STUDIA INFORMATICA 2003
Volume 24 Number 3 (48)

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

DYNAMICZNE PROJEKTOWANIE HURTOWNI DANYCH NA PODSTAWIE PYTAŃ ANALITYCZNYCH

Streszczenie. W tym artykule przedstawiono propozycję innego podejścia do problemu projektowania hurtowni danych. Bazując na tym podejściu oraz zaproponowanej metodzie dynamicznego rozszerzania schematu hurtowni, omówiono problem dynamicznego projektowania hurtowni danych, biorąc pod uwagę pytania analityczne formułowane przez użytkownika końcowego. W tym podejściu, każde nowe pytanie analityczne analizowane jest pod kątem możliwości jego realizacji. Jeśli nie może być wykonane, wówczas poddawane jest dalszej analizie w celu wyodrębnienia ewentualnych pytań pomocniczych lub cząstkowych (jednoprzebiegowych), tzn. takich pytań, których wyniki są danymi wejściowymi do nowego pytania analitycznego. Na podstawie tych wyodrębnionych pytań pomocniczych lub cząstkowych podejmowana jest decyzja o inkrementalnym, dynamicznym rozszerzaniu schematu za pomocą zaproponowanej metody.

DYNAMIC DATA WAREHOUSE DESIGN BASED ON ANALITICAL QUERIES

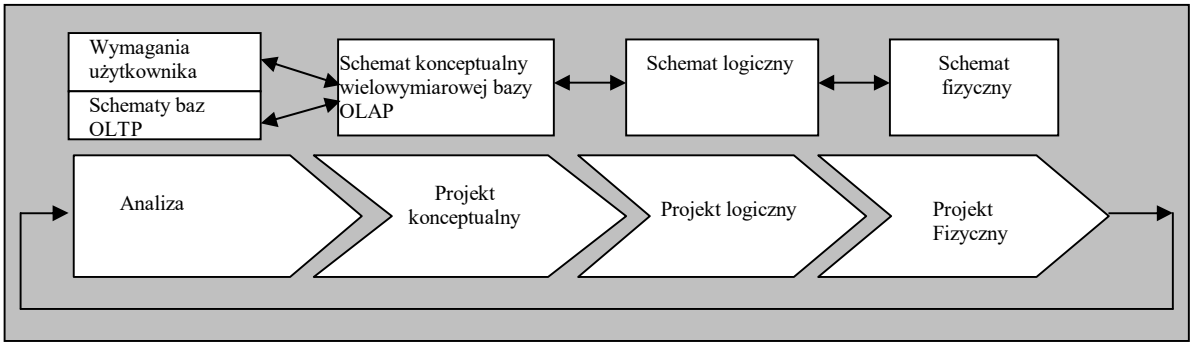
Summary. In this article – a proposal of different approach to data warehouse design problem has been presented. Based on this approach and dynamically extension data warehouse schema proposed method, a data warehouse design problem taking into account analytical queries formulated by end user has been discussed. In this approach every new analytical query is analyzed at an angle of it’s realizability. If it can not been executed then to isolate possible auxiliary or partially (one-route) queries, eg. such queries whose results are input data to a new analytical query to further analysis is submitted. Based on this isolated auxiliary or partially queries, a decision about incrementally and dynamically data warehouse schema extension is taking with the aid of proposed method.

1. Wprowadzenie

Pomimo tego, że projektowanie hurtowni danych wymaga technik kompletnie różnych [1, 2, 3] od tych, które zostały zaadoptowane z systemów transakcyjnych, nie podjęto jednak dotychczas [4, 5] znaczących wysiłków zmierzających do rozwinięcia spójnej metodologii projektowania hurtowni danych. Największe zainteresowanie środowisk bazodanowych i zespołów badawczych poświęcone jest wielowymiarowym modelom danych, zmaterializowanym widokom, wyborze indeksów oraz tym aspektom projektowym, które determinują wydajność hurtowni danych. Z drugiej strony, na poziomie projektowania konceptualnego szereg pozycji, między innymi specyfikacja wymagań, czy też konceptualne modele danych nie zostały należycie zbadane. Nie mniej jednak istnieje wiele prac poświęconych tej tematyce.

Wśród innych prac poruszających pozostałe aspekty OLAP-owych metod projektowania wspomnieć można prace dotyczące konstruowania sześcianów OLAP-owych [6], ewolucji wymiarów i schematów w heterogenicznych bazach danych [7, 8] oraz tworzenia hurtowni danych za pomocą serwera SQL [9].

W chwili obecnej, na podstawie publikacji [10, 11, 12, 13, 14] wynika, że tradycyjny proces projektowania schematu hurtowni danych przedstawić można schematycznie jako ciąg następujących po sobie i wzajemnie zazębiających się etapów, co pokazano na rys. 1.



Rys. 1. Proces projektowania środowiska hurtowni danych
Fig. 1. The data warehouse environment project process

W przedstawionym zamkniętym cyklu projektowym wyróżniamy cztery zasadnicze etapy projektowania środowiska hurtowni danych.

1. Etap analizy wymagań użytkownika oraz schematów transakcyjnych baz OLTP.
2. Etap projektu konceptualnego - mający do czynienia z wysokim poziomem reprezentacji świata rzeczywistego. Poprzez to, że powstający na tym etapie konceptualny schemat hurtowni danych jest zrozumiały dla końcowych użytkowników, możliwa jest weryfikacja ich wymagań, identyfikacja możliwych luk oraz przeprowadzenie analizy celów biznesowych. Rozwijany na tym etapie formalny i kompletny konceptualny model danych pozwala na wysoko poziomowe projektowanie encji i ich wzajemnych relacji reprezentowanych w

przyjazny dla użytkownika i niezależny od implementacyjnych kwestii sposób. Formalność i kompletność tego modelu oznacza, że może być poddany jednoznacznej transformacji do następnego schematu logicznego.

3. Etap projektu logicznego – czyli etap pośredni pomiędzy etapem konceptualnym i fizycznym, próbujący zbalansować paradygmat niezależności magazynowania (storage-independent) i naturalnej reprezentacji informacji w kategoriach komputerowo-zorientowanych koncepcji.
4. Etap fizycznego projektu schematu hurtowni danych – operujący na poziomie szczegółów reprezentujących informację w sprzęcie.

Ponieważ konceptualny model danych stanowi centralną część cyklu projektowego hurtowni danych, interesujące podejście do problemu rozwoju (ewolucji) schematu wielowymiarowej bazy danych zaproponowano w pracy [15]. Autorzy tej pracy zdefiniowali formalne podstawy, tj. wielowymiarowy model danych wraz z operacjami zmieniającymi/rozwijającymi, które mogą być użyte do implementacji narzędzi wspierających ewolucję schematu wielowymiarowej bazy danych na poziomie konceptualnym. Celem ich podejścia jest automatyczna propagacja wzdłuż cyklu projektowego do innych modeli danych wszelkich zmian, dokonanych przez projektanta na poziomie konceptualnym. W przypadku pojawienia się nowych wymagań użytkownika, proces projektowania środowiska hurtowni danych powtarzany jest cyklicznie.

Wśród innych konceptualnych modeli danych zaproponowanych przez [11, 16, 17, 18, 19, 20] i innych, na szczególną uwagę zasługuje [21] graficzny konceptualny model dla hurtowni danych (zwany Dimensional Fact Model) oraz zaproponowana pół-automatyczna metodologia określania jej schematu na podstawie istniejących diagramów ER (Entity Relationship) z zastanych systemów transakcyjnych. Również ciekawą, chociaż nie operującą na poziomie konceptualnym, jest zaproponowana przez [22] trójstopniowa metoda projektowania schematu hurtowni danych z tradycyjnych modeli ER. W innych pracach [5, 8, 12, 23, 24, 25] zaproponowano różne podejścia do półautomatycznego lub automatycznego tworzenia logicznych i fizycznych schematów hurtowni danych. Żadne z wymienionych podejść nie zawierało jednak mechanizmu do automatycznego wyszukiwania w systemach OLTP oraz początkowego określania miar i faktów. Początkowe określenie miar i faktów może okazać się najbardziej trudną częścią procesu projektowania. Zaproponowano więc różne ręczne podejścia w celu ich określania. Tak więc dopiero praca [14] stanowi pierwszy krok na drodze automatycznego tworzenia całego konceptualnego schematu na podstawie schematu transakcyjnych baz OLTP, łącznie z początkowym określeniem faktów i miar. W tej pracy zaproponowano algorytm do automatycznego tworzenia i oceny schematów konceptualnych na podstawie schematów baz transakcyjnych OLTP. Wynikowe schematy konceptualne będące danymi wyjściowymi tego

algorytmu stanowiły jednocześnie dane wejściowe następnego algorytmu ich oceny pod kątem spełniania wymagań użytkownika zdefiniowanych w postaci pytań OLAP. Wyselekcjonowany tą drogą schemat konceptualny poddawano dalszym ręcznym zmianom, bazując na wiedzy użytkownika i projektanta.

Mało zostało również powiedziane odnośnie projektów konceptualnych, biorąc pod uwagę wymagania użytkownika jako punkt początkowy. Na tym polu obok historycznie już pierwszej pracy [7] oraz innych [6, 15, 30] wyróżnia się praca [26]. 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 [11, 16], polegającego na rozszerzeniu modelu ER dla wielowymiarowego paradygmatu [3, 11], 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 omawiali procesu projektowania hurtowni danych, 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.

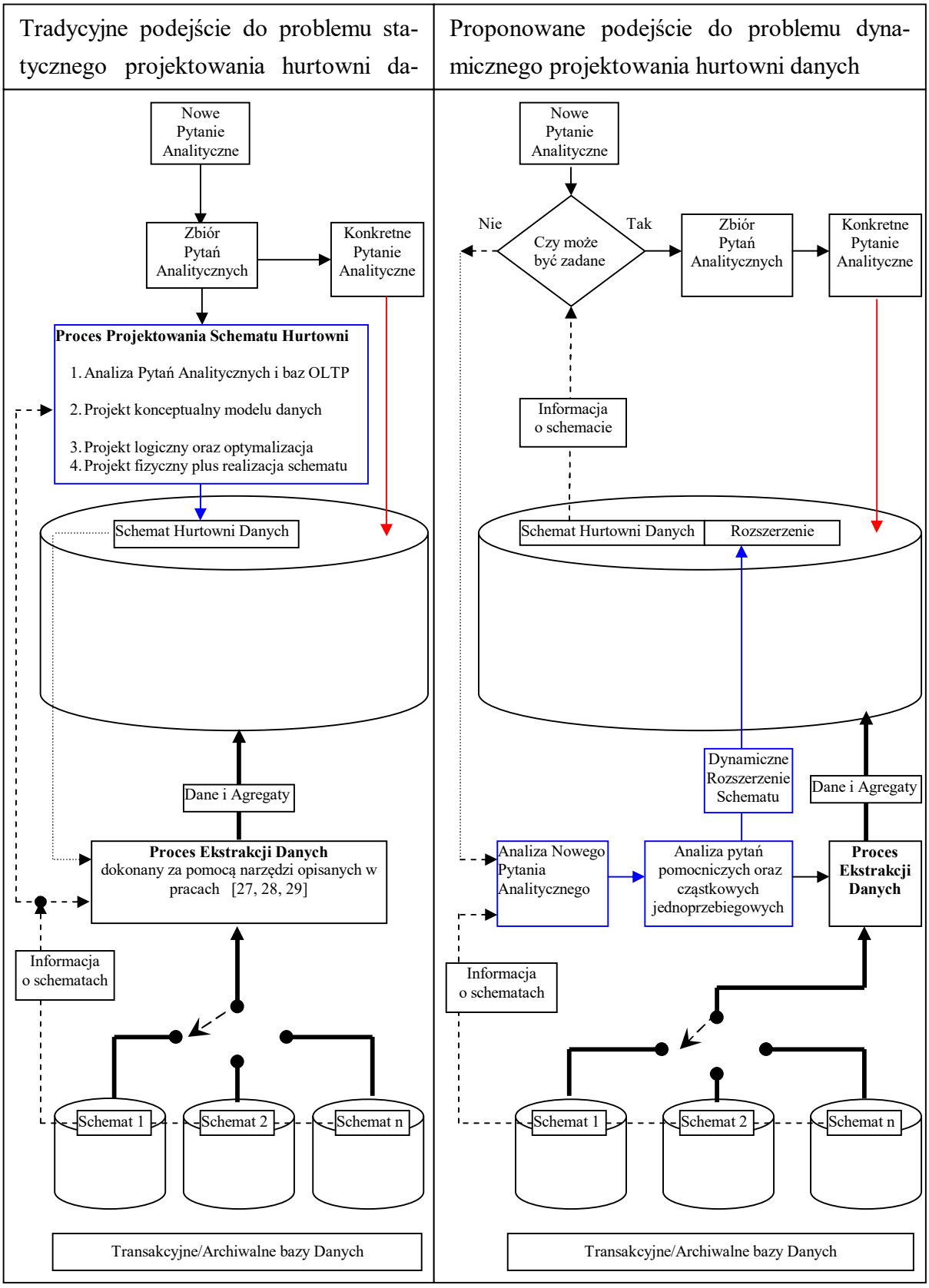
Niezależnie od przyjętego konceptualnego modelu danych, stanowiącego centralną część zamkniętego cyklu projektowego hurtowni danych, problemem zasadniczym przy definiowa-

niu i rozwoju schematu wielowymiarowej jest osiągnięcie pewnych celów w wielowymiarowym modelowania danych. W przeciwieństwie do jednego z głównych celów jakim jest normalizacja, do której dąży się przy projektowaniu złożonych operacyjnych baz danych OLTP, celem wielowymiarowego modelowania danych w tak określonym procesie projektowania, jest stworzenie takiej struktury wielowymiarowej bazy danych OLAP, która jest łatwa do zrozumienia i wykorzystania przez użytkownika końcowego, kierującego do niej pytania analityczne. Drugim celem jest maksymalizacja efektywności wykonania tych pytań. Cele te osiąga się przez minimalizację liczby tablic i łączących je relacji, przez co redukuje się złożoność wielowymiarowej bazy danych oraz minimalizuje się liczbę złączeń wymaganych do realizacji pytania analitycznego. W celu realizacji tak określonych celów podczas projektowania wielowymiarowej bazy danych, najlepszym rozwiązaniem jest zdefiniowanie jej schematu w postaci gwiazdy [3, 22, 34].

2. Problem dynamicznego projektowania hurtowni danych

Mając na uwadze przedstawiony we wprowadzeniu aktualny stan wiedzy odnoszący się do tradycyjnego podejścia do problemu statycznego projektowania, budowy hurtowni oraz ekstrakcji danych [27, 28, 29], dokonano krótkiej syntezy tych informacji, którą schematycznie zaprezentowano na rys. 2. Na podstawie tej syntezy zaproponowano inne podejście do problemu dynamicznego projektowania i budowy hurtowni danych biorąc pod uwagę wymagania użytkownika końcowego, w szczególności pytania analityczne pojawiające się *ad-hoc*. Jest to podejście przeciwne do tradycyjnego podejścia projektowania hurtowni danych a w szczególności do obowiązującej w nim naczelnej zasady [22], która zakłada, że schemat hurtowni powinien bezpośrednio wynikać z wcześniej określonego zbioru pytań analitycznych. Zbiór ten powinien obejmować wszystkie spodziewane typy pytań analitycznych mogące być zadane przez użytkownika. Jest on zatem niezbędny do zaprojektowania ‘właściwie’ określonego schematu w tradycyjnym podejściu do projektowania hurtowni danych. Pod pojęciem ‘właściwie’ określony schemat hurtowni danych rozumie się taki jej schemat, który umożliwia realizację większości pytań analitycznych użytkownika.

W tym miejscu należy wspomnieć, że oprócz tradycyjnego podejścia do projektowania schematu hurtowni danych na podstawie wymagań użytkownika końcowego zmaterializowanych w postaci zbioru pytań analitycznych, istnieją również inne. Punktem wyjścia w tych podejściach projektowania schematu hurtowni danych są schematy ER z zastanych transakcyjnych systemów OLTP a nie zbiór pytań analitycznych. Przykładem takiego podejścia mogą być prace [14, 17, 21, 22].



Rys. 2. Tradycyjne i proponowane podejście do problemu projektowania hurtowni danych
Fig. 2. The traditional and proposed data warehouse design problem approach

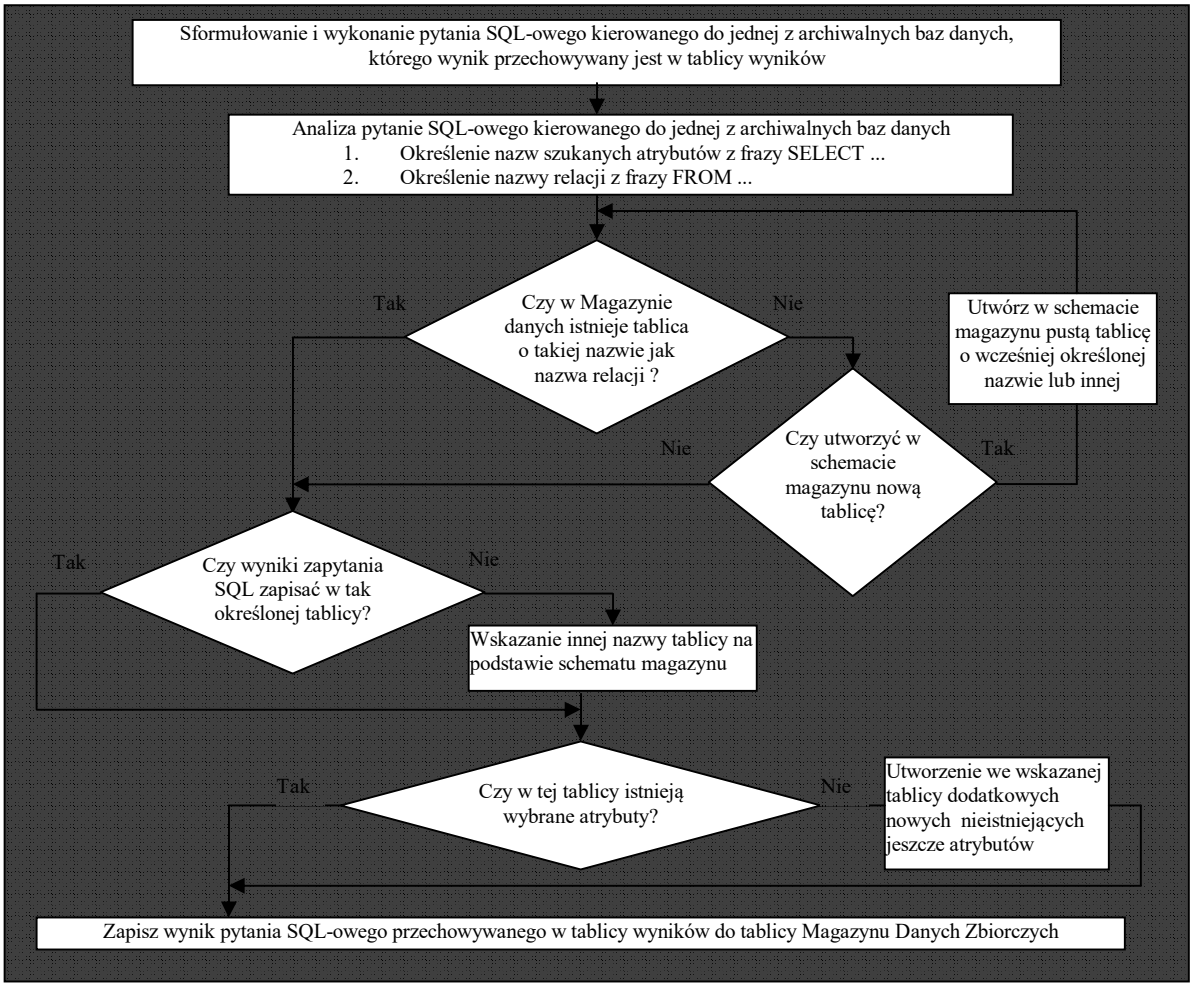
Tak więc wysoka nieprzewidywalność i zmienność w czasie wymagań analitycznych użytkownika końcowego skutkuje niestabilnymi podstawami projektowymi, co stanowi poważny problem i jedną z głównych wad podejścia tradycyjnego. Problem ten jest już obecny w najnowszej literaturze, jak choćby podejście zaproponowane w pracy [31, 32]. W podejściu zaprezentowanym w pracy [32], hurtownia danych postrzegana jest jako zbiór zmaterializowanych widoków nad zbiorem relacji podstawowych. Autorzy tego podejścia koncentrują się na problemie inkrementalnego projektowania hurtowni danych w przypadku pojawiających nowych pytań, za pomocą zaprojektowanego inkrementalnego algorytmu wyboru i materializacji zbioru nowych widoków. Reasumując, zaproponowane i omawiane w niniejszej pracy podejście może okazać się uzasadnione również w sytuacji, w której wiedza na temat zbioru pytań analitycznych na etapie projektowania jest ograniczona oraz jeśli nie wiadomo, kiedy pojawią się nowe pytania analityczne. W tym podejściu wykorzystano zaproponowaną już wcześniej metodę [27] dynamicznego rozszerzania schematu hurtowni, z której wyeliminowano jedną z jej najistotniejszych wad. Polegała ona na tym, że wygenerowana postać schematu magazynu danych była całkowicie zależna od schematu zastanych archiwalnych baz danych. Obecnie, dzięki nowej implementacji mechanizmu dynamicznego rozszerzania schematu działającego według zmienionego algorytmu, przedstawionego na rys. 3, możliwe staje się utworzenie poprawnej struktury Magazynu Danych Zbiorczych typu gwiazdy lub płatka śniegu w każdym przypadku. Pomimo wyeliminowania ww. wady, zaproponowana metoda obarczona jest jeszcze innymi wadami, z których najważniejsze to:

1) wygenerowany schemat magazynu danych pozbawiony jest tego, co człowiek wnosi do procesu projektowania struktury hurtowni, czyli optymalizacji struktury danych,
2) przedstawiona metoda obejmuje tylko przypadki jednoprzebiegowe, tj. gdy pytanie SQL-owe kierowane do archiwalnych baz danych generuje oczekiwane wyniki (dane zagregowane); nie zaimplementowano mechanizmu umożliwiającego pozyskanie wyników oczekiwanych po kilku przebiegach.
3) aktualną implementację tej metody ograniczono do typu zastanych archiwalnych baz danych, tj. dBase; możliwa jest implementacja tej metody do obsługi innych baz danych, np. Gupta, Btrieve, Paradox, Microsoft Access, SQL Serwer czy Oracle.

Dzięki jednak swej prostocie, w przypadku tworzenia małych hurtowni tematycznych na podstawie istniejących zastanych przemysłowych systemów informacyjnych, zaproponowana metodę dynamicznego rozszerzania schematu hurtowni może okazać się przydatna.

W zaproponowanym podejściu oparto się również na wspomnianym we wprowadzeniu modelu MAC [26] i stowarzyszonych z nim pojęć, w celu dokonania analizy wpływu pytań analitycznych na postać dynamicznie tworzonego/rozszerzanego schematu hurtowni danych. W szczególności, wykorzystano wprowadzoną w tym modelu koncepcję ścieżek analizy. Ba-

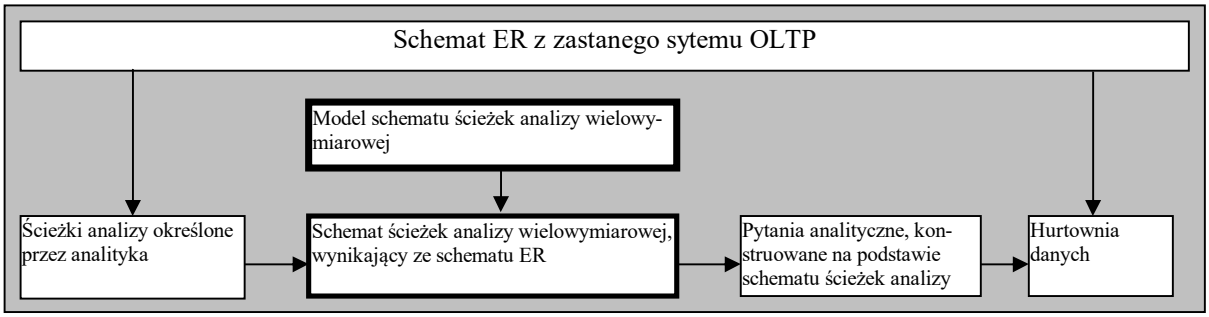
zując na tej koncepcji zaproponowano i zdefiniowano formalny model schematu ścieżek analizy wielowymiarowej, który wykorzystano do określenia konkretnego schematu ścieżek analizy wielowymiarowej wynikającego z zastanego systemu OLTP.



Rys. 3. Algorytm mechanizmu dynamicznego rozszerzania schematu magazynu danych
Fig. 3. The dynamic algorithm schema data warehouse extension mechanism

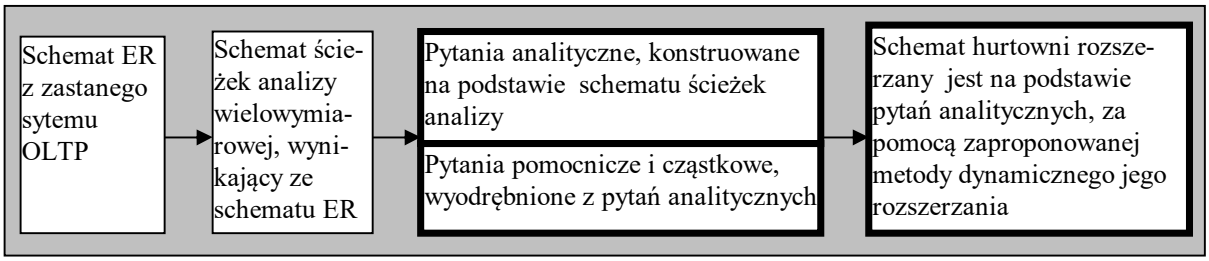
Schemat ten stanowił podstawę konstrukcji właściwych, z punktu widzenia zastanych systemów OLTP, pytań analitycznych. Pytania analityczne użytkownika formułowano w oparciu o kombinacje różnych ścieżek analizy określonych w schemacie ścieżek analizy wielowymiarowej, co schematycznie pokazano na rys. 4, zapewniając tym samym potencjalną możliwość realizacji takich pytań. Reasumując, ścieżki analizy określano na podstawie zastanego modelu danych bazy OLTP, które następnie przekształcano wykorzystując ww. model, do odpowiedniego schematu ścieżek analizy wielowymiarowej. Zaproponowane podejście zapewnia ewolucję 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. Podejście to zapewnia również wy-

eliminowanie wstępnie zdefiniowanego na podstawie wymagań użytkownika zbioru pytań analitycznych.



Rys 4. Konstrukcja pytań analitycznych na podstawie schematu ścieżek analizy
Fig 4. The analytical queries construction based on analytical paths schema

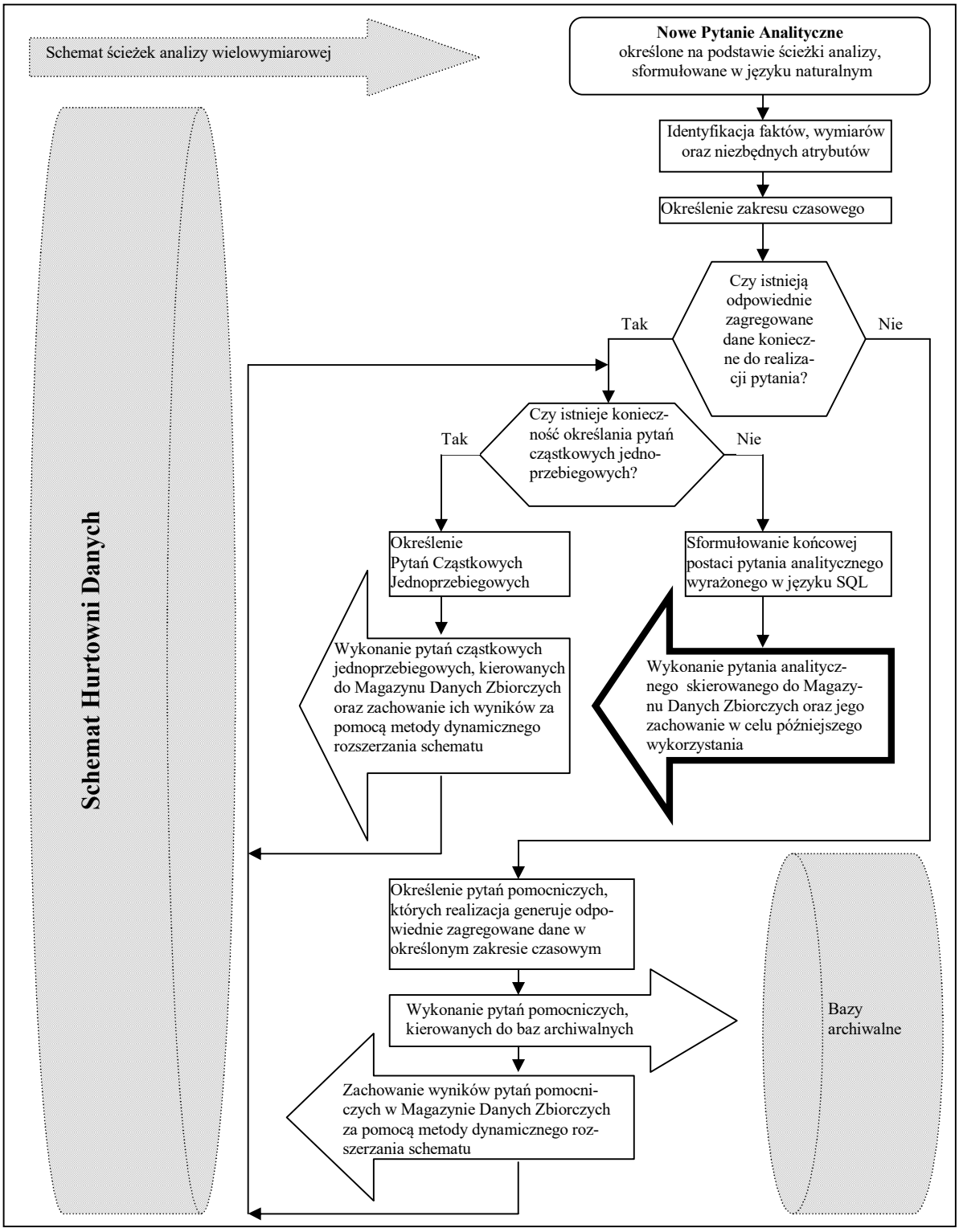
Tak więc w zaproponowanym podejściu, każde nowo pojawiające się pytanie analityczne poddawane jest analizie pod kątem możliwości jego zrealizowania. Jeśli nie może być zrealizowane, wówczas poddawane jest dalszej analizie w celu wyodrębnienia ewentualnych pytań pomocniczych lub cząstkowych (jednoprzebiegowych), tzn. takich, których wyniki są danymi wejściowymi do nowego pytania analitycznego. W trakcie realizacji pytań cząstkowych podejmowano decyzję o inkrementalnym, dynamicznym rozszerzaniu schematu hurtowni danych za pomocą zaproponowanej metody, co schematycznie pokazano na wcześniej zaprezentowanym rys. 2. Tak więc, w wyniku zaproponowanego podejścia, proces projektowania hurtowni danych można schematycznie przedstawić w postaci ciągu zdarzeń, co pokazano na rys. 5.



Rys 5. Koncepcja projektowania hurtowni danych na podstawie pytań analitycznych
Fig 5. The concept of the data warehouse project based on analytical queries

Pierwszym krokiem podczas wstępnej analizy nowego pytania analitycznego za pomocą algorytmu pokazanego na rys. 6 jest określenie, które z poszukiwanych informacji odnoszą się do danych ilościowych (faktów) a które do danych kwalifikujących (wymiarów). Celem drugiego kroku jest określenie zakresu czasowego tego pytania. W trzecim kroku określa się, czy w Magazynie Danych Zbiorczych istnieją odpowiednie zagregowane dane konieczne do zrealizowania nowego pytania analitycznego, czyli czy istnieje konieczność formułowania dodatkowych pytań cząstkowych jednoprzebiegowych generujących odpowiednio zagregowane dane.

W zaprezentowanym podejściu, przedstawiony algorytm analizy pytań analitycznych prezentuje ogólny sposób postępowania z każdym kolejnym nowym pytaniem analitycznym wyrażonym w języku naturalnym.



Rys. 6. Algorytm analizy pytań analitycznych
Fig. 6. The analytical questions analyse algorithm

Nie zaimplementowano na jego podstawie żadnych mechanizmów do automatycznego podejmowania decyzji o możliwości (lub nie) zadania pytania analitycznego do hurtowni. Nie zaimplementowano również żadnych mechanizmów automatycznie generujących pytania cząstkowe lub pomocnicze. W przyjętym rozwiązaniu pytania te konstruowano *ad-hoc* na podstawie logicznej analizy pytania analitycznego, schematu ścieżek analizy wielowymiarowej oraz znajomości schematów ER zastanych archiwalnych systemów OLTP. Ponadto przyjęto, że początkowy schemat hurtowni danych określono za pomocą metody opisanej w publikacji [22], 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. Reasumując, biorąc powyższe pod uwagę, sformułowano problem badawczy będący przedmiotem niniejszej pracy, który brzmi następująco: możliwe jest dalsze dynamiczne rozszerzanie początkowego schematu hurtowni danych, zgodnie z zaproponowaną koncepcją projektowania hurtowni danych, na podstawie pytań analitycznych formułowanych na bazie schematu ścieżek analizy wielowymiarowej oraz schematu ER zastanych archiwalnych systemów OLTP. W celu wykazania istnienia rozwiązania tego problemu, analizie poddano wpływ niektórych przykładowych pytań analitycznych na początkową postać schematu, należących do jednej z trzech klas pytań:

klasa 1) pytania należące do klasy pytań zwiększających wymiary hurtowni,

klasa 2) pytania należące do klasy pytań zwiększających liczbę ścieżek analizy w ramach danego wymiaru,

klasa 3) pytania należące do klasy pytań zwiększających liczbę poziomów w ramach danego wymiaru.

2.1. 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 [5, 33, 35, 36, 37] 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, ...} [33, 35]. Krawędź e_k utożsamia się z nieuporządkowaną 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 [36] dowolną nieuporządkowaną parę {e_i, e_j} taką, że ((e_i, e_j) ∈E)□ ((e_j, e_i) ∈E).

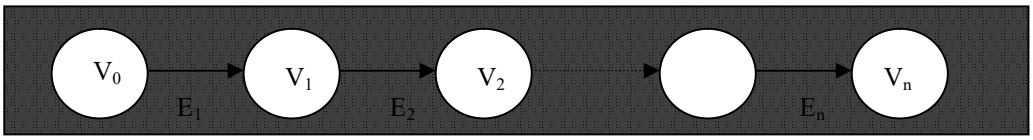
Definicja 3. Ukierunkowanym (zorientowanym) [33, 37] grafem nazywamy taki graf G=(V,E), w którym E jest zbiorem takich uporządkowanych par (e_i, e_j) dla których krawędź łącząca dwa węzły v_i, v_j posiada określony kierunek.

Definicja 4. Ścieżką w ukierunkowanym grafie G=(V, E) nazywamy [37] 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 [33].

Definicja 6. Ukierunkowanym acyklicznym grafem nazywamy taki graf, w którym istnieje tylko jedna acykliczna ścieżka pomiędzy dowolną parą węzłów [33].

Definicja 7. Niech **g**=(V, E) będzie ukierunkowanym acyklicznym grafem [5], gdzie V jest zbiorem węzłów, natomiast E jest zbiorem krawędzi. Mówimy, że **g** – co pokazano na rys. 7 - jest quasi-drzewem z korzeniem w V_0∈V, jeśli każdy wierzchołek V_j∈V może być osiągnięty z v_0 za pomocą przynajmniej jednej ukierunkowanej ścieżki. Oznaczmy przez s_{ij}⊆**g** ukierunkowaną ścieżkę (jeśli istnieje) rozpoczynającą się w V_i i kończącą się w V_j. Oznaczmy dalej przez sub(V_i)⊂**g** quasi drzewo zakorzenione w węźle V_i≠V_0.



Rys. 7. Przykład ukierunkowanego acyklicznego grafu zakorzenionego w V_0
Fig. 7. The example of the directed acyclic graph rooted in V_0

Korzystając z przytoczonych definicji oraz formalizmu zaproponowanego przez [5], poniżej przedstawiono formalną definicję pojęcia schematu ścieżek analizy wielowymiarowej.

Definicja 8. Schemat ścieżek analizy wielowymiarowej $\mathbf{S}_{śaw}$=(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 wyra-

żenia 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 $x,y \in <1, r>$, 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.

Przykład 1, Dla pewnej miary $M_1 \in M$ określono dwa wymiary analizy, dla których określono następujące zbiory poziomów analizy, tj.:

wymiar W_1, składający się z dwóch ścieżek analizy, tj.: $P_{11}=\{p_{111} + p_{112} + p_{113} + p_{114}\}$

$P_{12}=\{p_{121} + p_{122} + p_{123}\}$

oraz wymiar W_2 składający się z jednej ścieżki analizy, tj.: $P_{21}=\{p_{211} + p_{212} + p_{213}\}$.

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

dla wymiaru W_1, ścieżka 1:

S_{11}={ (p_{111} , p_{112}), (p_{111} , p_{113}), (p_{111} , p_{114}), (p_{112} , p_{113}), (p_{112} , p_{114}), (p_{113} , p_{114}) },

dla wymiaru W_1, ścieżka 2:

S_{12}={ (p_{121} , p_{122}), (p_{121} , p_{123}), (p_{122} , p_{123}) },

dla wymiaru W_2, ścieżka 1:

S_{21}={ (p_{211} , p_{212}), (p_{211} , p_{213}), (p_{212} , p_{213}) },

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

Dla wymiaru W_1, ścieżka 1, mamy: $s_{11,\,12}$ = (p_{111} , p_{112})

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

$s_{11,\,14}$ = (p_{111} , p_{114})

$s_{11,\,23}$ = (p_{112} , p_{113})

$s_{11,\,24}$ = (p_{112} , p_{114})

$s_{11,\,34}$ = (p_{113} , p_{114}),

zatem S_{11} = { $s_{11,\,12}$, $s_{11,\,13}$, $s_{11,\,14}$, $s_{11,\,23}$, $s_{11,\,24}$, $s_{11,\,34}$ }.

Dla wymiaru W_1, ścieżka 2, mamy: $s_{12,\,12}$ = (p_{121} , p_{122})

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

$s_{12,\,23}$ = (p_{122} , p_{123}),

zatem S_{12} = { $s_{12,12}$, $s_{12,\,13}$, $s_{12,\,23}$ }.

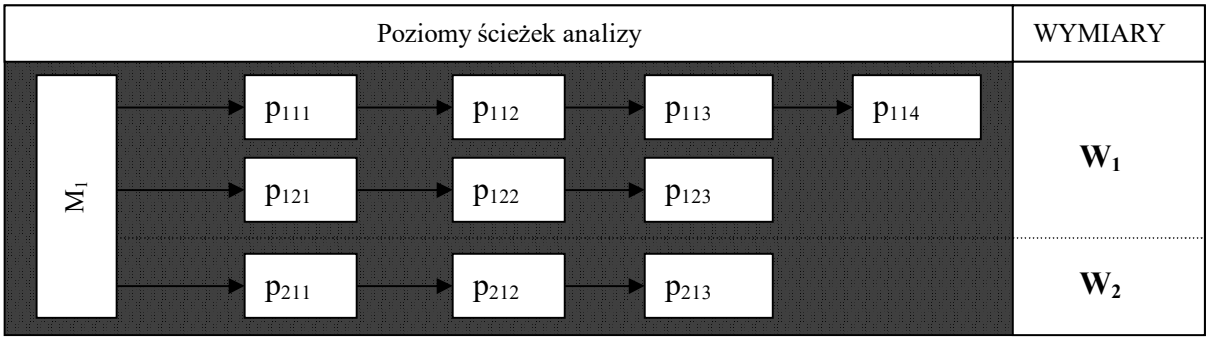
Dla wymiaru W_2, ścieżka 1, mamy: $s_{21,\,12}$ = (p_{211} , p_{212})

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

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

zatem S_{21}={ $s_{21,\,12}$, $s_{21,\,13}$, $s_{21,\,23}$ }.

Mamy zatem, że dla każdego zbioru $S_{ij} \in S$= {S_{11} + S_{12} + S_{21}}, 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 w tym przykładzie, grupa powiązanych danych $\mathbf{S}_{saw}$=(M, W, P, S) stanowi schemat ścieżek analizy wielowymiarowej, który graficznie przedstawiono na rys. 8.



Rys. 8. Przykład schematu ścieżek analizy wielowymiarowej

Fig. 8. The multidimensional analytical schema paths exam-

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.

2.1.1. 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 $p_0 \in P$.

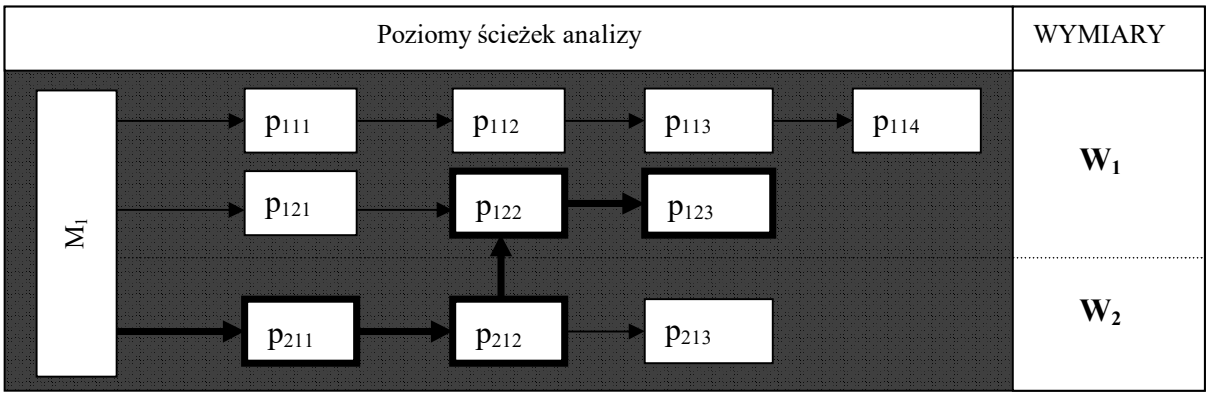
Definicja 10. Dla danego schematu ścieżek analizy wielowymiarowej $\mathbf{S}_{śaw}=(M, W, P, S)$, złożoną skrośną ścieżką analizy nazywa się taką ukierunkowaną ścieżkę analizy, która stanowi sumę dwóch lub więcej ukierunkowanych ścieżek analizy spełniających następujące warunki:

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 $p_0 \in P$.

Przykład 2. Dla schematu ścieżek analizy wielowymiarowej określonego w przykładzie 1 przyjęto, że $s_{ij,vx} = s_{21,\,12}$ oraz $s_{mn,yz} = s_{12,\,23}$. Dla tak określonych ukierunkowanych ścieżek analizy, złożoną ścieżką skrośną, którą pokazano na rys. 9, jest ścieżka $s_{complex} = s_{21,12} + s_{12,23}$.

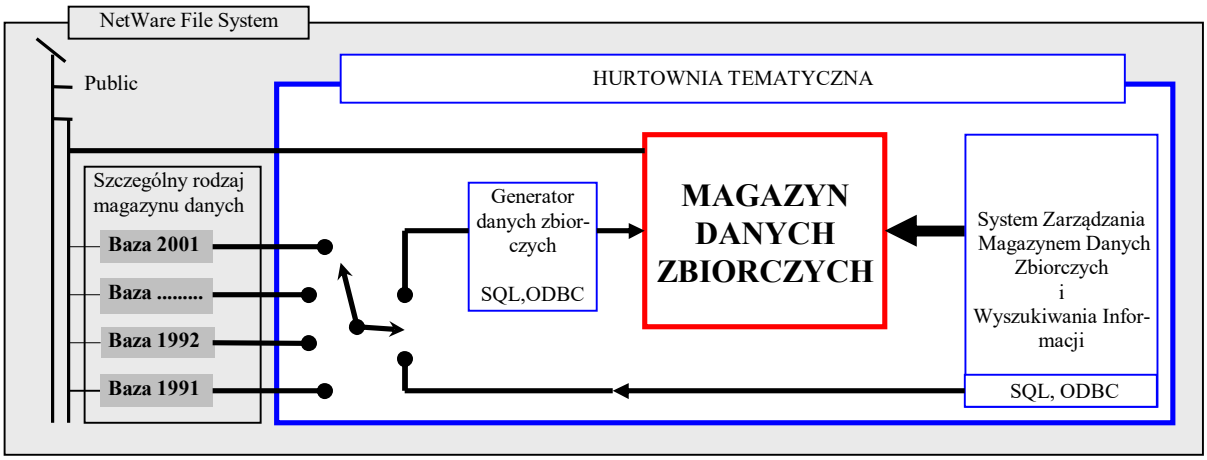


Rys. 9. Przykład analitycznej ścieżki skrośnej złożonej
Fig. 9. The complex analytical through path example

3. Realizacja dynamicznie rozszerzanego schematu hurtowni danych na podstawie pytań analitycznych

Bazując na algorytmie zaproponowanego podejścia do problemu dynamicznego projektowania hurtowni danych (przedstawionego już wcześniej na rys. 2) na podstawie analizy pytań analitycznych za pomocą wspomnianego już algorytmu (przedstawionego na rys. 3), w dalszej części pracy przedstawiono niektóre pytania analityczne, ich analizę i ich wpływ na postać dynamicznie rozszerzanego początkowego schematu Magazynem Danych Zbiorczych w przykładowej tematycznej hurtowni danych. Pytania te konstruowano na podstawie schematu ścieżek analizy wielowymiarowej, określonego na podstawie zastanego modelu danych obecnie eksploatowanego w ZTS "Nitron" S.A. przemysłowego systemu informacyjnego o nazwie "*Wyroby Gotowe*". Z tego systemu istnieje potrzeba uzyskania informacji zbiorczych 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 syste-

mów baz danych wspomagających przetwarzanie analityczne. Postanowiono zatem wykorzystać to podobieństwo do budowy prostej hurtowni tematycznej, której architekturę przedstawiono na rys. 10.

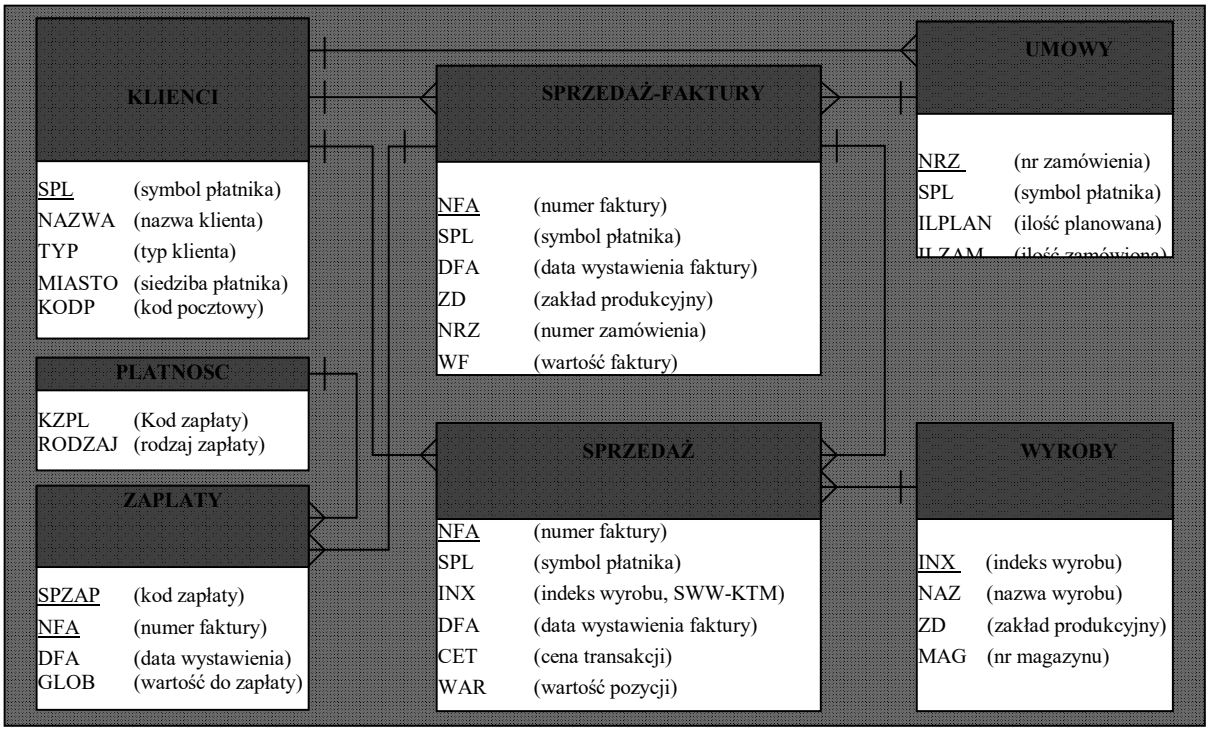


Rys. 10. Przykład architektury prostej hurtowni tematycznej
Fig. 10. The simple data marts architecture example

W związku z tym, że wspomniany 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 te kierowano je do Magazynu Danych Zbiorczych za pomocą przykładowego systemu zarządzania tym magazynem [27], który zrealizowano wykorzystując pakiet SQLWindows Application Development Module, stworzony przez amerykańską firmę komputerową CENTURA (dawniej GUPTA). Pakiet ten jest w pełni obiektowym narzędziem (4GL) opartym na wstępnie zdefiniowanych klasach obiektowych z wszystkimi korzyściami wynikającymi z programowania obiektowego, między innymi dziedziczenie i polimorfizm. Ponadto, dzięki mechanizmowi ODBC, pakiet ten umożliwia za pomocą sterowników własnych dostęp do baz danych typu Btrieve, dBase, Paradox. Korzystając ze sterowników obcych, umożliwia dostęp do innych baz danych. Przykładowo, za pomocą sterownika Oracle ODBC Driver, możliwe jest uzyskanie dostępu do bazy danych Oracle ver. 8.0 dostarczanej wraz systemem operacyjnym Novell NetWare 4.2.

3.1. Schemat ER archiwalnych baz danych z zastanego systemu informacyjnego

Jak już wspomniano, przemysłowy system informacyjny o nazwie "*Wyroby Gotowe*" jest aplikacją użytkową współpracującą z plikami relacyjnej bazy danych, wraz z jej archiwalnymi kopiami pochodzącymi z zamkniętych okresów obliczeniowych za lata 1991-2001. Archiwalne kopie baz danych z zamkniętych okresów obliczeniowych zeskładowano w oddzielnych katalogach systemu plików pewnego serwera sieciowego. W dalszej części pracy przyjęto, że istnieje schemat ER (przedstawiony na rys. 11) łączący fragmenty niektórych relacje w poszczególnych składowych archiwalnych bazach szczególnego magazynu danych.

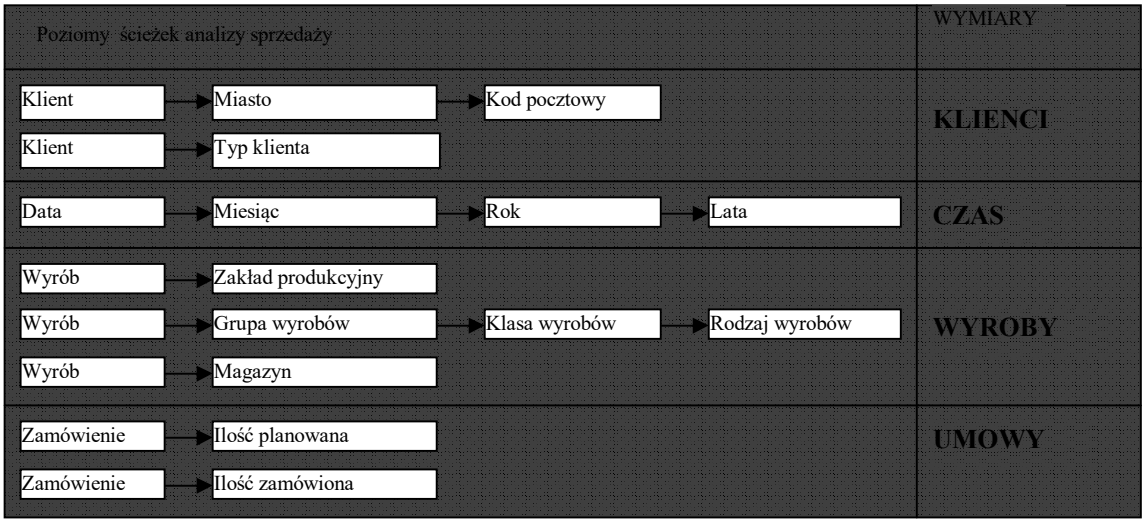


Rys. 11. Fragmenty schematów relacji z operacyjnych archiwalnych baz danych
Fig. 11. The relational schema fragments from archival operational databases

Tak więc 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. Jest on nieodzownym elementem zaproponowanego podejścia do problemu dynamicznego projektowania hurtowni danych. Stanowi podstawę do określenia schematu ścieżek wielowymiarowej analizy sprzedaży.

3.2. Ścieżki wielowymiarowej analizy sprzedaży

Biorąc pod uwagę przedstawione już 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. 12. Przy ich określaniu skorzystano z wprowadzonego przez [26] pojęcia ścieżek analizy. Innymi słowy, ścieżki analizy sprzedaży wynikają z zastanego modelu danych bazy OLTP, natomiast pytania analityczne konstruowano w oparciu o kombinacje różnych ścieżek analizy sprzedaży.



Rys. 12. Przykłady ścieżek analizy sprzedaży
Fig. 12. The analytical sales paths examples

Pokazane na powyższym rysunku przykładowe ścieżki analizy zgrupowano w cztery wymiary. Najbardziej szczegółowy poziom każdego wymiaru odpowiada podstawowym własnościom sprzedawanych produktów, tak jak to zarejestrowano w systemie transakcyjnym. Przykładowo, na poziomie 'Wyrób' w wymiarze WYROBY użyto symbolu wyrobu (SWW-KTM). Każdą ścieżkę analizy skonstruowano z dwóch lub więcej poziomów. W tradycyjnym podejściu, projektant aplikacji OLAP-owych definiuje schemat wymiarów i ich hierarchii przeważnie na etapie projektowania. W zaproponowanym podejściu, oprócz pytań konstruowanych na podstawie ścieżek analizy mogą pojawiać się pewne pytania pomocnicze typu *ad-hoc*, które mogą wymagać zdefiniowania ich własnych ścieżek analizy. Tak więc przy projektowaniu Magazynu Danych Zbiorczych, wymiary i poziomy ścieżek analizy definiowano na podstawie potrzeb procesu wielowymiarowej analizy informacji, wykorzystując przedstawiony model schematu ścieżek analizy wielowymiarowej. Przedstawione przykładowe ścieżki analizy sprzedaży przekształcono dalej, na podstawie formalnie zaproponowanego modelu schematu ście-

żek analizy wielowymiarowej, do odpowiedniego schematu ścieżek wielowymiarowej analizy sprzedaży.

3.3. Schemat ścieżek wielowymiarowej analizy sprzedaży

Przyjęte i opisane w poprzednim paragrafie na podstawie zastanego modelu danych bazy OLTP ścieżki analizy, przekształcono je dalej do odpowiedniego schematu ścieżek wielowymiarowej analizy sprzedaży. Dla zaprezentowanych przykładowych ścieżek analizy, przykładowy 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 przykładowe 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 $S_{32} = \{s_{31,12}, s_{32,13}, s_{32,14}, s_{32,23}, s_{32,24}, s_{32,34}\}$

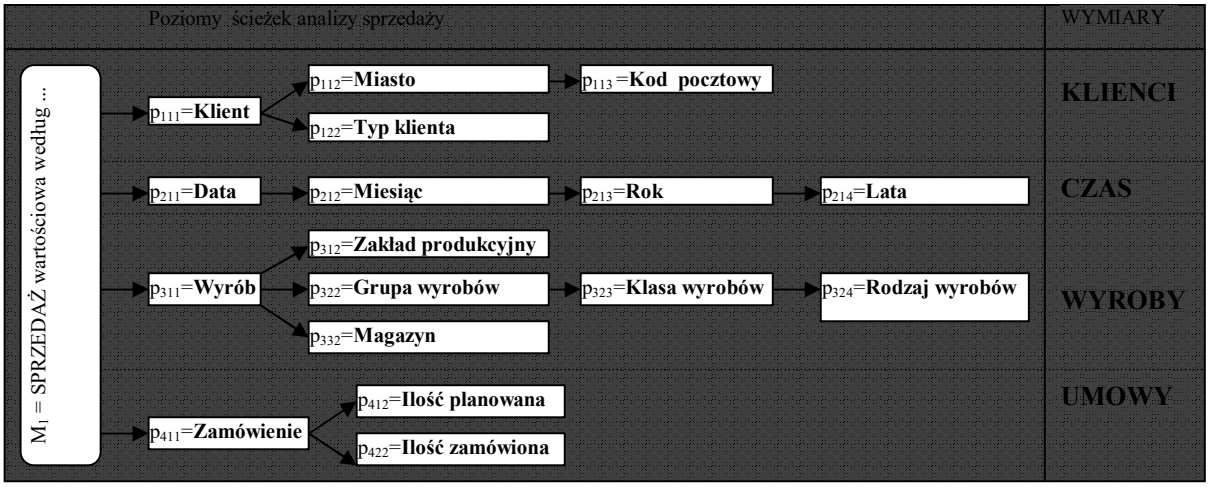
ścieżka 3, $s_{33,12} = (p_{331}, p_{332})$ zatem $S_{33} = \{s_{33,12}\}$.

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 $S_{42} = \{s_{42,12}\}$.

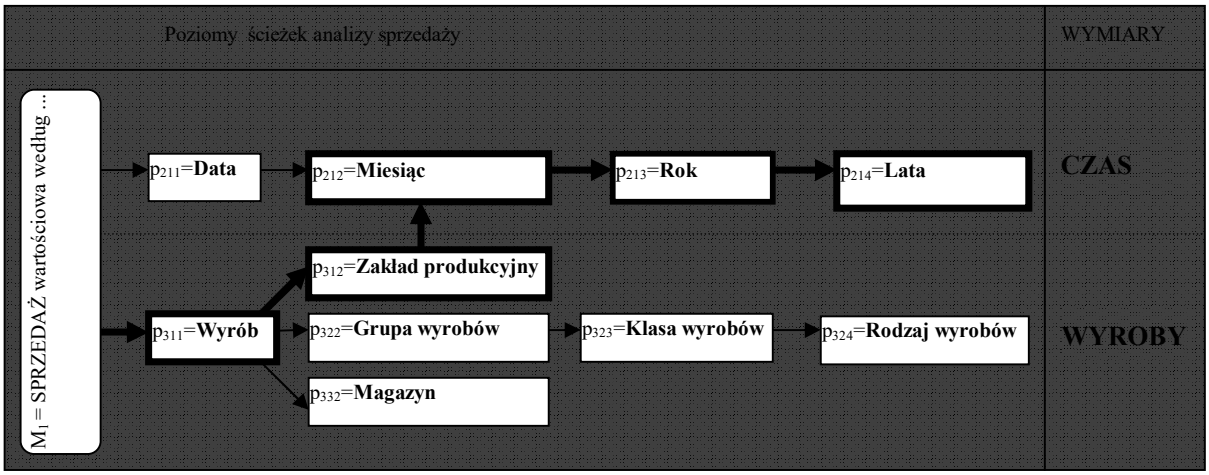
Uwzględniając, że $p_{111} = p_{121}$ = {Klient} oraz $p_{311} = p_{321} = p_{331}$ = {Wyrób}, otrzymano wynikowy schemat ścieżek wielowymiarowej analizy sprzedaży, który przedstawiono na rys. 13.



Rys. 13. Przykład schematu ścieżek wielowymiarowej analizy sprzedaży
Fig. 13. The multidimensional analytical sales paths schema example

3.3.1. Ścieżki skrośne w schemacie ścieżek wielowymiarowej analizy sprzedaży

Na podstawie otrzymanego przykładowego 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. 14.



Rys. 14. Przykład złożonej skrośnej ścieżki analizy sprzedaży
Fig. 14. The complex analytical through path sales example

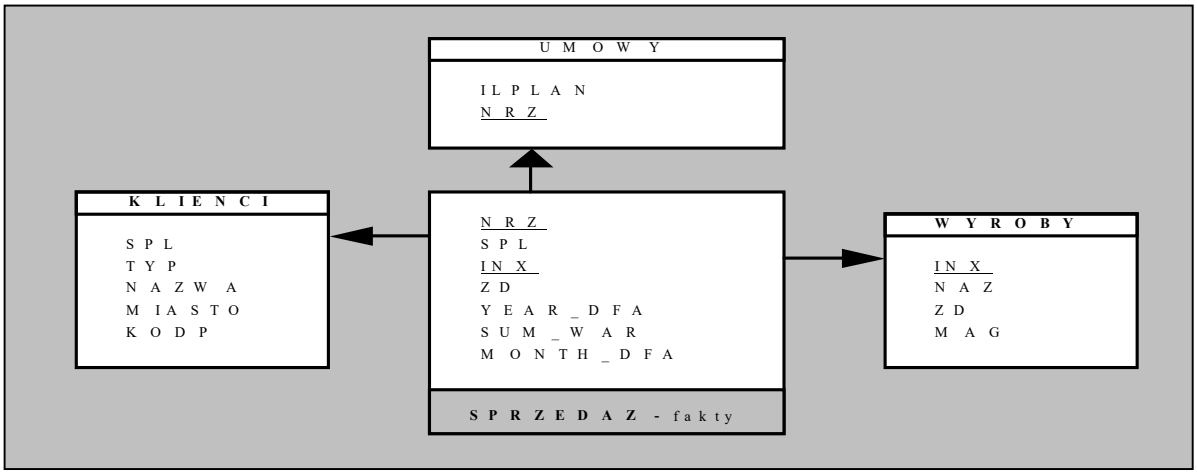
3.4. Początkowy schemat Magazynu Danych Zbiorczych

Jak już wspomniano, początkowy schemat tematycznej hurtowni danych określono za pomocą metody opisanej w publikacji [22], wykorzystującej tradycyjny model ER do projektowania hurtowni danych na podstawie przemysłowych modeli danych. W przypadku projektowania schematu hurtowni danych typu gwiazda, która może być łatwo wyprowadzona z zastanego modelu ER, tablica faktów formowana jest na podstawie encji transakcyjnych. Natomiast tablice określające wymiary tworzone są dla każdej encji komponentowej poprzez denormalizację hierarchicznie powiązanych encji klasyfikujących. 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, natomiast do encji klasyfikujących relacja PLATNOSC.

Tak więc 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 for-

mownia tablic wymiarów na podstawie encji komponentowych wykorzystano relacje KLIENCI, WYROBY oraz UMOWY. Relacji PLATNOSC należącej również do encji komponentowych nie brano na tym etapie pod uwagę z tego względu, iż wymiaru PŁATNOŚCI nie uwzględniono w schemacie ścieżek analizy wielowymiarowej.

Reasumując, otrzymany na podstawie tej metody początkowy schemat przykładowego Magazynu Danych Zbiorczych przedstawiony na rys. 15 jest schematem typu gwiazda, w którym poziomy wymiaru czasu (lata, rok, miesiące) przechowywane są w relacji faktów.



Rys. 15. Przykład początkowego schematu Magazynu Danych Zbiorczych
Fig. 15. The initial data warehouse schema example

W końcu Magazyn Danych Zbiorczych zasilono odpowiednimi, na różnych poziomach zagregowanymi danymi, pochodzącymi z archiwalnych kopii danych z zamkniętych okresów obliczeniowych.

3.4.1. Ocena początkowego schematu Magazynu Danych Zbiorczych

Jak widać z przedstawionego już rys. 15, początkowa postać schematu Magazynu Danych Zbiorczych otrzymanego za pomocą metody [22] jest taka sama, jaką uzyskano za pomocą metody [27]. Trudno jest ocenić, czy na podstawie tak określonego schematu możliwe jest zrealizowanie większości pytań analitycznych i tym samym stwierdzić, że jest on 'właściwie' określony. Za wyjątkiem pytań typu Q3, których na tym etapie rozwoju schematu Magazynu Danych Zbiorczych nie można formułować, ponieważ określony wcześniej schemat ścieżek analizy wielowymiarowej dotyczy tylko jednej miary (tj. sprzedaż według ...), niemniej jednak pozwala on na zrealizowanie każdego innego pytania analitycznego należącego do jednej z poniższych trzech klas powszechnie zadawanych OLAP-owych pytań analitycznych. Klasy te w jawny sposób określono na podstawie analizy przykładowych pytań analitycznych zawartych w pracy [26]. 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.

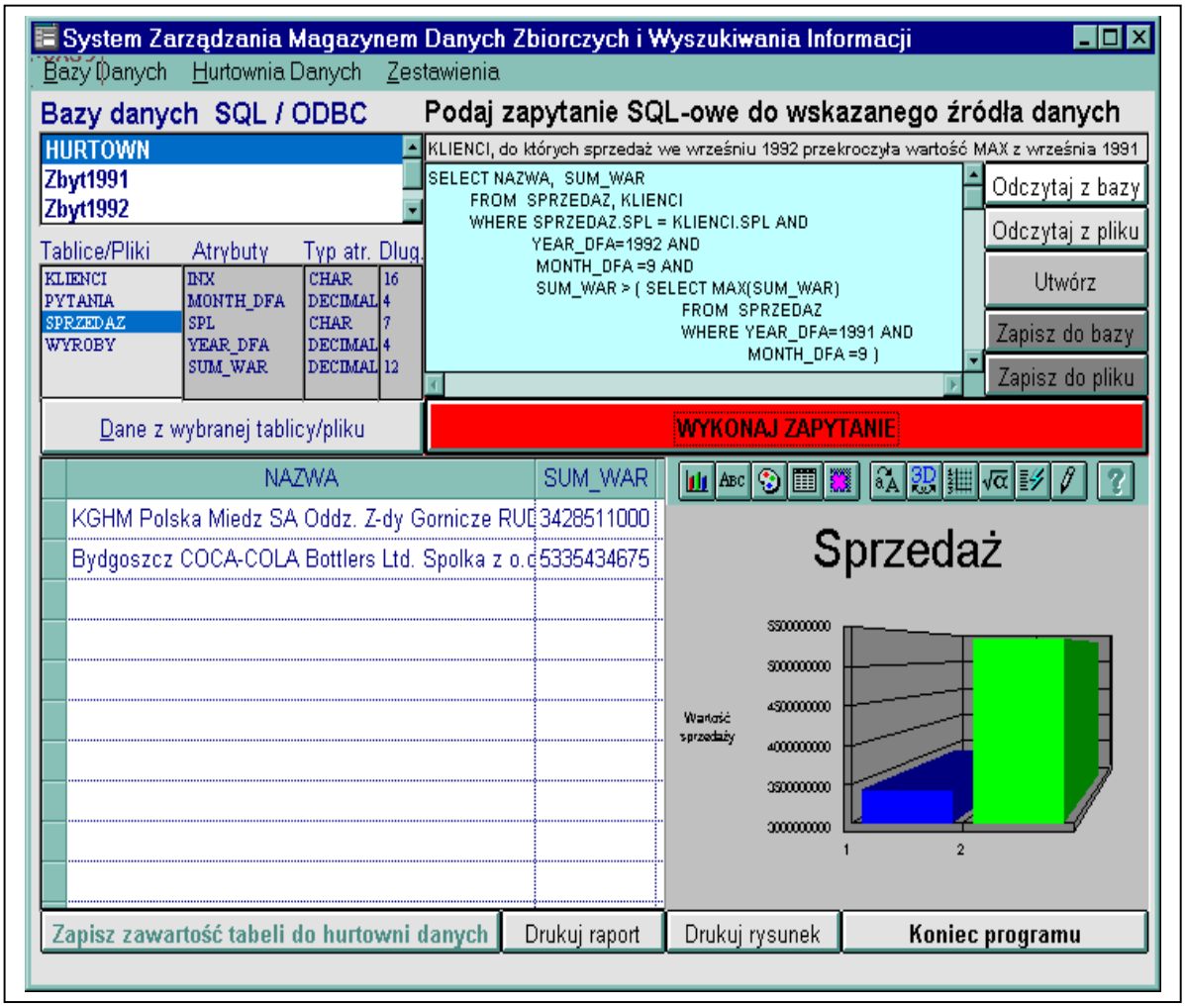
C) Klasa pytań dokonujących selekcji opartej na wcześniej zagregowanych danych na różnych poziomach; są to pytania zagnieżdżone typu Q4, dotyczące jednej miary względem jednej ścieżki kilku wymiarów, w których zagnieżdżony operator selekcji bazuje na wcześniej zagregowanych danych.

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 Magazynu Danych Zbiorczych za pomocą wcześniej już wspomnianego przykładowego systemu zarządzania Magazynem Danych Zbiorczych i Wyszukiwania Informacji. Pytanie to wyrażone w języku naturalnym brzmi następująco:

'*Wyszukaj klientów, do których sprzedaż we wrześniu 1992 r. przekroczyła wartość maksymalną z września 1991*'.

Sformułowano go na podstawie złożonej ścieżki skrośnej tj. *Klient->Miesiąc->Rok->Lata*. Analiza tego pytania prowadzona według przedstawionego już algorytmu analizy pytania analitycznego (rys. 6) 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, które skierowano wprost do Magazynu Danych Zbiorczych. Przyjmuje ono poniższą postać, którego wynik przedstawiono na rys. 16.

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

Fig 16. Question: *Find the customers to which september's sales in 1992 exceeded september's sales in 1991 year*

3.5. Dynamiczne rozszerzanie schematu Magazynu Danych Zbiorczych na podstawie niektórych przykładowych pytań analitycznych

Jak wspomniano w rozdziale 2, określony metodą [22] na podstawie zastanego przemysłowego systemu informacyjnego początkowy schemat tematycznej hurtowni danych, dynamicznie rozszerzano zgodnie z zaproponowaną koncepcją projektowania hurtowni danych na podstawie pytań analitycznych. Pytania te formułowano wykorzystując wcześniej określony schemat ścieżek analizy wielowymiarowej. Jak już wspomniano, w tej koncepcji wykorzystano zaproponowaną już wcześniej metodę [27] dynamicznego rozszerzania schematu hurtowni. Jak widać z przedstawionego już algorytmu mechanizmu dynamicznego rozszerzania sche-

matu magazynu danych (rys. 3), za pomocą dodatkowych bloków warunkowych umożliwiono dokonywanie wyboru tablicy z Magazynu Danych Zbiorczych, w której zachowywano zagregowane dane stanowiące wynik zadanego pytania SQL-owego. Tak więc celu wykazania istnienia rozwiązania sformułowanego w rozdziale 2 problemu badawczego, analizie poddano wpływ niektórych przykładowych pytań analitycznych na początkową postać schematu, należących do jednej z trzech klas pytań:

klasa 1) pytania należące do klasy pytań zwiększających wymiary hurtowni,

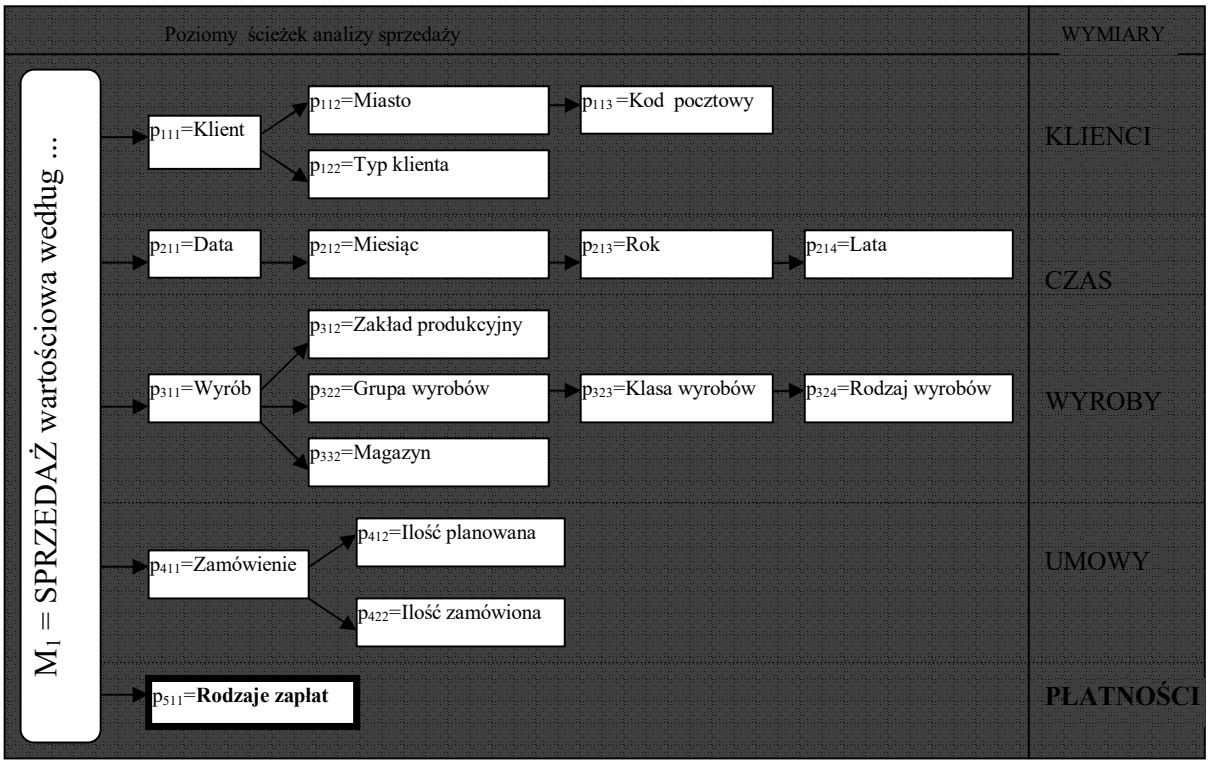
klasa 2) pytania należące do klasy pytań zwiększających liczbę ścieżek analizy w ramach danego wymiaru,

klasa 3) pytania należące do klasy pytań zwiększających liczbę poziomów w ramach danego wymiaru.

Poniżej przedstawiono przykłady pytań analitycznych, na podstawie których w dynamiczny sposób rozszerzano początkowy schemat Magazynu Danych Zbiorczych.

3.5.1. Wpływ pytań analitycznych należących do klasy pytań zwiększających liczbę wymiarów na postać początkowego schematu Magazynu Danych Zbiorczych

I Pytanie analityczne: '*Podaj grudniową sprzedaż według rodzajów płatności w latach 1993-1994*'. Pytanie to stanowi dobrą ilustrację sytuacji, w której zaproponowane i omawiane w niniejszej pracy podejście okazuje się uzasadnione. Sytuacja taka może być spowodowana różnymi względami, najczęściej takimi, w której wiedza na temat zbioru pytań analitycznych na etapie projektowania jest ograniczona lub jeśli nie wiadomo, kiedy pojawią się nowe pytania analityczne sformułowane przez kierownictwo firmy. Jest to przykład pytania typu Q1 z klasy A, należącego równocześnie do pierwszej klasy pytań zwiększających ilość wymiarów przykładowego Magazynu Danych Zbiorczych. Wymiarem w tym wypadku stanowią PŁATNOŚCI za sprzedany towar. Ponieważ jest to pytanie analityczne, którego nie skonstruowano w oparciu wcześniej określony schemat ścieżek analizy wielowymiarowej, należy zatem tak go rozszerzyć, aby uwzględniał zaistniałą sytuację. Rozszerzono go zatem o brakujący wymiar i ścieżkę analizy. Schemat ten po rozszerzeniu przybrał postać, którą przedstawiono na rys. 20. Tak więc ścieżką analizy właściwą do konstrukcji ww. pytania analitycznego jest złożona ścieżka skrośna postaci: *Rodzaje zapłaty->Miesiąc->Rok->Lata*. Ponieważ w tym pytaniu żądano pewnych informacji zbiorczych, których w Magazynie Danych Zbiorczych nie ma, zatem ww. pytanie analityczne nie mogło być do niego na tym etapie kierowane. Analiza tego pytania prowadzona według przedstawionego już algorytmu analizy pytań analitycznych (tj. bloku odpowiedzialnego za identyfikację faktów, wymiarów oraz niezbędnych atrybutów) pozwala stwierdzić, że pytanie to dotyczy zagregowanych danych dotyczących sposobów realizacji należności za sprzedane towary w latach 1993-1994, czyli sprzedaż wartościowa według...rodzajów zapłat.



Rys. 20. Nowy wymiar PŁATNOŚCI w rozszerzonym schemacie ścieżek analizy
Fig. 20. The new dimension PŁATNOŚCI in the extended analytical paths schema

Selekcji żądanych informacji należało dokonać względem nieistniejącego jeszcze w schemacie Magazynu Danych Zbiorczych wymiaru PŁATNOŚCI w funkcji wymiaru CZAS-u. Aby pytanie analityczne mogło być zrealizowane, zachodziła najpierw konieczność określenia i wykonania następującego pytania pomocniczego: '*Jak kształtowała się wartość sprzedaży według poszczególnych rodzajów płatności w poszczególnych miesiącach w latach 1991-2001*'. Pytanie to wyrażono w składni języka SQL, które przyjęło następującą postać:

```
SELECT YEAR(DFA), MONTH(DFA), SPZAP, SUM(GLOB)
    FROM  ZAPLATY
    GROUP BY  YEAR(DFA), MONTH(DFA), SPZAP
```

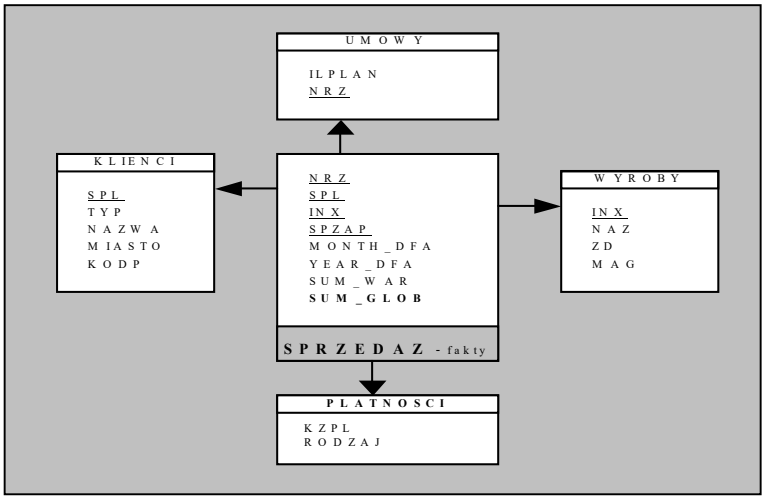
Jak wcześniej już wspomniano, przy określaniu początkowego schematu magazynu nie uwzględniono relacji ZAPLATY należącej do encji komponentowych. Ponieważ zawiera ona istotne, z punktu widzenia realizacji tego pytania analitycznego informacje, dlatego też ta relacja jest adresatem kierowanych do niej ww. pytania pomocniczego we wszystkich archiwalnych bazach danych, tj. z lat 1991-2001. Wykonanie powyższego pytania za pośrednictwem przykładowego systemu zarządzania Magazynem Danych Zbiorczych a następnie zapisanie otrzymanego rezultatu w tablicy faktów w Magazynu Danych Zbiorczych, spowodowało podczas zapisywania wyników jej dynamiczne rozszerzenie za pomocą

zaproponowanej metody. Przyjęła ona ostatecznie postać SPRZEDAZ(INX, SPL, SPZAP, YEAR_DFA, MONTH_DFA, SUM_WAR, SUM_GLOB).

Jak już wspomniano wymiary względem których należało dokonać selekcji potrzebnych informacji to PŁATNOŚCI w funkcji CZAS-u. Jak do tej pory wymiaru PŁATNOŚCI w Magazynie Danych Zbiorczych jeszcze nie określono. Tak więc, aby to pytanie analityczne mogło być zrealizowane, zachodzi konieczność określenia pytania pomocniczego postaci:

Wybierz poprawne dane o rodzajach płatności:

SELECT KZPL, RODZAJ FROM PLATNOSC, które skierowano najpierw do poszczególnych operacyjnych archiwalnych baz danych a następnie zapisano otrzymany rezultat w Magazynie Danych Zbiorczych. Podczas zapisywania wyników dynamicznie rozszerzono za pomocą zaproponowanej metody jego schemat o nową tablicę - PLATNOSC(KZPL, RODZAJ). Reasumując, rozszerzony w powyższy sposób początkowy schmat Magazynu Danych Zbiorczych posiada obecnie strukturę gwiazdy, co schematycznie przedstawiono na rys. 21.



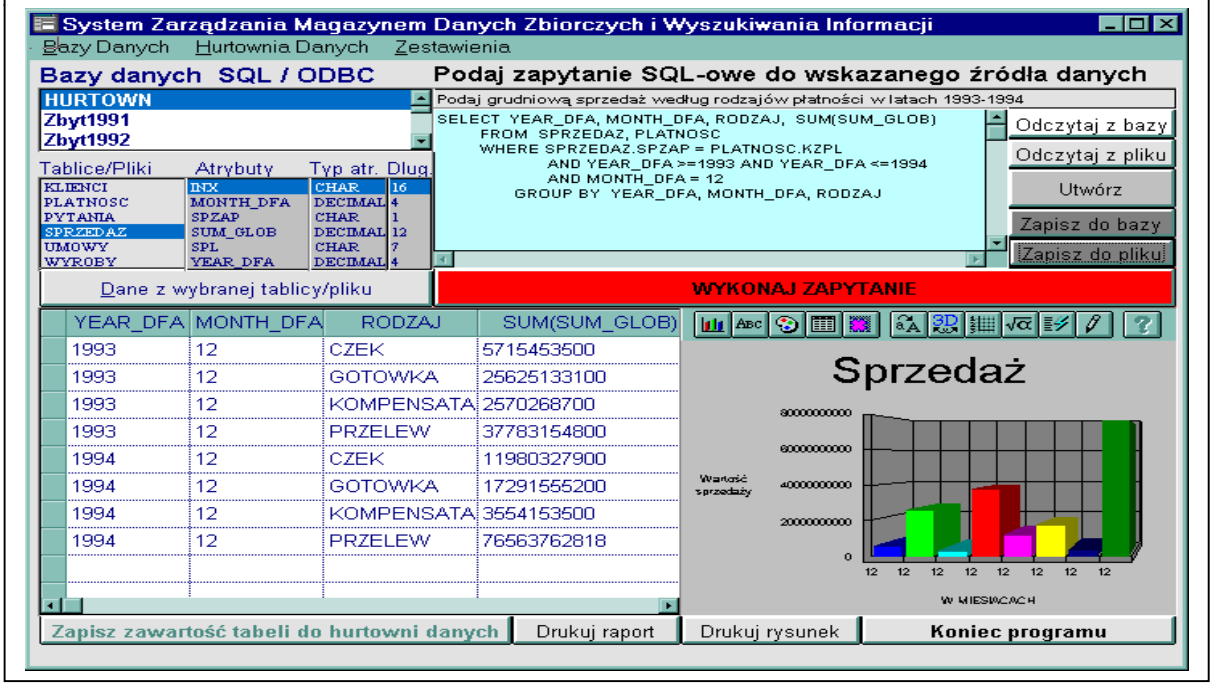
Rys. 21. Nowy wymiar PLATNOSCI schemacie Magazynu Danych Zbiorczych

Fig. 21. The new dimension PLATNOSCI in the data warehouse schema

Zatem pytanie analityczne '*Podaj grudniową sprzedaż według rodzajów płatności w latach 1993-1994*' wyrażone w języku SQL przyjmuje poniższą postać:

```
SELECT  YEAR_DFA, MONTH_DFA, RODZAJ,  SUM(SUM_GLOB)
      FROM  SPRZEDAZ, PLATNOSC
      WHERE SPRZEDAZ.SPZAP = PLATNOSC.KZPL
            AND YEAR_DFA >=1993 AND YEAR_DFA <=1994
            AND MONTH_DFA = 12
   GROUP BY  YEAR_DFA, MONTH_DFA, RODZAJ
```

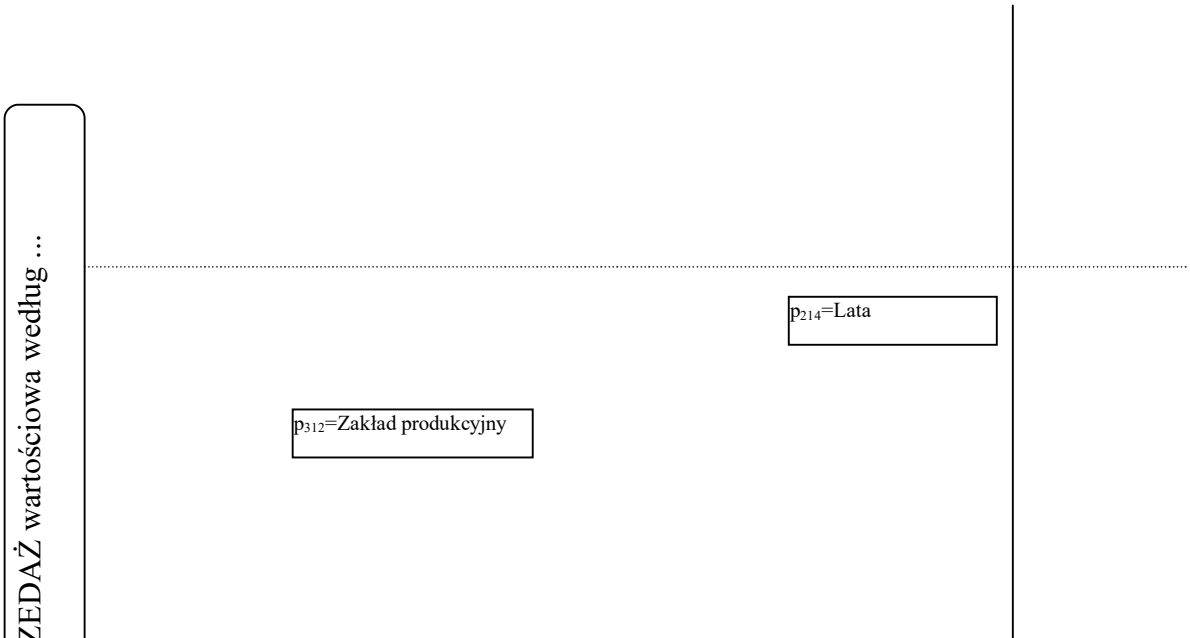
Pytanie to skierowano do Magazynu Danych Zbiorczych za pośrednictwem systemu zarządzania Magazynem Danych Zbiorczych, które wygenerowało wynik przedstawiony na rys. 22.

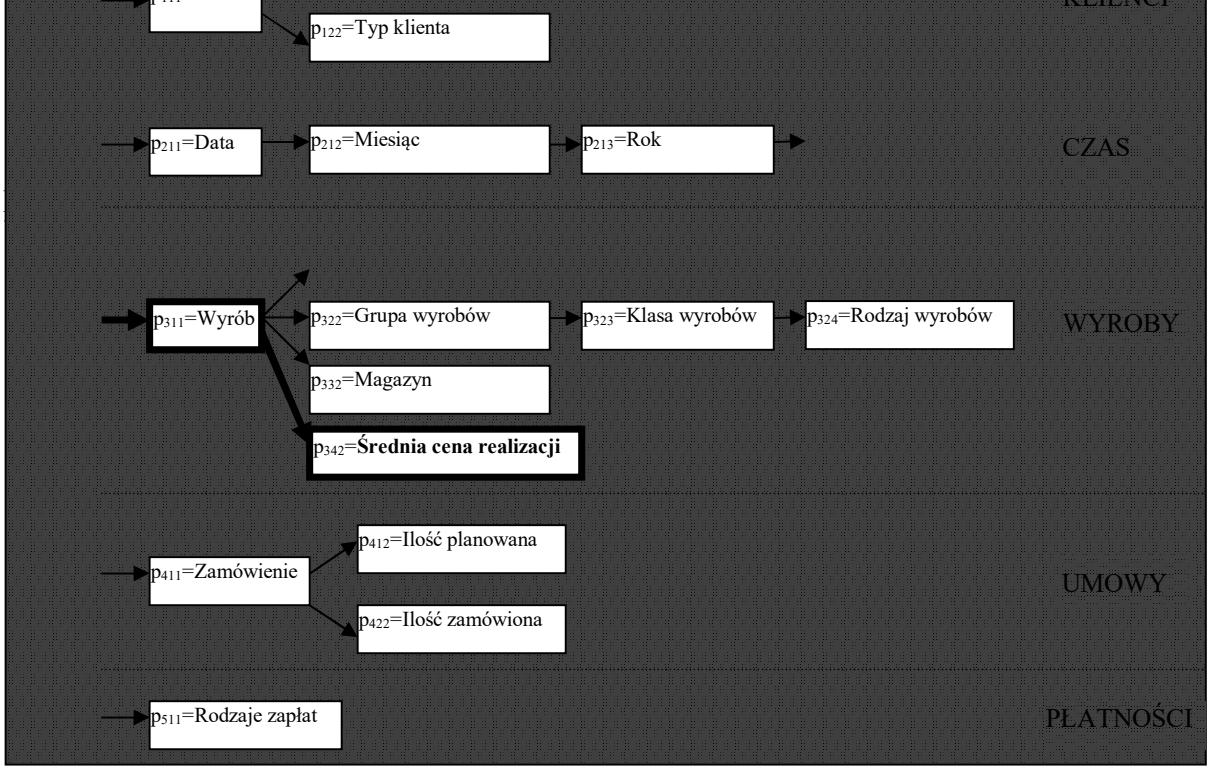


Rys 22. Pytanie: *Podaj grudniową sprzedaż według rodzajów płatności w latach 1993-1994*
Fig 22. Question: *Give the december's sales according to payment kind in 1993-1994 years*

3.5.2. Wpływ pytań analitycznych z klasy pytań zwiększających liczbę ścieżek analizy w ramach pewnego wymiaru na postać schematu Magazynu Danych Zbiorczych

II Pytanie analityczne: '*Podaj marcową średnią cenę realizacji transakcji sprzedaży wyrobu X w latach 1991-1994*'. Jest to przykład pytania typu Q1 z klasy A, należącego równocześnie do drugiej klasy pytań zwiększających liczbę ścieżek analizy w ramach pewnego wymiaru. Ponieważ jest to pytanie analityczne, którego nie skonstruowano (podobnie jak w pytaniu I) w oparciu o wcześniej określony schemat ścieżek analizy wielowymiarowej, zatem należy go tak rozszerzyć, aby uwzględniał zaistniałą sytuację. Rozszerzono go zatem o brakującą w wymiarze WYROBY ścieżkę analizy, którego nową postać, przedstawiono na rys. 23.





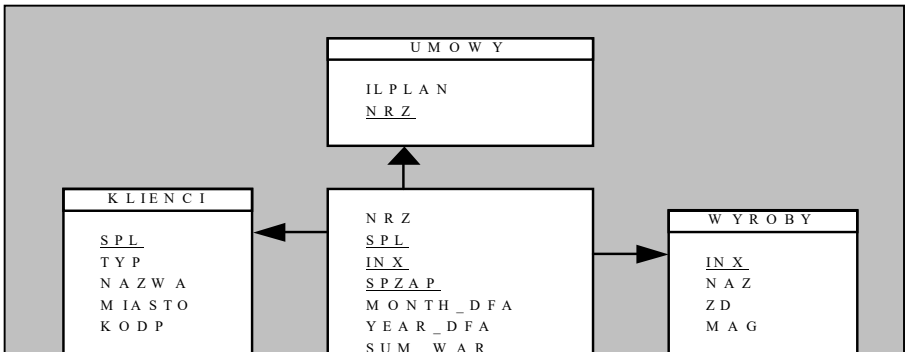
Rys. 23. Nowa ścieżka analizy w rozszerzonym schemacie ścieżek analizy
Fig. 23. The new analytical path in the extended analytical paths schema

Tak więc ścieżką analizy właściwą do konstrukcji ww. pytania analitycznego jest złożona ścieżka skrośna postaci: *Wyrob->Średnia cena realizacji->Miesiąc->Rok->Lata*.

Analiza tego pytania prowadzona według przedstawionego już algorytmu analizy pytań analitycznych pozwala stwierdzić, że pytanie to stanowi podzbiór z pewnego zbioru zawierającego zagregowane dane dotyczących średnich cen realizacji transakcji sprzedaży wyrobów w poszczególnych miesiącach a latach 1991-2001. Zatem aby ww. pytanie analityczne mogło być zrealizowane, zachodzi najpierw konieczność określenia i wykonania następującego pytania pomocniczego: '*Podaj średnie ceny realizacji transakcji sprzedaży wyrobów w poszczególnych miesiącach roku...*'

```
SELECT   YEAR(DFA),  MONTH(DFA), SPRZEDAZ.INX,  AVG(CET)
      FROM SPRZEDAZ
      GROUP BY   YEAR(DFA), MONTH(DFA), SPRZEDAZ.INX
```

Wykonanie powyższego pytania skierowanego do poszczególnych archiwalnych baz danych z lat 1991-2001 roku za pośrednictwem systemu zarządzania Magazynem Danych Zbiorczych a następnie zapisanie otrzymanego rezultatu w tablicy faktów w Magazynu Danych Zbiorczych, spowoduje dynamiczne jej rozszerzenie za pomocą zaproponowanej metody o nowy atrybut AVG_CET. Przyjmuje ona teraz postać następującą: SPRZEDAZ(INX, SPL, SPZAP, YEAR_DFA, MONTH_DFA, SUM_WAR, SUM_GLOB, AVG_CET). W ten sposób dokonano dalszego, rozszerzenia schmatu Magazynu Danych Zbiorczych, którego schemat przedstawiono na rys. 24.



Rys. 24. Rozszerzona tablica faktów w schemacie magazynu danych
Fig. 24. The extended fact table in the data warehouse schema

Tak więc, rozszerzony Magazyn Danych Zbiorczych posiada już odpowiedni schemat i zawiera odpowiednio zagregowane informacje niezbędne do poprawnego skonstruowania wspomnianego II pytania analitycznego. I tak poszukiwana w tym pytaniu miara z tablicy faktów to *sprzedaż według... średnich cen realizacji*, którą reprezentuje atrybut AVG_CET. W pytaniu tym poszukiwany jest pewien wyrób, którego nazwa reprezentowana są przez atrybut NAZ (*Nazwa Klienta* z tablicy WYROBY). Wymiary względem których dokonuje się selekcji to WYROBY oraz CZAS (zawarty w relacji faktów). W magazynie znajdują się zagregowane dotyczące średnich cen realizacji transakcji sprzedaży wyrobów w poszczególnych miesiącach z lat 1991-1994. Zatem w końcowym pytaniu analitycznym wyrażonym w języku SQL zaangażowane będą atrybuty grupujące YEAR_DFA, MONTH_DFA, NAZ oraz AVG_CET. Poszukiwane *marcowe średnie ceny realizacji transakcji sprzedaży wyrobu X w latach 1991-1994* stanowią odpowiedź na postawione II pytanie analityczne skierowane wprost do Magazynu Danych Zbiorczych. Pytanie to wyrażone w języku SQL przyjmuje ostatecznie poniższą postać:

```
SELECT  YEAR_DFA, MONTH_DFA, WYROBY.NAZ, AVG_CET
      FROM  SPRZEDAZ, WYROBY
      WHERE  SPRZEDAZ.INX = WYROBY.INX
            AND SPRZEDAZ.INX = '1361-113-100-001'
            AND YEAR_DFA  >= 1991 AND YEAR_DFA  <= 1994
            AND MONTH_DFA = 3
       GROUP BY  YEAR_DFA, MONTH_DFA, NAZ, AVG_CET
```

Po skierowaniu go do Magazynu Danych Zbiorczych za pośrednictwem systemu zarządzania Magazynem Danych Zbiorczych generuje wynik, który przedstawiono na rys. 25.

Rys 25. Pytanie: Jaka była marcową średnią cenę realizacji transakcji sprzedaży wyrobu X w latach 1991-1994
Fig 25. Question: What was the march's average transaction realization sales prize of the X product in 1993-1994 years

Reasumując, powyżej przedstawiono analizę pytania analitycznego, należącego do drugiej klasy pytań zwiększających liczbę ścieżek analizy w ramach pewnego wymiaru. Jego wpływ na postać schamatu Magazynu Danych Zbiorczych zaznaczył się rozszerzeniem tablicy faktów o nowy atrybut. W ramach dyskusji nad pytaniami analitycznymi należącymi do drugiej klasy pytań zwiększających liczbę ścieżek analizy w ramach pewnego wymiaru, poniżej przedstawiono analizę innego pytania. Wpływ tego pytania na postać schematu magazynu zaznaczył się w inny sposób. Realizacja tego pytania spowodowała rozszerzenie dotychczas otrzymanego schematu Magazynu Danych Zbiorczych do postaci, którą zakwalifikować można do postaci typu płatka śniegu.

Pytanie analityczne IIa: '*Jaka była grudniowa sprzedaż zakładu X w latach 1991-1993*'

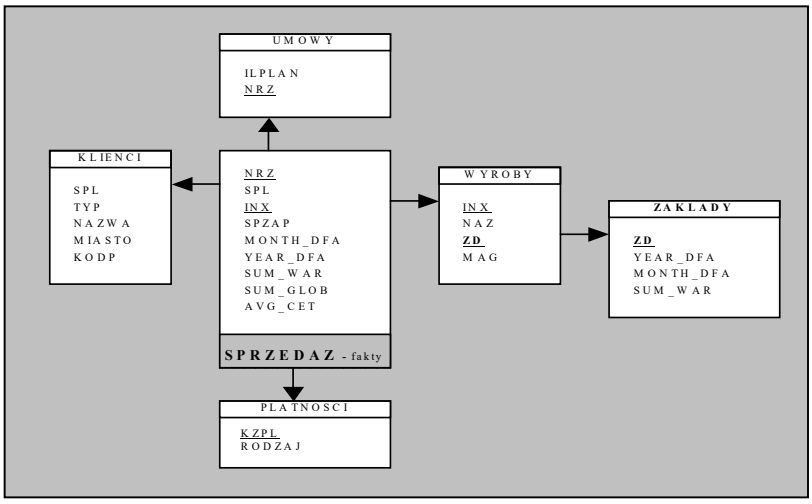
Jest to przykład pytania typu Q1 z klasy A, należącego równocześnie do drugiej klasy pytań zwiększających liczbę poziomów w ścieżkach analizy w ramach pewnego wymiaru. W przeciwieństwie do pytania II jest to pytanie analityczne, które skonstruowano w oparciu o wcześniej określony schemat ścieżek analizy wielowymiarowej tj. w oparciu o złożoną ścieżkę skrośną postaci: *Wyrob->Zakład produkcyjny->Miesiąc->Rok->Lata*. W pytaniu tym poszukiwane są zakłady produkcyjne reprezentowane przez atrybut ZD (Zakład). Wymiary względem których dokonuje się selekcji to WYROBY oraz CZAS. Dalej można stwierdzić, że pyta-

nie dotyczy zagregowanych danych dotyczących sprzedaży poszczególnych zakładów w kolejnych miesiącach na przestrzeni lat 1991-1993. Zatem, aby udzielić poprawnej odpowiedzi na ww. pytanie wymagane jest zrealizowanie pytania pomocniczego tj.: '*Jaka była wartość sprzedaży poszczególnych zakładów w poszczególnych latach*'. Pytanie to wyrażone w języku SQL przyjmuje następującą postać:

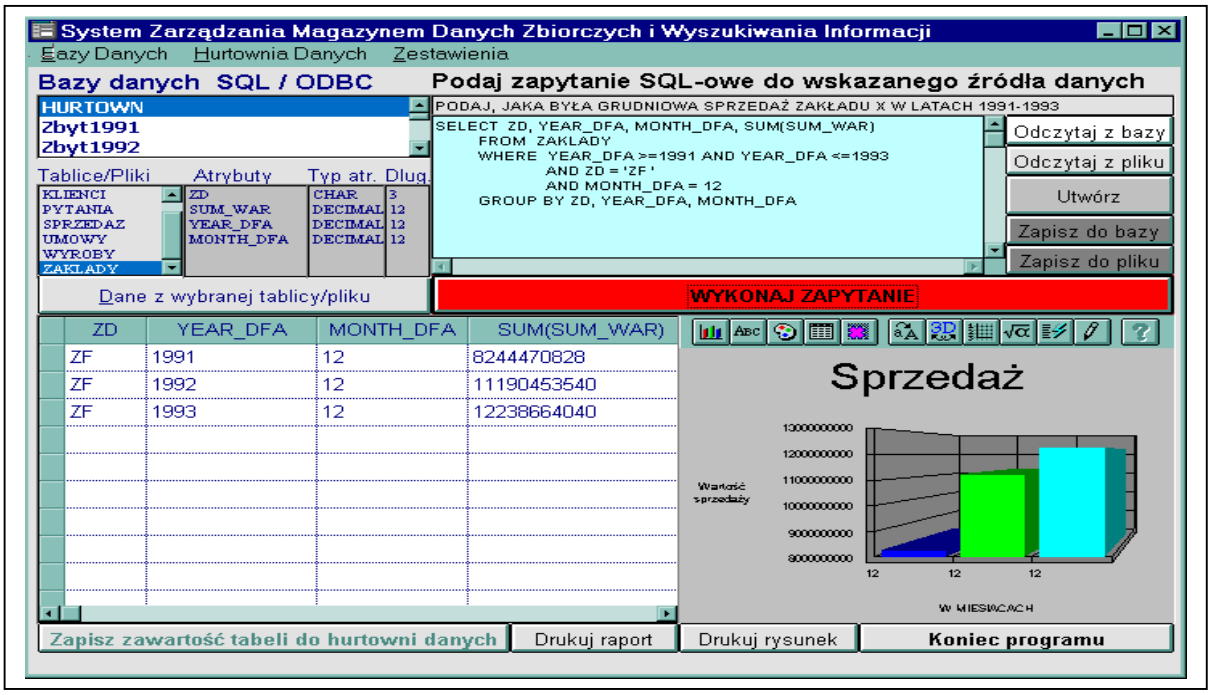
```
SELECT  YEAR(DFA),  MONTH(DFA),  ZD, SUM(WAR)
        FROM  SPRZEDAZ
        GROUP BY  YEAR(DFA), MONTH(DFA), ZD
```

Skierowano go za pośrednictwem systemu zarządzania Magazynem Danych zbiorczych do poszczególnych archiwalnych baz danych z lat 1991-2001. Wyników tego pytania nie zapisano jednak w tablicy faktów Magazynu Danych Zbiorczych. Korzystając z wcześniej zaprezentowanego algorytmu mechanizmu dynamicznego rozszerzania schematu magazynu danych, wyniki zapisano w nowo utworzonej tablicy o nazwie ZAKLADY. Za pomocą tego mechanizmu jej schemat określono następująco: ZAKŁADY(YEAR_DFA, MONTH_DFA, ZD, SUM_WAR). W ten sposób dokonano dalszego rozszerzenia schmatu Magazynu Danych Zbiorczych, którego schemat typu płatka śniegu przedstawiono na rys. 26. Przy tak określonym schemacie Magazynu Danych Zbiorczych pytanie analityczne '*Jaka była grudniowa sprzedaż zakładu X w latach 1991-1993*' wyrażone w języku SQL może przyjąć poniższą postać, którego wynik przedstawiono na rys. 27.

```
SELECT  ZD, YEAR_DFA, MONTH_DFA, SUM(SUM_WAR)
        FROM  ZAKLADY
        WHERE  YEAR_DFA >=1991 AND YEAR_DFA <=1993
                AND ZD = 'ZF'
                AND MONTH_DFA = 12
        GROUP BY ZD, YEAR_DFA, MONTH_DFA
```



Rys. 26. Schemat magazynu danych typu płatek śniegu
Fig. 26. The snow flake schema type of the data warehouse



Rys. 27. Pytanie: Jaka była grudniowa sprzedaż zakładu X w latach 1991-1993
Fig. 27. Question: Wat was the december's sales X factory in 1991-1993 years

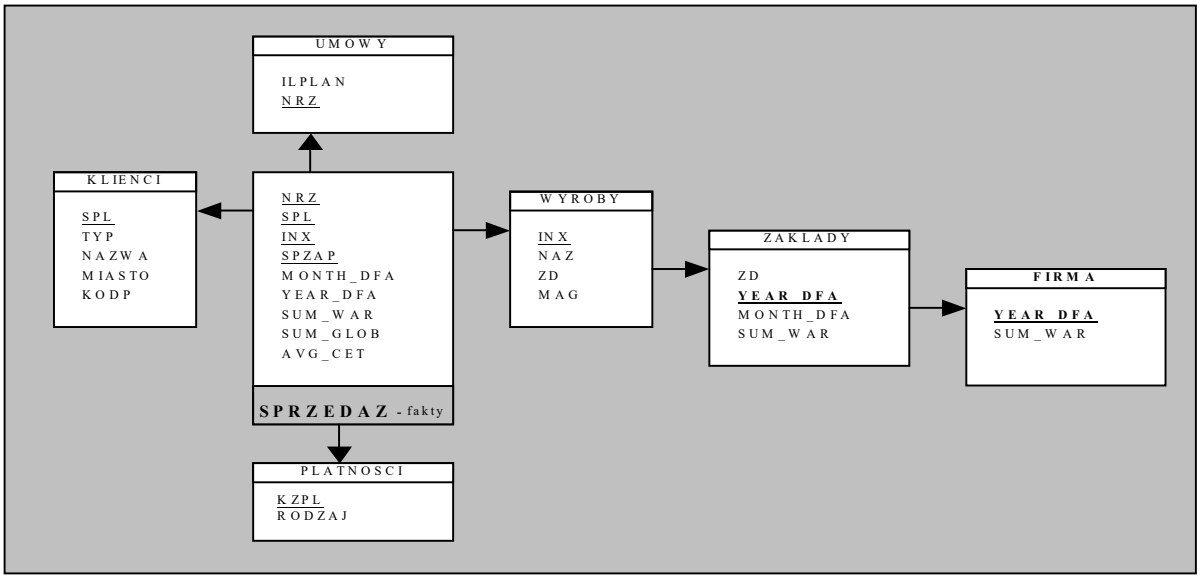
3.5.2. Wpływ pytań analitycznych, z klasy pytań zwiększających liczbę poziomów w pewnej ścieżce analizy pewnego wymiaru, na postać schematu Magazynu Danych Zbiorczych

III Pytanie analityczne: '*Podaj udział sprzedży majowej w stosunku do globalnej w latach 1991-1993*'. Jest to przykład pytania typu Q1 z klasy A, należącego równocześnie do trzeciej klasy pytań zwiększających liczbę poziomów w pewnej ścieżce analizy pewnego wymiaru. W przeciwieństwie do pytania IIa jest to pytanie analityczne, dla konstrukcji którego dokonano dalszego rozszerzenia schematu ścieżek analizy wielowymiarowej. Dodano do niego nową ścieżkę złożoną ścieżkę prostą postaci: *Wyrob->Zakład produkcyjny->firma*, która wraz ze ścieżką prostą z wymiaru CZAS daje złożoną skrośną ścieżkę postaci: *Wyrob->Zakład produkcyjny->firma->Miesiąc->Rok->Lata*. Dalej można stwierdzić, że pytanie dotyczy zagregowanych danych dotyczących sprzedaży globalnej na przestrzeni lat 1991-1993. Zatem, aby udzielić poprawnej odpowiedzi na ww. pytanie wymagane jest zrealizowanie pytania pomocniczego: '*Jaka była wartość sprzedaży w poszczególnych latach'*. Wyrażone w języku SQL przyjmuje następującą postać:

```
SELECT  YEAR(DFA),  SUM(WAR)
    FROM  SPRZEDAZ
```

GROUP BY YEAR(DFA)

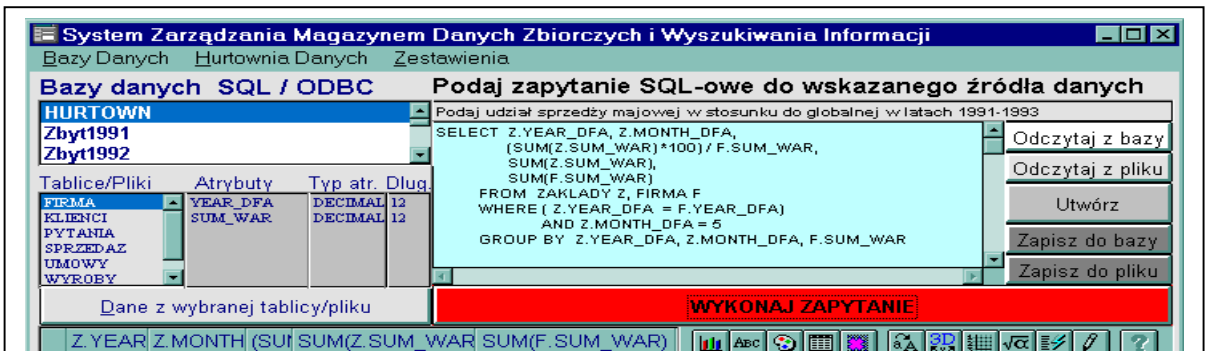
Pytanie to skierowano za pośrednictwem systemu zarządzania Magazynem Danych zbiorczych do poszczególnych archiwalnych baz danych z lat 1991-2001. Wyników tego pytania, podobnie jak w pytaniu IIa, nie zapisano w tablicy faktów Magazynu Danych Zbiorczych, lecz w dynamiczny sposób utworzonej tablicy FIRMA(YEAR_DFA, SUM_WAR). W ten sposób dokonano dalszego rozszerzonia schmatu Magazynu Danych Zbiorczych, którego schemat typu płatka śniegu przedstawiono na rys. 28.



Rys. 28. Rozszerzony schemat magazynu danych typu płatek śniegu
Fig. 28. The extended snow flake schema type of the data warehouse

Przy tak określonym schemacie Magazynu Danych Zbiorczych pytanie analityczne '*Podaj udział sprzedży majowej w stosunku do globalnej w latach 1991-1993*' wyrażone w języku SQL może przyjąć poniższą postać, którego wynik przedstawiono na rys. 29.

```
SELECT  Z.YEAR_DFA, Z.MONTH_DFA,
        (SUM(Z.SUM_WAR) *100) / F.SUM_WAR,
        SUM(Z.SUM_WAR),
        SUM(F.SUM_WAR)
    FROM  ZAKLADY Z, FIRMA F
    WHERE ( Z.YEAR_DFA  = F.YEAR_DFA)
          AND Z.MONTH_DFA = 5
    GROUP BY  Z.YEAR_DFA, Z.MONTH_DFA, F.SUM_WAR
```



Rys. 29. Pytanie: *Podaj udział sprzedży majowej w stosunku do globalnej w latach 1991-1993*
Fig. 29. Question: *Give the may's participation sale relative to global in 1991-1993 years*

3.6. Wnioski i ocena wpływu niektórych pytań analitycznych na postać dynamicznie rozszerzanego początkowego schematu Magazynu Danych Zbiorczych

W pracy analizie poddano wpływ niektórych pytań analitycznych na postać schematu Magazynu Danych Zbiorczych, które konstruowano na podstawie ścieżek analizy ze schematu ścieżek analizy wielowymiarowej. Badano wpływ pytań należących do klasy pytań zwiększających liczbę wymiarów hurtowni, zwiększających liczbę ścieżek analizy w ramach pewnego wymiaru oraz pytań zwiększających liczbę poziomów w pewnej ścieżce analizy pewnego wymiaru. Dzięki zastosowaniu zmodyfikowanego algorytmu mechanizmu dynamicznego rozszerzania schematu magazynu danych uzyskano możliwość swobodnego kształtowania postaci dynamicznie rozszerzanego schematu Magazynu Danych Zbiorczych. Wpływ przykładowych pytań analitycznych (IIa oraz III) należących do powyższych klas na postać tego schematu był taki, że początkowy typ schematu Magazynu Danych Zbiorczych określony jako gwiazda doprowadzono do schematu postaci płatka śniegu. Zademonstrowano w ten sposób możliwość swobodnego kształtowania postaci dynamicznie rozszerzanego schematu Magazynu Danych Zbiorczych. Wniosek zasadniczy wynikający z przeprowadzonej dyskusji jest następujący: pytania analityczne należące do klasy pytań zwiększających liczbę ścieżek analizy w ramach pewnego wymiaru lub też pytania analityczne należące do klasy zwiększających liczbę poziomów w pewnej ścieżce analizy pewnego wymiaru ze schematu ścieżek analizy wielowymiarowej, mogą lecz nie muszą skutkować dynamicznym rozszerzeniem schematu Magazynu Danych Zbiorczych w postaci nowej tablicy.

3.6.1. Wpływ pytań analitycznych na powstanie ewentualnych zależności typu wiele-do-wielu pomiędzy tablicą faktów i tablicami wymiarów

Często się zdarza, że powszechnie akceptowane schematy hurtowni danych typu gwiazda zawierające relacje typu jeden-do-wielu, mogą zawierać również pomiędzy tablicą faktów i pewną tablicą wymiarów relacje typu wiele-do-wielu. Istnienie relacji tego typu generuje kilka trudnych problemów. Wśród nich wyróżnić można utratę prostoty struktury typu gwiazda, wzrost stopnia złożoności formułowanych pytań analitycznych, spadek efektywności wykonywanych pytań spowodowanych wprowadzeniem większej ilości złączeń. Jednym z rozwiązań tych problemów może być wprowadzenie do schematu hurtowni danych [34] tablicy łączącej. Jest ona podobna do encji pośredniczącej [38], powiązanej za pomocą relacji jeden-do-wielu ze znormalizowanymi encjami początkowo zawierających relacje typu wiele-do-wielu.

W pracy badano wpływ pytań analitycznych na postać dynamicznie tworzonego schematu Magazynu Danych Zbiorczych, które konstruowano na podstawie określonych ścieżek analizy ze schematu ścieżek analizy wielowymiarowej. Ponieważ relacje łączące poziomy w tym schemacie są relacjami grupowania/klasyfikowania typu jeden-do-wielu, zatem wynikowa postać schematu Magazynu Danych Zbiorczych również zawiera relacje typu jeden-do-wielu.

4. Podsumowanie

Jak wspomniano we wprowadzeniu, projektowanie hurtowni danych wymaga technik kompletnie różnych od tych, które zostały zaadoptowane z systemów transakcyjnych. W artykule przedstawiono aktualny przegląd oraz syntezę stanu wiedzy odnoszącego się do tradycyjnego podejścia do problemu statycznego projektowania, budowy hurtowni oraz ekstrakcji danych. Na podstawie przedstawionej syntezy zaproponowano inne podejście do problemu dynamicznego projektowania i budowy hurtowni danych, biorąc pod uwagę pojawiające się w różnym czasie nowe pytania analityczne. Podejście to przedstawiono w kontekście najnowszych prac badawczych dostępnych w Internecie. W artykule przedstawiono aktualny i wyczerpujący stan prac badawczych związanych z projektowaniem hurtowni danych, w szczególności na poziomie konceptualnym. W zaproponowanym podejściu do problemu dynamicznego projektowania hurtowni danych, wykorzystano zaproponowaną metodę dynamicznego rozszerzania schematu hurtowni. Jest ona obarczona istotnymi wadami, niemniej jednak w przypadku tworzenia małych hurtowni tematycznych na podstawie istniejących zastanych przemysłowych systemów informacyjnych, może okazać się przydatna dzięki swej prostocie oraz wykorzystaniu standardu ODBC. Ponadto, w zaproponowanym podejściu, zastosowano metodę, która zapewnia ewolucję schematu Magazynu Danych Zbiorczych w sytuacji, gdy

pojawiają się nowe pytania analityczne. Zalety zaproponowanego podejścia dynamicznego projektowania hurtowni danych są następujące:

1) wyeliminowanie zbioru pytań analitycznych określonych przez użytkownika końcowego, niezbędnego do zaprojektowania właściwego schematu hurtowni danych,
2) pytania analityczne użytkownika mogą być formułowane *ad-hoc* w oparciu o kombinacje wcześniej określonych na podstawie zastanego modelu danych bazy OLTP ścieżek analizy, zapewniając tym samym możliwość realizacji takich pytań.

Natomiast podstawową wadą zaproponowanego podejścia jest konieczność określenia ścieżek analizy na podstawie wymagań analityka oraz schematu ER (o ile istnieje) 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. W przeciwieństwie jednak do rozwiązania tradycyjnego w którym zakłada się, że schemat hurtowni powinien bezpośrednio wynikać z wcześniej określonego zbioru pytań analitycznych, zaproponowane rozwiązanie jest bardziej elastyczne, ponieważ na etapie projektowania hurtowni danych zbędna staje się znajomość pełnego zbioru pytań analitycznych. Znacznie łatwiej dla projektanta współpracującego z analitykiem jest określenie potencjalnych ścieżek analizy, niż kompletnego zbioru potencjalnych pytań analitycznych.

LITERATURA

1. Kimbal R.: A Dimensional Modeling Manifesto. DBMS, August 1997, http://www.dbmsmag.com/9708d15.html.
2. Kimbal R.: Is ER Modeling Hazardous to DSS. DBMS – June 1995 – Data Warehouse Architect, http://www.dbmsmag.com/9610d05.html.
3. 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.
4. Golfarelli M., Rizzi S.: Designing the Data Warehouse: Key Steps and Crucial Issues. In the Journal of the Computer Science and Information Management, Vol.2, N.3, 1999, http://citeseer.nj.nec.com/cache/papers/cs/7966/ftp:zSzzftpdb.deis.unibo.itzSzpubzSzstefanozSzjcsim99.pdf/golfarelli99designing.pdf.

5. 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.
6. 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.
7. Kimbal R.: Slowly Changing Dimensions. http://www.dbmsmag.com/9604d05.html.
8. Brien P.: Schema Evolution in Heterogeneous Database Architectures, A Schema Transf ormation Approach. http://citeseer.nj.nec/cache/papers/cs/23007/http:zSzzSzwww.dcs. bbk.ac.ukzSz~apzSzpubszSztr310300.pdf/schema-evolution-inheterogenous.pdf.
9. Sorensen J., Alnor K.: Creating a Data Warehouse using SQL Server. In Proceedings of the International Workshop on Design and Management of Data Warehouses (DMDW'99), Heidelberg, Germany, 14-15.6 1999, http://sunsite.informatik. rwth-aachen.de/Publications/CEUR-WS/Vol-19/paper10.pdf
10. Vassiliadis P., Sellis T.: A Survey on Logical Models for OLAP Databases. Technical Report DWQ NTUA 301, accepted for publications at SIGMOD RECORD, http://www.cs.toronto.edu/~mendel/dwbib.html.
11. 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~syste m42zSzpublicationszSzdwdm98.pdf/extending-the-e-r.pdf.
12. Hüsemann B., Lechtenbörger J., Vossen G.: Conceptual Data Warehouse Design. In Proceedings of the International Workshop on Design and Management of Data Warehouses (DMDW'2000), Stockholm, Sweden, June 5 6, 2000, http:sunsite.informatik.rw th-aachen.de/Publications/CEUR-WS/Vol-28/paper6.pdf.
13. Lechtenbörger J., Vossen G.: Multidimensional Normal Forms for Data Warehouse Design. http://citeseer.nec.com/cache/papers/cs/22985/http:zSzzSzdbms.unimuenster.dezS zpublicationszSzdownloadszSzmnf.pdfmultidimensional-normal-forms-for.pdf.
14. Phipps C., Davis K.: Automating Data Warehouse Conceptual Schema Design and Evaluation. In Proceedings of 4th International Workshop DMDW'2002, Toronto, Canada, May 27, 2002, http://sunsite.informatik.rwth-aachen.de/Publications/CEUR-WS/Vol-58/phipps-davis.pdf.
15. Blaschka M, Sapia C., Höfling G.: On Schema Evolution in Multidimensional Databases. http://citeseer.nj.nec/cache/papers/cs/21200/http:zSzzSzwww.forwiss.dezSz~system42z SzpublicationszSzdawak_camera.pdf/blashka98schema.pdf

16. 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
17. Boehnlein M, Ende A.: Deriving Initial Data Warehouse Structures from the Conceptual Data Models of the Underlying Operational Information Systems. 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/dolap99_Boeh.pdf
18. Hahn K., Sapia C., Blaschka M.: Automatically Generating OLAP Schemata from Conceptual Graphical Models. In Proceedings of the ACM Third International Workshop on Data Warehousing and OLAP, DOLAP2000, Washington, DC, USA, November 2000, http://www.cis.drexel.edu/faculty/song/DOLAP2000/wrkshpproc/Hahn_110.pdf.
19. Franconi E., Sattler U.: A Data Warehouse Conceptual Data Model for Multidimensional Aggregation. In Proceedings of the International Workshop DMDW'99, Haidelberg, Germany, June 14-15, 1999, http://sunsite.informatik.rwth-aachen.de/Publications/CEUR-WS/Vol19/paper13.pdf.
20. Li C., Wang X.: A Data Model for Supporting On-Line Analytical Processing. In Proceedings of the Conference on Information and Knowledge Management, November 1996, 81-88, http://www.cs.toronto.edu/~mendel/dwbib.html.
21. Golfarelli M., Maio D., Rizzi S.: Conceptual Design of Data Warehouses from E/R Schemas. In the Proceedings of the Hawaii International Conference On System Sciences, January 6-7, 1998, Kona, Hawaii.
22. 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.
23. Schouten H.: Analysis and Design of Data Warehouses. In Proceedings of the International Workshop DMDW'99, Haidelberg, Germany, June 14 15, 1999, http://sunsite.informatik.rwth-aachen.de/Publications/CEUR-WS/Vol-19/paper5.pdf.
24. Theodoratos D., Sellis T.: Designing Data Warehouses. Data and Knowledge Engineering, 31, 3, Oct. 1999, pp 279 301, http://www.cs.toronto.edu/~mendel/dwbib.html.
25. Theodoratos D., Sellis T.: Data Warehouse Configuration. In Proceedings of the 23rd VLDB Conference (VLDB'97), Athens, Greece, August 1997, http://www.dblab.ntua.gr/~dwq/vldb_dwq.pdf.

26. 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
27. Bok Z..: Integracja relacyjnych baz danych w zastanych przemysłowych systemach informatycznych, Studia Informatica, Volume 23, Nr 4, Politechnika Śląska, Gliwice, 2002.
28. Kimbal R.: Mastering Data Extraction. DBMS – June 1995 – Data Warehouse Architect, http://www.dbmsmag.com/9606d05.html.
29. Haisten M.: Designing a Data Warehouse. http://www.damanconsulting.com/solutions/data_org/whitepaper/designing.pdf.
30. 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
31. Kotidis Y., Roussopoulos N.: DynaMat: A Dynamic View Management System for Data Warehouses. In Proc. ACM/SIGMOD'99, http://www.cs.toronto.edu/~mendel/dwbib.html.
32. Theodoratos D., Sellis T.: Dynamic Data Warehouse Design. In Proceedings of the International Conference on Data Warehousing and Knowledge Discovery, (DaWa'99), Florence Italy, August 1999, Springer LNCS 1676, pp. 1 10, http://www.cs.toronto.edu/~mendel/dwbib.html.
33. McGuff F.: Designing the Perfect Data Warehouse. http://members/aol.com/fmcguff/dwmodel/frtext.htm.
34. Song I., Rowen W.: An Analysis of Many-to-Many Relationships Between facts and Dimension Tables in Dimensional Modeling. http://sunsite.informatik.rwth-aachen.de/Publications/CEUR-WS/Vol-39/paper6.pdf
35. 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.
36. Ignasiak E.: Teoria grafów i planowanie sieciowe. Państwowe Wydawnictwo Ekonomiczne, Warszawa 1982.
37. Jankowski B.: Grafy, algorytmy w Pascalu. Wydawnictwo "Mikom", Warszawa 1988, ISBN 83-7158-077-0.
38. Muller R.J.: Bazy danych, język UML w modelowaniu danych. ISBN 83-7279-000-0, wydawnictwo MIKOM, Warszawa, luty 2000.

Recenzent:

Wpłynęło do Redakcji 31 stycznia 2003 r.

Abstract

In this article – a proposal of different approach to data warehouse design problem has been presented. Based on this approach and dynamically extension data warehouse schema proposed method, a data warehouse design problem taking into account analytical queries formulated by end user has been discussed. In this approach every new analytical query is analyzed at an angle of it's realizability. If it can not been executed then to isolate possible auxiliary or partially (one-route) queries, eg. such queries whose results are input data to a new analytical query to further analysis is submitted. Based on this isolated auxiliary or partially queries, a decision about incrementally and dynamically data warehouse schema creation/extension is taking with the aid of proposed method. In proposed approach to data warehouse design problem, the Multidimensional Aggregation Cube data model and associated with it terms and concepts has been selected in order to accomplish analytical queries influence to the shape of dynamically extended warehouse data schema. In particular, based on introduced by MAC model author's analysis paths conception, a formal multidimensional schema analysis paths model was proposed, which was base to construction correct – from the legacy OLTP systems viewpoint - analytical queries.