

Data Warehouses

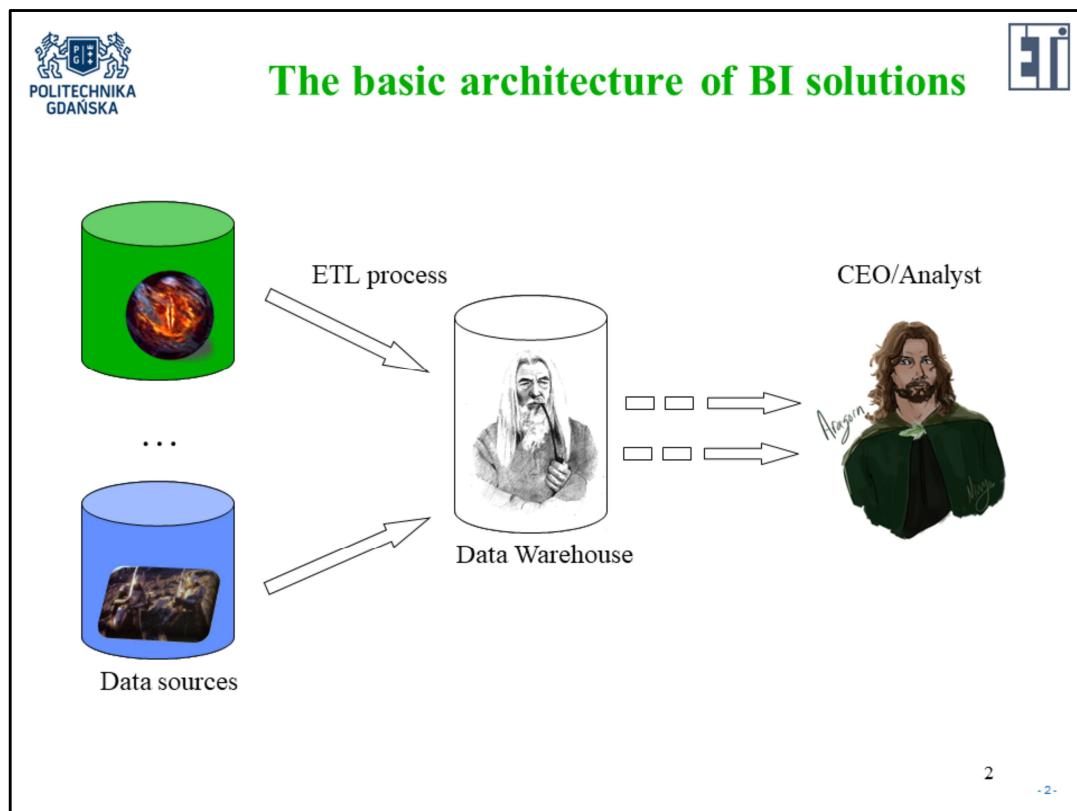
The role of data warehouse
in *Business Intelligence* systems



- 1 -

W ramach tego wykładu powiem trochę więcej na temat hurtowni danych.

The basic architecture of BI solutions



2

- 2 -

Przypomnijmy sobie architekturę systemu typu BI (pamiętamy, że skrót ten oznacza z angielskiego *Business Intelligence* czyli inteligencję biznesową). Mamy źródła danych, z których dane w procesie ETL (*Extract, Transform, Load*) są integrowane i ładowane do hurtowni danych. Następnie, systemy raportujące udostępniają kokpit menadżerskie analitykom. Informacje zawarte na kokpitach menadżerskich pomagają rozwiązywać problemy analityczne i podejmować decyzje biznesowe.

A **subject-oriented, integrated, time-variant and non-volatile collection of data** in support of management's decision making process.

Bill Inmon

A copy of transaction **data** specifically **structured for query and analysis**.

Ralph
Kimball

- 3 -

Przyjęły się dwie definicje pojęcia hurtownia danych. Pierwsza z nich (zdefiniowana przez Billa Inmona) zakłada, że jest to zbiór danych wykorzystywany w procesie podejmowania decyzji. Taki zbiór danych ma pewne cechy: jest trwałym, ukierunkowanym tematycznie, zintegrowanym i zależnym od czasu.

Definicja Ralphe Kimball'a nie definiuje bezpośrednio cech takiego zbioru, a jedynie jednoznacznie stwierdza, że taki zbiór danych musi być zapisany w strukturach dopasowanych do wykonywania zapytań analitycznych.

Collection of data

Data warehouse is a huge collection of data / database – (hundreds of GB, even TB).

Such database is optimised for **analytical** processing and not **transactional**.

- 4 -

Zazwyczaj hurtownia danych to zbiór danych zawierających setki GB lub TB danych. Warto zajrzeć do przedrostków SI (<http://www.jednostek.pl/przedrostki-si>). Czy wiedzą Państwo od czego pochodzi przedrostek tera? Nie należy jednak mylić hurtowni danych z modnym teraz hasłem Big Data. Hurtownie danych nie są dostosowane do analizowania danych typu Big Data (charakteryzujących się modelem 3(5)V). W ostatnich latach zaczęły pojawiać się rozwiązania hurtowni danych dla danych typu Big Data, ale na ten temat można więcej znaleźć pod hasłem Big Data Warehouses.

Subject-oriented

A data warehouse can be used to analyze a particular subject area, for example, "sales"

Time-variant

Historical data is kept in a data warehouse

Non-volatile

Once data is in the data warehouse, it will not change.

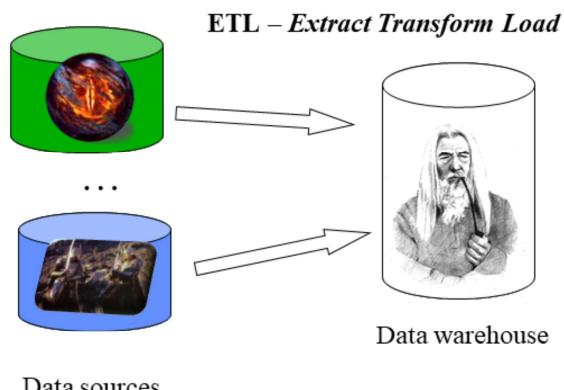
- 5 -

Pozostałe cechy hurtowni danych określają, że dane te są ukierunkowane tematycznie i odnoszą się do konkretnych zagadnień i procesów biznesowych w instytucjach. Ponadto, dane w hurtowniach danych zawsze przechowują historię i w ogólności dana raz zapisana w hurtowni danych nie powinna podlegać zmianie.

Features of a data warehouse (3)

Integrated

A data warehouse integrates data from multiple data sources

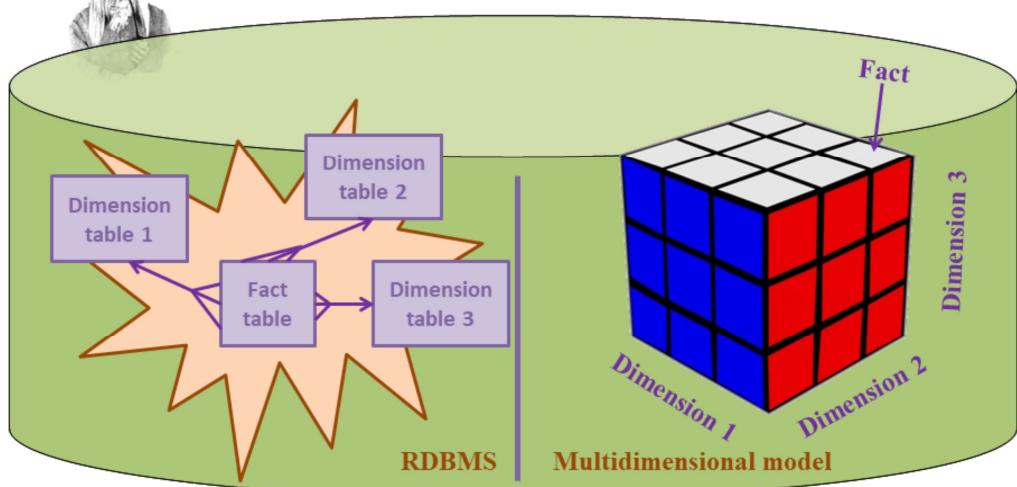


- 6 -

Dane w hurtowni danych są zintegrowane. Oznacza to, że z danych z różnych źródeł tworzony jest jeden zbiór danych (jedna wersja danych, na podstawie których będą prowadzone analizy). Więcej szczegółów na ten temat znajdą Państwo w wykładzie poświęconym procesowi ETL. Jak wielkim problem jest integracja danych świadczy fakt, że 80% pracy podczas wdrażania rozwiązań typu BI jest poświęcone zaprojektowaniu, implementacji i wdrożeniu procesu ETL.

Features of a data warehouse (4)

Specifically structured for query and analysis



- 7 -

Kiedy znamy już obie definicje hurtowni danych zastanówmy się, czym są struktury danych ukierunkowane na wykonywanie zapytań analitycznych (przecież o takich strukturach w swojej definicji mówi Ralph Kimball). Na struktury te patrzymy w dwojakim sposobie: z jednej strony jest to **specjalna struktura relacyjna** (najczęściej schemat gwiazdy składający się z tabel faktów i tabel wymiarów), z drugiej strony zaś struktura wielowymiarowa (inaczej **wielowymiarowy model danych**) będących kostką zbudowaną z faktów opisanych wymiarami.

Data in data warehouses are represented in the form of facts. **Fact** is a single event associated with the process and contains the measurement data (**measures**) associated with that event.

A **dimension** is a collection of reference information about a measurable event (fact). One dimension contains descriptive attributes i.e. **dimension attributes**. **Dimension members** are text labels describing facts.

- 8 -

W pierwszej kolejności omówimy wielowymiarowy model danych, gdyż właśnie dane w tym modelu są prezentowane użytkownikom systemów BI. Model wielowymiarowy jest reprezentowany w postaci faktów. **Fakt** to nic innego jak pewne zdarzenie, skojarzone z procesem biznesowym i generujące metryki liczbowe.

Przykłady:

Sprzedaż produktu to fakt, który generuje następujące przykładowe metryki liczbowe: ile danego produktu sprzedano, ile zapłacono i ile zapłacono podatku.

Przelot samolotu to fakt, który generuje następujące przykładowe metryki: ile osób zostało przewiezionych, ile benzyny zużyto, ile było obsługi, ile było miejsc wolnych i ile porcji jedzenia wydano.

Przeprowadzenie egzaminu to fakt, który generuje następujące przykładowe metryki liczbowe: jaką ocenę otrzymał student, ile czasu student pisał egzamin i ile punktów otrzymał student.

Te wygenerowane metryki liczbowe to nic innego jak **miary**. Pamiętajmy, to jedna z najważniejszych reguł w hurtowniach danych: **Miara jest wartością liczbową generowaną przez fakt**. Miara to liczba, miara jest liczbowa, miarę można policzyć, miara może mieć jednostkę miary itd...

Wprowadziliśmy już dwa ważne pojęcia faktu (czy na pewno zapamiętaliśmy, że to zdarzenie generujące metryki liczbowe ⓘ ?) i miary (czy na pewno zapamiętaliśmy, że miara jest wartością liczbową? ⓘ). Teraz wprowadzimy trzecie pojęcie tj. **wymiar**. Wymiar to informacja opisująca fakt.

Przykłady:

Sprzedaż produktu to fakt, który jest opisany takimi przykładowymi wymiarami jak: sklep, w którym dokonano sprzedaży; czy produkt, który zakupiono. Przelot samolotu to fakt, który jest opisany takimi przykładowymi wymiarami jak: model samolotu, pilot samolotu, linie lotnicze, lotnisko startowe, czy lotnisko docelowe.

Przeprowadzenie egzaminu to fakt, który jest opisany takimi przykładowymi wymiarami jak: student podchodzący do egzaminu, prowadzący, czy numer sali.

Ups... czy numer sali nie jest przypadkiem wartością liczbową? Dlaczego potraktowałam ten numer jako wartość opisową? To proste:

Przez wartość liczbową rozumiemy taką wartość, dla której mają sens operacje matematyczne. Numer sali mógłby być również dobrze określony jako A123.

Jak już na pewno pamiętamy czym jest miara (wartością liczbową opisującą fakt, prawda?) to zapamiętajmy również, że **wymiar jest wartością opisaną faktem**. Wymiar opisuje fakt, wymiar to rodzaj kategorii, wymiar ma zawsze charakter opisowy itd...

Wprowadziliśmy już trzy pojęcia faktu (czy na pewno zapamiętaliśmy, że to zdarzenie generujące metryki liczbowe ⓘ ?), miary (czy na pewno zapamiętaliśmy, że miara jest wartością liczbową? ⓘ) i wymiaru (czy na pewno zapamiętaliśmy, że wymiar ma charakter opisowy ⓘ?). Teraz wprowadzimy kolejne pojęcie tj. **atrybut wymiaru**.

Przykłady:

Sprzedaż produktu to fakt, który jest opisany przykładowym wymiarem sklep, który posiada następujące przykładowe cechy: nazwę; miasto, w którym sklep się znajduje czy ulicę, na której sklep się znajduje.

Przelot samolotu to fakt, który jest opisany przykładowym wymiarem lotnisko startowe, który posiada następujące przykładowe cechy: kod IATA lotniska czy miasto, w którym dane lotnisko się znajduje.

Przeprowadzenie egzaminu to fakt, który jest opisany przykładowym wymiarem egzaminowany student, który posiada następujące przykładowe cechy: imię i nazwisko studenta, PESEL studenta, numer indeksu studenta, narodowość studenta.

Cecha opisowa wymiaru to nic innego, jak właśnie **atrybuty wymiaru**.

Wprowadziliśmy już cztery pojęcia faktu (czy na pewno zapamiętaliśmy, że to zdarzenie generujące metryki liczbowe ⓘ ?), miary (czy na pewno zapamiętaliśmy, że miara jest wartością liczbową? ⓘ), wymiaru (czy na pewno zapamiętaliśmy, że wymiar ma charakter opisowy ⓘ?) i atrybutu wymiaru (czy na pewno zapamiętaliśmy, że atrybut wymiaru to cecha opisowa wymiaru ⓘ). Teraz wprowadzimy ostatnie – piąte pojęcie tj. **element wymiaru**.

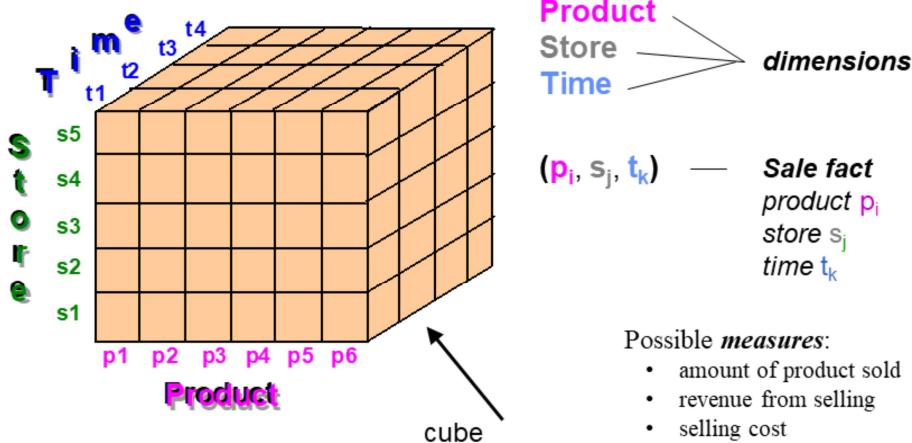
Przykłady:

Sprzedaż produktu to fakt, który jest opisany przykładowym wymiarem sklep, a więc sprzedaż produktu odbywa się przykładowo w sklepie Biedronka w Gdańsku przy ulicy Dąbrowszczaków.

Przelot samolotu to fakt, który jest opisany przykładowym wymiarem lotnisko startowe, a więc samolot przykładowo startuje z Lotniska im. Lecha Wałęsy w Gdańsku, o kodzie IATA GDN.

Przeprowadzenie egzaminu to fakt, który jest opisany przykładowym wymiarem egzaminowany student, a więc przeprowadzono egzamin przykładowo dla studenta o polskiej narodowości Adama Niegódkiego, o numerze PESEL 99121709098, który posiada numer indeksu 1238990. Konkretny sklep jest elementem wymiaru sklep, konkretnie lotnisko jest elementem wymiaru lotnisko startowe, zaś konkretny student jest elementem wymiaru student.

Example of dimensional model



- 9 -

Wiedząc już czym jest fakt miara i wymiar wyobraźmy sobie zbiór faktów, które układamy w formie kostki.

Najpierw przyjrzyjmy się prostej kostce, która jest przedstawiona na slajdzie. Składa się ona z faktów, które są zobrazowane „tymi małymi” sześcianami. Inaczej mówiąc jeden pojedynczy fakt (tutaj sprzedaż produktu w danym sklepie w danym czasie) to sześciąan. Cała przykładowa kostka to prostopadłościan, którego trzy wymiary oznaczają wymiar produktu, sklepu i czasu.

Tutaj przechodzimy do bardzo ważnej cechy faktu. Mianowicie:

Każdy fakt jest jednoznacznie identyfikowany elementami wymiaru!

Czy umieją Państwo wymienić elementy wymiaru produkt? Jeśli nie, podpowiadam: p1, p2, p3, p4 i p5.

Teraz już bez trudu wymienią Państwo elementy wymiaru sklep i czas.

Zdefiniowana przed chwilą cecha faktu (każdy fakt jest jednoznacznie identyfikowany elementami wymiaru) oznacza:

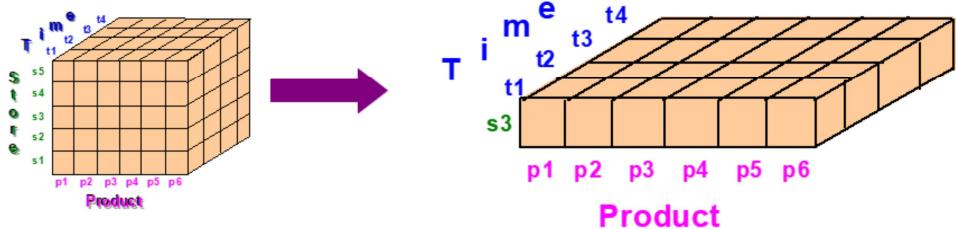
- liczba faktów w prezentowanej kostce jest nie większa niż:
 $[liczba\ elementów\ wymiaru\ sklep] * [liczba\ elementów\ wymiaru\ produkt] * [liczba\ elementów\ wymiaru\ czas] = 5 * 6 * 4 = 120$
- istnieje dokładnie jeden fakt o takich samych wszystkich elementach wymiaru
- znając wszystkie elementy wymiaru możemy jednoznacznie zidentyfikować fakt, który te elementy wymiaru opisują.

Wewnątrz każdego sześcianu – faktu znajdują się miary, które ten fakt generuje.

Teraz proszę przez chwilę się zastanowić, dlaczego do zobrazowania pojęcia kostki wykorzystujemy model 3-wymiarowy?

Examples of multidimensional analysis

Slicing and Dicing



Efekt:

(p_i , s_3 , t_k) – all sale facts in store s_3

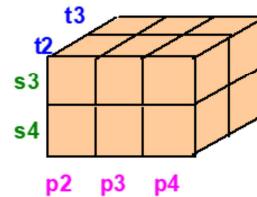
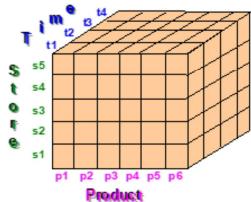
- 10 -

Jaka jest najważniejsza operacja, którą możemy wykonać na kostce. Możemy wyciąć tylko pewien podzbiór faktów. Tutaj wycięłam wszystkie fakty sprzedaży, które odbyły się w sklepie s3.

Examples of multidimensional analysis



Slicing and Dicing



Efekt:

(p_i, s_j, t_k) – sale facts in stores s_3 i s_4
products p_2, p_3, p_4 in time t_2, t_3

- 11 -

A tutaj te które odbyły się w sklepach s_3 lub s_4 , dotyczyły produktów p_2, p_3 i p_4 i odbywały się w czasie t_2 lub t_3 .

For the set of facts it is possible to count
aggregations.

Measures have **aggregate functions** assigned.

The aggregation function determines what mathematical operation is performed on the measured values.

The correct aggregate function for measures: the amount of product sold, sales revenue, and selling cost is the **SUM** function.

- 12 -

Ale po co wycinać tylko podzbiór faktów? Odpowiedź na to kryje się w miarach generowanych przez fakt. Większość analiz w hurtowniach danych polega na:

1. Wycięciu podzbioru faktów
2. Policzeniu agregatów dla tak wyciętego zbioru faktów.

Agregat – co to takiego? Każdy fakt generuje pewne metryki. W naszej przykładowej kostce weźmy pierwszą z nich: to ilość sprzedanego produktu.

1. Wycinamy podzbiór faktów: tylko te fakty sprzedaży, które odbyły się w sklepie s3.
2. Liczymy ilość sprzedanych produktów (nie ważne jakich) w sklepie s3 – sumując wartość miary ilość sprzedanego produktu po wszystkich wyciętych faktach.

Dochodzimy do definicji pojęcia agregat:

Agregat to zagregowana wartość pewnej miary.

Korzystając ze słownika synonimów: zagregowany to połączony, ułożony w całość, zebrany w całość, zebrany w jedno.

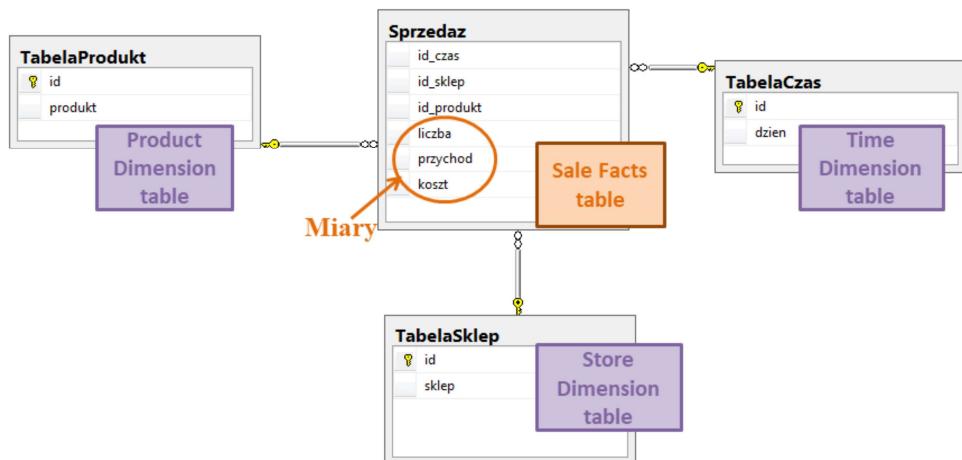
A jak z wielu wartości uzyskać jedną? Musimy wszystkie te wartości poddać jakiejś operacji matematycznej.

Mogemy te wartości zsumować, mogemy policzyć ich wartość maksymalną, minimalną, czy np.: średnią.

Ta operacja matematyczna, którą chcemy wykonać, aby uzyskać jedną wartość z wielu wartości tej samej miary nazywamy **funkcją agregującą**. W naszym przykładzie ilość sprzedanego produktu ma przypisaną sumę jako funkcję agregującą.

Co należy zapamiętać: **Miary generowane przez fakt mają mieć przypisane funkcje agregujące**. Później pokażę przykład, w którym złamie tę zasadę, ale nie wszystko na raz.

Star schema



- 13 -

Omówiliśmy podstawy modelu wielowymiarowego. Teraz przypominamy sobie, że istnieje również specjalna struktura relacyjna służąca do przechowywania danych w hurtowniach danych. Tą strukturę relacyjną nazywamy schematem gwiazdy.

Przyjęła się nazwa schematu gwiazdy, gdyż zawsze środkiem tego schematu znajduje się tabela faktów, zaś od niej odchodzą promienie w postaci tabel wymiarów i zdefiniowanych związków pomiędzy tabelą faktów i wymiarów.

Tabela faktów przechowuje fakty. **Ogólna zasada głosi, że znajdują się w niej wartości miar i klucze obce do tabel wymiarów będące implementacją związków n:1 pomiędzy tabelą faktów, a tabelą wymiarów.**

Tabele wymiarów przechowują elementy wymiarów.

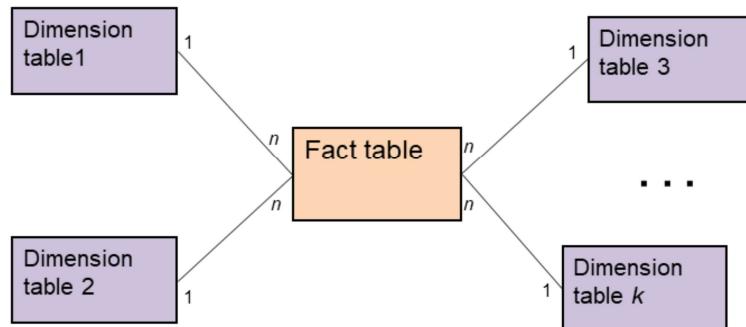
Na rysunku widzą Państwo przykładowy schemat gwiazdy dla wcześniej zdefiniowanej kostki.

TabelaProdukt będzie miała 6 krotek (o wartościach atrybutu produkt odpowiednio: p1, p2, p3, p4, p5 i p6).

TabelaCzas będzie miała 4 krotki (o wartościach atrybutu dzień: t1, t2, t3 i t4).

TabelaSklep będzie miała 5 krotek (o wartościach atrybutu sklep: s1, s2, s3, s4 i s5).

Star schema

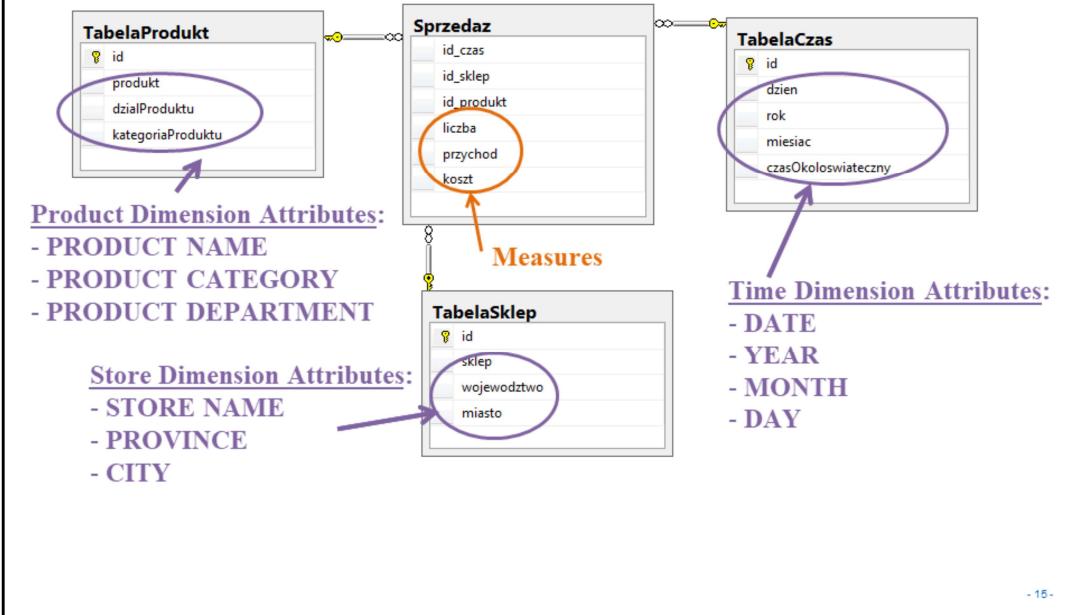


Fact table with *numeric attributes* (**measures**)
Dimension tables with *descriptive attributes* (**dimension members**)

- 14 -

A tutaj generyczny schemat gwiazdy.

Dimensional model versus star schema (1)

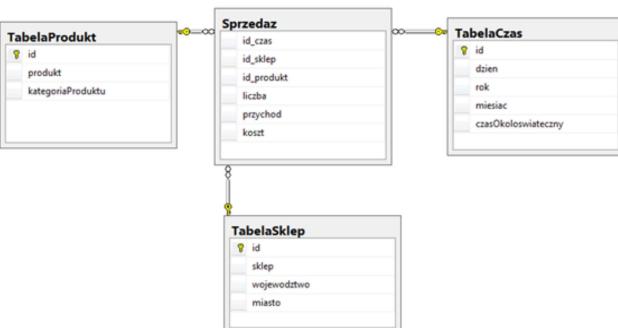


- 15 -

Jeszcze raz schemat gwiazdy dla naszego przykładowego modelu wielowymiarowego. Zauważamy, że w odróżnieniu od poprzedniego schematu gwiazdy (tego ze slajdu 13) mamy tu zdefiniowanych więcej atrybutów wymiaru.

Slajd obrazuje odwzorowanie pomiędzy modelem wielowymiarowym, a jego reprezentacją relacyjną w postaci schematu gwiazdy.

Hierarchical dimensions



Dimensions:

- PRODUCT
- TIME
- STORE

Hierarchical Dimensions :

- | | | |
|--|---|---|
| <ul style="list-style-type: none"> - TIME: • YEAR •• MONTH ••• DAY | <ul style="list-style-type: none"> - LOCATION OF THE STORE • WOJEWÓDZTWO •• MIASTO ••• NAZWA SKLEPU | <ul style="list-style-type: none"> - PRODUCT • PRODUCT DEPARTMENT •• PRODUCT CATEGRORY ••• PRODUCT NAME |
|--|---|---|

- 16 -

Ostatnie pojęcie, o którym wcześniej nie wspominaliśmy to **wymiar hierarchiczny**. Wymiar hierarchiczny to taki wymiar, którego atrybuty wymiaru tworzą hierarchię.

Typowym wymiarem hierarchicznym jest wymiar czasu – data ma tzw. hierarchię naturalną. Wiadomo, że jest rok, w roku są miesiące, a w ramach miesięcy są konkretne dni. Innym przykładem wymiaru hierarchicznego jest lokalizacja, jest to również hierarchia naturalna. Państwo jest podzielone na województwa, w województwach znajdują się miasta, zaś w miastach konkretne sklepy.

Mogą również istnieć atrybuty hierarchiczne nie wynikające z naturalnej hierarchii. W naszym przykładzie może to być wymiar produkt.

Wyobraźmy sobie, że każde konkretne ubranie jest podzielone wg dwóch kategorii:

1. Związana z angielską nazwą (*product department*): Odzież damska, męska i dziecięca.
2. Związana z rodzajem ubrania (*product category*): spodnie, skarpety, kurtki.

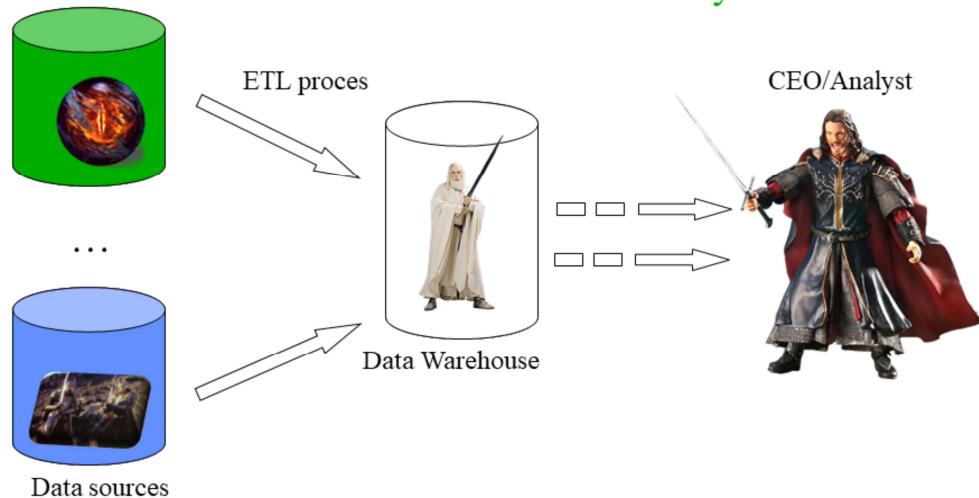
Możemy utworzyć dwie hierarchie: Product Department → Product Category → Product Name

Product Category → Product Department → Product Name

Każda z nich jest jednakowo poprawna i jej użyteczność zależy jedynie od potrzeb biznesowych.

Architecture once again

Basic architecture of BI Systems



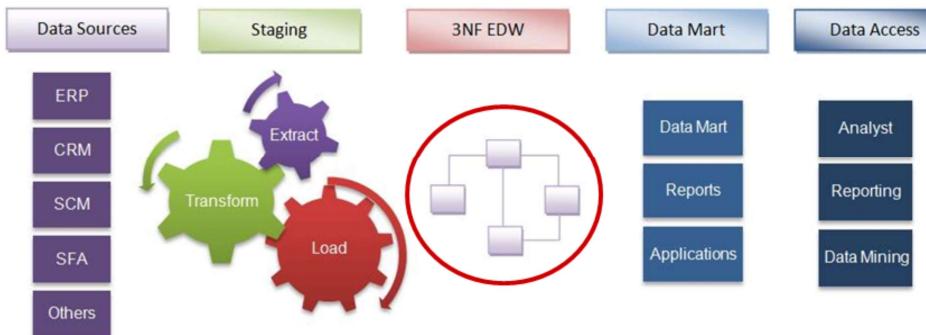
- 17 -

Wróćmy jeszcze na chwilę do architektury. Jak pamiętamy mamy dwie definicje hurtowni danych. Mamy również dwie architektury odpowiadające wizji Pana Kimballa i Pana Inmona.

Bill Inmon definition

A data warehouse is a **subject-oriented, integrated, time-variant** and **non-volatile** collection of data in support of **management's decision making process**.

Bill Inmon



<http://searchbusinessintelligence.techtarget.in/tip/Inmon-vs-Kimball-Which-approach-is-suitable-for-your-data-warehouse>

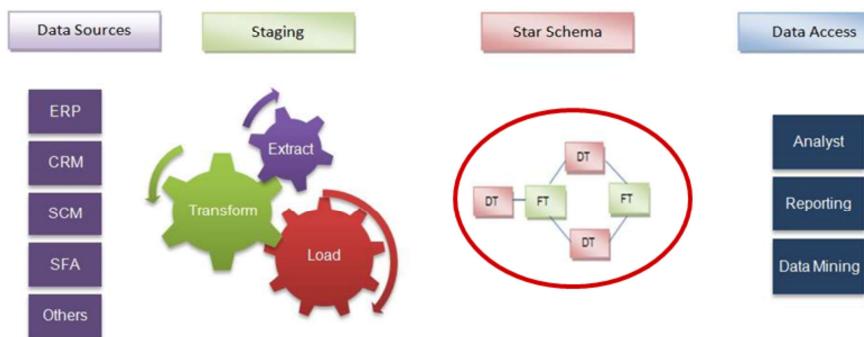
- 18 -

Ten slajd przedstawia architekturę Pana Inmona.

Ralph Kimball definition

A data warehouse is a copy of transaction data **specifically structured for query and analysis.**

Ralph
Kimball



DT – Dimension Table
FT – Fact Table

<http://searchbusinessintelligence.techtarget.in/tip/Inmon-vs-Kimball-Which-approach-is-suitable-for-your-data-warehouse>

- 19 -

Ten ząś Pana Kimballa.

Teraz czas na chwilę mojego lenistwa. Wolno? Nie wolno? Na szczęście ja decyduję. Znalazłam bardzo dobre opracowanie na ten temat. Niech Państwo zajrzą:
<http://tdan.com/data-warehouse-design-inmon-versus-kimball/20300>.

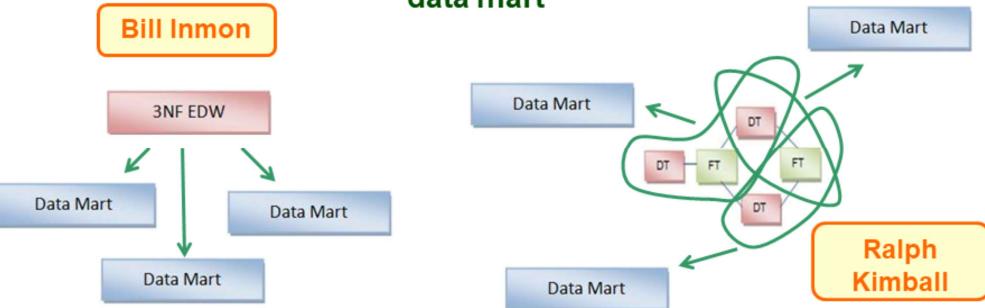
Jeśli byłby jakiś problem z linkiem proszę dać znać. Znajdę wtedy coś innego. Obiecuję – równie interesującego.

Data mart

Subject oriented data warehouse

A data mart is a repository of data that is designed to serve a particular community of knowledge workers.

↓
data mart



- 20 -

To teraz już naprawdę na koniec.

Spotkają się Państwo jeszcze z pojęciem *Data mart*. *Data mart* to nic innego jak taka mała hurtownia danych: minihurtownia lub hurtownia tematyczna. Dlaczego mała, bo ukierunkowana tylko na jeden temat (jeden dział firmy) np.: HR. Hurtownia danych jest dla całej organizacji, minihurtownia dla wybranego działu.

Nie wszyscy o wszystkim muszą wiedzieć! Nie wszyscy wszystkiego potrzebują!

What every student should know now...

1. Two definitions of data warehouse and differences between them.
2. Understand multidimensional model.
3. Understand star schema.
4. Inmon's and Kimball's Architectures and differences between them.



- 21 -