

Zasady projektowania hurtowni danych

Przykład hurtowni danych dla systemu NFZ



- 1 -

W poprzednim wykładzie wyjaśniliśmy ogólne zasady budowania hurtowni danych. Wykład, który teraz rozpoczynamy, przedstawia szczegółowe zasady projektowania hurtowni danych, a dokładnie schematu gwiazdy i schematu płatka śniegu, dla różnych typów zdarzeń. Wykład jest prowadzony w formie analizy przypadku użycia.

Szpital - definicja problemu

Zaprojektować hurtownię danych dla Narodowego Funduszu Zdrowia.

Hurtownia ma umożliwiać analizowanie hospitalizacji w polskich szpitalach, które podpisały umowę z NFZ na świadczenie usług medycznych.

Analizy mają dotyczyć aspektów finansowych oraz zdrowotnych.

Hurtownia ma bazować na danych źródłowych zbieranych przez szpitale oraz na danych zawartych w centralnej bazie NFZ.



- 2 -

A o jaki przypadek użycia chodzi?

Rodzaje zdarzeń – rodzaje faktów

Zdarzenia dyskretne

Zdarzenia występujące w jednym punkcie czasowym. Zazwyczaj odnoszą się do pojedynczych transakcji w systemach operacyjnych.

Wykonanie pojedynczej procedury leczniczej NFZ

Transaction Fact Table

Zdarzenia powtarzające się

Zdarzenia występujące co określony przedział czasowy. Zazwyczaj są to zdarzenia reasumujące zdarzenia dyskretne.

Miesięczne zestawienie wykonanych procedur NFZ

Periodic Snapshot

Zdarzenia rozwijające się

Zdarzenia występujące w dłuższym przedziale czasowym. Zazwyczaj stanowią serię zdarzeń dyskretnych.

Pobyt pacjenta w szpitalu

Accumulating Snapshot

- 3 -

Do tej pory mówiliśmy, że fakt to w rzeczywistości pewne zdarzenie biznesowe generujące metryki liczbowe, czyli miary.

Teraz okazuje się, że te zdarzenia biznesowe, które później przekładają się na fakty, można podzielić na trzy różne typy: **zdarzenia dyskretne**, **zdarzenia powtarzające się** i **zdarzenia rozwijające się**.

Wykonanie pojedynczej procedury leczniczej NFZ

Faktem będzie **wykonanie pojedynczej procedury NFZ**. Jest to nasze biznesowe zdarzenie dyskretne.

Z opisu problemu wynika, że miary będą dotyczyć kwestii finansowych i kwestii związanych z leczeniem.

- **Miary:**

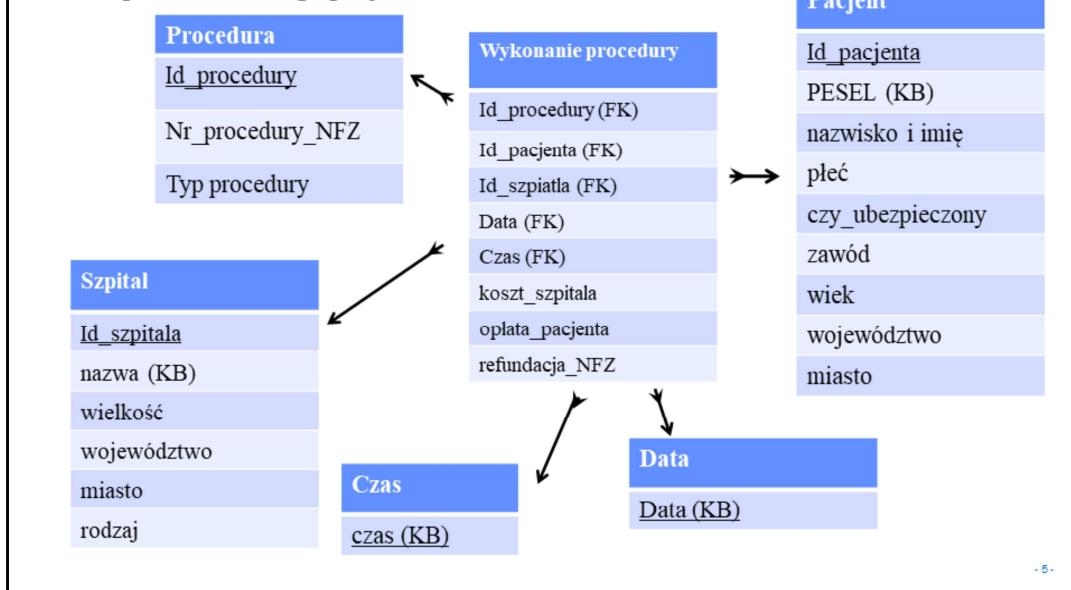
- koszt szpitala
 - opłata wniesiona przez pacjenta
 - refundacja NFZ
 - (W) strata =
$$\text{koszt szpitala} - \text{opłata wniesiona przez pacjenta} - \text{refundacja NFZ}$$
 - ...

- 4 -

Skupmy się w pierwszej kolejności na zdarzeniach dyskretnych.

Wykonanie pojedynczej procedury leczniczej (1)

Faktem będzie wykonanie pojedynczej procedury NFZ, w danym dniu, w danym szpitalu dla danego pacjenta.



- 5 -

Dla tak zdefiniowanego faktu dyskretnego definiujemy tabelę faktów: Wykonanie procedury oraz 5 tabel wymiarów: Procedura, Szpital, Czas, Data i Pacjent. Zauważamy, że w tabeli faktów znajdują się tylko klucze obce do tabel wymiarów (oznaczone FK) oraz miary (koszt_szpitala, opłata_pacjenta, refundacja_NFZ).

Ziarnistość

Określenie ziarnistości jest kluczowym krokiem przy definiowaniu hurtowni danych!

Ziarnistość atomowa odnosi się do najniższego poziomu, na którym dane są przechwytywane przez dany proces biznesowy.

Nie wolno różnych ziarnistości mieszać w ramach tego samego faktu!!!

W przykładzie:

1. **Hurtownia danych** ma ziarnistość odnoszącą się do przeprowadzenia pojedynczej procedury w konkretnym szpitalu, w danym dniu, w danym czasie na konkretnym pacjencie.
2. **Hurtownia danych** ma ziarnistość atomową.

- 6 -

Kiedy definiujemy fakt najważniejsze jest określenie jego ziarnistości. Określenie ziarnistości to nic innego jak jednoznaczne zdefiniowanie wymiarów danego faktu. Posłużymy się przykładem. Definiujemy nieformalnie fakt zdawania egzaminu. Generujemy dla niego miary: *liczba studentów podchodzących do egzaminu i ocena z egzaminu danego studenta*. Na pierwszy rzut oka wszystko wygląda poprawnie. Fakt odnosi się do rzeczywistego zdarzenia biznesowego i generowane przez ten fakt miary są liczbowe i podlegają agregacji.

Jak się Państwo domyślają, jednak coś jest źle.

Mianowicie, musimy doprecyzować fakt (zdefiniujmy go – z uwzględnieniem ziarnistości, czyli z dokładnością do wymiarów w dwojakim sposobie):

1. Przeprowadzenie egzaminu w danym czasie, z danego przedmiotu.
2. Przeprowadzenie egzaminu w danym czasie, z danego przedmiotu dla danego studenta.

Miara *liczba studentów podchodzących do egzaminu* jest poprawna zarówno w pierwszym, jak i w drugim przypadku. W pierwszym przypadku oznaczać będzie faktyczną liczbę studentów podchodzących do egzaminu (dla jednego faktu będzie to 115, dla drugiego 120, a jeszcze innego 110). W drugim przypadku dla każdego faktu miara ta będzie posiadała dokładnie tą samą wartość. Równą ile? Już się Państwo pewnie domyślili – zawsze 1.

Ocena z egzaminu danego studenta jest miarą poprawną tylko w przypadku drugiej definicji faktu. Moglibyśmy zdefiniować nową miarę *średnia ocena z egzaminu* i taka miara byłaby poprawna dla definicji pierwszej.

Różnica jest ewidentna: pierwsza definicja odnosi się do przeprowadzenia egzaminu w odniesieniu do całej grupy studentów, druga zaś z większą szczegółowością do przeprowadzenia egzaminu dla konkretnego studenta.

Bądźmy uważni przy projektowaniu hurtowni danych. Często potrzebne są fakty na różnym poziomie ziarnistości.

Techniki projektowania tabeli faktów (1)

**Tabela faktów zawiera ogromną liczbę krotek!
Krotki powinny zajmować mało miejsca!
Każda krotka powinna zawierać tylko wartości numeryczne!**

Mamy trzy rodzaje miar:

- Miary **addytywne** (mogą być sumowane po wszystkich wymiarach, w przykładzie **wszystkie miary z przykładu**),
- Miary **pół-addytywne** (po pewnych wymiarach mogą być sumowane, a po pewnych nie, przykład: **stan na rachunku** po wykonaniu transakcji bankowej – nie może być sumowany po czasie, może być sumowany po typie rachunku lub wieku właściciela rachunku),
- Miary **nie-addytywne** (nie mogą być sumowane po żadnych wymiarach, przykład: **zysk procentowy** ze sprzedaży).

- 7 -

Kiedy mówimy o tabeli faktów, musimy pamiętać, że liczba faktów jest zawsze nieproporcjonalnie większa od liczby elementów wymiarów. Wynika stąd, że krotki w tabeli faktów powinny zajmować mało miejsca. Stąd zaś wynika, że tabela faktów powinna zawierać tylko wartości numeryczne. Zgadza się to z zasadami, które podałem wcześniej: zazwyczaj tabela faktów zawiera tylko klucze obce i miary.

Wiedzą również Państwo, że miary mają skojarzone ze sobą funkcje agregujące. Właśnie ze względu na możliwość przypisania funkcji SUM miary zostały podzielone na trzy rodzaje.

1. Miary addytywne – bez względu na to jak przeprowadzimy operację wycinania, zawsze dla uzyskanego zbioru faktów możemy zsumować wartości miary addytywnej.
2. Miary pół-addytywne – skorzystajmy z przykładu pokazanego na slajdzie. Wykonujemy operację bankową, jest to fakt, który generuje miarę: stan na rachunku bankowym. Wycinamy fakty, w taki sposób, że bierzemy stan na rachunku w danym dniu o danej porze dla wszystkich klientów poniżej 30 roku życia. Zadajemy pytanie: ile łącznie Ci klienci mają pieniądze zdeponowanych w banku. Przypisując funkcję SUM mierze stan na rachunku bankowym bez problemu możemy wyliczyć agregat. Przeprowadźmy wycinanie, w taki sposób, że bierzemy wszystkie fakty transakcji bankowych dla Pana Jana Kowalskiego. Teraz również dokonujemy sumowania miary stan na rachunku bankowym. Jakie znaczenie biznesowe ma wartość, którą uzyskaliśmy? Nie jestem najlepsza z ekonomii, ale wydaje mi się, że żadnego.
3. Miary nie-addytywne – bez względu na to jak przeprowadzimy operację wycinania, nigdy dla uzyskanego zbioru faktów nie możemy zsumować wartości miary nie-addytywnej.

W tabeli faktów miary MOGĄ przyjmować wartości NULL!

Kluczem głównym tabeli faktów jest złożenie kluczy obcych.

W tabeli faktów klucze obce NIE MOGĄ przyjmować wartości NULL – naruszenie integralności referencyjnej kluczy.

W przykładzie, gdy nie znamy pacjenta to dodajemy sztuczny wiersz w tabeli wymiaru Pacjent o wartości „Nieznany”.

- 8 -

W poprzednim wykładzie powiedzieliśmy sobie, że każdy fakt jest jednoznacznie definiowany elementami wymiarów. Założenie to narzuca wymóg, że klucze obce w tabeli faktów nigdy nie przyjmują wartości NULL. Z założenia tego wynika również fakt, że klucz główny w tabeli faktów jest złożeniem kluczy obcych – nie musi być on jednak jawnie definiowany.

Co w takim razie zrobić, jeżeli z jakiś przyczyn (rzeczywistość jest nie do przewidzenia), któryś z elementów wymiarów nie jest znany. W takiej sytuacji tworzymy „sztucznie” element wymiaru oznaczający nieznane.

Techniki projektowania tabel wymiarów (1)

Tabela wymiarów zawiera liczbę krotek zdecydowanie mniejszą od liczby krotek w tabeli faktów!

Krotki mogą zajmować dużo miejsca!

Każda krotka powinna zawierać poza kluczami tylko wartości opisowe!

W przykładzie, wielkość szpitala i wiek pacjenta nie są wartościami numerycznymi tylko kategoriami – przedziałami liczbowymi określającymi wielkość szpitala i wiek pacjenta.

Co zrobimy jeżeli potrzebujemy odpowiedzi na zapytanie:
Jaka jest średnia wielkość szpitala lub średni wiek pacjenta, który...

- 9 -

W odróżnieniu od tabeli faktów tabele wymiarów zawierają wartości opisowe, więc krotki są długie. Ale na pocieszenie, liczba krotek jest dużo mniejsza.

Dotykamy teraz dość ważnego zagadnienia. Uważny student potrafi w tej chwili przytoczyć definicję atrybutu wymiaru (no!, ale bez zaglądania do poprzedniego wykładu ☺) Powiedzieliśmy sobie, że atrybuty wymiaru to cechy opisowe wymiaru.

No więc rozważmy wiek pacjenta. Czym jest wiek naszego pacjenta? Miarą, czy atrybutem wymiaru? Ratunku! Pomocy! Przecież jest cechą elementu wymiaru, a nie miarą generowaną przez fakt?

A niby dlaczego nie miarą generowaną przez fakt. Czy jest to wartość liczbową? Tak. Czy jest generowana przez fakt w momencie wykonania procedury medycznej? Drugi raz tak. Czy możemy przypisać wiekowi funkcję agregującą? Trzeci raz tak (średnia).

Przykładowe pytanie analityczne:

Podaj średni wiek pacjenta, dla którego jest wykonywana operacja usunięcia wyrostka robaczkowego?

Wszystko jest OK!

A może jednak nie. A co jeżeli chcę zapytać:

Ile kosztowały operacje usunięcia wyrostka robaczkowego w Polsce na pacjentach poniżej 30 roku życia w 2016 roku?

I co teraz? Dla tak zadanego pytania analitycznego miarą jest koszt procedury medycznej, a cechą opisową jest wiek pacjenta. Dlaczego cechą opisową, a nie liczbową? Zazwyczaj w hurtowniach danych nie ma potrzeby analizować danych z dokładnością do konkretnego wieku pacjenta. W zakresie zainteresowań sfery biznesowej będą przedziały wieku pacjenta (np.: pacjenci do 30, 30-60, po 60). Kiedy zamienimy liczbę na kategorię już nikt nie będzie miał wątpliwości, że jest to atrybut wymiaru.

Znowu wszystko jest OK!

Wykonanie pojedynczej procedury leczniczej (2)

Wykonanie procedury

Id_pacjenta

Id_szpitala

data

czas

koszt_szpitala

opłata_pacjenta

refundacja_NFZ

wielkość_szpitala

wiek_pacjenta

- 10 -

Taką charakterystykę ma zarówno wielkość szpitala, jak i wiek pacjenta. Jako kategoria atrybuty te znalazły się w tabelach wymiarów, zaś jako miary zostały również dołożone do tabeli faktów.

Techniki projektowania tabel wymiarów (2)

W każdej tabeli wymiaru znajduje się jeden klucz główny!

Każda tabela wymiaru (ewentualnie poza tabelą daty/czasu) zawiera **klucz główny surogatowy**. Klucz surogatowy jest to klucz generowany od wartości 1, zwiększany o 1. Przykładowo Id_pacjenta w tabeli wymiaru Pacjent. Klucz główny tabel wymiarów jest przechowywany w tabeli faktów.

Dodatkowo w tabelach wymiaru powinny znajdować się **klucze biznesowe**. Klucze biznesowe nie są kluczami głównymi. Przykładowo PESEL w tabeli Pacjent.

- 11 -

Kolejna zasada projektowania hurtowni danych narzuca, aby prawie (można z tego wykluczyć czas i datę) każda tabela wymiaru miała klucz surogatowy. Klucz surogatowy to klucz generowany przez relacyjną bazę danych. Na ten moment podam tylko jedno uzasadnienie. Jest to fakt, że klucze te są przechowywane w tabeli faktów, a więc muszą być jak najprostsze. Jest jeszcze jeden powód, ale o nim przy wymiarach zmieniających się w czasie.

Ponadto, z pewnością zauważycie Państwo, że przy niektórych atrybutach wymiarów jest oznaczenie KB. KB oznacza klucz biznesowy, czyli atrybut dzięki któremu rozpoznajemy elementy wymiarów w świecie rzeczywistym.

Techniki projektowania tabel wymiarów (3)

Elementy wymiarów mają być wartościami samoopisującymi się.

Nigdy nie kodujemy wartości. W systemach typu BI nie ma dedykowanych pod hurtownię danych aplikacji. Korzysta się z istniejących rozwiązań raportujących.

Kolumna płeć w tabeli wymiaru Pacjent przyjmuje wartości: kobieta lub mężczyzna. Nigdy K lub M albo 0 lub 1.

- 12 -

Ta zasada nie wymaga komentarza. Prezesa z grubą książką kodów trudno sobie wyobrazić.

Techniki projektowania tabel wymiarów (4)

Dane w tabeli wymiaru Data, Czas zawierają poza datą wartości opisowe opisujące daną datę lub czas.

Dane w tabelach wymiaru Data i Czas są generowane i tabele to powinny być wypełnione przed wykonaniem procesu ETL.

- 13 -

Kilka zasad na temat tabel wymiarów dla czasu i daty.

1. Jeśli potrzebujemy i czasu i daty to zawsze tworzymy dwie tabele wymiarów
2. Tabela daty i czasu nie musi mieć kluczy surogatowych (choć oczywiście może, jeśli komuś tak łatwiej)
3. Tabela czasu ma wszystkie momenty czasowe na wymaganym poziomie szczegółowości (np.: po kolej wszystkie sekundy doby)
4. Tabela daty ma krótki odpowiadające wszystkim dniom w określonym zakresie lat
5. Tabela daty i czasu ma mieć bardzo dużo atrybutów wymiarów, data i czas mogą być opisane w dowolny możliwy, potrzebny biznesowo sposób.
6. Data i Czas to wymiary hierarchiczne

Wykonanie pojedynczej procedury leczniczej (3)

Czas	Data
<u>Czas (KB)</u>	<u>Data (KB)</u>
godzina	rok
pora_dnia	miesiąc
	dzień
	sezon
	dzień_pracujący
	dzień_tygodnia
	święta

- 14 -

Przykład tabel wymiarów daty i czasu dla projektowanej hurtowni danych.

Techniki projektowania tabel wymiarów (5)

Wolno zmieniające się wymiary (ang. *Slowly Changing Dimensions*) określają te wymiary które zmieniają się w czasie.

SCD 1 – wartości są nadpisywane, brak historii, analizy mogą być zakłamane.

SCD 2 – nowe krotki jeżeli jakaś wartość ulega zmianie, klucz biznesowy nie zmienia wartości, jest generowany nowy klucz surogatowy. Dodatkowo w tabeli przechowywane są Data wstawienia krotki, Data kiedy krotka straciła aktualność i opcjonalnie pole określające czy krotka jest aktualna, czy też nie.

SCD 3 – obie wartości stara i nowa przechowywane są w tej samej krotce.

- 15 -

Kolejny aspekt projektowania tabel wymiarów dotyczy takich wymiarów, dla których wartości atrybutów wymiarów mogą zmieniać się w czasie. Takie wymiary nazywamy **wymiarami wolnozmieniającymi się**.

Wykonanie pojedynczej procedury leczniczej (4)

Jeden pacjent może w szpitalu pojawić się kilkakrotnie. Nie zawsze musi być w tym samym przedziale wiekowym.

Ten sam szpital, nie zawsze będzie miał taką samą wielkość, może się rozbudować...

Szpital
<u>Id_szpitala</u>
nazwa (KB)
wielkość
województwo
miasto
rodzaj
<u>data_wstawienia</u>
<u>data_aktualizacji</u>

Pacjent
<u>Id_pacjenta</u>
PESEL (KB)
nazwisko i imię
pleć
<u>czy_ubezpieczony</u>
zawód
wiek
województwo
miasto
<u>data_wstawienia</u>
<u>data_aktualizacji</u>

- 16 -

W przykładzie SCD 2 jest zaimplementowany w tabeli Szpital i Pacjent. Zauważmy, że jeżeli pacjent będzie kilkakrotnie poddawany procedurom medycznym, a jego kategoria wiekowa się zmieni to w hurtowni danych będzie kilka krotek tego samego pacjenta. Krotki te będą miały ten sam klucz biznesowy (PESEL) inny klucz surogatowy i inną wartość atrybutu wiek. To jest właśnie ten drugi powód, dla którego generujemy klucz surogatowy.

Wykonanie pojedynczej procedury leczniczej (5)

Dodatkowo:

- procedura medyczna została wykonana w sali o zadanym numerze (nr_sali), sal w szpitalu jest od 50 – 500 w zależności od szpitala;
- pacjent w czasie wykonania procedury był/lub nie znieczulany (typ_znieczulenie), mamy około 50 różnych typów znieczuleń;
- w czasie wykonywania procedury medycznej była z pacjentem osoba towarzysząca (osoba_twarzyszaca);
- procedura wymagała wzywania dodatkowej pomocy (dodatkowa_pomoc);
- ...

Inne
<u>Id_inne</u>
nr_sali
typ_znieczulenia
osoba_twarzyszaca
dodatkowa_pomoc
...

Liczba możliwych krotek = $500 \times 50 \times 2 \times 2 \times \dots$

- 17 -

Kiedy analizujemy fakt i jego wymiary, często zdarza się, że pozostaje nam pewna grupa cech, których nie można zdefiniować jako atrybuty wymiarów. Takich cech nie można pogrupować w logicznie powiązane ze sobą wymiary. W takiej sytuacji tworzy się dodatkową tabelę wymiarów np.: Inne, gdzie takie cechy się znajdują.

Wstawienie wszystkich cech (nie powiązanych logicznie) do tabeli Inne niesie za sobą ryzyko bardzo dużej liczby krotek w tej tabeli. Wynika to właśnie z faktu, że atrybuty tej tabeli nie są ze sobą logicznie powiązane. Łamie to ustaloną zasadę, że tabela wymiarów może mieć długie krotki, ale ich liczba nie powinna być zbyt duża.

Techniki projektowania tabel wymiarów (6)

Tworzymy tabelę wymiarów Inne dla atrybutów wymiarów ze sobą funkcjonalnie nie powiązanych.

Optymalizacja:

- wstawiamy tylko krótki faktycznie występujące, a nie wszystkie możliwe
- w przypadku zbyt dużej liczby krotek w tabeli Inne tworzymy tabelę Inne 1 i Inne 2.

W przykładzie tabela wymiaru Inne.

- 18 -

W jaki sposób rozwiązać ten problem?

Wykonanie pojedynczej procedury leczniczej (6)

Procedury medyczne są wykonywane w ramach jednego leczenia/pobytu.

Jest nadawany unikalny identyfikator takiego leczenia: [SYMBOL SZPITALA_NR_LECZENIA].

Poza tą informacją nie ma dodatkowych informacji o leczeniu pacjenta (uproszczenie).

Wykonanie procedury

Id_pacjenta
Id_szpitala
data
data_oplacenja_pacjent
data_oplacenja_NFZ
czas
Id_inne
Nr_leczenia
koszt_szpitala
oplata_pacjenta
wielkość_szpitala
wiek_pacjenta
refundacja_NFZ
ile_dni_do_opłacenia_pacjent
ile_dni_do_opłacenia_NFZ

- 19 -

Kolejne zagadnienie dotyczy problemu wymiarów zdegenerowanych. Czasami z faktem skojarzona jest cecha, której wartość stanowi krótki ciąg znaków lub liczb i wartość ta powtarza się jedynie dla kilku/kilkunastu faktów. W takiej sytuacji wartość ta nie trafia do wymiaru Inne, ani nie jest tworzona dla niej oddzielna tabela wymiarów, ale jest wpisywana bezpośrednio do tablicy faktów. W naszym przykładzie jest to Nr_leczenia. Inny rzeczywisty przykład to nr faktury dla faktu sprzedaży konkretnego produktu (więcej niż jeden produkt może być zakupiony na tą samą fakturę).

Należy pamiętać, że wymiary zdegenerowane to jedyne wartości, które nie są kluczami obcymi ani miarami, a które są zapisywane w tabeli faktów.

Należy pamiętać również o tym, że dotąd identyfikowaliśmy fakt samymi kluczami obcymi. W przypadku, kiedy są wymiary zdegenerowane fakt jest identyfikowany złożeniem kluczy obcych i wartościami wymiarów zdegenerowanych.

Techniki projektowania tabel wymiarów (7)

Wymiar zdegenerowany to wymiar, który nie posiada pogrupowanych do tej samej kategorii logicznej innych atrybutów niż klucz główny.

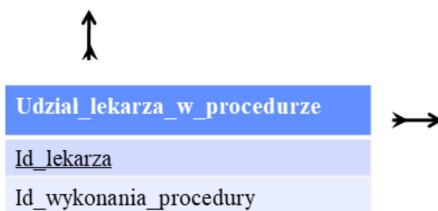
Wymiar zdegenerowany jest umieszczony w tabeli faktów z potwierdzeniem, że nie zawiera skojarzonej z nim tabeli wymiarów.

W przykładzie wymiarem zdegenerowanym jest nr_Leczenia.

Wykonanie pojedynczej procedury leczniczej (8)

Jedna procedura medyczna może być wykonywana przez kilku lekarzy.

Lekarz
<u>Id_lekarza</u>
Nr_zezwolenia (KB)
nazwisko i imię
specjalizacja



Wykonanie procedury
<u>Id_wykonania_procedury</u>
Id_pacjenta
Id_szpitala
data
data_opłacenia_pacjent
data_opłacenia_NFZ
czas
Id_inne
Nr_leczenia
koszt_szpitala
opłata_pacjenta
wielkość_szpitala
wiek_pacjenta
refundacja_NFZ
ile_dni_do_opłacenia_pacjent
ile_dni_do_opłacenia_NFZ

- 21 -

Do tej pory nie mówiliśmy jeszcze o sytuacji, w której wymiar W jest w związku wiele do wiele z faktem F.

Przykład: Procedura medyczna jest wykonywana przez więcej niż jednego lekarza. Jeden lekarz wykonuje więcej niż jedną procedurę.

Postępujemy w tym przypadku analogicznie, jak w relacyjnych bazach danych. Związek wiele do wiele implementujemy wstawiając nową tabelę z dwoma kluczami obcymi do tabeli faktów F i tabeli wymiarów W.

Zakładaliśmy, że tabela faktów nie musi mieć jawnie zdefiniowanego klucza głównego. W tym przypadku jednak niezbędne jest wygenerowanie tego klucza dla tabeli faktów.

Dodając tabelę Udział_lekarza_w_procedurze „zepsuliśmy” nasz schemat gwiazdy. I co teraz? Teraz już nawet trudno powiedzieć, co jest tabelą faktów, a co tabelą wymiarów.

Techniki projektowania tabel wymiarów (8)

W przypadku gdy grupy logiczne atrybutów opisujących dany fakt są powiązane ze sobą związkiem wiele do wiele, związek ten zostaje zaimplementowany zgodnie z zasadami określonymi dla relacyjnych baz danych.

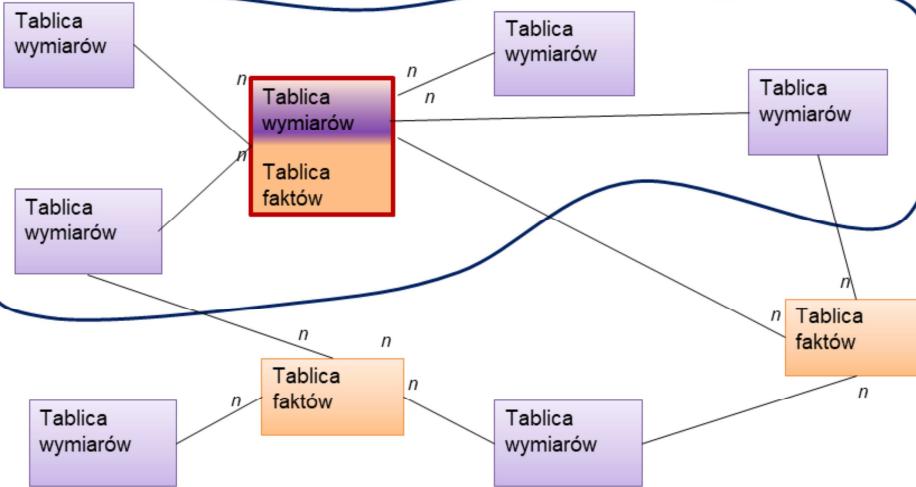
Nowa dodana tabela to tabela faktów (najczęściej bez miar).

W przykładzie mamy konstelację (dwie tabele faktów).

- 22 -

Konstelacja (1)

Gwiazda I



W przypadku gdy mamy więcej niż jedną tabelę faktów mamy schemat, który nazywamy konstelacją.

- 23 -

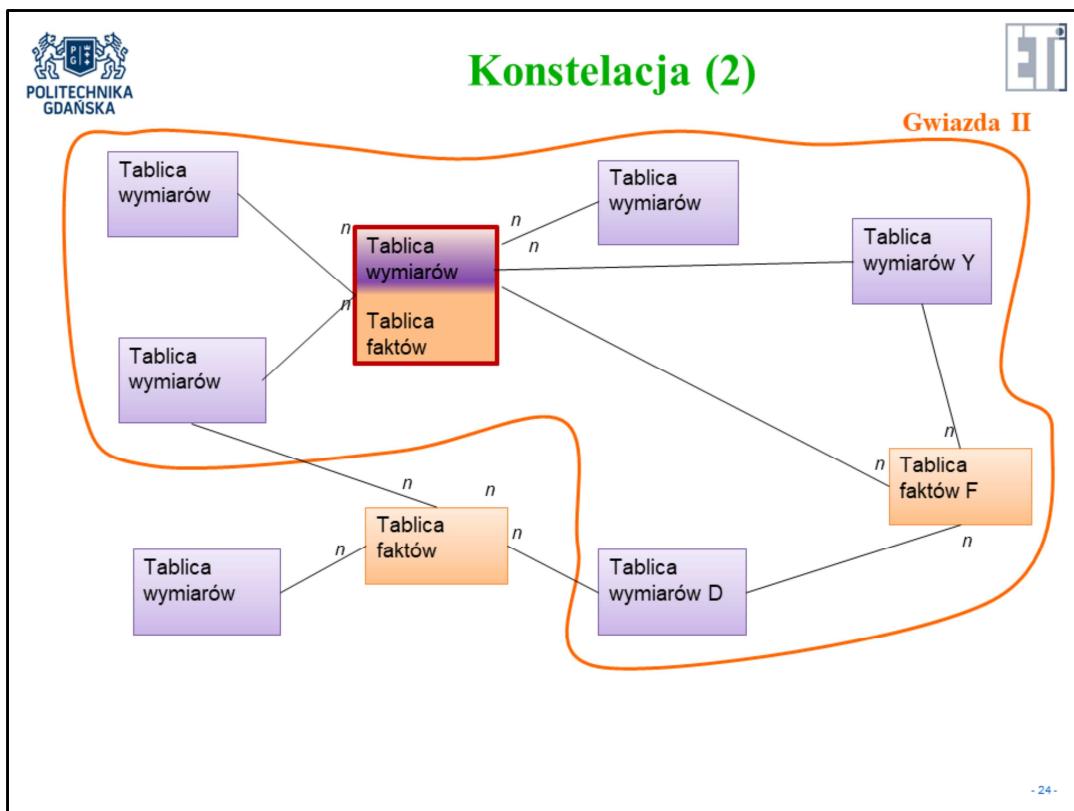
Tablice faktów i wymiarów mogą być ze sobą w różny sposób połączone.

Spójrzmy na slajd i skupmy naszą uwagę na granatowej pętli. Otacza ona cztery tabele wymiarów i jedną tabelę faktów (oznaczoną również jako tabela wymiarów - ale na razie się tym nie martwimy).

Jest to zwykły schemat gwiazdy.

Konstelacja (2)

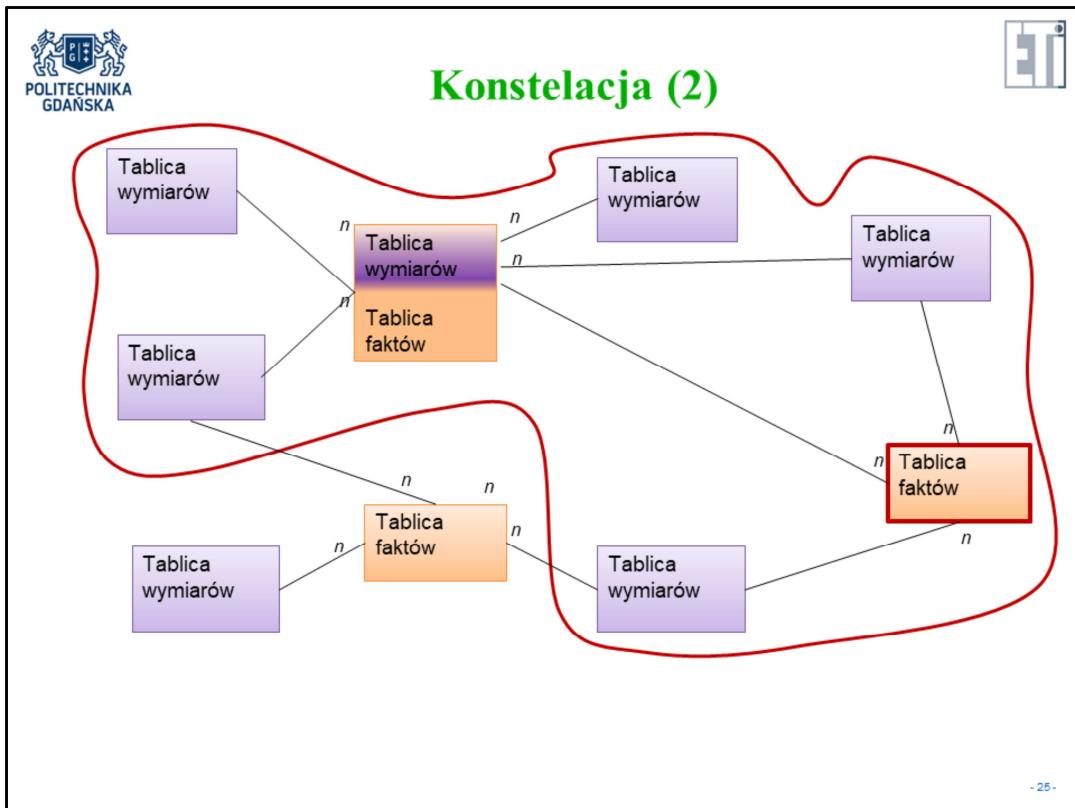
Gwiazda II



- 24 -

Gwiazdę I rozszerzyliśmy o dodatkową tabelę wymiarów D połączoną związkiem wiele do wiele (implementacja związku przez tabelę faktów F). Ponadto tablica wymiarów Y jest teraz podłączona dwoma logicznymi związkami z tabelą wymiarów (związkiem bezpośrednim i związkiem n:n implementowanym przy użyciu tabeli faktów F).

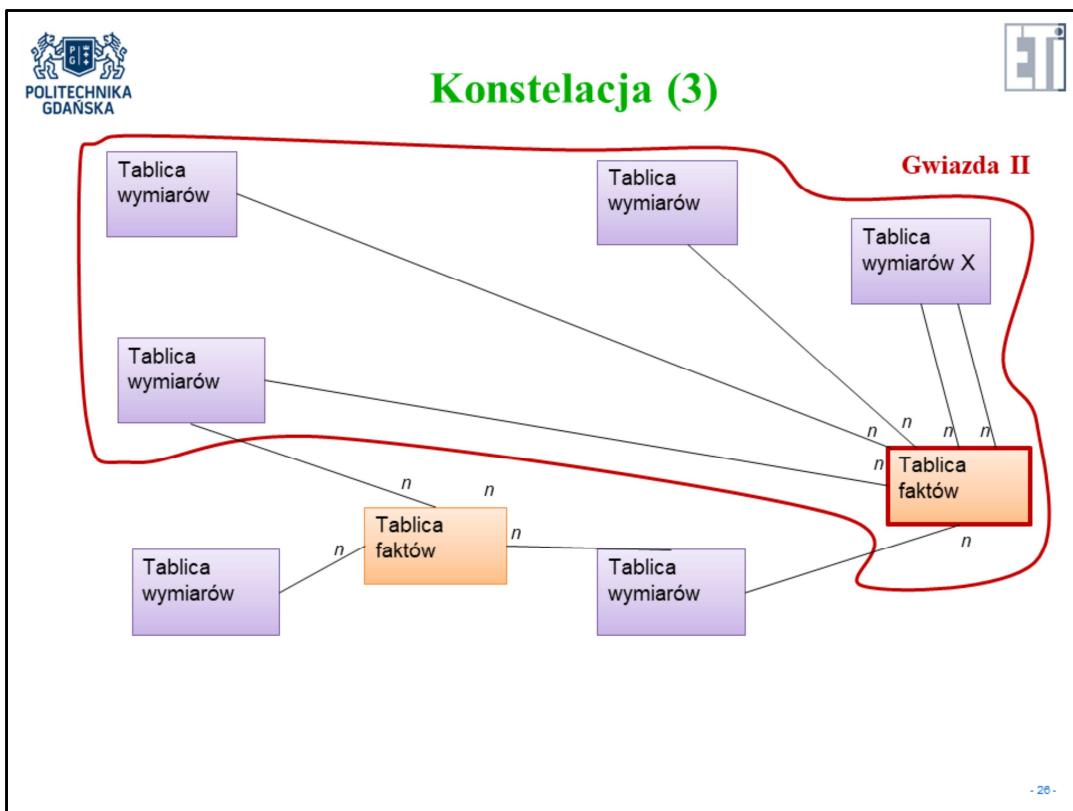
Konstelacja (2)



- 25 -

Teraz przyjrzyjmy się pętli czerwonej. Mamy tablicę faktów połączoną z tablicą faktów z Gwiazdy I (granatowej). Jednak tablica ta w czerwonej pętli to nic innego, jak tablica wymiarów. Ma klucz surogatowy i nie posiada żadnych wartości opisowych – atrybutów wymiarów, poza ewentualnie wymiarem zdegenerowanym.

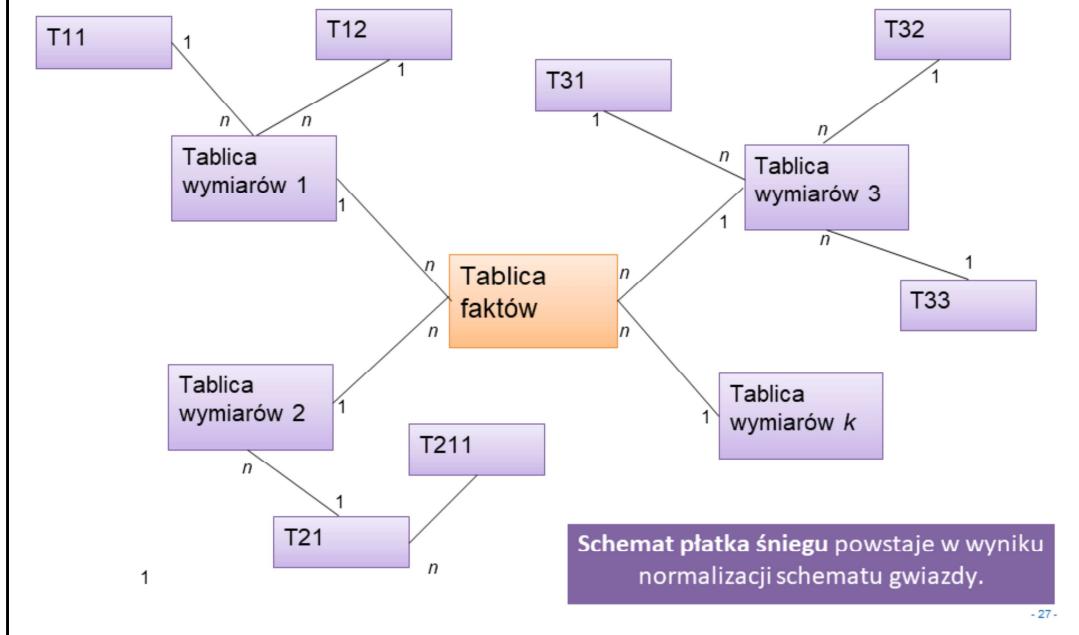
Konstelacja (3)



- 26 -

Schemat z poprzedniego slajdu jest pod względem informacji identyczny z tutaj przedstawionym (pod warunkiem przepisania wymiarów zdegenerowanych - z tabeli dwukolorowej do tabeli faktów). Zauważmy również, że pojawił się dodatkowy związek pomiędzy tabelą faktów i tabelą wymiarów X. Jest to ten sam związek, który na poprzednim slajdzie przechodzi przez tabelę dwukolorową. Nie może pozostać tylko jeden związek, gdyż logicznie związków te oznaczają co innego.

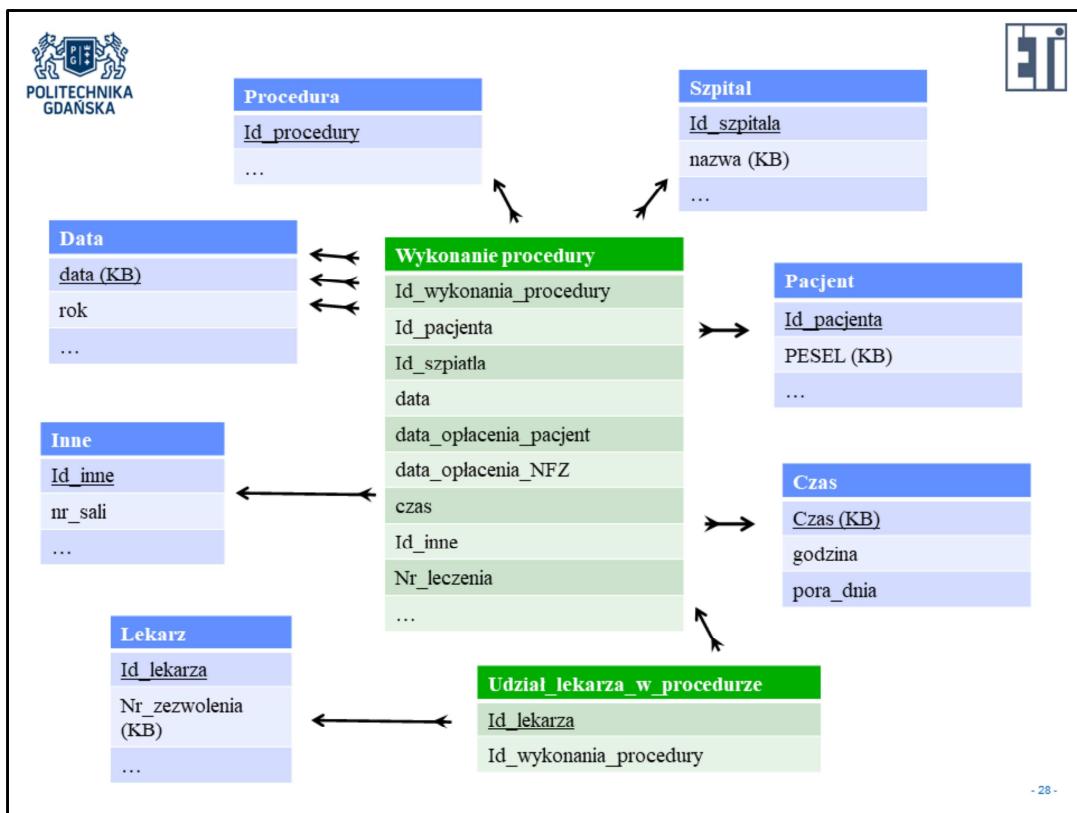
Płatek śniegu



Uważny student zauważycie, że na slajdzie 25 schemat otoczony czerwoną pętlą nie jest podpisany jako Gwiazda II. Uczyniłem to dopiero później na slajdzie 25. Wynika to z faktu, że schemat ze slajdu 25 to schemat płatka śniegu. Różni się od schematu gwiazdy tym, że jest częściowo znormalizowany. Każdy schemat płatka śniegu może zostać zdenormalizowany do schematu gwiazdy.

Zapamiętajmy:

Schemat gwiazdy jest zawsze poprawny!!! Schemat płatka śniegu ma listę wad! Jakie dokładnie opowiem przy okazji omawiania serwerów OLAP.



Na slajdzie zobrazowana jest konstelacja dla hurtowni danych – przypadek użycia NFZ.

Zdarzenia powtarzające się

Zdarzenia powtarzające się
Zdarzenia występujące co określony przedział czasowy.
Zazwyczaj są to zdarzenia reasumujące zdarzenia dyskretne.

Miesięczne zestawienie wykonanych procedur NFZ

Codzienny inwentarz leków

Miesięczna statystyka pacjentów

Cotygodniowe zestawienie wykonanych operacji

- 29 -

Należy pamiętać, że zdarzenia powtarzające się to nic innego, jak zdarzenie odzwierciedlające stan na pewną chwilę, w odniesieniu do ścisłe określonego przedziału czasowego.

Zdarzenia powtarzające się

1. Możliwość zamodelowania zdarzenia na poziomie dyskretnym.

Praktyczne? Niepraktyczne?

2. Nie wszystkie metryki da się wyrazić przy ziarnistości zdarzenia dyskretnego.



Jak określić zużycie maści w ramach jednej aplikacji?

- 30 -

Dlaczego wprowadzono takie fakty? Jeżeli zdarzenie powtarzające się reasumuje zdarzenia dyskretne to dlaczego nie ustawić dla poszczególnych miar funkcji agregujących i nie wyliczać potrzebnych wartości?
Zazwyczaj takie rozwiązanie jest niepraktyczne, ze względu na czas wykonywania obliczeń. Istnieje jednak jeszcze inny powód. Mianowicie, nie dla każdej miary da się wyliczyć jej wartość na poziomie zdarzeń dyskretnego. Przykładem jest zużycie maści czy pasty do zębów.

Zdarzenie powtarzające się

Codzienny inwentarz leków

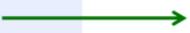
Faktem będzie całodzienna inwentaryzacja leków.



- 31 -

Na slajdzie przedstawiono uproszczony schemat hurtowni danych dla faktu codziennej inwentaryzacji leków.

Inwentarz leków
Data
Id_szpitala
Id_leku
Ilość na stanie
Ilość dostarczonych
Koszt zakupu
Ilość zużytych
Koszt zużytych



Miara pół-addytywna

- 32 -

Przy zdarzeniach powtarzających się należy zwrócić uwagę na fakt, że często pojawiają się miary określające stan po inventaryzacji. Wszystkie tego typu miary są pół-addytywne.

Zdarzenia rozwijające się

Zdarzenia rozwijające się

Zdarzenia występujące w dłuższym przedziale czasowym. Zazwyczaj stanowią serię zdarzeń dyskretnych.

Pobyt pacjenta w szpitalu

Accumulating Snapshot

Faktem będzie **pobyt pacjenta w szpitalu**.



- 33 -

Ostatni – trzeci typ zdarzeń odnosi się do zdarzeń rozwijających się. Typową cechą tych zdarzeń jest więcej niż jedna data.

Techniki projektowania tabel wymiarów

Tabele wymiarów pełniące różne role (ang. *Role Playing Dimensions*).

Tabela faktów jest powiązana dwa lub więcej razy z tą samą tabelą wymiarów (więcej niż jeden związek). Każdy związek oznacza inną rolę.

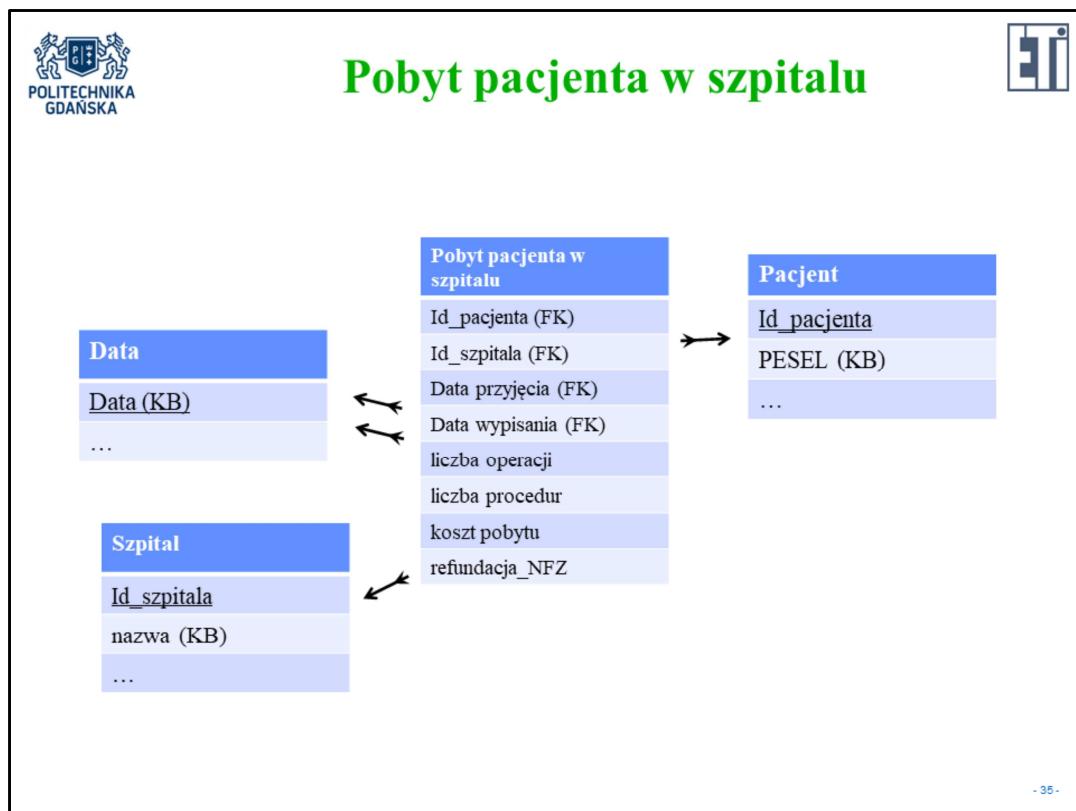
W przykładzie mamy dwa związki:

1. data przyjęcia do szpitala
2. data wypisania ze szpitala

- 34 -

Warto zwrócić uwagę, że jeżeli zdarzenie jest opisane więcej niż jedną datą, nie oznacza to tworzenia nowych tabel wymiarów. Tworzone są nowe związki.

Pobyt pacjenta w szpitalu



Przykładowo. Pobyt pacjenta zaczyna się w dniu przyjęcia do szpitala, a kończy się datą wypisania ze szpitala. W tabeli faktów pojawiają się dwa klucze obce do tabeli wymiarów Data reprezentujące związki data przyjęcia do szpitala i data wypisania ze szpitala.

Takie momenty czasowe, które występują w każdym pojedynczym zdarzeniu reprezentującym fakt pobytu pacjenta w szpitalu nazywamy kamieniami milowymi. Każde zdarzenie rozwijające się może być podzielone kamieniami milowymi na kilka etapów. Po zakończenie etapu (kamień milowy) są znane wartości pewnych metryk.

W przykładzie mamy dwa kamienie milowe:

Przyjęcie do szpitala

Wypisanie ze szpitala

Przyjęcie do szpitala generuje tylko jedną miarę: liczba pobytów (dla każdego faktu równą 1).

Wypisanie ze szpitala generuje wartości dla pozostałych miar.

Zdarzenia rozwijające się cd.

Id pacjenta	Id szpitala	Data przyjęcia	Data wypisania	Liczba operacji	Liczba procedur	Koszt pobytu	Refundacja
1	3	2012.08.09	0	null	null	null	null

Id pacjenta	Id szpitala	Data przyjęcia	Data wypisania	Liczba operacji	Liczba procedur	Koszt pobytu	Refundacja
1	3	2012.08.09	2012.08.30	1	12	1200,-	1200,-

Zdarzenia rozwijające się mają zazwyczaj również cechy zdarzeń powtarzających się !

- 36 -

Na slajdzie została pokazana zawartość tabeli faktów (dla jednego zdarzenia pobytu pacjenta w szpitalu) dla dwóch kamieni milowych (przyjęcia i wypisania ze szpitala).

Typy zdarzeń - porównanie

	Dyskretne	Powtarzające się	Rozwijające się
Okresowość	Dyskretny punkt w czasie	Cykliczne, powtarzające się co określony przedział czasowy	Niezdeterminowany okres czasu dla danego przepływu wydarzeń

- 37 -

Zakończmy ten krótki wykład porównaniem trzech typów zdarzeń.

Pierwsza z cech to **okresowość**. Przez tę cechę rozumiemy kiedy to zdarzenie występuje, jak długo trwa i do jakiego czasu się odnosi.

Zdarzenia dyskretne występują w pewnym dyskretnym momencie czasowym, a jego cechy odnoszą się do tego konkretnego momentu czasowego.

Zdarzenie powtarzające się występuje również w pewnym dyskretnym momencie czasowym, ale ściśle określonym. Wiadomo kiedy było zdarzenie poprzednie i kiedy nastąpi zdarzenie kolejne.

Zdarzenie rozwijające się rozpoczyna się w jednym dyskretnym momencie czasowym i trwa do następnego momentu czasowego. Czas trwania takiego zdarzenia jest niezdeterminowany.

Typy zdarzeń - porównanie

	Dyskretne	Powtarzające się	Rozwijające się
Okresowość	Dyskretny punkt w czasie	Cykliczne, powtarzające się co określony przedział czasowy	Niezdeterminowany okres czasu dla danego przepływu wydarzeń
Ziarnistość	1 wiersz na zdarzenie dyskretne	1 wiersz na zdarzenie reasumujące	1 wiersz na cały przepływ wydarzeń

- 38 -

Druga z cech to **ziarnistość**. Przez ziarnistość rozumiemy z jaką dokładnością zapisywane są fakty.

Typy zdarzeń - porównanie

	Dyskretne	Powtarzające się	Rozwijające się
Okresowość	Dyskretny punkt w czasie	Cykliczne, powtarzające się co określony przedział czasowy	Niezdeterminowany okres czasu dla danego przepływu wydarzeń
Ziarnistość	1 wiersz na zdarzenie dyskretne	1 wiersz na zdarzenie reasumujące	1 wiersz na cały przepływ wydarzeń
Daty wymiarów	Data transakcji - zdarzenia dyskretnego	Data utworzenia podsumowania	Wiele dat dla poszczególnych kamieni milowych danego przepływu

- 39 -

Kolejna cecha to **daty wymiarów** określająca ile takich dat jest zdefiniowanych dla pojedynczego zdarzenia.

Typy zdarzeń - porównanie

	Dyskretne	Powtarzające się	Rozwijające się
Okresowość	Dyskretny punkt w czasie	Cykliczne, powtarzające się co określony przedział czasowy	Niezeterminowany okres czasu dla danego przepływu wydarzeń
Ziarnistość	1 wiersz na zdarzenie dyskretne	1 wiersz na zdarzenie reasumujące	1 wiersz na cały przepływ wydarzeń
Daty wymiarów	Data transakcji - zdarzenia dyskretnego	Data utworzenia podsumowania	Wiele dat dla poszczególnych kamieni milowych danego przepływu
Miary	Transakcyjne	Kumulacyjne dla danego przedziału czasowego	Związane z poszczególnymi kamieniami milowymi

- 40 -

Cecha **miary** definiuje do jakiego okresu czasowego odnoszą się wartości miar. W przypadku zdarzeń dyskretnych jest to zawsze pojedynczy moment czasowy. W przypadku zdarzeń powtarzających się miary odnoszą się do momentu wystąpienia zdarzenia powtarzającego (stany) oraz do przedziału czasowego, do którego odnosi się całe zdarzenie. W zdarzeniach rozwijających się miary odnoszą się do kamieni milowych określonych dla zdarzeń rozwijających się.

Typy zdarzeń - porównanie

	Dyskretne	Powtarzające się	Rozwijające się
Okresowość	Dyskretny punkt w czasie	Cykliczne, powtarzające się co określony przedział czasowy	Niezdeterminowany okres czasu dla danego przepływu wydarzeń
Ziarnistość	1 wiersz na zdarzenie dyskretne	1 wiersz na zdarzenie reasumujące	1 wiersz na cały przepływ wydarzeń
Daty wymiarów	Data transakcji - zdarzenia dyskretnego	Data utworzenia podsumowania	Wiele dat dla poszczególnych kamieni milowych danego przepływu
Miary	Transakcyjne	Kumulacyjne dla danego przedziału czasowego	Związane z poszczególnymi kamieniami milowymi
Gęstość krotek w tabeli faktów	Gęste lub rzadkie, w zależności od aktywności	Gęste	Gęste lub rzadkie, w zależności od przepływu

- 41 -

Co rozumiemy, przez gęstość krotek w tabeli faktów? Przypomnijmy sobie, że mając zdefiniowane 3 wymiary odpowiednio o 5, 6 i 10 elementach wymiarów możemy w kostce zapisać 300 faktów. Możemy ich jednak zapisać również mniej niż 300. Krotki są gęste w tabeli faktów jeżeli ich liczba jest bliska 300, rzadkie jeżeli jest ich zdecydowanie mniej.

Typy zdarzeń - porównanie

	Dyskretne	Powtarzające się	Rozwijające się
Okresowość	Dyskretny punkt w czasie	Cykliczne, powtarzające się co określony przedział czasowy	Niezdeterminowany okres czasu dla danego przepływu wydarzeń
Ziarnistość	1 wiersz na zdarzenie dyskretne	1 wiersz na zdarzenie reasumujące	1 wiersz na cały przepływ wydarzeń
Daty wymiarów	Data transakcji - zdarzenia dyskretnego	Data utworzenia podsumowania	Wiele dat dla poszczególnych kamieni milowych danego przepływu
Miary	Transakcyjne	Kumulacyjne dla danego przedziału czasowego	Związane z poszczególnymi kamieniami milowymi
Gęstość krotek w tabeli faktów	Gęste lub rzadkie, w zależności od aktywności	Gęste	Gęste lub rzadkie, w zależności od przepływu
Aktualizacja tabeli faktów	Aktualizacja tylko błędnych danych	Aktualizacja tylko błędnych danych	Aktualizacja w momencie pojawienia się aktywności w ramach przepływu

- 42 -

Aktualizacja tabeli faktów oznacza kiedy dane dotyczące pojedynczego faktu są wpisywane do hurtowni danych.

Co każdy student potrafić powinien...

Zaprojektować hurtownię danych dla różnych typów zdarzeń.

