# Wprowadzenie do Hurtowni Danych

Mariusz Rafało
mrafalo@sgh.waw.pl

# WPROWADZENIE DO HURTOWNI DANYCH

# Co to jest hurtownia danych?

- „Hurtownia danych jest zbiorem danych
  - zorientowanych tematycznie,
  - zintegrowanych,
  - przeznaczonych tylko do odczytu,
  - wersjonowanych czasem,
  - zorganizowanych dla wspierania celów zarządczych.”

- William H. Inmon

# Systemy wspomagania decyzji

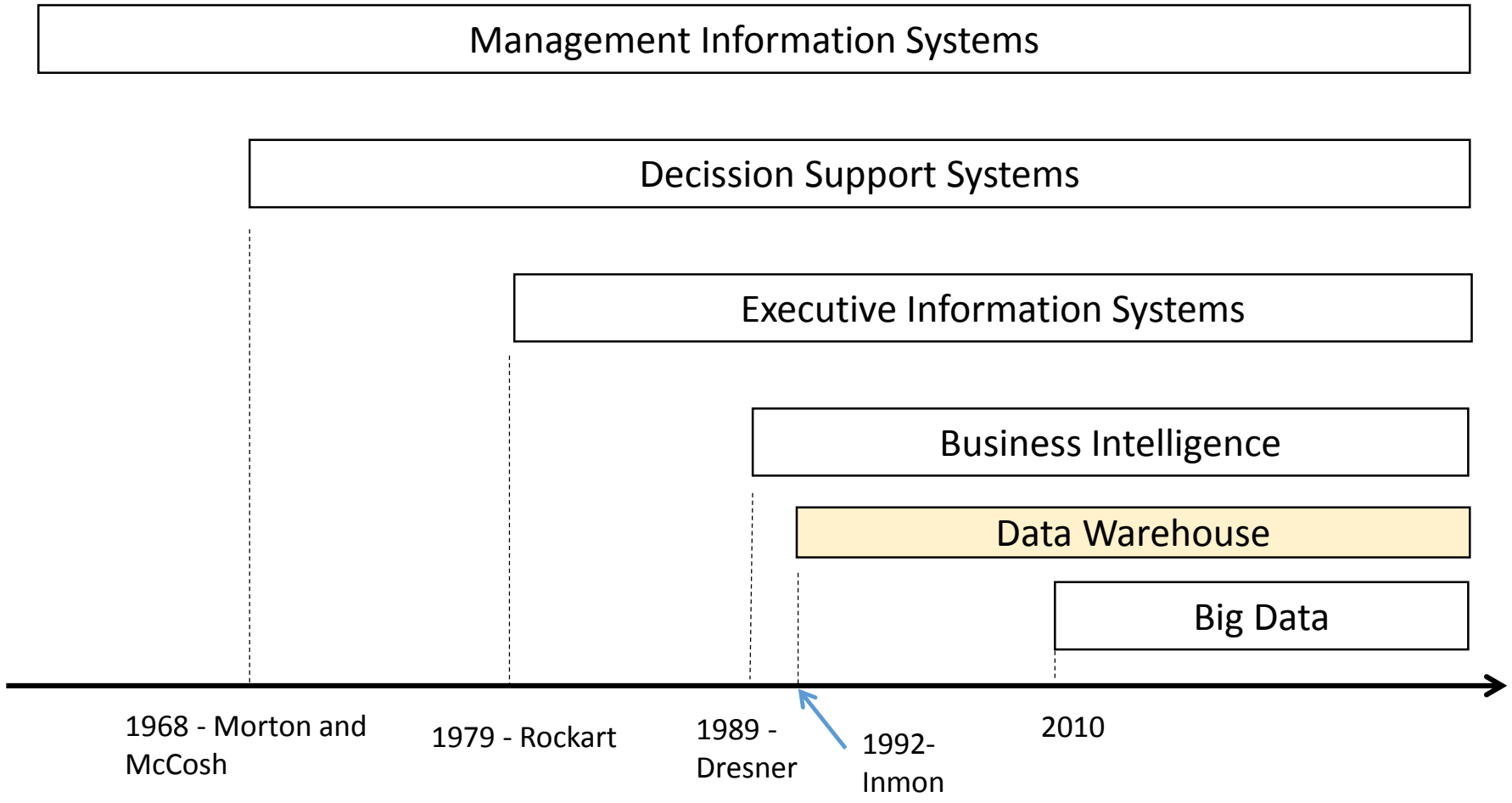
- System wspomagania decyzji – SWD *(Decision Support System – DSS)* - System, którego zadaniem jest dostarczenie użytkownikowi informacji umożliwiających przeanalizowanie sytuacji i podjęcie decyzji.

# Business Intelligence (BI)

- Business Intelligence includes concepts and methodologies for improvement of business decisions using facts and information from supporting systems

H.Dresner, 1989

# Systemy wspomagania decyzji



# Business Intelligence a Hurtownia Danych

- Business Intelligence czerpie wiedzę z systemów funkcjonujących w przedsiębiorstwie, następnie wykorzystuje tą wiedzę do wspomagania decyzji
- Hurtowania danych jest miejscem przechowującym dane z systemów funkcjonujących w przedsiębiorstwie
- Business Intelligence nie musi korzystać z Hurtowni Danych
- Hurtownia Danych jest zawsze elementem Business Intelligence

# Specyfika danych w hurtowni danych(1)

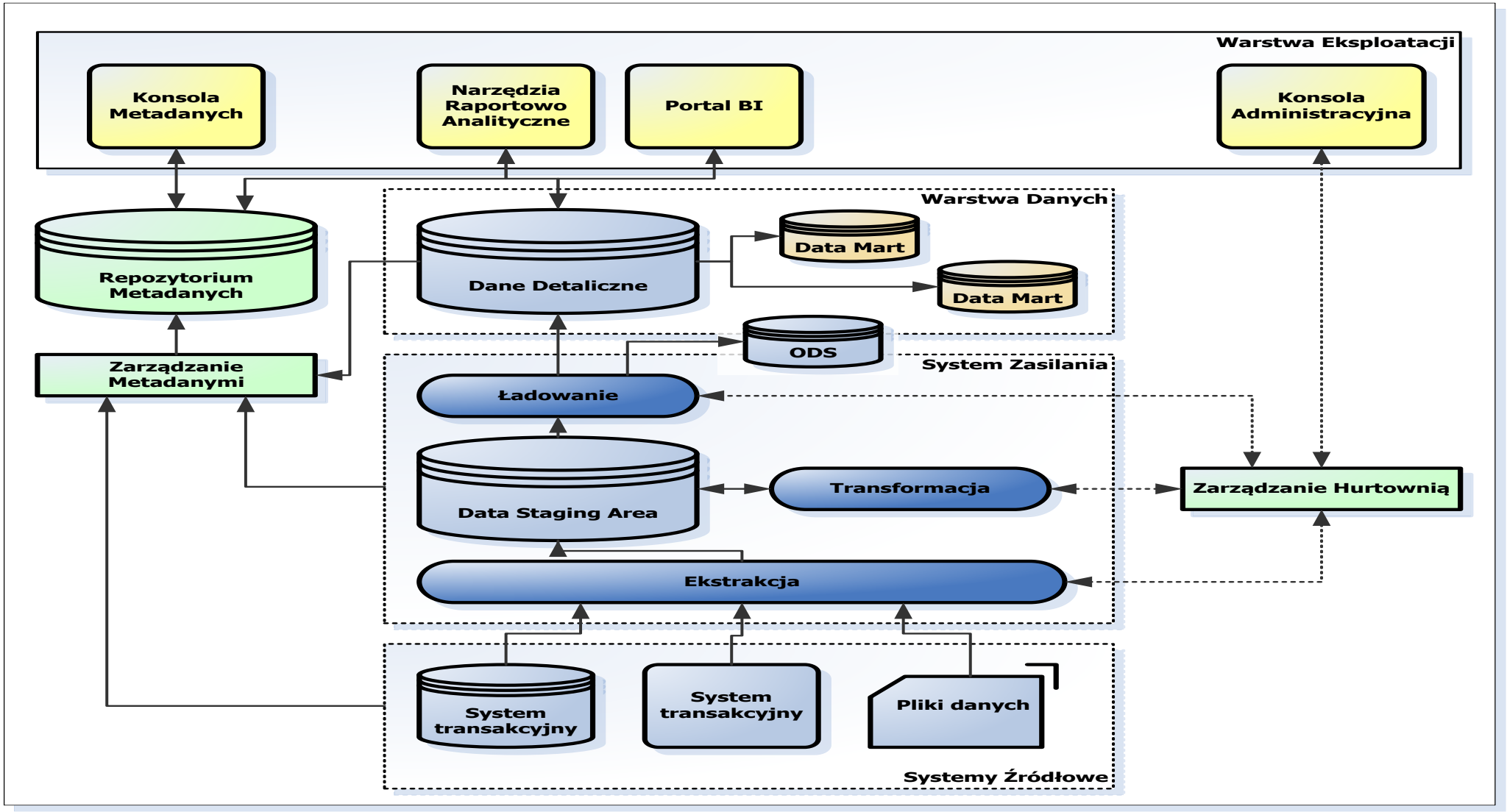
- Dane muszą być zorientowane tematycznie
  - dane są zorganizowane według tematów mających kluczowe znaczenie dla organizacji, a nie według funkcjonalności, czy też podziału organizacyjnego
  - wynika to z faktu, że analiza funkcjonowania organizacji i jej otoczenia wymaga globalnego spojrzenia na dane
- Dane muszą być zintegrowane
  - dane gromadzone w hurtowni danych integrują  informacje o poszczególnych tematach pochodzące z wielu źródeł, tak aby dawać w miarę pełny opis sytuacji
- Dane muszą być nieulotne
  - dane przechowywane w hurtowni są przeznaczone tylko do odczytu
  - nie usuwamy danych historycznych

# Specyfika danych w hurtowni danych(2)

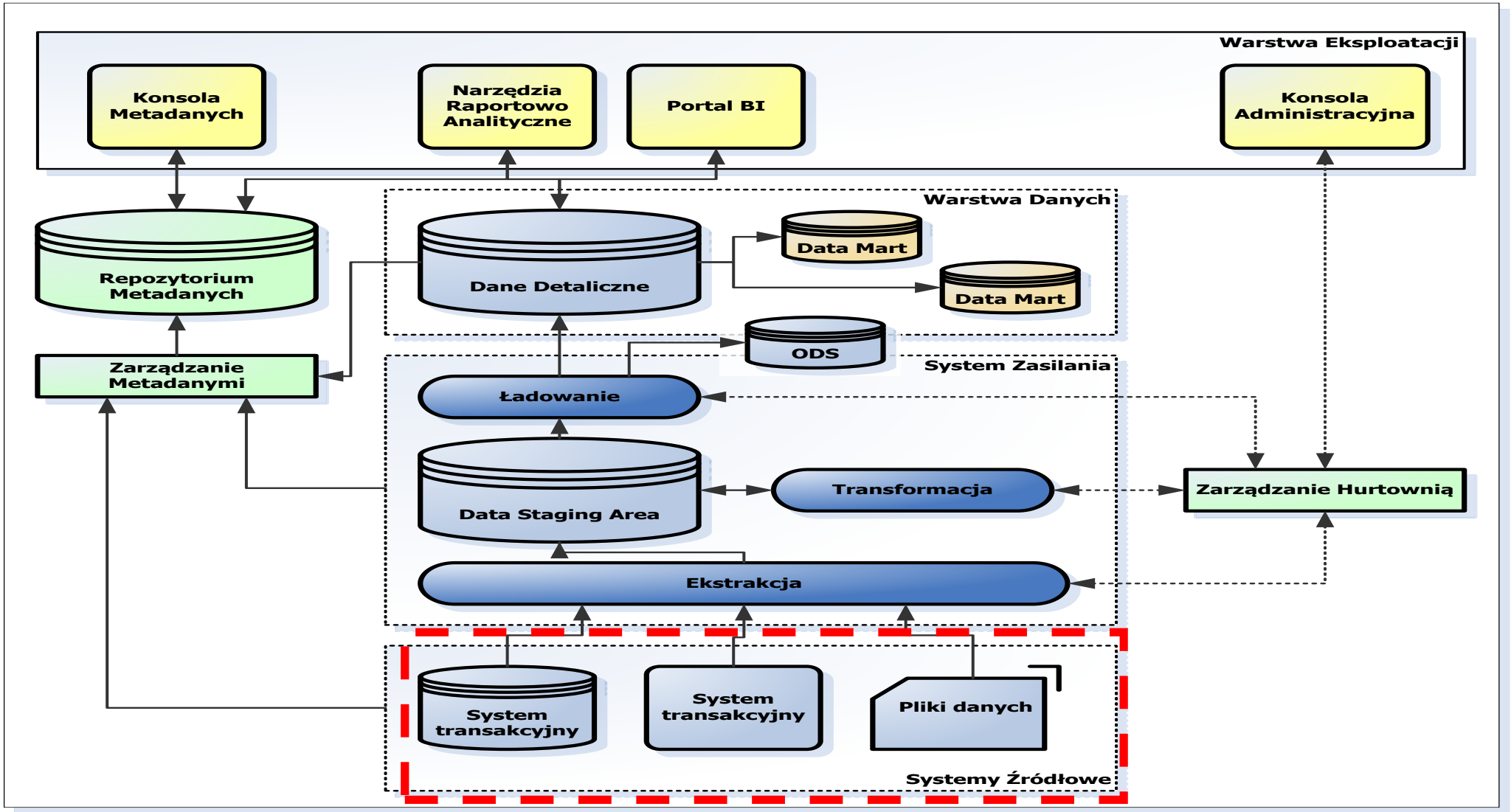
- Dane muszą być wersjonowane wg czasu powstania
  - zawartość hurtowni danych stanowią nie tylko dane obrazujące obecną sytuację, ale również - lub przede wszystkim - dane historyczne
  - gromadzenia danych historycznych umożliwia badanie nie tylko sytuacji obecnej, ale również trendów, zmian otoczenia, pozwala odpowiedzieć na pytanie „dlaczego?” oraz dokonywać prognoz
- Dane muszą być zorganizowane pod kątem działalności zarządczej, analitycznej
  - działalność zarządcza i analityczna polega między innymi na analizowaniu sytuacji, trendów, zmian, wzorców oraz prognozowaniu

# ARCHITEKTURA HURTOWNI DANYCH

# Architektura HD



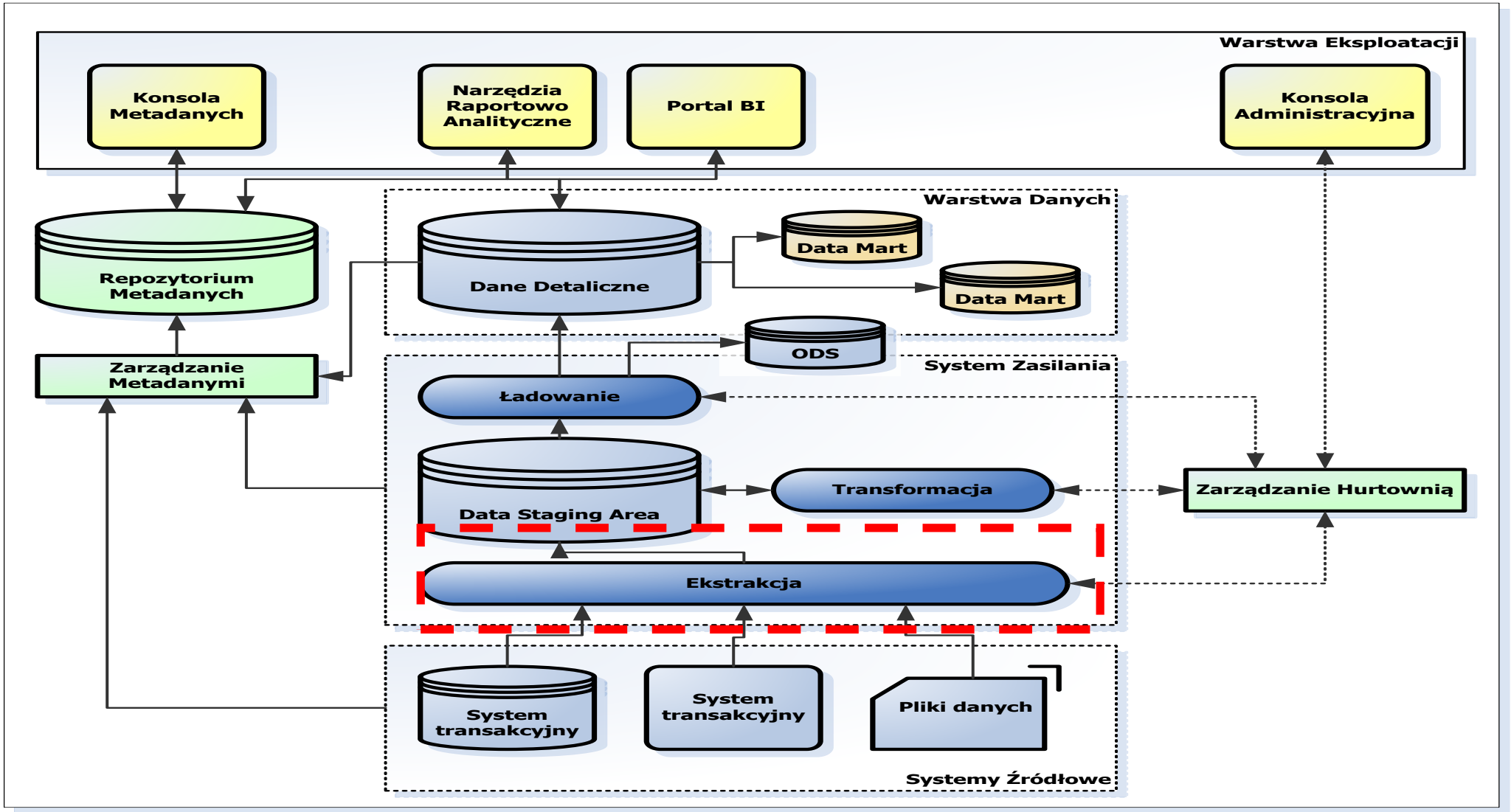
# Systemy źródłowe



# Charakterystyka danych operacyjnych (źródłowych)

- Ściśle określone zapytania dotyczące danych na detalicznym poziomie
- Często brak danych historycznych
- Brak organizacji tematycznej
- Brak integracji pojęć pomiędzy systemami
- Sztuczne klucze
- Typowo dane są silnie znormalizowane

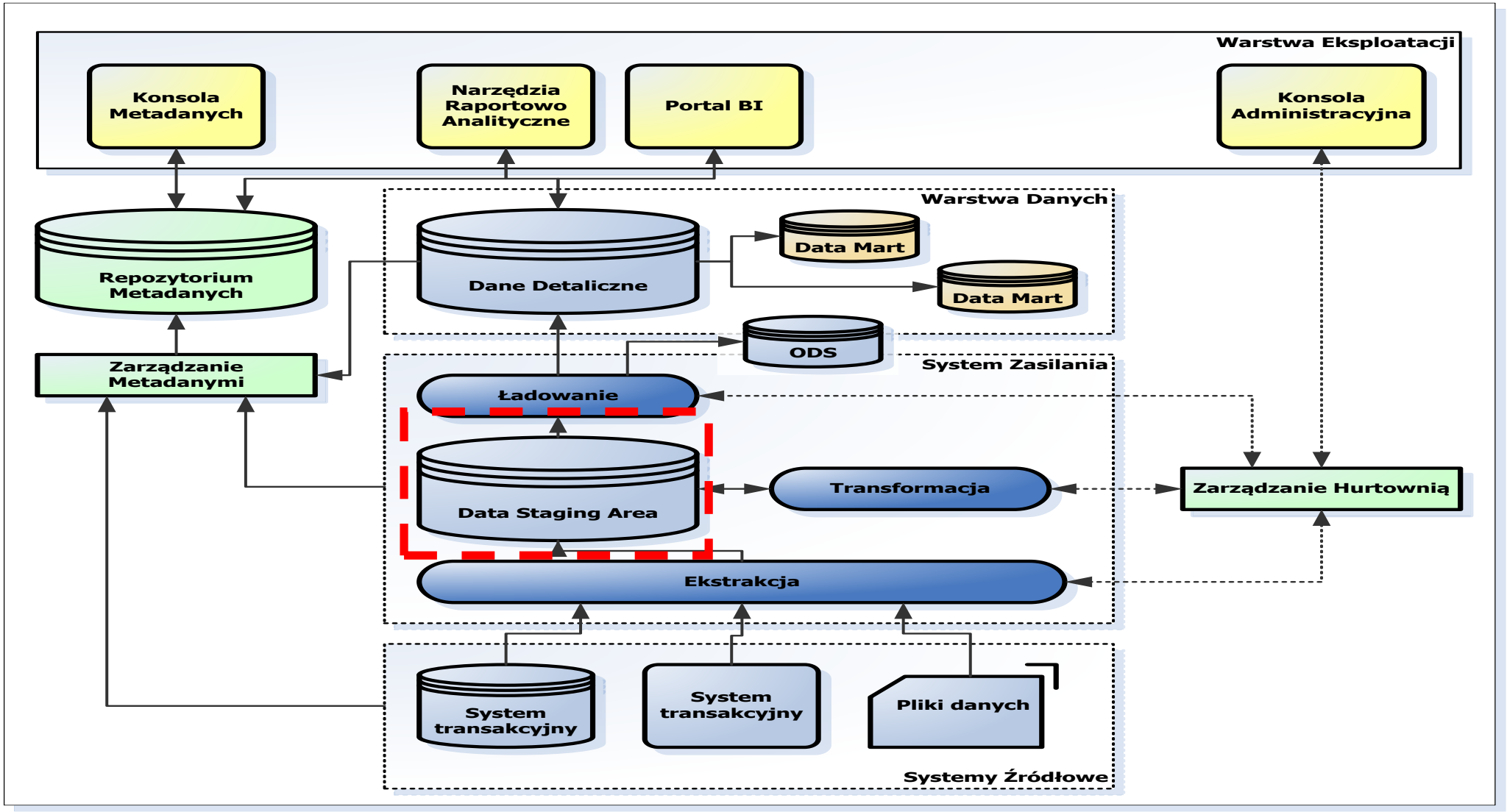
# Ekstrakcja


Warstwa Eksploatacji
Konsola Metadanych
Narzędzia Raportowo Analityczne
Portal BI
Konsola Administracyjna
Warstwa Danych
Repozytorium Metadanych
Dane Detaliczne
Data Mart
Data Mart
ODS
Zarządzanie Metadanymi
System Zasilania
Ładowanie
Data Staging Area
Transformacja
Zarządzanie Hurtownią
Ekstrakcja
System transakcyjny
System transakcyjny
Pliki danych
Systemy Źródłowe

# Ekstrakcja danych

- Pobieranie danych z systemów źródłowych
  - dane trafiają do obszaru danych tymczasowych (ang. Data Staging Area)
- Różne modele
  - „pull” – odpytywanie systemów źródłowych
  - „push” – systemy źródłowe same wysyłają dane
- Ekstrakcja
  - Pełna
  - Przyrostowa
- Metody określania przyrostu danych
  - Timestamp
  - Sekwencja
  - Replikacja
  - Log
  - Migawka
  - Dedykowane rozwiązania Change Data Capture(CDC)

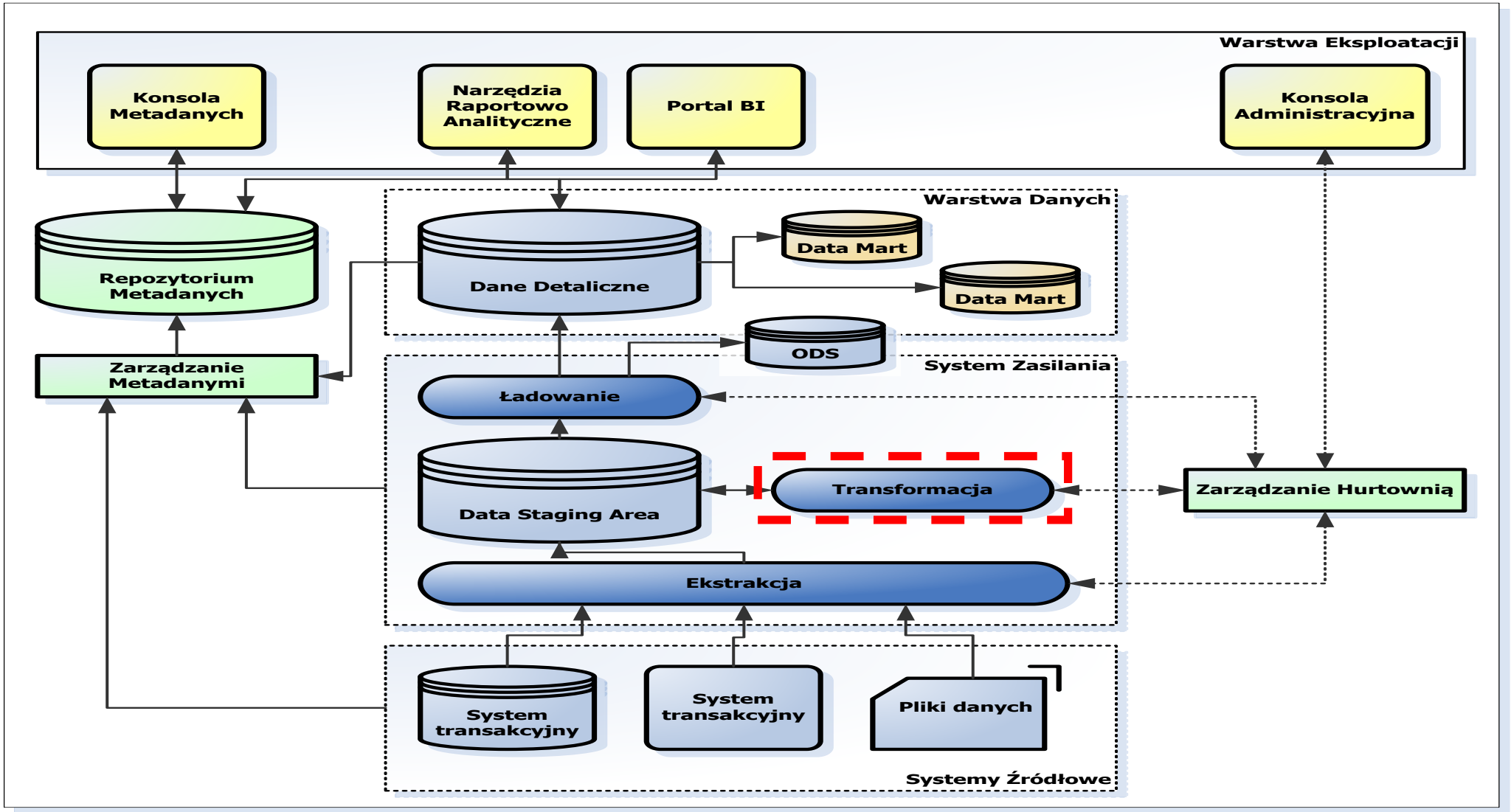
# Obszar danych tymczasowych



# Obszar Danych Tymczasowych (Data Staging Area)

- Gromadzi dane przetwarzane w procesie zasilania
  - Dane pobrane z systemów źródłowych (przyrost)
  - Dane pomocnicze wykorzystywane w zasilaniu
    - Słowniki
    - Tabele „look-up”
    - Parametry
- W tym obszarze działają procesy
  - Transformacji
  - Czyszczenia
  - Poprawy jakości

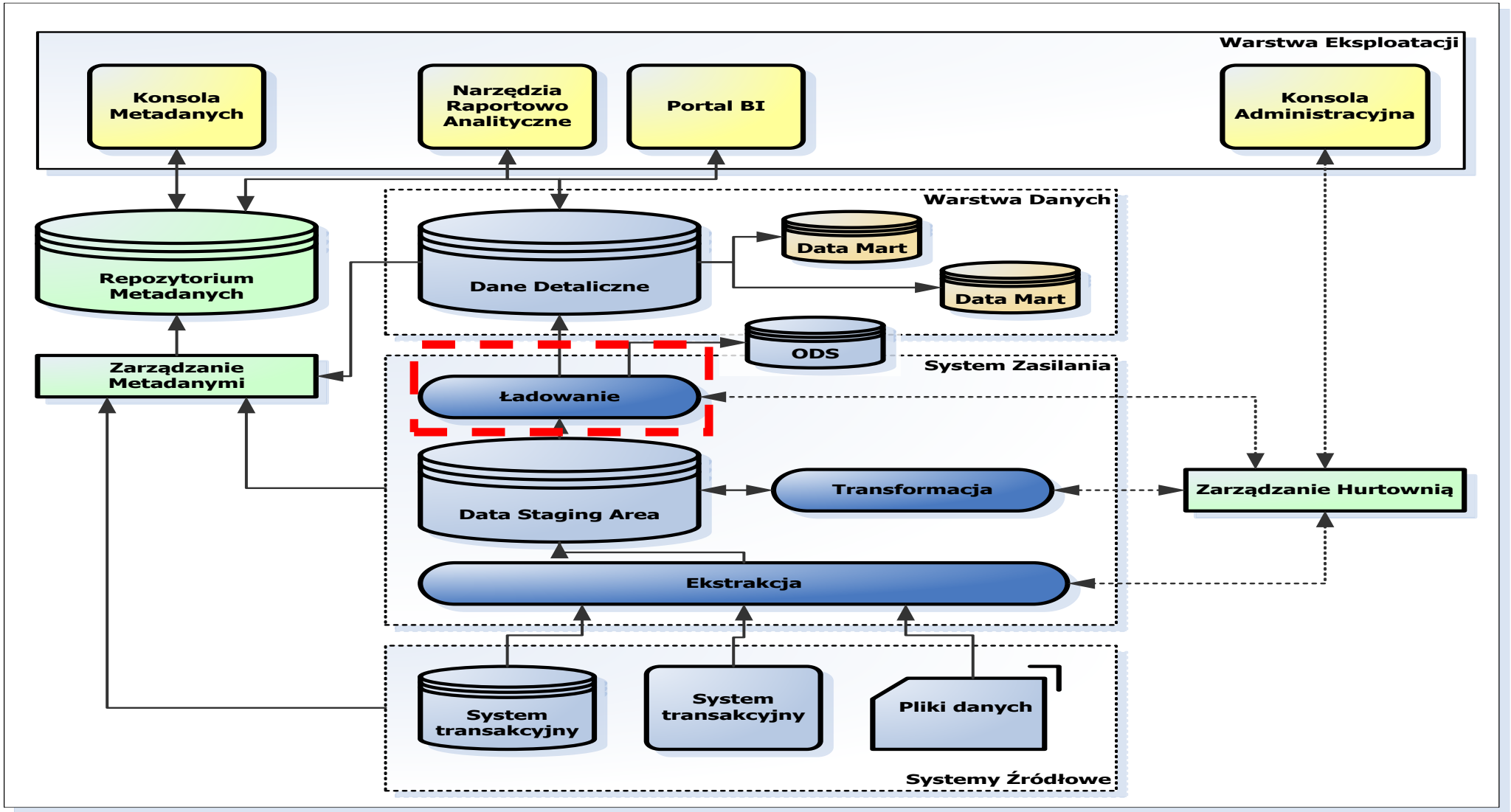
# Transformacja



# Transformacja danych

- Przekształcanie danych do postaci pożądanej w Hurtowni Danych
  - Czyszczenie (braki, błędy, uzupełnianie)
  - Usuwanie duplikatów
  - Łączenie
  - Integracja
  - Kalkulacje
  - Agregacje
  - Standaryzacja
  - …

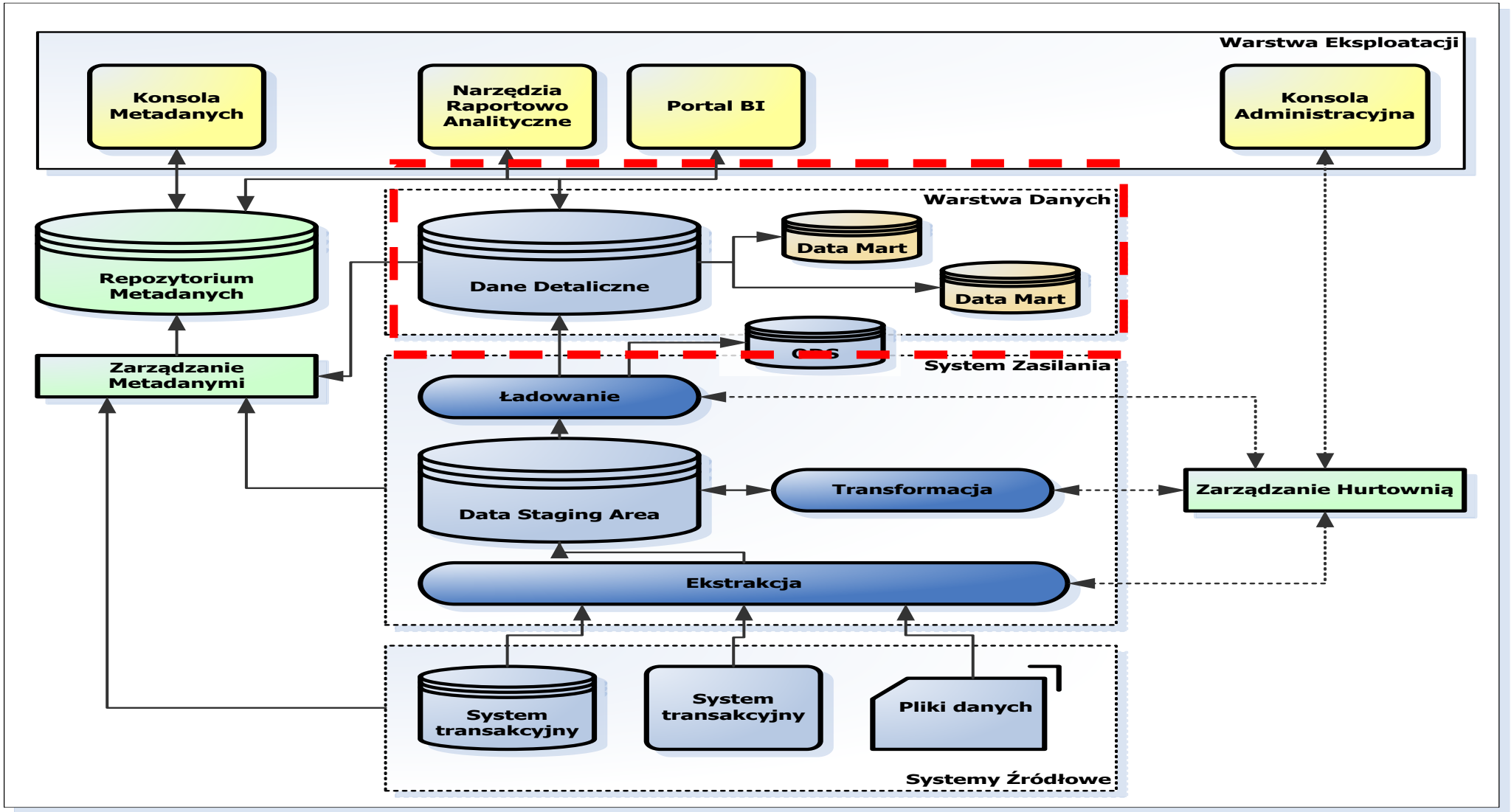
# Ładowanie


Warstwa Eksploatacji
Konsola Metadanych
Narzędzia Raportowo Analityczne
Portal BI
Konsola Administracyjna
Warstwa Danych
Repozytorium Metadanych
Dane Detaliczne
Data Mart
Data Mart
ODS
Zarządzanie Metadanymi
System Zasilania
Ładowanie
Data Staging Area
Transformacja
Zarządzanie Hurtownią
Ekstrakcja
System transakcyjny
System transakcyjny
Pliki danych
Systemy Źródłowe

# Ładowanie danych

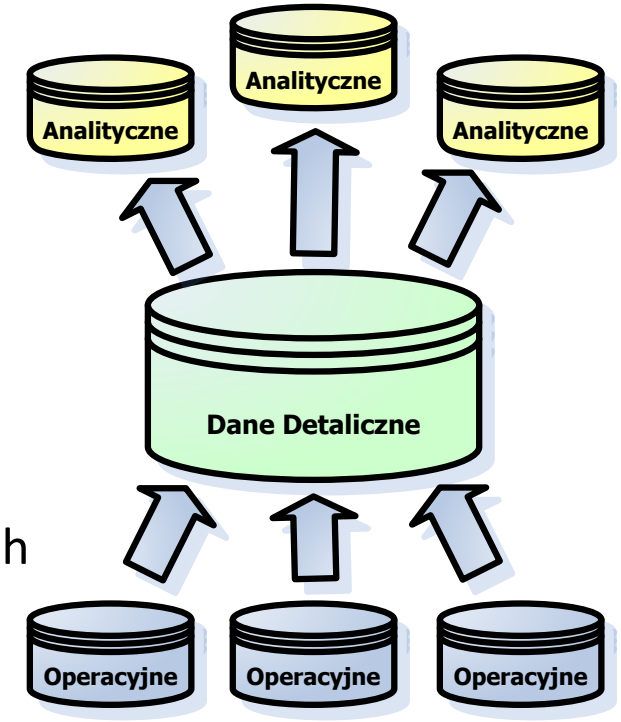
- Ładowanie gotowych danych do baz danych Hurtowni Danych
- Wykorzystanie specjalnych trybów ładowania danych
  - Bulk
  - Ładowanie do różnych partycji
  - Ładowanie on-line / off-line
  - Ładowanie równoległe
- Ekstrakcja, transformacja i ładowanie danych są elementami procesu ETL(Extract, Transform, Load)

# Warstwa danych

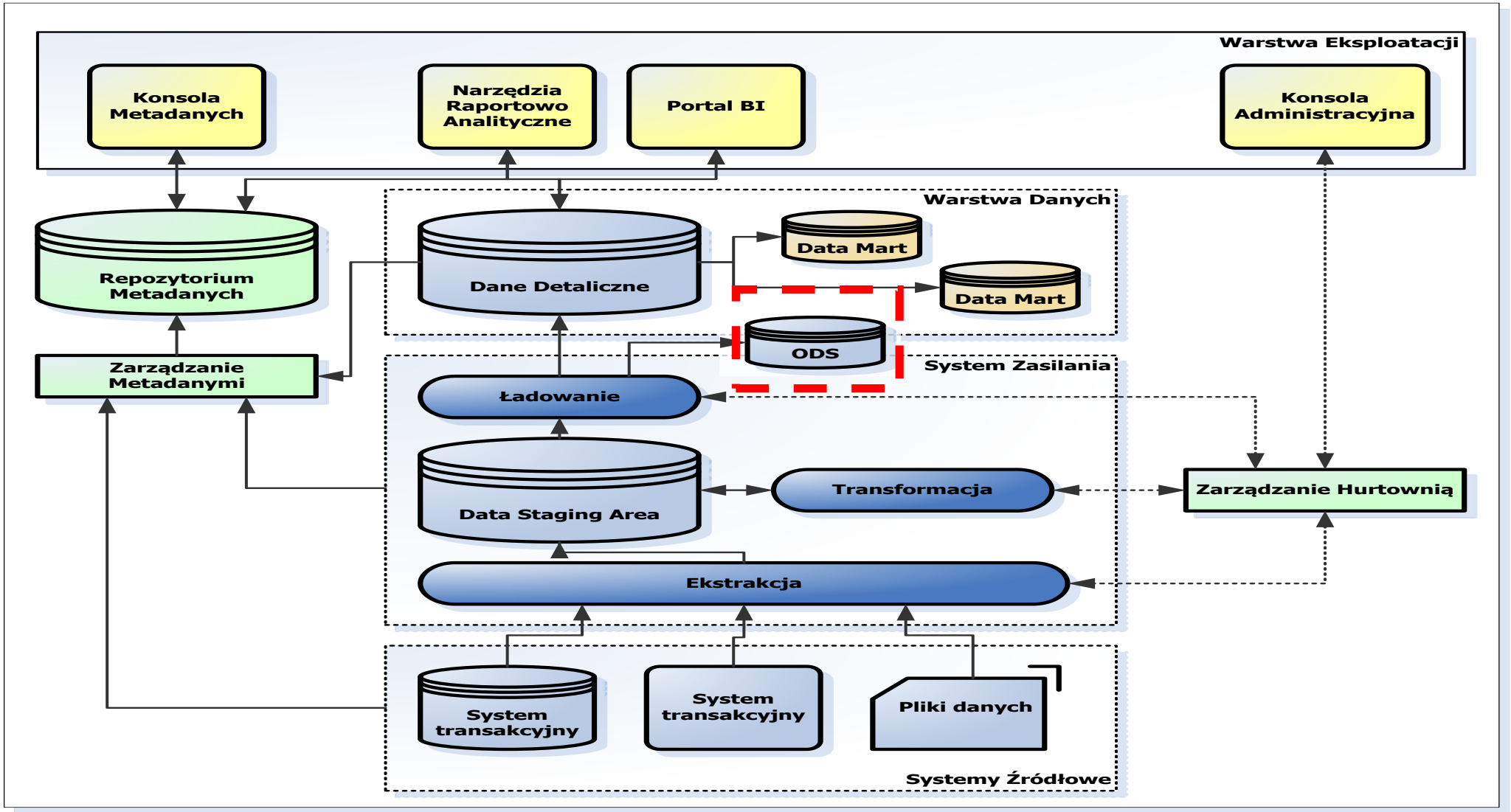


# Warstwa Danych

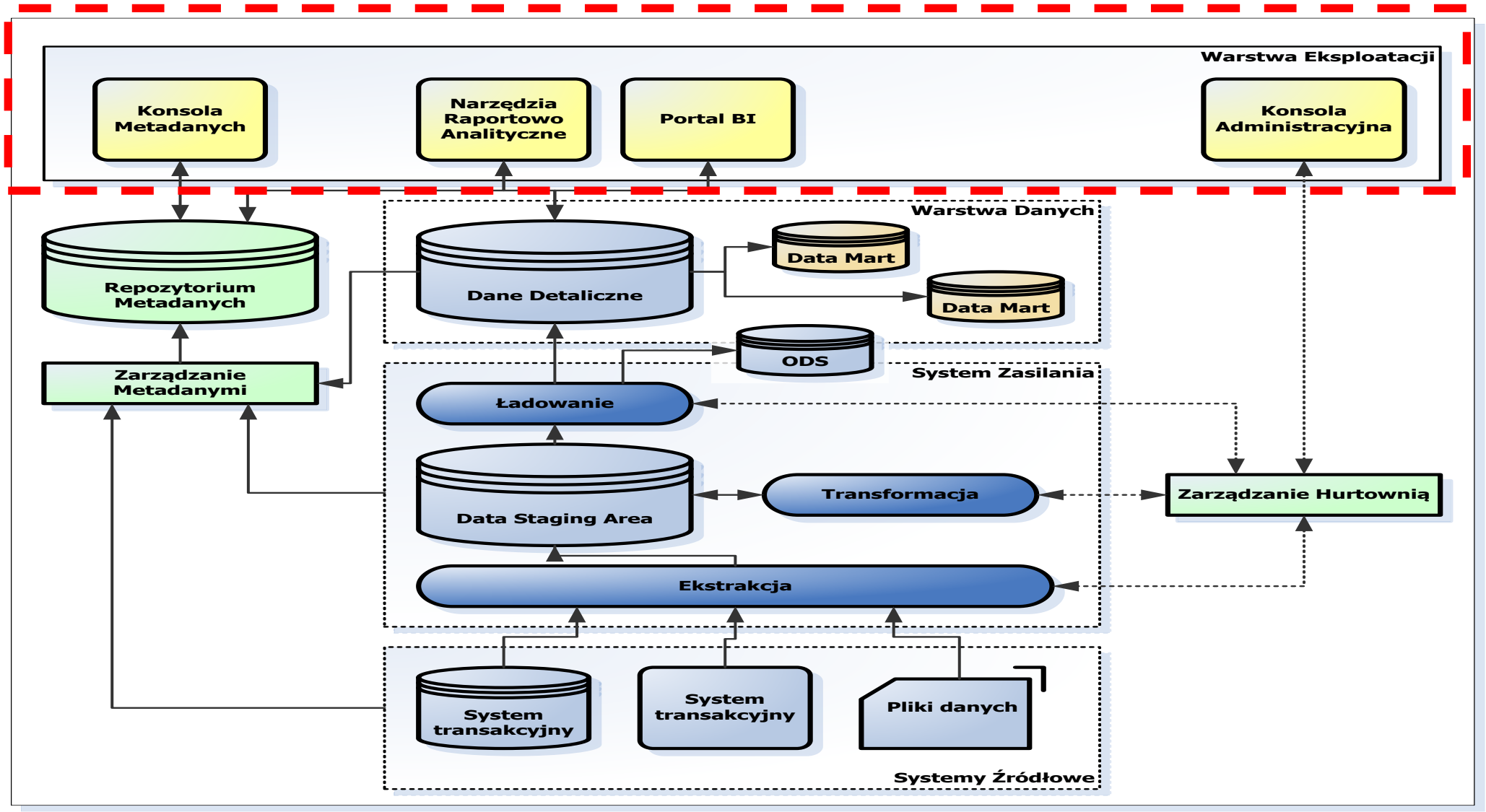
- Gromadzi dane Hurtowni Danych
  - Czasem nazywana warstwą „prezentacji” lub „dostępu”, gdyż udostępnia dane klientom
- Zazwyczaj występują dwie warstwy danych
  - Dane detaliczne
    - zintegrowane i zorganizowane tematycznie
    - na detalicznym poziomie
    - gromadzone w modelach znormalizowanych
    - nacisk na elastyczność modelu
  - Dane analityczne (zagregowane)
    - zbudowane poprzez przekształcenie danych detalicznych
    - często zagregowane
    - występuje redundancja danych
    - dostosowane dla konkretnych analiz



# Operacyjny magazyn danych - Operational Data Store



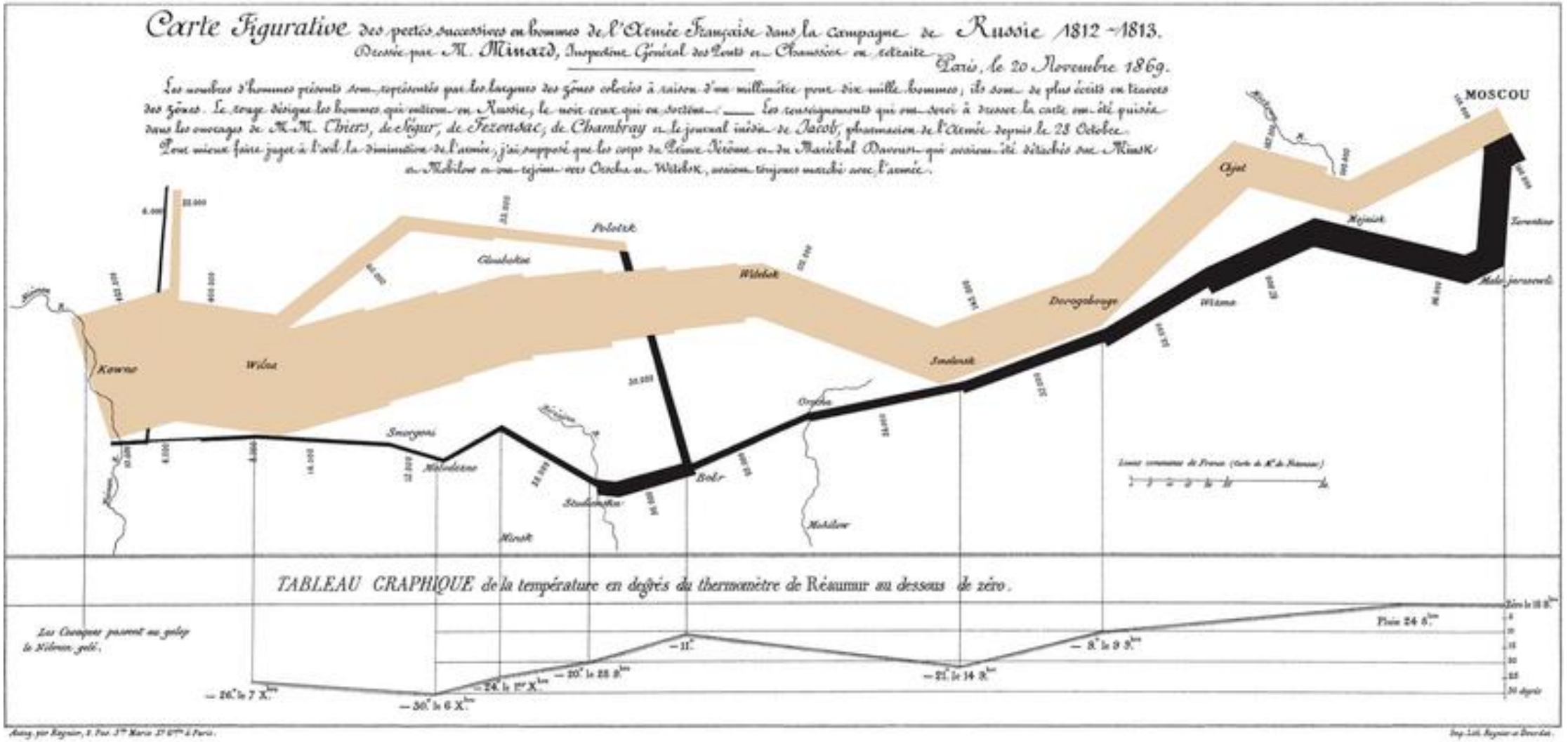
# Warstwa eksploatacyjna



# Warstwa eksploatacyjna

- Udostępnia narzędzia i usługi dostępu do danych gromadzonych w Hurtowni Danych
  - Narzędzia raportowo-analityczne
  - Portale BI
  - Narzędzia dla administratora danych
  - Narzędzia dostępu do danych z poziomu innych systemów

# Raportowanie i wizualizacja (XIXw)

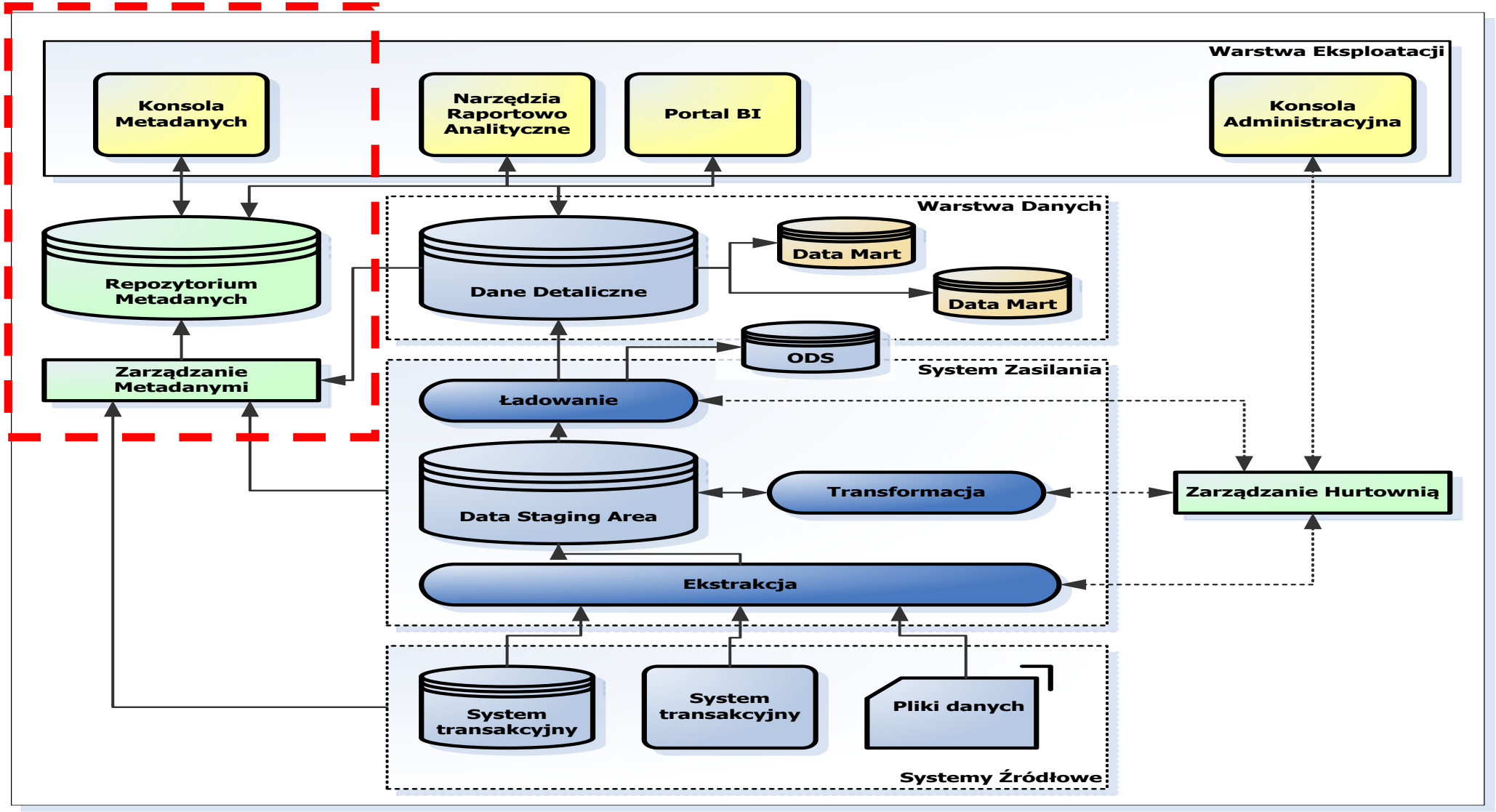



http://en.wikipedia.org/wiki/Charles_Joseph_Minard

# Raportowanie i wizualizacja (XXIw)


https://www.youtube.com/watch?v=UqJ1x2hAK70

# Metadane

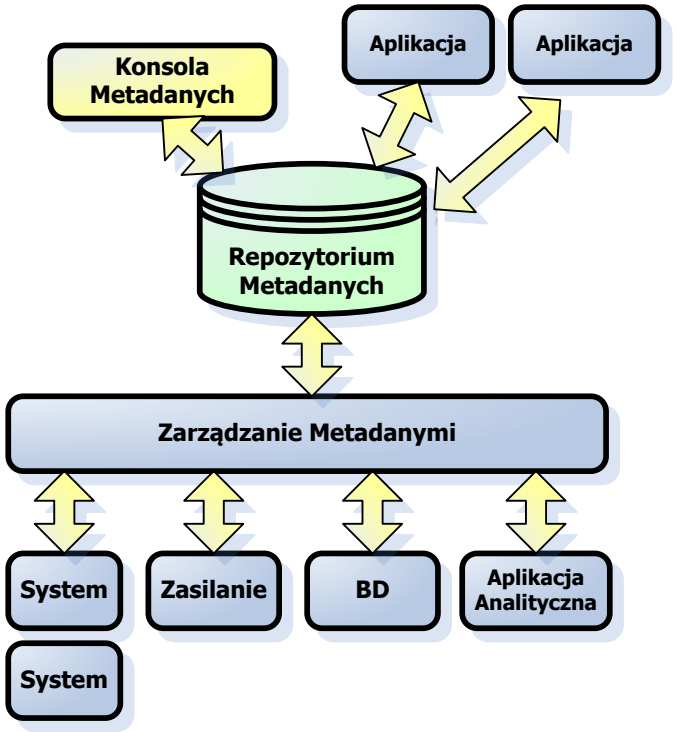

Warstwa Eksploatacji
Konsola Metadanych
Narzędzia Raportowo Analityczne
Portal BI
Konsola Administracyjna
Warstwa Danych
Repozytorium Metadanych
Dane Detaliczne
Data Mart
Data Mart
ODS
Zarządzanie Metadanymi
System Zasilania
Ładowanie
Data Staging Area
Transformacja
Zarządzanie Hurtownią
Ekstrakcja
System transakcyjny
System transakcyjny
Pliki danych
Systemy Źródłowe

# Metadane Hurtowni Danych

- Metadane, to „dane o danych”
  - Opisują gromadzone dane oraz procesy realizowane i wspierane przez Hurtownię Danych

- Metadane są używane:
  - Podczas eksploatacji Hurtowni Danych
  - Podczas utrzymania i rozwoju

- Metadane dzielimy na:
  - Biznesowe
  - Techniczne
  - Procesowe



# Metadane Biznesowe

- Definicyjne
  - Opis zawartość Hurtowni z punktu widzenia użytkownika
  - Informacje o tym jak używać danych, jaki jest ich kontekst
  - Kto jest właścicielem danych
- Nawigacyjne
  - Opis gdzie szukać danych
  - Jakie są związki pomiędzy danymi
- Jakość
  - Opis jakości danych
- Pochodzenie
  - Opis skąd pochodzą dane, jakim podlegają przemianom

# Metadane Techniczne

- Źródła danych
- Modele danych w Hurtowni
- Metody i procesy zasilania, transformacji i czyszczenia
  - Opis przetwarzania danych
- Mapowanie danych źródłowych
- System bezpieczeństwa
- Audyt wykorzystania(Data Lineage)

# Metadane Procesowe

- Historia przetwarzania danych
  - Kto, co, kiedy, dlaczego?
  - Wynik i efekt przetwarzania
- Logi przetwarzania
  - Błędy
  - Statystyki
  - Dane dotyczące wskaźników wydajności i wykorzystania infrastruktury
- Historia zmian modeli danych Hurtowni Danych
- Historia zmian uprawnień i użycia Hurtowni Danych
  - Obciążenie przez systemy zewnętrzne i użytkowników

Dziękuję za uwagę