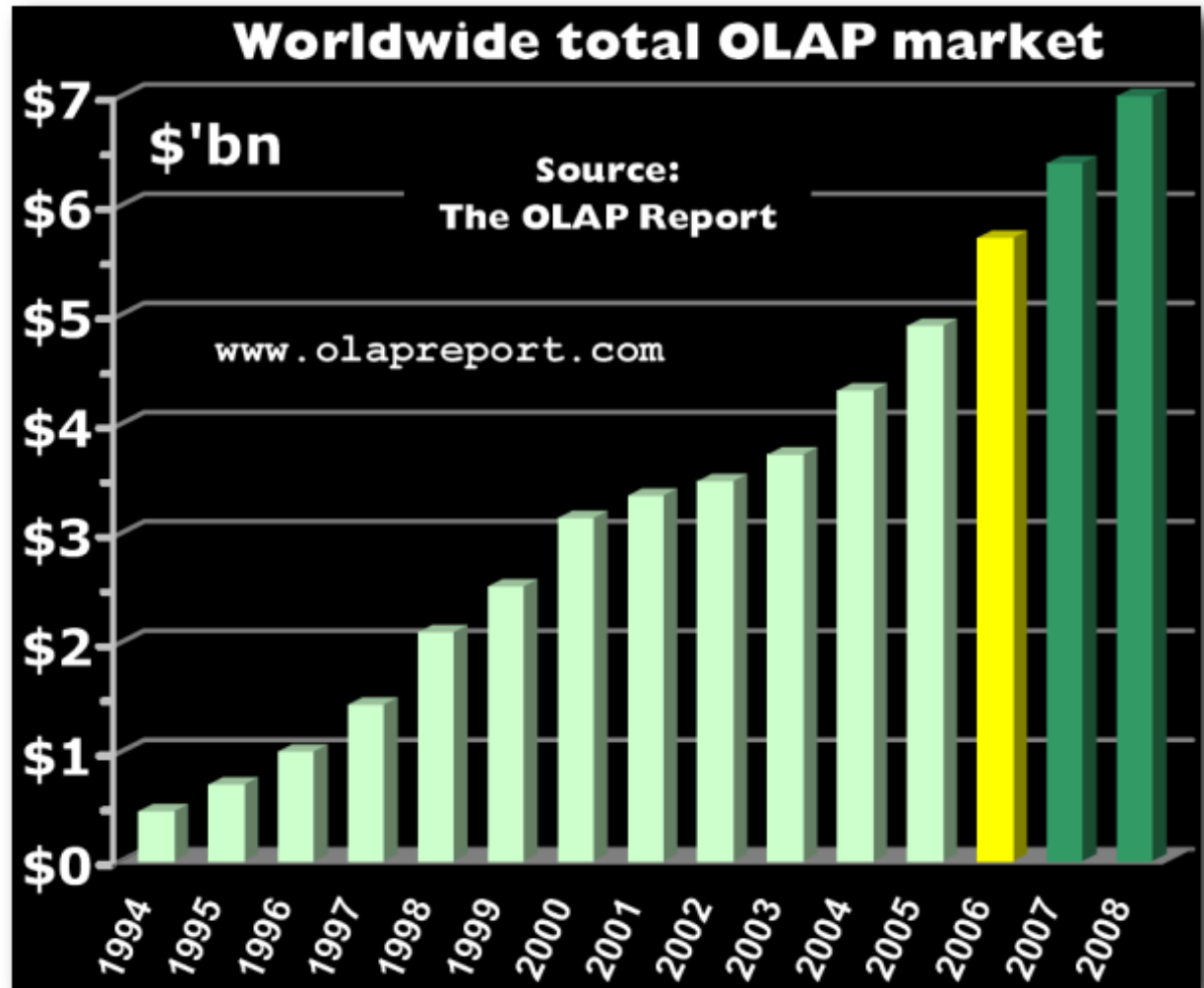
- 1. Zapewnienie jednolitego dostępu do wszystkich danych gromadzonych w ramach przedsiębiorstwa**
- 2. Dostarczenie technologii (platformy) przetwarzania analitycznego - **technologii OLAP****
 - wykonywanie **zaawansowanych analiz**, wspomagających zarządzanie przedsiębiorstwem, np.
 - analiza trendów sprzedaży
 - analiza nakładów reklamowych i zysków
 - analiza ruchu telefonicznego
 - **eksploracja danych**
 - **analiza rozwiązań alternatywnych (what-if analysis)**
 - symulowanie i przewidywanie przyszłości w MD

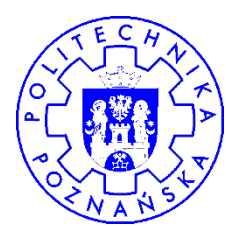


Technologia OLAP

⇒ Błyskawicznie rozwijający się rynek badawczy i technologiczny

- $9.9 * 10^9$ \$ w 2008 (METAGROUP)





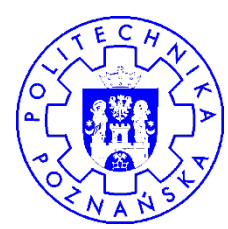
OLTP a OLAP

	OLTP	OLAP
użytkownik	"zwykły"	analityk
funkcja	bieżące operacje, kluczowe dla działania firmy	wspomaganie decyzji
dane	bieżące, elementarne	elementarne, zagregowane, historyczne
aplikacje	powtarzalność działań	ad hoc
dostęp	odczyt/zapis	odczyt
transakcja	krótka	długa (godziny)
l. przetwarzanych rek.	kilka, kilkadziesiąt	miliony lub więcej
l. użytkowników	kilkudzies., tysiące, setki tys.	kilku, kilkunastu
DB size	setki GB	dziesiątki TB
metric	przepustowość (l. transakcji w jednostce czasu)	czas odpowiedzi

Rozmiary HD

- ⇒ **Polska: HD Era GSM powyżej 30TB**
- ⇒ **Około 80% HD powyżej 1TB** (dane z XI 2007 wg. DMReview 17.04.08)
- ⇒ **Wall-Mart: powyżej 500TB (2005 r)**
- ⇒ **Amazon: powyżej 15TB (2005 r)**
- ⇒ **CERN Hadron Collider: 3TB dziennie (przewidywane)**
- ⇒ **NASA EOSDIS: 1000TB rocznie**





Projekt Systemu HD

(wg. Metodyki R. Kimball)

