STUDIA INFORMATICA 2003
Volume 24 Number 5 (48)

Zygmunt BOK
Zakłady Tworzyw Sztucznych ”Nitron” S.A.

MODEL SCHEMATU ŚCIEŻEK ANALIZY WIELOWYMIAROWEJ

Streszczenie. W tym artykule przedstawiono formalny model schematu ścieżek analizy wielowymiarowej. W oparciu o ten model oraz schemat ER zastanego systemu informacyjnego określono przykładowy schemat ścieżek analizy wielowymiarowej. Na jego podstawie konstruowano właściwe pytania analityczne, które kierowano do przykładowej tematycznej hurtowni danych. Pytania analityczne formułowano w oparciu o kombinacje różnych ścieżek analizy określonych w schemacie ścieżek analizy wielowymiarowej. Umożliwiono w ten sposób potencjalną możliwość ich realizacji.

MULTIDIMENSIONAL ANALYTICAL PATHS SCHEMA MODEL

Summary. In this article a formal multidimensional analytical paths schema model was presented. Based on this model as well as ER schema from legacy informational OLTP system a multidimensional analytical paths schema example was determined. On his basis an analytical questions was constructed which was directed to data warehouse example. Based on combination different analytical paths determined in multidimensional analytical paths schema an analytical example question was formulated. In this way a potential possibility of their execution was made possible.

1. Wprowadzenie

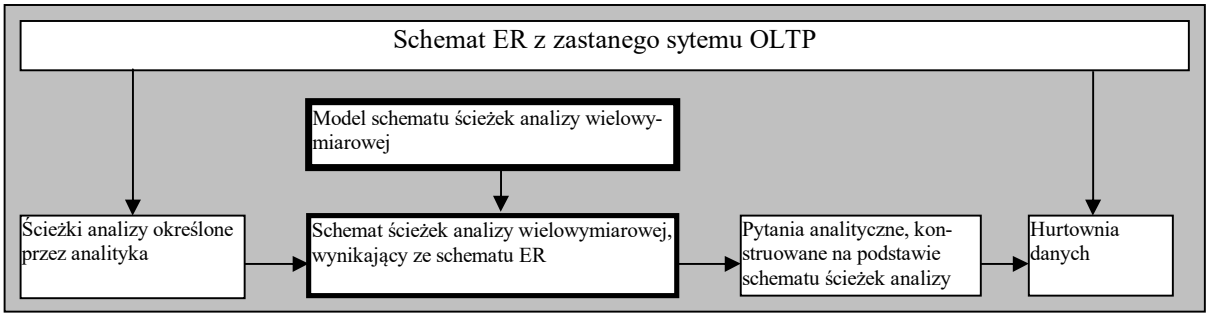
W chwili obecnej na poziomie projektowania konceptualnego hurtowni danych istnieje szereg pozycji, które nie zostały należycie zbadane. Mało zostało również powiedziane odnośnie projektów konceptualnych, biorąc pod uwagę wymagania użytkownika jako punkt początkowy. Na tym polu obok historycznie już pierwszej pracy [1] oraz innych [2, 3, 4] wyróżnia się praca [5]. Poświęcona jest ona problemowi modelowania danych używanych w analizie wielowymiarowej na poziomie konceptualnym. Autorzy tej pracy postrzegając ten problem z

perspektywy użytkownika końcowego, opisują zbiór wymagań niezbędnych dla modelowania konceptualnych scenariuszy OLAP-owych świata rzeczywistego. Bazując na tych wymaganiach zdefiniowali nowy konceptualny wielowymiarowy model danych MAC (Multidimensional Aggregation Cube data model) zdolny objąć swoim zasięgiem i wyrazić statyczne właściwości rozpatrywanych informacji. Zaproponowali oni nieco inne podejście do definicji użytecznego konceptualnego modelu danych, w którym informacja użyta w analizie wielowymiarowej jest podstawowym obiektem ich pojęć modelujących. Informacja użyta w procesie analizy stanowi zagregowane dane na różnych poziomach agregacji lub kombinacji tych poziomów. Takie podejście stanowi przeciwieństwo do podejścia zaproponowanego przez [6, 7], polegającego na rozszerzeniu modelu ER dla wielowymiarowego paradygmatu [6, 8], koncentrujące się na reprezentacji szczegółowych danych źródłowych (source-detailed data). Z przeprowadzonej przez autorów modelu MAC dyskusji nad wymaganiami dotyczącymi konceptualnego modelu danych odpowiedniego do analizy wielowymiarowej wynika, że powinien on umożliwić definiowanie poziomów wymiaru, relacji grupowania/klasyfikowania (tzn. relacje łączące poziomy) oraz ścieżek analizy. Zaproponowany model danych MAC jest użytkowo-centrycznym (user-centric) konceptualnym modelem danych, zapewniającym wysoki poziom ekspersji oraz intuicyjną metodologię modelowania informacji użytej w wielowymiarowej analizie. Model MAC opisuje dane za pomocą pojęć takich jak: poziomy wymiaru, relacji zwijania, ścieżek wymiaru, sześcianów i atrybutów, które znaczeniowo bliskie są sposobowi postrzegania informacji przez OLAP-owych użytkowników. Pomimo, że autorzy modelu MAC, nie wyprowadzali schematu hurtowni z zaproponowanego modelu konceptualnego, jak również nie rozpatrywali sposobu jej ładowania informacjami pochodzącymi z systemów transakcyjnych OLTP, niemniej jednak zaproponowany w ich pracy model MAC jest odpowiedni dla użytkowników hurtowni danych, którzy dokonują analizy informacji za pomocą aplikacji OLAP-owych, w szczególności pytań analitycznych pojawiających się *ad-hoc*.

2. Opis problemu

W celu ułatwienia konstrukcji pytań analitycznych *ad-hoc* kierowanych do hurtowni danych, postanowiono wykorzystać wprowadzoną w modelu MAC koncepcję ścieżek analizy. Bazując na tej koncepcji, zaproponowano i zdefiniowano formalny model schematu ścieżek analizy wielowymiarowej, który wykorzystano do określenia konkretnego schematu ścieżek analizy wielowymiarowej. Schemat ten stanowił podstawę konstrukcji pytań analitycznych *ad-hoc*, kierowanych do przykładowej tematycznej hurtowni danych. Pytania analityczne użytkownika formułowano w oparciu o kombinacje różnych ścieżek analizy określonych w sche-

macie ścieżek analizy wielowymiarowej, co schematycznie pokazano na rys. 1, zapewniając tym samym potencjalną możliwość realizacji takich pytań. Założono, że zarówno schemat ścieżek analizy wielowymiarowej oraz schemat przykładowej tematycznej hurtowni danych określono na podstawie schematu ER z zastanego systemu informacyjnego OLTP.



Rys 1. Konstrukcja pytań analitycznych na podstawie schematu ścieżek analizy
Fig 1. The analytical queries construction based on analytical paths schema

Ponadto przyjęto, że początkowy schemat tematycznej hurtowni danych do którego kierowano te pytania określono za pomocą metody opisanej w publikacji [9], wykorzystującej tradycyjny model ER do projektowania hurtowni danych na podstawie przemysłowych modeli danych. Cechą charakterystyczną tej metody jest zastosowanie denormalizacji do wstępnie pogrupowanych encji z zastanego modelu ER przemysłowego systemu informacyjnego według przyjętych przez autorów trzech kategorii klasyfikujących, tj. kategorii encji transakcyjnych, klasyfikacyjnych oraz komponentowych. Poprzez zastosowanie operatora agregacji w stosunku do encji transakcyjnych możliwe jest utworzenie nowych encji zawierających zagregowane dane. W ten sposób, za pomocą tej metody można w łatwy sposób określić różne typy schematów hurtowi danych. W przypadku projektowania schamatu hurtowni danych typu gwiazda, tablica faktów formowana jest na podstawie encji transakcyjnych, natomiast tablice określające wymiary tworzone są dla każdej encji komponentowej poprzez denormalizację hierarchicznie powiązanych encji klasyfikujących.

3. Model schematu ścieżek analizy wielowymiarowej

W celu sformalizowania wprowadzonego przez autorów modelu MAC pojęcia ścieżek analizy, zaproponowano wprowadzenie pojęcia schematu ścieżek analizy wielowymiarowej. Dla jego formalnego zdefiniowania wprowadzono za [10, 11, 12, 13, 14] poniższe definicje.

Definicja 1. Grafem G=(V, E) nazywamy sieć składającą się ze zbioru węzłów $V=\{v_1, v_2, ...\}$ oraz zbioru krawędzi $E=\{e_1, e_2, ...\}$ [10, 11]. Krawędź e_k utożsamia się z nie-

uporządkowaną parą węzłów (v_i, v_j). Węzły v_i, v_j związane z krawędzią e_k nazywa się węzłami końcowymi krawędzi e_k.

Definicja 2. Krawędzią grafu G=(V, E) nazywamy [12] dowolną nieuporządkowaną parę {e_i, e_j} taką, że ((e_i, e_j) ∈E)□ ((e_j, e_i) ∈E).

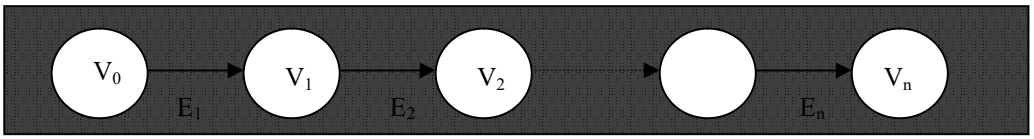
Definicja 3. Ukierunkowanym (zorientowanym) [10, 13] grafem nazywamy taki graf G=(V,E), w którym E jest zbiorem takich uporządkowanych par (e_i, e_j) dla których krawędź łącząca dwa węzły v_i, v_j posiada określony kierunek.

Definicja 4. Ścieżką w ukierunkowanym grafie G=(V, E) nazywamy [13] ciąg krawędzi (e_1,e_2), (e_2,e_3),..., (e_{n-1},e_n).

Definicja 5. Ścieżką acykliczną nazywamy taką ścieżkę, którą można przemierzyć (pokonać) tylko w jeden sposób [10].

Definicja 6. Ukierunkowanym acyklicznym grafem nazywamy taki graf, w którym istnieje tylko jedna acykliczna ścieżka pomiędzy dowolną parą węzłów [10].

Definicja 7. Niech **g**=(V, E) będzie ukierunkowanym acyklicznym grafem [14], gdzie V jest zbiorem węzłów, natomiast E jest zbiorem krawędzi. Mówimy, że **g** – co pokazano na rys. 2 - jest quasi-drzewem z korzeniem w $V_0 \in V$, jeśli każdy wierzchołek $V_j \in V$ może być osiągnięty z v_0 za pomocą przynajmniej jednej ukierunkowanej ścieżki. Oznaczmy przez $s_{ij} \subseteq \mathbf{g}$ ukierunkowaną ścieżkę (jeśli istnieje) rozpoczynającą się w V_i i kończącą się w V_j. Oznaczmy dalej przez sub(V_i)$\subset$**g** quasi drzewo zakorzenione w węźle $V_i \neq V_0$.



Rys. 2. Przykład ukierunkowanego acyklicznego grafu zakorzenionego w V_0
Fig. 2. The example of the directed acyclic graph rooted in V_0

Korzystając z przytoczonych definicji oraz formalizmu zaproponowanego przez [15], poniżej przedstawiono formalną definicję pojęcia schematu ścieżek analizy wielowymiarowej.

Definicja 8. Schemat ścieżek analizy wielowymiarowej $\mathbf{S}_{saw}$=(M, W, P, S) stanowi grupa powiązanych danych, gdzie:

M – jest zbiorem miar. Każda miara $M_n \in M = \sum_{n=1}^{m} \{M_n\}$, definiowana jest przez wyrażenia numeryczne pochodzące z systemów informacyjnych,

W – jest zbiorem wymiarów w analizie wielowymiarowej, tj. $W = \sum_{i=1}^{w} \{W_i\}$,

P – jest zbiorem wszystkich poziomów analizy, tj. $P = \sum_{i=1}^{w} \sum_{j} \{P_{ij}\}$,

gdzie $P_{ij} = \sum_{r} \{p_{ijr}\}$ jest zbiorem poziomów w ścieżkach analizy pewnego wymiaru $W_i \in W$, w którym i – oznacza numer wymiaru, j - oznacza numer ścieżki analizy, natomiast r – oznacza numer poziomu analizy,

S – jest zbiorem wszystkich uporządkowanych podzbiorów, każdy składający się ze zbioru uporządkowanych par, tj. $S = \sum_{i=1}^{w} \sum_{j} \{S_{ij}\}$, gdzie $S_{ij} = \sum_{u} (p_{ijx}, p_{ijy})_u$ jest zbiorem uporządkowanych par, w którym i – oznacza numer wymiaru, j - oznacza numer ścieżki analizy, u – oznacza takie uporządkowane pary w których x<y oraz x,y∈<1, r>, natomiast r – oznacza ilość poziomów analizy. Uporządkowane pary, określające ukierunkowane ścieżki analizy $s_{ij,xy}=(p_{ijx}, p_{ijy})$, modelują relacje typu wiele do jednego. Za pomocą ukierunkowanych ścieżek analizy, poziom analizy p_{ijy} może być osiągnięty, wychodząc od poziomu analizy p_{ijx}, gdzie: $p_{ijy} \in \{p_0\} \cup P_{ij} = \{p_0\} + \{p_{ij1}\} + \{p_{ij2}\} + ... + \{p_{ijy}\}$

$p_{ijx} \in P_{ij} = \{p_{ij1}\} + \{p_{ij2}\} + ... + \{p_{ijx}\}$

Jeśli zatem dla dowolnego wymiaru $W_i \in W$, każdy zbiór $S_{ij} \in S$ jest takim zbiorem, że graf **g**(V, E), gdzie: $V = \{p_0\} \cup P_{ij}$

$E = S_{ij}$ (j - oznacza numer ścieżki analizy w ramach i-tego wymiaru)

jest ukierunkowanym, acyklicznym grafem zakorzenionym w $p_0 \in P$ takim, że każdy poziom analizy $p_{ijx} \in P_{ij}$ znajdujący się na j-tej ścieżce i-tego wymiaru może być osiągnięty wychodząc z poziomu analizy p_0 za pomocą przynajmniej jednej ukierunkowanej ścieżki, wówczas grupa powiązanych danych $\mathbf{S}_{śaw}$=(M, W, P, S) stanowi schemat ścieżek analizy wielowymiarowej.

Jak widać z powyższego rysunku poziomy analizy połączone są ze sobą za pomocą relacji grupujących lub klasyfikujących. Ścieżki reprezentują sekwencję uzasadnionych operacji grupujących, które mogą być wykonywane podczas analizy wielowymiarowej.

3.1. Ścieżki skrośne w schemacie ścieżek analizy wielowymiarowej

Wykorzystując wprowadzone definicją 6 pojęcie quasi-drzewa zakorzenionego w węźle $V_i \neq V_0$ oznaczonego symbolem sub(V_i) oraz formalnie zdefiniowanego pojęcia schematu ścieżek analizy wielowymiarowej, do dalszych rozważań wprowadza się pojęcie ścieżki analizy skrośnej w schemacie ścieżek analizy wielowymiarowej.

Definicja 9. Dla danego schematu ścieżek analizy wielowymiarowej $\mathbf{S}_{śaw}$=(M, W, P, S), prostą skrośną ścieżką analizy nazywa się taką ukierunkowaną ścieżkę analizy, która stanowi sumę dwóch ukierunkowanych ścieżek analizy spełniających następujące warunki:

1º pierwsza z ukierunkowanych ścieżek $s_{ij,\,vx}=(p_{ijv}, p_{ijx})$ pewnego wymiaru $W_i \in W$ zakorzeniona jest w $p_0 \in P_{ij}$, gdzie i – oznacza numer wymiaru, j - oznacza numer ścieżki analizy, natomiast v, x, y – oznaczają numery poziomów analizy,

2º druga z ukierunkowanych ścieżek $s_{ij,\,yz}=(p_{ijy}, p_{ijz})$ z tego samego wymiaru jest quasi-drzewem sub(P_{ij}) zakorzenionym w węźle $P_{ij} \neq p_0$,

3º ścieżka $s_{ij,\,vz}=(p_{ijv}, p_{ijx}) + (p_{ijy}, p_{ijz})$ jest quasi-drzewem z korzeniem w węźle $p_0 \in P$.

Definicja 10. Dla danego schematu ścieżek analizy wielowymiarowej $\mathbf{S}_{śaw}$=(M, W, P, S), złożoną skrośną ścieżką analizy nazywa się taką ukierunkowaną ścieżkę analizy, która stanowi sumę dwóch lub więcej ukierunkowanych ścieżek analizy spełniających następujące warunki:

1º pierwsza z ukierunkowanych ścieżek $s_{ij,\,vx}=(p_{ijv}, p_{ijx})$ pewnego wymiaru zakorzeniona jest w $p_0 \in P_{ij}$, gdzie i – oznacza numer wymiaru, j - oznacza numer ścieżki analizy, natomiast v, x, y – oznaczają numery poziomów analizy,

2º druga lub dalsze z ukierunkowanych ścieżek $s_{mn,\,yz}=(p_{mny}, p_{mnz})$ z innych wymiarów są quasi-drzewemi sub(P_{mn}) zakorzenione są w węzłach $P_{mn} \neq p_0$,

3º ścieżka $s_{complex}=(p_{ijv}, p_{ijx}) +...+ (p_{mny}, p_{mnz})$, jest quasi-drzewem z korzeniem w węźle $p_0 \in P$.

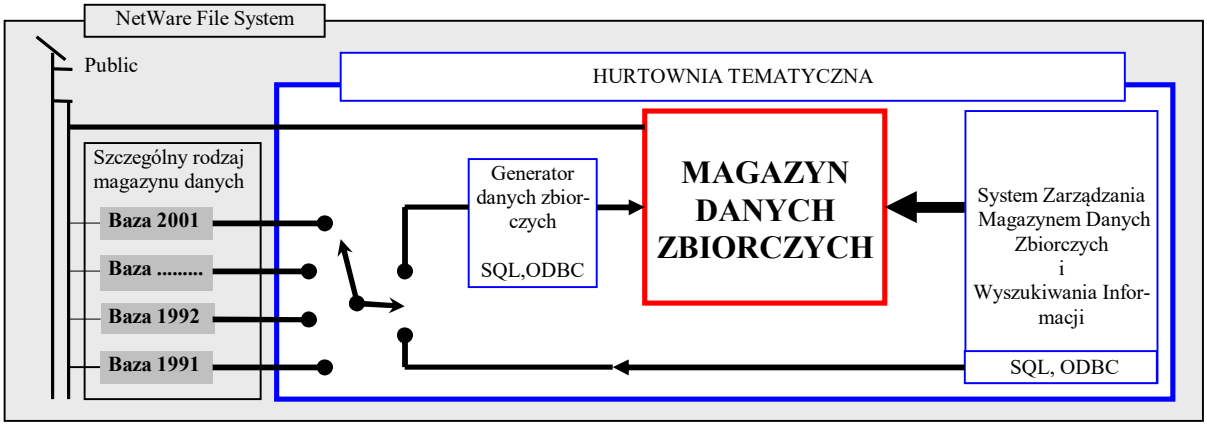
4. Realizacja niektórych pytań analitycznych *ad-hoc*, konstruowanych na podstawie schematu ścieżek analizy wielowymiarowej

Bazując na wprowadzonym modelu ścieżek analizy wielowymiarowej, pytania analityczne *ad-hoc* kierowane do przykładowej tematycznej hurtowni danych, konstruowano na podstawie schematu ścieżek analizy wielowymiarowej. Schemat ścieżek analizy wielowymiarowej określono na podstawie zaproponowanego formalnego modelu ścieżek analizy wielowymiarowej oraz zastanego modelu danych obecnie eksploatowanego w ZTS ”Nitron” S.A. przemysłowego systemu informacyjnego o nazwie ”*Wyroby Gotowe*”.

4.1. Architektura przykładowej tematycznej hurtowni danych

Z wspomnianego zastanego systemu informacyjnego istnieje potrzeba uzyskania informacji zbiorczych na podstawie danych zawartych w relacyjnych archiwalnych bazach danych dotyczących poprzednich zamkniętych okresów obliczeniowych. Zbiór archiwalnych baz danych z lat 1991-2001 zeskładowanych w oddzielnych katalogach systemu plików pewnego serwera sieciowego, potraktowano jako podstawowe repozytorium informacji lub inaczej jako pewny

szczególny rodzaj magazynu danych, który w pewnym sensie podobny jest do wydzielonych systemów baz danych wspomagających przetwarzanie analityczne. Architektura tych systemów zakłada bowiem pełną izolację przetwarzania operacyjnego i analitycznego. Informacje powstające w operacyjnych bazach danych tych systemów są replikowane i fizycznie składowane w pewnym magazynie danych do późniejszego przetwarzania analitycznego. Ponieważ w opisywanym przypadku istnieje pełna izolacja pomiędzy bazami archiwalnymi i operacyjnymi, jak również to, że nie ma potrzeby replikowania danych archiwalnych do osobnego magazynu danych, zatem w tym właśnie sensie wyżej wymieniony szczególny rodzaj magazynu danych podobny jest do wydzielonych systemów baz danych wspomagających przetwarzanie analityczne. Postanowiono zatem wykorzystać to podobieństwo do zbudowania prostej tematycznej hurtowni tematycznej, której architekturę przedstawiono na rys. 3.



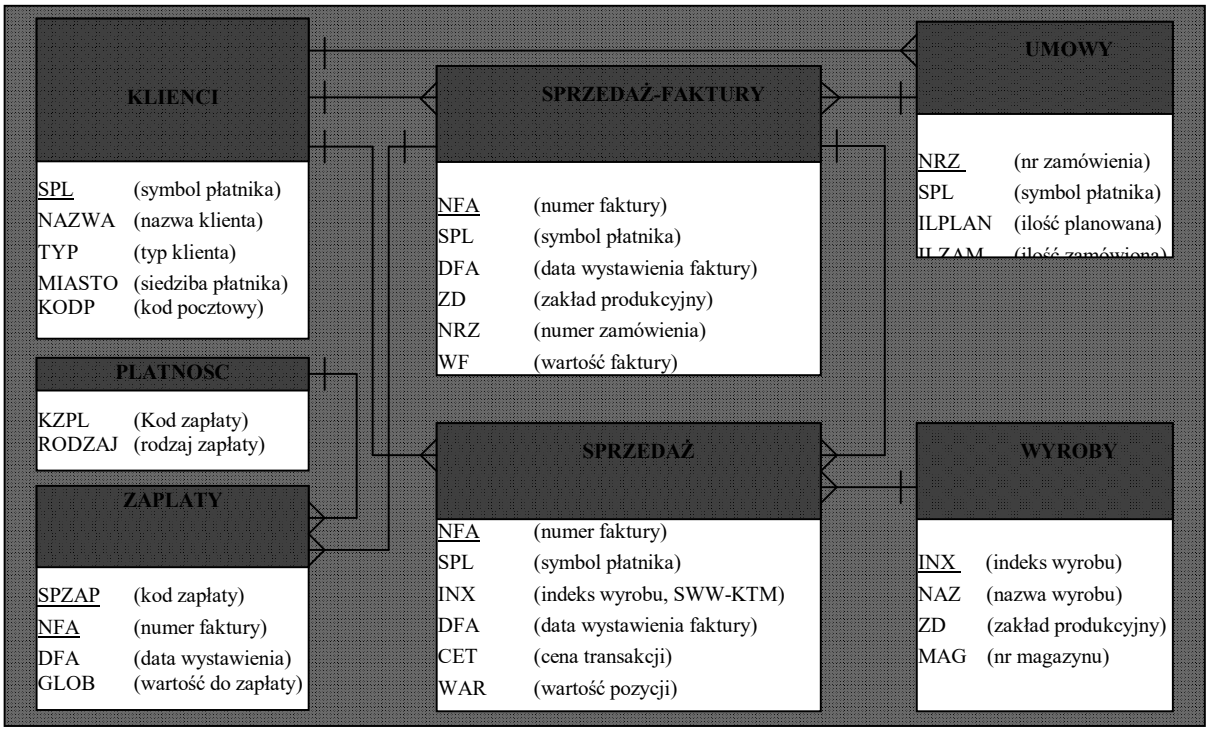
Rys. 3. Przykład architektury prostej hurtowni tematycznej
Fig. 3. The simple data marts architecture example

Jak widać z tego rysunku, szczególny rodzaj magazynu danych stanowi zbiór archiwalnych relacyjnych baz danych $\boldsymbol{r} = \{r_{ij}\}$ o schematach $\boldsymbol{R}=\{R_{ij}\}$. W tym zbiorze wskaźnikiem $i\in\{1,...,n\}$ oznaczano poszczególne archiwalne bazy danych, natomiast wskaźnikiem $j\in\{1,...,m\}$ kolejne jej relacje. Ponieważ w tym zbiorze dla każdego ustalonego j zachodzi $R_{1j}= R_{2j}=...=R_{nj}$, zatem odpowiednie relacje w poszczególnych bazach składowych opisywanego zbioru baz danych mają taki sam schemat. W związku z tym, że szczególny rodzaj magazynu nie zawiera informacji zbiorczych zagregowanych na różnych poziomach niezbędnych do analitycznego przetwarzania, konieczne stało się zatem zaprojektowanie Magazynu Danych Zbiorczych, którego celem będzie przechowywanie informacji zbiorczych pochodzących ze szczególnego rodzaju magazynu danych. Jego schemat powinien być tak określony, aby umożliwiał realizację większości potencjalnych pytań analitycznych. Niestety, o pytaniach tych wiadomo tylko tyle, że powinny rozszerzać zbiór standardowych zestawień, predefiniowanych w aplikacji obsługującej bazy archiwalne. Horyzont czasowy tych zestawień sięga jednego roku, tj. dzień, tydzień, miesiąc, kwartał, rok. Innymi słowy, otrzymywane odpowiedzi na pytania

analityczne kierowane do Magazynu Danych Zbiorczych w wymiarze czasu powinny obejmować zagregowane dane dotyczące kilku lat. Pytania analityczne konstruowano w oparciu o schemat ścieżek analizy wielowymiarowej i kierowano do Magazynu Danych Zbiorczych za pomocą przykładowego systemu zarządzania tym magazynem [16].

4.2. Schemat ER archiwalnych baz danych z zastanego systemu informacyjnego

Jak już wspomniano, szczególny rodzaj magazynu danych stanowi zbiór archiwalnych relacyjnych baz danych. W dalszej części pracy przyjęto, że istnieje schemat ER łączący fragmenty niektórych relacje w poszczególnych bazach składowych szczególnego rodzaju magazynu danych, który przedstawiono na rys. 4. Stanowił on podstawę do określenia początkowego schematu ścieżek wielowymiarowej analizy sprzedaży.

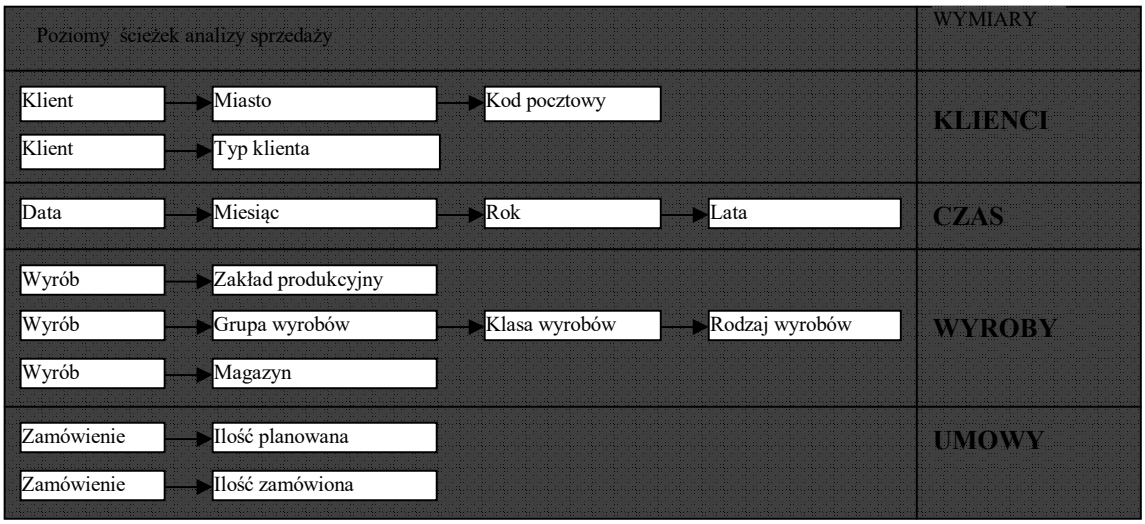


Rys. 4. Fragmenty schematów relacji z operacyjnych archiwalnych baz danych
Fig. 4. The relational schema fragments from archival operational databases

4.3. Ścieżki wielowymiarowej analizy sprzedaży

Biorąc pod uwagę przedstawione fragmenty schematów relacji z operacyjnych archiwalnych baz danych, przyjęto, że przykładowe pytania analityczne dotyczyły sprzedaży wyrobów gotowych względem hierarchii różnych wymiarów, tj. ścieżek analizy sprzedaży, które przedstawiono na rys. 5. Przy ich określaniu skorzystano z wprowadzonego przez [5] pojęcia ścieżek analizy. Przyjęto założenie, że początkowy zbiór ścieżek analizy sprzedaży definiowano na

podstawie zastanego modelu danych bazy OLTP oraz wstępnych potrzeb analityka w zakresie wielowymiarowej analizy informacji.



Rys. 5. Początkowy zbiór ścieżek analizy sprzedaży
Fig. 5. The initial set analytical sales paths

Pokazany na powyższym rysunku początkowy zbiór ścieżek analizy zgrupowano w cztery wymiary. Najbardziej szczegółowy poziom każdego wymiaru odpowiada podstawowym własnościom sprzedawanych produktów, tak jak to zarejestrowano w systemie transakcyjnym. Przedstawiony początkowy zbiór ścieżek analizy sprzedaży przekształcono dalej, na podstawie formalnie zaproponowanego modelu schematu ścieżek analizy wielowymiarowej, do odpowiedniego początkowego schematu ścieżek wielowymiarowej analizy sprzedaży.

4.4. Początkowy schemat ścieżek wielowymiarowej analizy sprzedaży

Przyjęty i opisany w poprzednim paragrafie na podstawie zastanego modelu danych bazy OLTP początkowy zbiór ścieżek analizy, przekształcono dalej do odpowiedniego początkowego schematu ścieżek wielowymiarowej analizy sprzedaży.

Dla zaprezentowanego początkowego zbioru ścieżek analizy sprzedaży, początkowy schemat ścieżek analizy wielowymiarowej, na podstawie definicji 8, stanowi grupę powiązanych danych $\mathbf{S}_{przykł}$ = (M, W, P, S), gdzie M = { Sprzedaż wartościowa według..., }, tzn. miara 'Sprzedaż wartościowa według...' definiowana jest i reprezentowana przez atrybut WAR z relacji SPRZEDAŻ z przedstawionego wcześniej schematu ER zastanego systemu informacyjnego, natomiast W = {KLIENCI, CZAS, WYROBY, UMOWY}. Dla tak określonych wymiarów analizy sprzedaży określono poniższe ścieżki analizy. Dla wymiaru W_1 = {KLIENCI} określono dwie ścieżki analizy, tj.:

$P_{11}=\{p_{111} + p_{112} + p_{113}\}$, gdzie: p_{111} = {Klient}

p_{112} = {Miasto}
p_{113} = {Kod pocztowy}

$P_{12}=\{p_{121} + p_{122}\}$, gdzie p_{121} = {Klient}
p_{122} = {Typ klienta}.

Dla wymiaru W_2 = {CZAS} określono jedną ścieżkę analizy, tj.:

$P_{21}=\{p_{211} + p_{212} + p_{213} + p_{214}\}$, gdzie p_{211} = {Data}
p_{212} = {Miesiąc}
p_{213} = {Rok}.
p_{214} = {Lata}.

Dla wymiaru W_3 = {WYROBY} określono trzy ścieżki analizy, tj.:

$P_{31}=\{p_{311} + p_{312}\}$, gdzie p_{311} = {Wyrób}
p_{312} = {Zakład produkcyjny}

$P_{32}=\{p_{321} + p_{322}, + p_{323} + p_{324}\}$, gdzie p_{321} = {Wyrób}
p_{322} = {Grupa wyrobów}
p_{323} = {Klasa wyrobów}
p_{324} = {Rodzaj wyrobów}

$P_{33}=\{p_{331} + p_{332}\}$, gdzie p_{331} = {Wyrób}
p_{332} = {Magazyn}.

Dla wymiaru W_4 = {UMOWY} określono dwie ścieżki analizy, tj.:

$P_{41}=\{p_{411} + p_{412}\}$, gdzie p_{411} = {Zamówienie}
p_{412} = {Ilość planowana}

$P_{42}=\{p_{421} + p_{422}\}$, gdzie p_{421} = {Zamówienie}
p_{422} = {Ilość zamówiona}.

Dla tak określonych zbiorów poziomów analizy w poszczególnych wymiarach, wynikają następujące zbiory uporządkowanych par, tj.:

dla wymiaru W_1 = {KLIENCI}
ścieżka 1: $S_{11}=\{(p_{111}, p_{112}), (p_{111}, p_{113}), (p_{112}, p_{113})\}$
ścieżka 2: $S_{12}=\{(p_{121}, p_{122}), (p_{121}, p_{123})\}$,

dla wymiaru W_2 = {CZAS}
ścieżka 1: S_{21} = {(p_{211}, p_{212}), (p_{211}, p_{213}), (p_{211}, p_{214}), (p_{212}, p_{213}), (p_{212}, p_{214}), (p_{213}, p_{214})},

dla wymiaru W_3 = {WYROBY}
ścieżka 1: S_{31} = {(p_{311}, p_{312}) }
ścieżka 2: S_{32} = {(p_{321}, p_{322}), (p_{321}, p_{323}), (p_{321}, p_{324}), (p_{322}, p_{323}), (p_{322}, p_{324}), (p_{323}, p_{324})}
ścieżka 3: S_{33} = {(p_{331}, p_{332})},

dla wymiaru W_4 = {UMOWY}
ścieżka 1: S_{41} = {(p_{411}, p_{412})}

ścieżka 2: $S_{42} = \{(p_{421}, p_{422})\}$,

które wyznaczają następujące zbiory ukierunkowanych ścieżki analizy $s_{ij, xy}=(p_{ijx}, p_{ijy})$.

I tak, dla wymiaru W_1 = {KLIENCI} mamy następujące ukierunkowane ścieżki analizy:

ścieżka 1, $s_{11, 12} = (p_{111}, p_{112})$

$s_{11, 13} = (p_{111}, p_{113})$

$s_{11, 23} = (p_{112}, p_{113})$ zatem $S_{11} = \{s_{11, 12}, s_{11, 13}, s_{11, 23}\}$,

ścieżka 2, $s_{12, 12} = (p_{121}, p_{122})$

$s_{12, 13} = (p_{121}, p_{123})$ zatem $S_{12} = \{s_{12, 12}, s_{12, 13}\}$.

Dla wymiaru W_2 = {CZAS}, ukierunkowane ścieżki analizy przedstawiają się następująco:

ścieżka 1, $s_{21, 12} = (p_{211}, p_{212})$

$s_{21, 13} = (p_{211}, p_{213})$

$s_{21, 14} = (p_{211}, p_{214})$

$s_{21, 23} = (p_{212}, p_{213})$

$s_{21, 24} = (p_{212}, p_{214})$

$s_{21, 34} = (p_{213}, p_{214})$ zatem $S_{21} = \{s_{21, 12}, s_{21, 13}, s_{21, 14}, s_{21, 23}, s_{21, 24}, s_{21, 34}\}$.

Dla wymiaru W_3 = {WYROBY}, mamy:

ścieżka 1, $s_{31, 12} = (p_{311}, p_{312})$ zatem $S_{31} = \{s_{31, 12}\}$.

ścieżka 2, $s_{32, 12} = (p_{321}, p_{322})$

$s_{32, 13} = (p_{321}, p_{323})$

$s_{32, 14} = (p_{321}, p_{324})$

$s_{32, 23} = (p_{322}, p_{323})$

$s_{32, 24} = (p_{322}, p_{324})$

$s_{32, 34} = (p_{323}, p_{324})$ zatem $S_{32} = \{s_{31, 12}, s_{32, 13}, s_{32, 14}, s_{32, 23}, s_{32, 24}, s_{32, 34}\}$

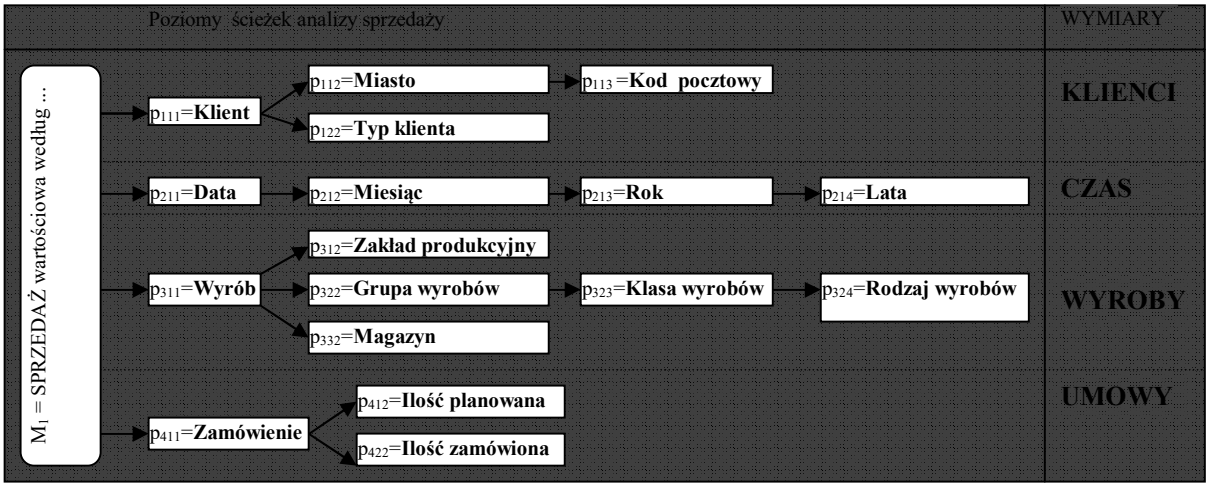
ścieżka 3, $s_{33, 12} = (p_{331}, p_{332})$ zatem $S_{33} = \{s_{33, 12}\}$.

Dla wymiaru W_4 = {UMOWY}, mamy:

ścieżka 1, $s_{41, 12} = (p_{411}, p_{412})$ zatem $S_{41} = \{s_{31, 12}\}$

ścieżka 2, $s_{42, 12} = (p_{421}, p_{422})$ zatem $S_{42} = \{s_{42, 12}\}$.

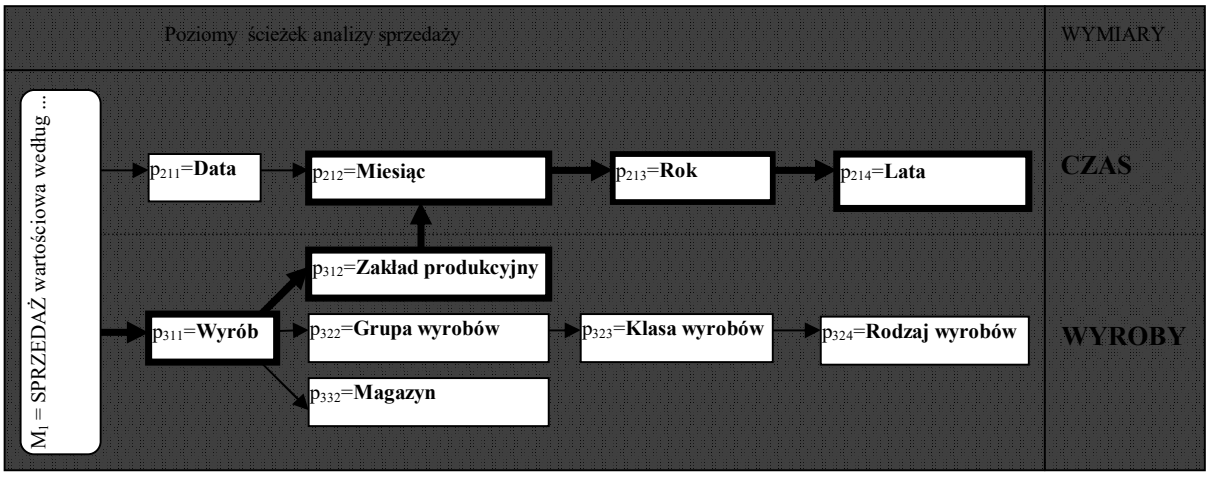
Mamy zatem, że dla każdego zbioru $S_{ij} \in S = \{S_{11} + S_{12} + S_{21} + S_{31} + S_{32} + S_{33} + S_{41} + S_{42}\}$, graf **g**(V, E), gdzie $V = \{p_0\} \cup P_{ij}$, natomiast $E = S_{ij}$, jest ukierunkowanym, acyklicznym grafem zakorzenionym w $p_0 \in P$, takim, że każdy poziom analizy $p_{ijx} \in P_{ij}$ znajdujący się na j-tej ścieżce i-tego wymiaru może być osiągnięty wychodząc z poziomu analizy p_0 za pomocą przynajmniej jednej ukierunkowanej ścieżki. Tak więc grupa powiązanych danych $\mathbf{S}_{śaw}$=(M,W,P,S) stanowi schemat ścieżek analizy wielowymiarowej. Uwzględniając, że $p_{111} = p_{121}$ = {Klient} oraz $p_{311} = p_{321} = p_{331}$ = {Wyrób}, otrzymano początkowy schemat ścieżek wielowymiarowej analizy sprzedaży, który przedstawiono na rys. 6. Stanowi on podstawę do konstruowania pytań analitycznych w oparciu o kombinacje różnych ścieżek analizy.



Rys. 6. Początkowy schemat ścieżek wielowymiarowej analizy sprzedaży
Fig. 6. The initial multidimensional analytical sales paths schema

4.4.1. Ścieżki skrośne w schemacie ścieżek wielowymiarowej analizy sprzedaży

Na podstawie otrzymanego początkowego schematu ścieżek wielowymiarowej analizy sprzedaży, możliwe są do zestawienia inne, tj. proste i złożone skrośne ścieżki analizy sprzedaży. Jeśli dla takiego schematu przyjąć, że $s_{ij,vx} = s_{31,\,12}$ oraz $s_{mn,yz} = s_{21,\,24}$, wówczas dla tak określonych ukierunkowanych ścieżek analizy, złożoną ścieżką skrośną wielowymiarowej analizy sprzedaży, jest ścieżka $s_{complex} = s_{31,\,12} + s_{21,\,24}$ lub prościej *Wyrób->Zakład Produkcyjny->Miesiąc->Rok->Lata*, którą pokazano na rys. 7.



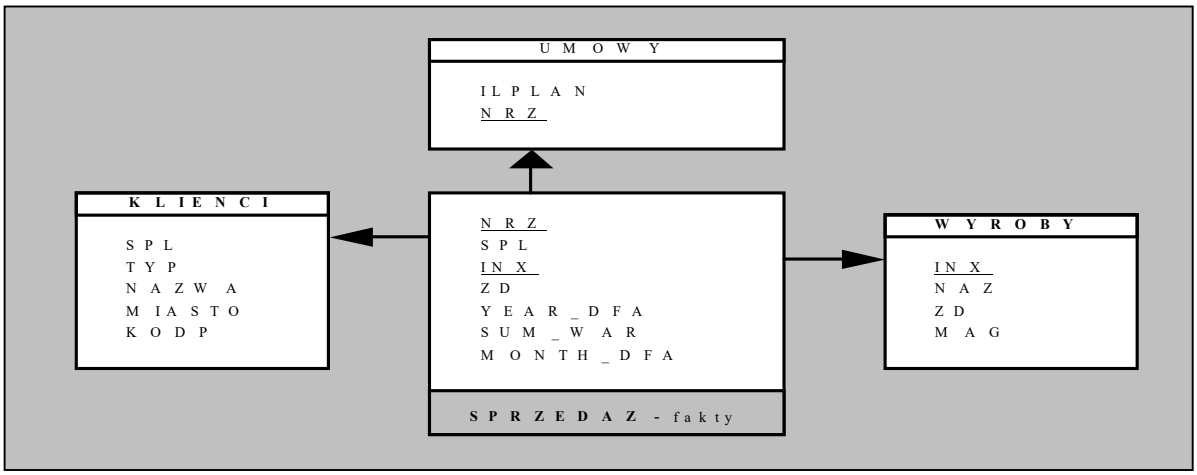
Rys. 7. Przykład złożonej skrośnej ścieżki analizy sprzedaży
Fig. 7. The complex analytical through path sales example

4.5. Początkowy schemat Magazynu Danych Zbiorczych

Jak już wspomniano, początkowy schemat tematycznej hurtowni danych określono za pomocą metody [9], wykorzystującej tradycyjny model ER do projektowania hurtowni danych na podstawie przemysłowych modeli danych. Założono również, że początkowy zbiór ścieżek analizy sprzedaży definiowano na podstawie zastanego modelu danych bazy OLTP oraz wstępnych potrzeb analityka w zakresie wielowymiarowej analizy informacji. Przyjęto zatem dalej, że początkowy schemat tematycznej hurtowni danych określi się za pomocą w/w metody, wykorzystującej tylko te relacje, które z punktu widzenia analityka zawierają istotne informacje.

Tak więc, dla wspomnianego już przykładowego schematu ER zawierającego fragmenty niektórych relacji z operacyjnych archiwalnych baz danych, do encji transakcyjnych należą relacje SPRZEDAŻ-FAKTURY, SPRZEDAŻ oraz ZAPLATY. Do encji komponentowych należą zaś relacje KLIENCI, WYROBY oraz UMOWY. Zatem, na podstawie cytowanej metody, w tym konkretnym przypadku możliwe do utworzenia są dwa schematy hurtowni danych typu gwiazda, w których relację faktów tworzą encje transakcyjne.

Do dalszych prac i analiz wybrano ten ze schematów, w którym relacja faktów jest formowana na podstawie relacji SPRZEDAŻ. Atrybuty grupujące i agregujące w relacji faktów tj. SPRZEDAŻ-fakty utworzono na podstawie istniejących atrybutów z relacji SPRZEDAŻ.
I tak, do atrybutów grupujących w relacji faktów należą SPL (Symbol Płatnika), INX (Indeks Wyrobu), MONTH_DFA (Miesiąc wystawienia faktury) oraz YEAR_DFA (Rok wystawienia faktury). Dwa ostatnie atrybuty utworzono na bazie atrybutu DFA (Data Faktury) z relacji SPRZEDAŻ. Do atrybutów agregujących należy SUM_WAR (Suma Wartości) utworzony na bazie atrybutu WAR (Wartość Pozycji) z relacji SPRZEDAŻ. Do formownia tablic wymiarów na podstawie encji komponentowych wykorzystano relacje KLIENCI, WYROBY oraz UMOWY.
Należy wspomnieć, że relacji PLATNOSC należącej również do encji klasyfikujących i komponentowych nie brano na tym etapie pod uwagę z tego względu, iż wymiaru PŁATNOŚCI nie uwzględniono w początkowym schemacie ścieżek analizy wielowymiarowej, który jak założono, uwzględniał wstępne potrzeby analityka w zakresie wielowymiarowej analizy informacji.
Reasumując, otrzymany na podstawie tej metody początkowy schemat przykładowego Magazynu Danych Zbiorczych przedstawiony na rys. 8 jest schematem typu gwiazda, w którym poziomy wymiaru czasu (lata, rok, miesiące) przechowywane są w relacji faktów. W końcu Magazyn Danych Zbiorczych zasilono odpowiednimi, na różnych poziomach zagregowanymi danymi, pochodzącymi z archiwalnych kopii danych z zamkniętych okresów obliczeniowych.



Rys. 8. Przykład początkowego schematu Magazynu Danych Zbiorczych
Fig. 8. The initial data warehouse schema example

4.6. Klasy powszechnie zadawanych pytań analitycznych

Powszechnie obecnie zadawane i kierowane są do hurtowni danych pytania analityczne można pogrupować w pewne klasy. Klasy te w jawny sposób określono na podstawie analizy przykładowych pytań analitycznych zawartych w pracy [5]. Wyróżniono w ten sposób trzy klasy pytań analitycznych.

A) Klasa pytań dotyczących jednej miary zawierających:
   - pytania typu Q1, dotyczące jednej miary względem jednej ścieżki z dwóch wymiarów,
   - pytania typu Q2, dotyczące jednej miary względem dwóch ścieżek z jednego wymiaru.

B) Klasa pytań dotyczących dwóch miar, czyli
   - pytania typu Q3, dotyczące dwóch miar względem jednej ścieżki z dwóch wymiarów.

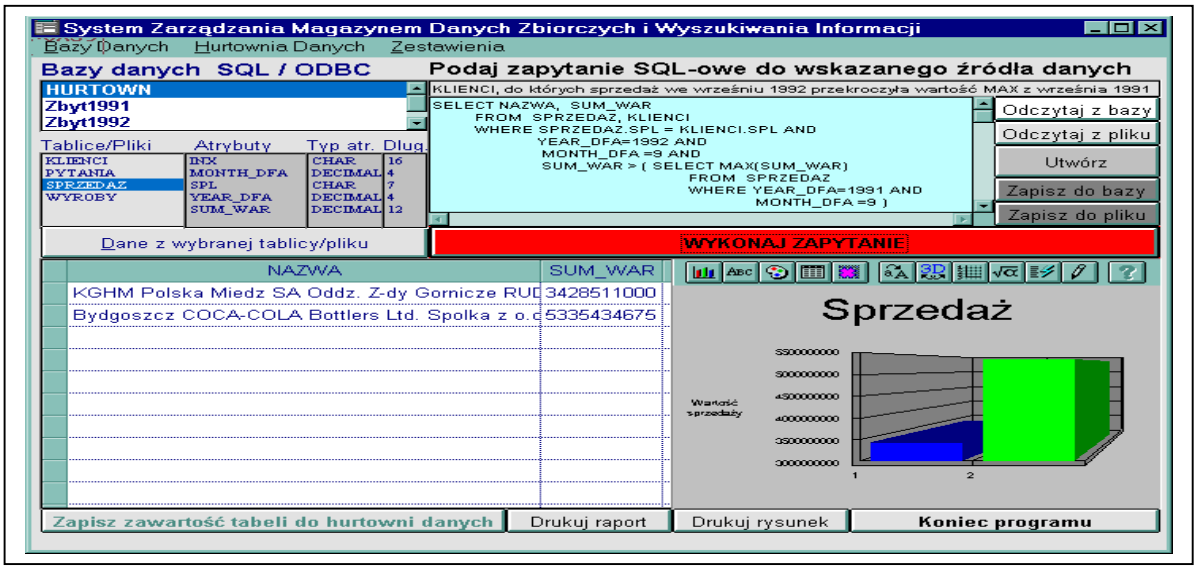
C) Klasa pytań dokonujących selekcji opartej na wcześniej zagregowanych danych na różnych poziomach; są to pytania zagnieżdżone typu Q4, dotyczące jednej miary względem jednej ścieżki kilku wymiarów, w których zagnieżdżony operator selekcji bazuje na wcześniej zagregowanych danych.

4.7. Realizacja przykładowego pytania analitycznego, skonstruowanego na podstawie schematu ścieżek analizy wielowymiarowej

Za wyjątkiem pytań typu Q3, których nie można formułować i kierować do przykładowego Magazynu Danych Zbiorczych, ponieważ określony wcześniej początkowy schemat ścieżek analizy wielowymiarowej dotyczy tylko jednej miary (tj. sprzedaż wartościowa według ...), niemniej jednak pozwala on na zrealizowanie innych pytań analitycznych należących do klas A lub C. Jako ilustrację zaproponowanego rozwiązania pokonującego trudności pojawiające się przy konstrukcji właściwych nowych pytań analitycznych, poniżej przedstawiono

przykład typowego i zarazem bardziej złożonego pytania analitycznego typu Q4 należącego do klasy C, dokonującego selekcji opartej na wcześniej zagregowanych danych na różnych poziomach względem wymiaru czasu. Skierowano go do przykładowego magazynu za pomocą wcześniej już wspomnianego przykładowego systemu zarządzania Magazynem Danych Zbiorczych i Wyszukiwania Informacji. Sformułowano go na podstawie złożonej ścieżki skrośnej tj. *Klient->Miesiąc->Rok->Lata*. Pytanie to wyrażone w języku naturalnym brzmi następująco: '*Wyszukaj klientów, do których sprzedaż we wrześniu 1992 przekroczyła wartość maksymalną z września 1991*'. Analiza tego pytania pozwala stwierdzić, że poszukiwana miara w tablicy faktów to *sprzedaż wartościowa według...*, którą reprezentuje atrybut SUM_WAR. W pytaniu tym poszukiwani są pewni klienci, których nazwy reprezentowane są przez atrybut NAZWA (Nazwa Klienta). Wymiary względem których dokonuje się selekcji to KLIENCI oraz CZAS (zawarty w relacji faktów). Dalej można stwierdzić, że pytanie to dotyczy zagregowanych danych odnoszących się do sprzedaży, mających swoje źródło w pogrupowanych dokumentach sprzedaży czyli fakturach, które wystawiano w poszczególnych miesiącach w latach 1991-1992. Zatem w końcowym pytaniu analitycznym wyrażonym w języku SQL zaangażowano atrybuty grupujące YEAR_DFA oraz MONTH_DFA. Poszukiwani klienci stanowią zatem odpowiedź na poniższe pytanie, którego wynik przedstawiono na rys. 9.

```
SELECT NAZWA, SUM_WAR
      FROM SPRZEDAZ, KLIENCI
      WHERE SPRZEDAZ.SPL = KLIENCI.SPL AND
            YEAR_DFA=1992 AND MONTH_DFA =9 AND
            SUM_WAR > ( SELECT MAX(SUM_WAR) FROM  SPRZEDAZ
                        WHERE YEAR_DFA=1991 AND  MONTH_DFA =9 )
```



Rys 9. Pytanie: *Wyszukaj klientów, do których sprzedaż we wrześniu 1992 przekroczyła wartość maksymalną z września 1991 roku*

Fig 9. Question: *Find the customers to which september's sales in 1992 exceeded september's sales in 1991 year*

5. Podsumowanie

Zaproponowany formalny model schematu ścieżek analizy wielowymiarowej może wspomóc nie tylko konstrukcję właściwych pytań analitycznych ale również proces ewolucji schematu hurtowni danych w sytuacji, gdy pojawiają się nowe pytania analityczne. Mogą to być pytania formułowane *ad-hoc* przez kierownictwo firmy lub też inne pytania analityczne o których wiadomo tylko tyle, że powinny rozszerzać zbiór standardowych zestawień, predefiniowanych w aplikacji obsługującej zastane bazy danych. Na podstawie tego modelu wskazano na korzyści płynące z możliwości konstruowania pytań analitycznych użytkownika formułowanych *ad-hoc* w oparciu o kombinacje wcześniej określonych na podstawie zastanego modelu danych bazy OLTP ścieżek analizy, zapewniając tym samym potencjalną możliwość realizacji takich pytań. Natomiast podstawową wadą zaproponowanego podejścia jest konieczność określenia ścieżek analizy na podstawie wymagań analityka oraz schematu ER z zastanego systemu informacyjnego. W przypadku braku takiego schematu, niewłaściwe zrozumienie przez projektanta związków w modelu danych tego systemu, prowadzić może do ustalenia niewłaściwych ścieżek analizy, co w konsekwencji prowadzi do formułowania pytań analitycznych, których nie można zrealizować.

LITERATURA

1. Kimbal R.: Slowly Changing Dimensions. http://www.dbmsmag.com/9604d05.html.
2. Niemi T., Nummenmaa J., Thanisch P.: Constructing OLAP Cubes Based On Queries. In Proceedings of the ACM International Workshop on Data Warehousing and OLAP, DOLAP 2001, Atlanta, GA, USA, November 9, 2001, http://www.cis.drexel.edu/faculty/song/DOLAP2001/Niemi%20-%202.pdf.
3. Blaschka M, Sapia C., Höfling G.: On Schema Evolution in Multidimensional Databases. http://citeseer.nj.nec/cache/papers/cs/21200/http:zSzzSzwww.forwiss.dezSz~system42zSzpublicationszSzdawak_camera.pdf/blashka98schema.pdf
4. Cheung D., Zhou B., Kao B., Lu H., Lam T., Ting H.: Requirement Based Data Cube Schema Design. http://citeseer.nj.nec.com/cache/papers/cs/8567/http:zSzzSzpearl.cs.hku.hkzSzpublicationszSztechrepszSzdocumentzSzTR-99-04.pdf/requirement-base-data-cube.pdf
5. Tsois A., Karayannidis N.: MAC: Conceptual Data Modeling for OLAP. In Proceedings of the 3rt International Workshop DMDW'2001, Interlaken, Switzerland, June 4, 2001, http://sunsite.informatik.rwth-aachen.de/Publications/CEUR-WS/Vol-39/paper5.pdf

6. Sapia C., Blaschka M, Höfling G.: Extending the E/R Model for the Multidimensional Paradigm. http://citeseer.nj.nec.com/cache/papers/cs/21200/http:zSzzSzwww.forwiss.dezSz~system42zSzpublicationszSzdwdm98.pdf/extending-the-e-r.pdf.
7. Tryfona N, Busborg F, Christiansen J.: starER: A Conceptual Model for Data Warehouse Design. In Proceedings of the ACM Second International Workshop on Data Warehousing and OLAP, DOLAP'99, Kansas City, USA, November 6, 1999, http://www.cis.drexel.edu/faculty/song/DOLAP99/dolap99Nectl.pdf
8. Chaudhuri S., Dayal U.: An Overview of Data Warehousing and OLAP Technology. Appears in ACM Sigmod Record, March 1997, ftp://ftp.research.microsoft.com/users/-surajitc/sigrecord.pdf.
9. Moody D., Kortink M.: From Enteprise Models to Dimensional Models: A Methodology for Data Warhouse and Data Mart Design. In Proceedings of the International Workshop on Design and Management of Data Warehouses (DMDW2000), Stockholm, Sweden, June 5-6, 2000, http:sunsite.informatik.rwth-aachen.de/Publications/CEUR-WS/Vol-28/paper5.pdf.
10. McGuff F.: Designing the Perfect Data Warehouse. http://members/aol.com/fmcguff/dwmodel/frtext.htm.
11. Kacprzyk J., Stańczak W.: Teoria grafów i jej zastosowania w informatyce. PWN, Warszawa, 1980, Tł. z Graph theory with applications to engineering and computer science, ISBN 83-01-00544-0.
12. Ignasiak E.: Teoria grafów i planowanie sieciowe. Państwowe Wydawnictwo Ekonomiczne, Warszawa 1982.
13. Jankowski B.: Grafy, algorytmy w Pascalu. Wydawnictwo "Mikom", Warszawa 1988, ISBN 83-7158-077-0.
14. Kimbal R.: Is ER Modeling Hazardous to DSS. DBMS – June 1995 – Data Warehouse Architect, http://www.dbmsmag.com/9610d05.html.
15. Golfarelli M, Rizzi S.: A Methodological Framework for Data Warehouse Design. In Proceedings of the Acm International Workshop on Data Warehousing and OLAP, DOLAP98, Washington, D.C., USA, November 7, 1998.
16. Bok Z.: Integracja relacyjnych baz danych w zastanych przemysłowych systemach informatycznych, Studia Informatica, Volume 23, Nr 4, Politechnika Śląska, Gliwice, 2002.

Recenzent:

Wpłynęło do Redakcji 30 czerwca 2003 r.

Abstract

In this article a proposal of formal multidimensional analytical paths schema model based on analysis paths concept introduced by [5] was presented. Based on this model as well as ER schema from legacy informational OLTP system a multidimensional analytical paths schema example was determined. On his basis an analytical questions was constructed which was directed to data warehouse example. Based on combination different analytical paths determined in multidimensional analytical paths schema an Q4 analytical question type belonged to C class was formulated. In this way a potential possibility of their execution was made possible.