MAPA POJĘĆ – HURTOWNIE DANYCH, DATA ANALYTICS, BIG DATA

1. Hurtownia danych (Inmon, Kimball)

Bill Inmon:
Hurtownia danych jest zorientowaną tematycznie (ang. subject-oriented), zintegrowaną (ang. integrated), nieulotną (ang. nonvolatille), prezentującą wymiar czasowy (ang. time-variant) kolekcją danych wspierającą proces decyzyjny.

System hurtowni danych przechowuje, integruje dane wokół wybranego obszaru przedsiębiorstwa bardziej niż na obszarach aplikacyjnych różnorodne dane pochodzące z systemów źródłowych, zapewniając uspójnienie formatu danych. Dane w hurtowni są nieulotne co oznacza, że raz wprowadzone do niej informacje nie mają prawa ulec zmianie. Po załadowaniu dane można jedynie czytać. Nowe dane zawsze są dodawane do już istniejącego zbioru. Każda informacja ma nadany wymiar czasowy.
Cechą charakterystyczną modelu hurtowni danych w ujęciu Inmona tzw. „jedna wersja prawdy", która wiąże się z brakiem redundancji danych. Każda informacja ma taką samą wartość dla każdego użytkownika, który ją odczyta.

R. Kimball
Ralph Kimball, w odróżnieniu od Billa Inmona, koncentruje się na funkcjonalnym aspekcie:
Hurtownia danych jest, strukturalnie przystosowaną do wykonywania efektywnych zapytań i przeprowadzania analiz, kopią danych transakcyjnych.

Nacisk kładziony jest na sposób pozyskania danych niezależnie od wybranego typu architektury.
Liczą się funkcje dostępne dla użytkownika systemu, przyjazność interfejsu oraz czas uzyskania odpowiedzi na zapytanie. Zaproponowany model hurtowni to model wymiarowy – polega na określeniu tabel faktów i wymiarów, gdzie fakty zawierają metryki, a wymiary atrybuty. Odpowiedni dobór wymiarów do tablic faktów jest fundamentem powodzenia w tym podejściu.

	Bill Inmon	**Ralph Kimball**
Metodologia i architektura		
Ogólne podejście do metodologii	Top-down	Bottom-up
Architektura	Korporacyjna hurtownia danych obejmująca bazy departamentowe	Hurtownie tematyczne pojedynczych procesów biznesowych
Złożoność metody	skomplikowana	prosta
Odniesienie się do znanych metodologii	Metodyka spiralna	Proces czterech kroków, wywodzący się z metod RDBMS
Projekt fizyczny	Dokładny	Ogólny
Modelowanie danych		
Orientacja danych	Zorientowanie na dane lub na temat	Zorientowanie na proces
Narzędzia	Tradycyjne, związane z DBMS: ERD, DIS	Modelowanie wymiarowe, wywodzące się z modelowania relacyjnego
Zaangażowanie użytkowników	Niskie	Wysokie
Filozofia podejścia		
Główni uczestnicy	Profesjonaliści IT	Końcowi użytkownicy
Miejsce w organizacji	Integralna część korporacyjnej fabryki informacji (CIF)	Transformacja i przejęcie danych operacyjnych
Cel	Dostarczyć należyte	Dostarczyć rozwiązania,

	rozwiązanie techniczne oparte na sprawdzonych metodach i technologiach baz danych	które ułatwiają użytkownikom końcowym bezpośrednie wyszukiwanie danych z zachowaniem rozsądnego czasu reakcji

Problemy hurtowni danych

- niedoszacowanie zasobów do ładowania danych,
- niepoprawnie zaprojektowany proces ETL
- wysokie zapotrzebowanie na zasoby
- duże koszty utrzymania
- długi czas realizacji projektu (wyrażony w latach)
- złożoność integracji w obszarze technologicznym
- homogenizacja danych – w dużych systemach zauważane są tendencje do uproszczenia, nieuzasadnionego ujednolicenia danych, przez co następuje ich homogenizacja, która, przykładowo, może prowadzić do zatracenia cech charakterystycznych poszczególnych, różnych, atrybutów, pomimo, że obydwa mogą dotyczyć tego samego faktu, np. sprzedaż nieruchomości (wartość) oraz wynajem nieruchomości (także wartość),

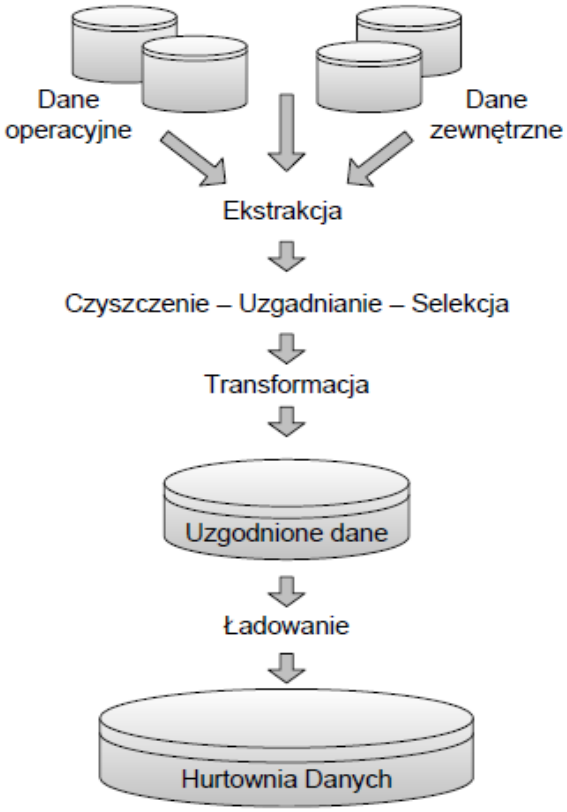
2. Hurtownia tematyczna (data mart)

Hurtownia tematyczna jest okrojoną hurtownią danych lub jej podzbiorem wspierającym realizację zadań konkretnego procesu biznesowego lub danego działu firmy. Zgromadzone w niej dane mogą mieć postać elementarną lub zagregowaną. Może ona działać jako samodzielny system lub jako połączenie do centralnego repozytorium danych. Zaletą hurtowni tematycznej jest jej przewaga w stosunku trudności budowy pełnej, hurtowni danych. Skala przedsięwzięcia jest mniejsza co za tym idzie ilość danych w niej zgromadzonych jest ograniczona, a jej budowa łatwiejsza. Hurtownia tematyczna koncentruje się wyłącznie na wymaganiach użytkownika związanego z pojedynczym działem firmy lub ściśle określonym procesem biznesowym co stwarza możliwość łatwiejszego zrozumienia i poruszania się po niej przez użytkowników.

3. Proces ETL

(Extraction, transformation, loading) – jest to proces pozyskiwania danych do hurtowni z systemów źródłowych. W klasycznym schemacie hurtowni danych można wyodrębnić cztery podprocesy

- ekstrakcja (ang. extraction),
- oczyszczanie (ang. cleansing),
- transformacja (ang. transformation) i
- ładowanie (ang. loading).

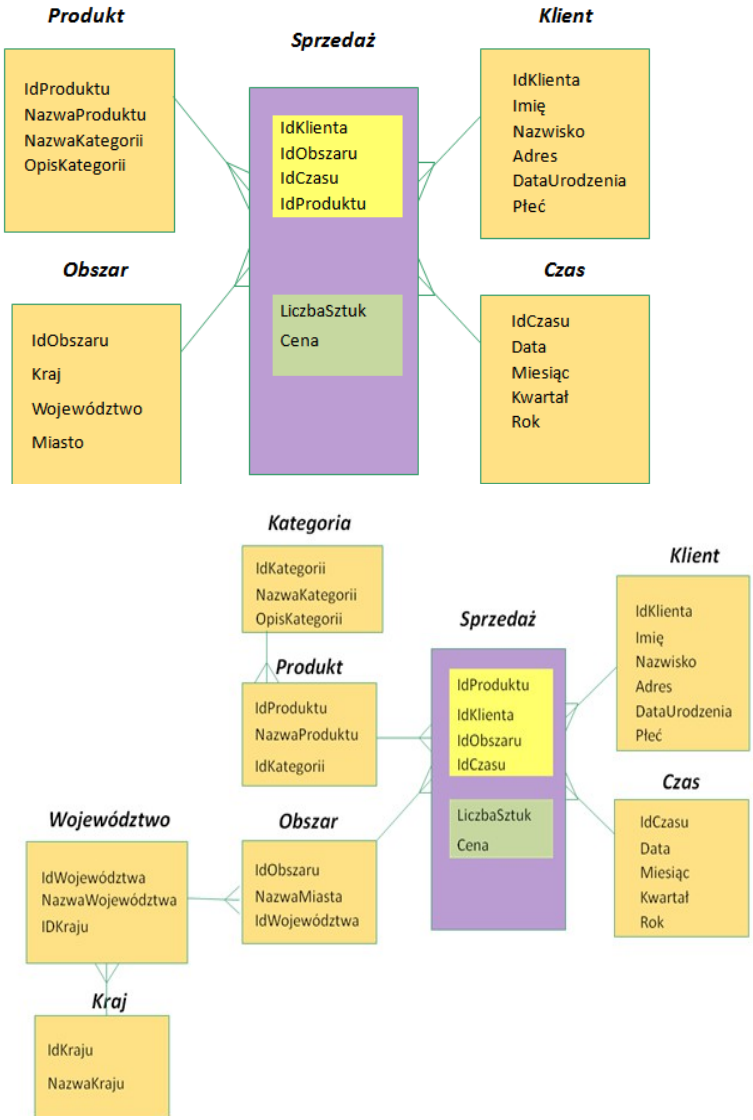


4. System OLTP

(On-line Transaction Processing) – transakcyjny system informatyczny wykorzystywany do dostarczania danych źródłowych do hurtowni danych. Dane zgromadzone w środowisku wielodostępowym są integralne. Efektywność systemu mierzy się liczbą transakcji w jednostce czasowej. Operacje wstawiania i uaktualniania danych są krótkotrwałe, inicjowane przez użytkownika końcowego. Zapytania charakteryzują się prostotą i krótkim czasem procesowania. Ilość miejsca potrzebna na przechowywanie danych jest stosunkowo niewielka. Dane są znormalizowane co wiąże się z dużą ilością tabel.

5. System OLAP

(On-line Analytical Processing) – analityczny system informatyczny wykorzystywany do analizowania danych zebranych w hurtowni danych (Data Mining). Efektywność systemu mierzy się czasem odpowiedzi na zapytanie. Celem danych zawierających informacje o aktualnym i historycznym stanie działań biznesowych, jest wspomaganie planowania, szukania rozwiązań oraz podejmowania decyzji. Uaktualnianie danych jest operacją okresową, długo trwającą najczęściej przeprowadzaną za pomocą plików wsadowych. Zapytania charakteryzują się dużą złożonością. Czas ich procesowania jest ściśle uzależniony od ilości danych – im więcej danych tym dłużej trwa proces. Potrzeba dużo miejsca do przechowywania danych ze względu na ich objętość (dane aktualne, agregowane i historyczne). Dane są zdenormalizowane co wiąże się z niewielką ilością tabel, stosuje się schemat gwiazdy lub płatka śniegu.



6. Data analysis

proces analizowania, oczyszczania, transformowania i modelowania danych, który ma na celu odkrycie użytecznych informacji, konstruktywnych wniosków wspierających podejmowanie decyzji. Data Analysis jest podzielone na eksploracyjną analizę danych (EDA), gdzie nowe cechy danych są odkrywane, weryfikującej analizy danych (CDA), gdzie istniejące hipotezy sprawdzane są pod kątem prawdziwość oraz na jakościową analizę danych (QDA).

7. Eksploracja danych

(Data mining) - Idea eksploracji danych polega na wykorzystaniu zasobów komputera do wyszukania niezauważalnych dla człowieka prawidłowości, wzorców
w danych zgromadzonych w hurtowniach danych.
Metody odkrywania wiedzy można podzielić na kategorie:

- Klasyfikacja
- Aproksymacja
- Zależności przyczynowe
- Podobieństwo

- Asocjacje

8. Business Intelligence

BI jest zestawem technik i narzędzi do transformacji surowych danych w użyteczne informacje dla celów biznesowych. Techniki BI są zdolne do manipulowania dużymi ilościami danych bez wyraźnej struktury w celu identyfikacji, rozwoju lub stworzenia nowych okazji biznesowych. Głównym celem jest przystępna interpretacja zbiorów danych oraz wypracowanie strategii opartej na nich. Techniki BI pozawala spojrzeć na działania biznesowe z różnych punktów widzenia: historycznego, bieżącego oraz przyszłościowego.
Do BI zalicza się również rozwiązania podejmujące decyzje na podstawie zadanych algorytmów postępowania.

•EIS - systemy powiadamiania kierownictwa (Executive Information Systems)
•DSS - systemy wspomagania decyzji (Decision Support Systems)
•MIS - Systemy wspomagania zarządzania (Management Information Systems)

9. Big data

Big data jest bardzo dużym zbiorem danych zgromadzonych np. w hurtowni danych, opisywanym w 3 wymiarach: duża ilość danych (volume), duża różnorodność typów danych (variety), duża prędkość procesowania danych (velocity). Od angielskich nazw wymiarów powstał model nazwany 3V. Model ten bywa rozszerzany o dwa kolejne wymiary: wartość biznesowa (value) i weryfikacja (veracity).

BIBLIOGRAFIA

1. „Architektury Hurtownii Danych” D.Dymek, W.Komnata, L.Kotulski, P.Szwed
2. http://whatis.techtarget.com/definition/3Vs
3. http://en.wikipedia.org/wiki/Business_intelligence
4. http://home.elka.pw.edu.pl/~mchoinsk/BI/
5. http://searchdatamanagement.techtarget.com/definition/data-analytics
6. http://en.wikipedia.org/wiki/Data_analysiss
7. https://edux.pjwstk.edu.pl/mat/246/lec/Rys3/Rysunek3_1.png
8. https://edux.pjwstk.edu.pl/mat/246/lec/Rys3/Rysunek3_2.png
9. http://datawarehouse4u.info/OLTPvsOLAP.html
10. http://datawarehouse4u.info/OLTPvsOLAP.html