STUDIA INFORMATICA 2001
Volume 22 Number 4 (46)

Katarzyna HARĘŻLAK, Stanisław KOZIELSKI, Sebastian SMERD
Politechnika Śląska, Instytut Informatyki

PROJEKT I IMPLEMENTACJA HURTOWNI DANYCH DZIEKANATU

Streszczenie. W pracy przedstawiono proces projektowania, a następnie realizacji hurtowni danych korzystającej z danych gromadzonych w systemie informatycznym DZIEKANAT. System ten zapewnia m.in. obsługę procesu dydaktycznego uczelni wyższej. W pracy zaproponowano dwupoziomową organizację hurtowni danych. Poziom podstawowy stanowi hurtownia danych o modelu ROLAP, gromadząca tylko dane szczegółowe. Wyższy poziom tworzy hurtownia o modelu MOLAP, mieszcząca dane zagregowane. Całość zrealizowano przy wykorzystaniu narzędzi programowych firmy Oracle.

DESIGN AND IMPLEMENTATION OF UNIVERSITY OFFICES DATA WAREHOUSE

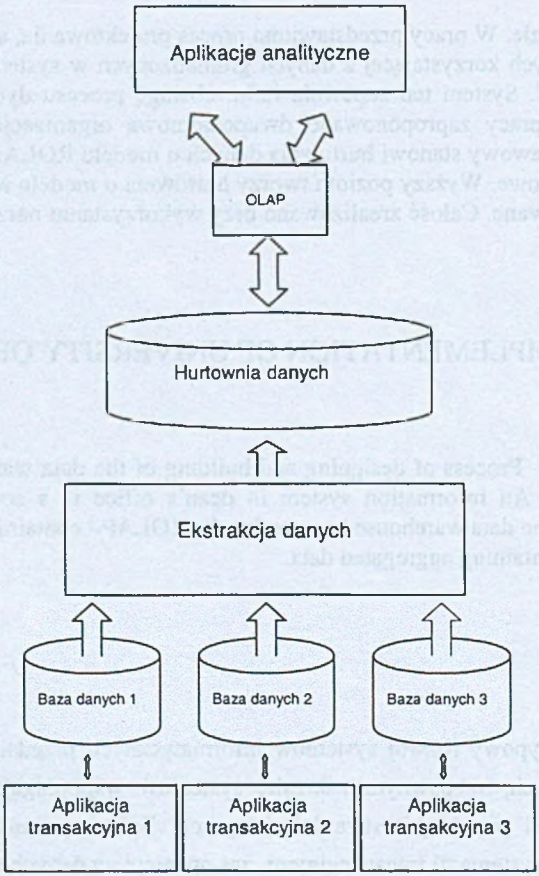
Summary. Process of designing and building of the data warehouse is presented in the paper. An information system in dean's office is a source of data to the warchouse. The data warehouse has two levels: ROLAP– containing detailed data and MOLAP – containing aggregated data.

1. Wstęp

Dotychczasowy typowy rozwój systemów informatycznych przedsiębiorstw polegał na tworzeniu niezależnych, w pewnym zakresie, systemów wspomagających poszczególne dziedziny działalności przedsiębiorstwa lub instytucji. Takie systemy oparte na bazach danych są nazywane systemami transakcyjnymi, zaś operacje na danych w tych systemach są określane jako transakcyjne przetwarzanie na bieżąco (ang. On–Line Transactional Processing – OLTP).

Rozwijane w ostatnich latach systemy wspomagania podejmowania decyzji wymagają globalnych analiz danych pochodzących z wielu lokalnych baz danych. Jak pokazały doświadczenia, stosowane dotychczas mechanizmy integracji baz danych nie zapewniają wystarczająco efektywnego dostępu do danych niezbędnego do wykonania takich analiz. Ponadto dane transakcyjne okresowo mogą być niedostępne na skutek blokad związanych z wykonywaniem transakcji.

Poszukiwanie sposobu pokonania tych ograniczeń doprowadziło do koncepcji hurtowni danych [7] jako centralnej składnicy kopii danych zawartych w bazach danych istniejących systemów informacyjnych przedsiębiorstwa. Podkreślić należy, że hurtownia danych nie jest celem, a jedynie etapem budowy systemu wspomagania decyzji. Ogólną strukturę takiego systemu prezentuje rys.1.



Rys. 1. Struktura systemu z hurtownią danych
Fig. 1. The structure of a system with a data warehouse

Dane gromadzone w lokalnych bazach danych systemów transakcyjnych są, w procesie ekstrakcji, przenoszone do hurtowni danych. Przetwarzanie danych w hurtowni na potrzeby wspomagania decyzji zyskało w literaturze ([6]) określenie „analitycznego przetwarzania na bieżąco" (ang. On-Line Analytical Processing – OLAP).

Podstawą takiego przetwarzania analitycznego jest model wielowymiarowej bazy danych. Zawartość bazy danych w omawianym modelu jest przedstawiona w postaci *wielowymiarowej macierzy* (hipersześcianu). Elementami macierzy (nazywanymi faktami lub miarami) są wielkości szczególnie istotne dla zarządzania przedsiębiorstwem (np. wartość/wielkość sprzedaży, liczba klientów, wielkość strat).

Wymiary macierzy są wielkościami opisującymi fakty, wyznaczającymi kierunki ich potencjalnych analiz (czas, struktura organizacyjna przedsiębiorstwa, asortyment towarów itp.). Wymiary mogą mieć złożoną, powiązaną wewnętrznie strukturę (np. wymiar czasu składa się z dni, miesięcy, kwartałów, lat). Takie elementy wymiarów tworzą zwykle hierarchię.

Forma danych w hurtowni w postaci wielowymiarowej macierzy jest bardzo wygodna, szczególnie dla analiz wymagających prezentacji danych zagregatyzowanych wzdłuż hierarchii wymiarów (np. sumaryczna wielkość sprzedaży w dniach, miesiącach, kwartałach itd.). Wielowymiarowy model bazy danych jest implementowany programowo w dwóch zasadniczych wariantach:

MOLAP (Multidimensional OLAP). W tym wariancie dane są przechowywane w hurtowni w strukturach wspomnianych wielowymiarowych macierzy. Oprogramowanie realizujące przetwarzanie analityczne (serwer OLAP) działa wprost na takich macierzach. Zapewnia to dużą szybkość przetwarzania, ale istniejące systemy tego typu mają ograniczoną skalowalność.

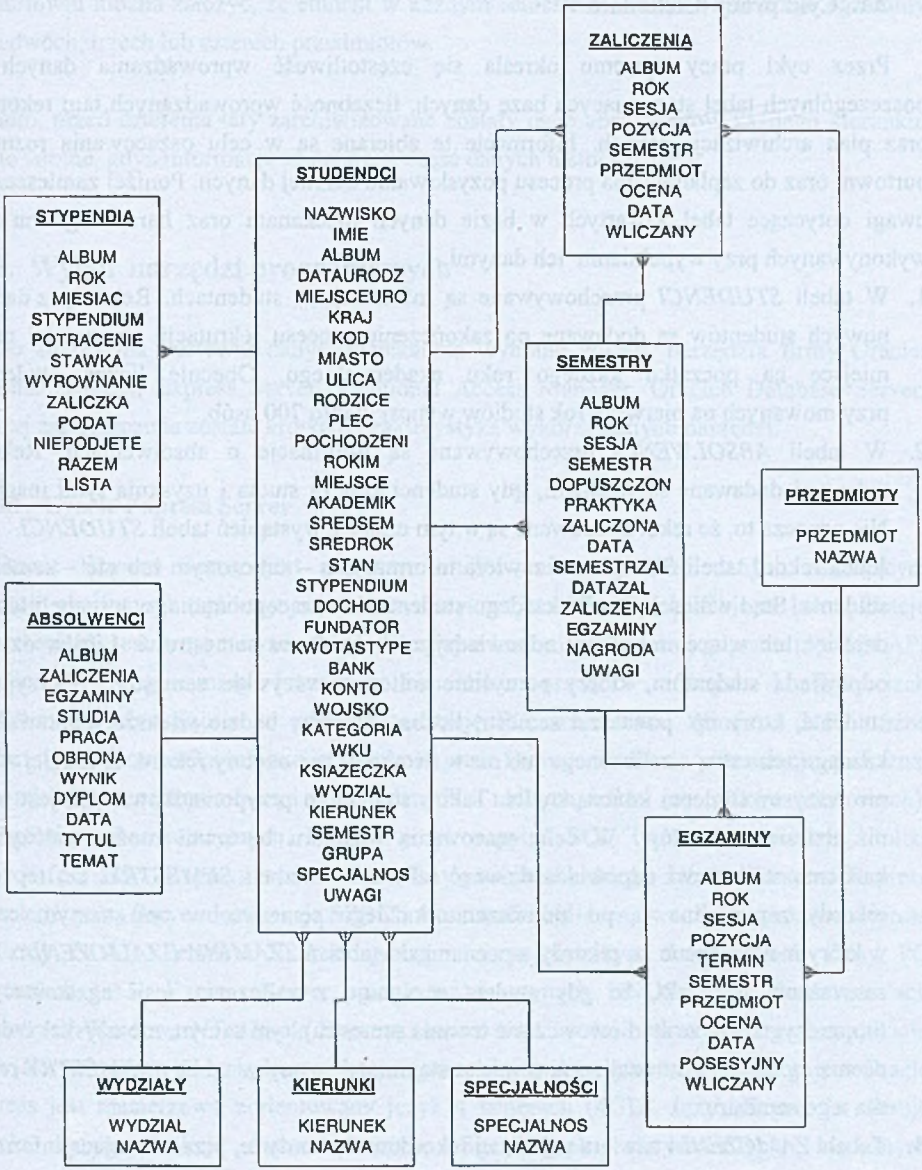
ROLAP (Relational OLAP). W tym rozwiązaniu hurtownię danych buduje się przy wykorzystaniu tradycyjnego systemu zarządzania relacyjną bazą danych. Przekształcenie modelu relacyjnego w model wielowymiarowy (udostępniany użytkownikowi) jest zadaniem serwera OLAP. Dla uzyskania odpowiedniej efektywności przetwarzania analitycznego (jednak wyraźnie niższej niż w systemach MOLAP) jest stosowana specyficzna organizacja tablic relacyjnej hurtowni danych. Zwykle jest stosowany tzw. *model gwiazdy,* z centralną tablicą faktów i promieniście otaczającymi ją tablicami wymiarów. Normalizacja tablic wymiarów prowadzi do tzw. *modelu płatka śniegu,* z hierarchicznie rozwiniętą strukturą tablic wymiarów.

Systemy ROLAP dominują na rynku narzędzi do hurtowni danych, zapewniają bowiem wysoką skalowalność i inne zalety systemów zarządzania relacyjnymi bazami danych.

2. Struktura bazy danych dziekanatu Wydziału Automatyki, Elektroniki i Informatyki

Baza danych systemu DZIEKANAT stanowi źródło danych dla projektowanej hurtowni. Tabele składające się na bazę danych tego systemu można podzielić na dwie grupy. Do pierwszej z nich należą te, które pełnią rolę słowników, zawierają stosunkowo mało rekordów i są bardzo rzadko aktualizowane. Obejmuje ona następujące tabele: *WYDZIAŁY, KIERUNKI, SPECJALNOŚCI, PRZEDMIOTY*. Drugą grupę stanowią tabele, których zawartość zmienia się regularnie w czasie. Do nich z kolei można zaliczyć tabele: *STUDENCI, ABSOLWENCI, SEMESTRY, EGZAMINY, ZALICZENIA* oraz *STYPENDIA*. Diagram encji bazy danych dziekanatu przedstawia rysunek 2.

Na bazę danych dziekanatu składa się siedem instancji (wystąpień) wszystkich wymienionych tabel. Cztery instancje zawierają dane czterech kierunków (jedna instancja zawiera dane jednego kierunku): „Informatyki", „Elektroniki", „Automatyki i Robotyki" oraz „Makrokierunku". Trzy pozostałe instancje zawierają dane studentów wyżej wymienionych kierunków (z wyjątkiem „Makrokierunku") studiów wieczorowych.



Rys. 2. Schemat bazy danych systemu DZIEKANAT
Fig. 2. The database schema of the system DEANERY

2.1. Cykl pracy dziekanatu

Przez cykl pracy systemu określa się częstotliwość wprowadzania danych do poszczególnych tabel stanowiących bazę danych, liczebność wprowadzanych tam rekordów oraz plan archiwizacji danych. Informacje te zbierane są w celu oszacowania rozmiaru hurtowni oraz do zaplanowania procesu pozyskiwania dla niej danych. Poniżej zamieszczono uwagi dotyczące tabel zawartych w bazie danych dziekanatu oraz harmonogramu prac wykonywanych przy wypełnianiu ich danymi.

1. W tabeli *STUDENCI* przechowywane są informacje o studentach. Rekordy z danymi nowych studentów są dodawane po zakończeniu procesu rekrutacji; najczęściej ma to miejsce na początku każdego roku akademickiego. Obecnie liczba studentów przyjmowanych na pierwszy rok studiów wynosi około 700 osób.
2. W tabeli *ABSOLWENCI* przechowywane są informacje o absolwentach. Rekordy z danymi dodawane są w chwili, gdy studenci kończą studia i uzyskują tytuł magistra. Nie oznacza to, że rekordy usuwane są w tym czasie z wystąpień tabeli *STUDENCI*.
3. Jeden rekord tabeli *SEMESTRY* zawiera informacje o – zaliczonym lub nie – semestrze studenta. Stąd wniosek, że dla każdego studenta, kończącego studia, w tej tabeli istnieje dziesięć lub więcej rekordów odpowiadających każdemu semestrowi. Liczba dziesięć odpowiada studentom, którzy pomyślnie zaliczają wszystkie semestry. W przypadku studenta, który np. powtarzał semestr, liczba rekordów będzie większa, ponieważ dla każdego semestru – zaliczonego lub nie – tworzony jest osobny rekord. Z drugiej strony nie wszyscy studenci kończą studia. Takim studentom przyporządkowanych jest mniej niż dziesięć rekordów. W celu szacowania rozmiaru hurtowni można założyć, że każdemu studentowi odpowiada dziesięć rekordów w tabeli *SEMESTRY*. Do tej tabeli rekordy zapisywane są po zakończeniu każdego semestru, w tym samym czasie, w którym wpisywane są rekordy z ocenami do tabel *EGZAMINY* i *ZALICZENIA*. Wart zauważenia jest fakt, że gdy student rezygnuje z zaliczania sesji egzaminacyjnej (np. zrezygnował ze studiów w czasie trwania semestru), tym samym nie uzyskał żadnych ocen z egzaminów lub zaliczeń, to nie zostanie utworzony w tablicy *SEMESTRY* rekord dla tego semestru.
4. Tabela *ZALICZENIA* zawiera najwięcej rekordów. Rekordy te, przechowujące informacje o zaliczanym przedmiocie (łącznie z oceną), wpisywane są po zakończeniu semestru. W celu szacowania rozmiaru hurtowni można założyć, że student w każdym semestrze akademickim zalicza około dziesięciu przedmiotów.
5. Liczba rekordów tabeli *EGZAMINY*, zawierającej informacje o egzaminie, jest mniejsza tylko od liczby rekordów tabeli *ZALICZENIA*. Dla każdego terminu – zdanego lub nie zdanego egzaminu – tworzony jest osobny rekord. Dane wpisywane są tak jak w poprzednich przypadkach po zakończeniu semestru. W celu szacowania rozmiaru

hurtowni można założyć, że student w każdym semestrze akademickim zdaje egzaminy z dwóch, trzech lub czterech przedmiotów.

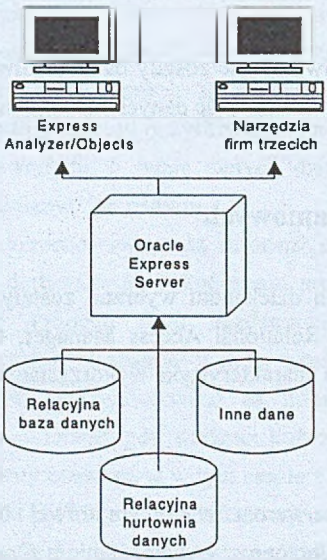
Ponadto, przed czterema laty zarchiwizowane zostały dane absolwentów każdego kierunku. Jest to istotne, gdyż informacje te stanowią część danych historycznych.

3. Wybór narzędzi programowych

Do stworzenia hurtowni danych dziekanatu wybrane zostały narzędzia firmy Oracle: Personal Express, Express Server, Relational Access Manager, Oracle8 Database Server. Poniżej zamieszczona została krótka charakterystyka wykorzystanych narzędzi.

3.1. Oracle Express Server

Oracle Express Server jest serwerem wielowymiarowej bazy danych zasilającym aplikacje analityczne (aplikacje stworzone z wykorzystaniem Oracle Objects lub aplikacje firm trzecich). Jest przygotowany do współpracy z innymi aplikacjami poprzez interfejs API (Application Programming Interface). Umożliwia pracę w różnych konfiguracjach klient/serwer, włączając w to pojedyncze stacje PC, środowisko Windows NT i różne platformy sprzętowe. Oprócz analiz Oracle Express Server umożliwia prezentację graficzną, komunikację, zarządzanie bazą i odczyt danych z różnych źródeł (np. plików tekstowych), szczególnie z baz relacyjnych. Poprzez język zapytań SQL i programy czytające pliki różnych formatów, Express może być zasilany prawie z każdego źródła. Personal Express i Oracle Express Server dostarczają SZBD i możliwości obliczeniowe dla produktów Express z rodziny OLAP. Personal Express pracuje na pojedynczych komputerach klasy PC i w sieciach LAN, Express Server pracuje na maszynach UNIX'owych, mainframe'ach i serwerach Windows NT. Oracle Express zawiera możliwości budowy i rozwoju aplikacji, takie jak: ładowanie danych, komunikacja i przechowywanie struktur. Środowiskiem Oracle Express jest macierzowo zorientowany język 4 generacji (4GL). Język Express'a oferuje szeroką gamę funkcji matematycznych, finansowych, statystycznych, logicznych i znakowych. Język 4GL umożliwia korzystanie z procedur przechowywanych w bazie oraz funkcji definiowanych przez użytkownika. 4GL obsługuje ponadto wielowymiarowy model danych. Rysunek 3 prezentuje architekturę systemu z hurtownią danych z serwerem Oracle Express Server.



Rys. 3. Architektura systemu z hurtownią danych zbudowana w oparciu o Oracle Express Server

Fig. 3. The architecture of the system with data warehouse built with usage of the Oracle Express Server

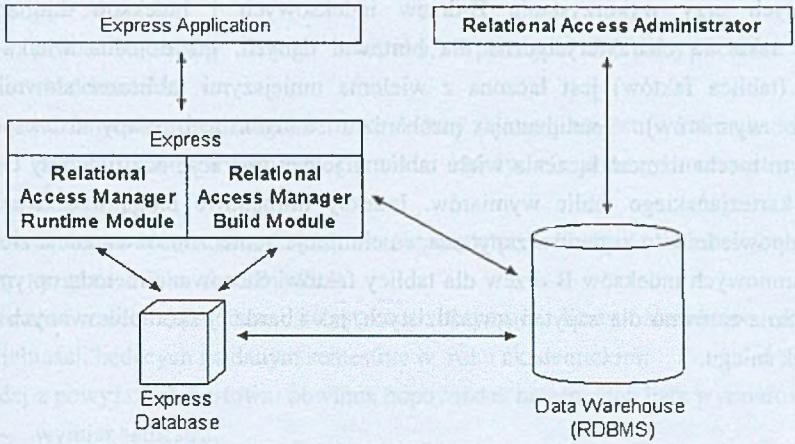
3.2. Relational Access Manager

Do łączności z bazami relacyjnymi wykorzystywane jest narzędzie Relational Access Manager (RAM). Łączy ono aplikacje Express'a z danymi znajdującymi się w relacyjnej hurtowni danych poprzez odpowiednie mapowanie tabel do obiektów wielowymiarowej bazy danych. Dzięki możliwości współpracy RAM'u z pakietem Express Objects dane zgromadzone w relacyjnej bazie danych mogą być przetwarzane i wyświetlane z użyciem określeń bazy wielowymiarowej bez konieczności pisania jakichkolwiek interfejsów.

Moduł RAM składa się z trzech następujących części:

- Relational Access Administrator – służy do definiowania modelu danych i skonfigurowania sposobu dostępu do relacyjnej bazy danych.
- Build module – zawiera program, który służy do zbudowania nowej bazy danych i aktualizowania istniejącej.
- Runtime module – zawiera narzędzia do pobierania danych z SZRBD.

Wymienione komponenty pozwalają zdefiniować model danych, który koresponduje z modelem w relacyjnej hurtowni danych oraz aktualizować go tak, aby odzwierciedlał zmiany w niej zachodzące. Rysunek 4 prezentuje budowę modułu RAM oraz jego związek z aplikacjami Express'a i hurtownią danych.



Rys. 4. Struktura modułu RAM
Fig. 4. The structure of the module RAM

3.3. Serwer bazy danych Oracle8

Hurtownie danych różnią się od systemów przetwarzania transakcji. Charakterystyczne dla hurtowni danych są niestandardowe, złożone zapytania obejmujące znaczne ilości danych. Dla wsparcia tych działań w bazie Oracle8 zastosowano liczne mechanizmy, w tym wydajny mechanizm kosztowej optymalizacji zapytań oraz skalowalną architekturę, które zapewniają właściwe wykorzystanie platformy sprzętowej. Oracle8 korzysta z licznych metod indeksowania i łączenia rekordów tak, aby szybko dostarczać wyniki zapytań. Złożone techniki kompresji wbudowane w serwer Oracle8 zapewniają efektywne przechowywanie i wykorzystanie indeksów bitmapowych.

Wysoką wydajność łączeń zapewnia mechanizm funkcji mieszającej, który jest stosowany w przypadku takich zapytań, w których nie mogą być wykorzystane indeksy. Takie nieprzewidywalne zapytania najczęściej występują właśnie w hurtowniach danych. Łączenia mieszające eliminują konieczność uprzedniego sortowania danych i działają w oparciu o tworzoną w pamięci serwera tablicę mieszającą. Wprowadzony w Oracle8 sposób deklaratywnego partycjonowania tablic ułatwia zarządzanie wielkimi bazami danych oraz zwiększa prędkość i zmniejsza koszt wykonywania niektórych zapytań.

Liczne techniki optymalizacji zapytań SQL wbudowane w serwer Oracle8 są całkowicie niewidoczne dla użytkowników baz danych. Optymalizator kosztowy dynamicznie dobiera najwydajniejszą dla danego zapytania ścieżkę dostępu i metody łączenia w oparciu o posiadane statystyki, takie jak informacje o rozmiarze i zawartości tablic.

W serwer Oracle8 wbudowano ponadto wydajny mechanizm optymalizacji zapytań gwiaździstych przy wykorzystaniu B-drzew indeksowych i indeksów bitmapowych. Zapytania takie są charakterystyczne dla hurtowni danych, gdzie jedna wielka tablica z danymi (tablica faktów) jest łączona z wieloma mniejszymi tablicami słownikowymi (tablicami wymiarów). Inteligentny mechanizm transformacji zapytań współdziała z wydajnym mechanizmem łączenia wielu tablic w jednej operacji, bez potrzeby tworzenia iloczynu kartezjańskiego tablic wymiarów. Indeksy bitmapowe mogą być dynamicznie łączone odpowiednio do kryteriów zapytania, co eliminuje konieczność tworzenia złożonych, wielokolumnowych indeksów B-drzew dla tablicy faktów. Stosowana metoda optymalizacji jest skuteczna zarówno dla zapytań gwiaździstych, jak i bardziej skomplikowanych zapytań typu płatek śniegu.

4. Projekt hurtowni dziekanatu

Dla hurtowni danych dziekanatu przyjęto organizację dwupoziomową. Poziom podstawowy tworzy hurtownia zbudowana w oparciu o system zarządzania relacyjną bazą danych (Oracle8), czyli hurtownia modelu ROLAP. Hurtownia ta zawiera tylko dane szczegółowe, pozyskiwane z bazy danych DZIEKANAT.

Wyższy poziom tworzy hurtownia zbudowana w oparciu o model wielowymiarowej bazy danych, czyli hurtownia o modelu MOLAP. Hurtownię tę stworzono z wykorzystaniem systemu Oracle Express. Zawiera ona dane zagregatyzowane, wyznaczone na podstawie danych szczegółowych z hurtowni ROLAP.

4.1. Projekt relacyjnej hurtowni danych

Stworzenie relacyjnej hurtowni danych dziekanatu zawierającej kompleksową informację o studentach uzyskano przez zaprojektowanie wielu dziedzinowych hurtowni danych. Każda dziedzinowa hurtownia zaprojektowana została w postaci gwiazdy oraz zawiera dane wyznaczające zakres informacji, do których można kierować zapytania. Ostateczna, globalna hurtownia danych powstała poprzez połączenie wszystkich dziedzinowych hurtowni w jeden schemat globalny. Połączenia dokonano z wykorzystaniem ujednoliconych tabel wymiarów.

Na globalną hurtownię składają się następujące dziedzinowe hurtownie danych:

1. Hurtownia zawierająca oceny z egzaminów uzyskane przez studentów.

2. Hurtownia zawierająca oceny z zaliczeń uzyskane przez studentów.
3. Hurtownia zawierająca kwoty stypendiów danego rodzaju, uzyskanych przez poszczególnych studentów w latach akademickich.
4. Hurtownia zawierająca dane statystyczne dotyczące liczby studentów poszczególnych kierunków, będących na danym semestrze w roku akademickim. Obejmuje ona następujące fakty:
   - liczba studentów danego kierunku, wpisanych na dany semestr w roku akademickim,
   - liczba studentów danego kierunku, którzy zaliczyli semestr w roku akademickim,
   - liczba studentów danego kierunku, którzy nie zaliczyli semestru w roku akademickim,
   - liczba studentów danego kierunku, którzy otrzymali wpis warunkowy na następny semestr w roku akademickim.
5. Hurtownia zawierająca dane statystyczne dotyczące liczby studentów określonych specjalności, będących na danym semestrze w roku akademickim.

Każdej z powyższych hurtowni powinna odpowiadać następująca lista wymiarów:

- wymiar semestru,
- wymiar roku akademickiego,
- wymiar studenta,
- wymiar wydziału,
- wymiar kierunku,
- wymiar specjalności,
- wymiar przedmiotu,
- wymiar terminu egzaminu,
- wymiar rodzaju stypendium.

Aby zobrazować zależności pomiędzy hurtowniami dziedzinowymi i wymiarami, posłużono się tabelą (rys. 5). Znacznik w komórce będącej na przecięciu odpowiedniego wiersza i kolumny oznacza, że dana hurtownia dziedzinowa może być wymiarowana przez odpowiedni wymiar.

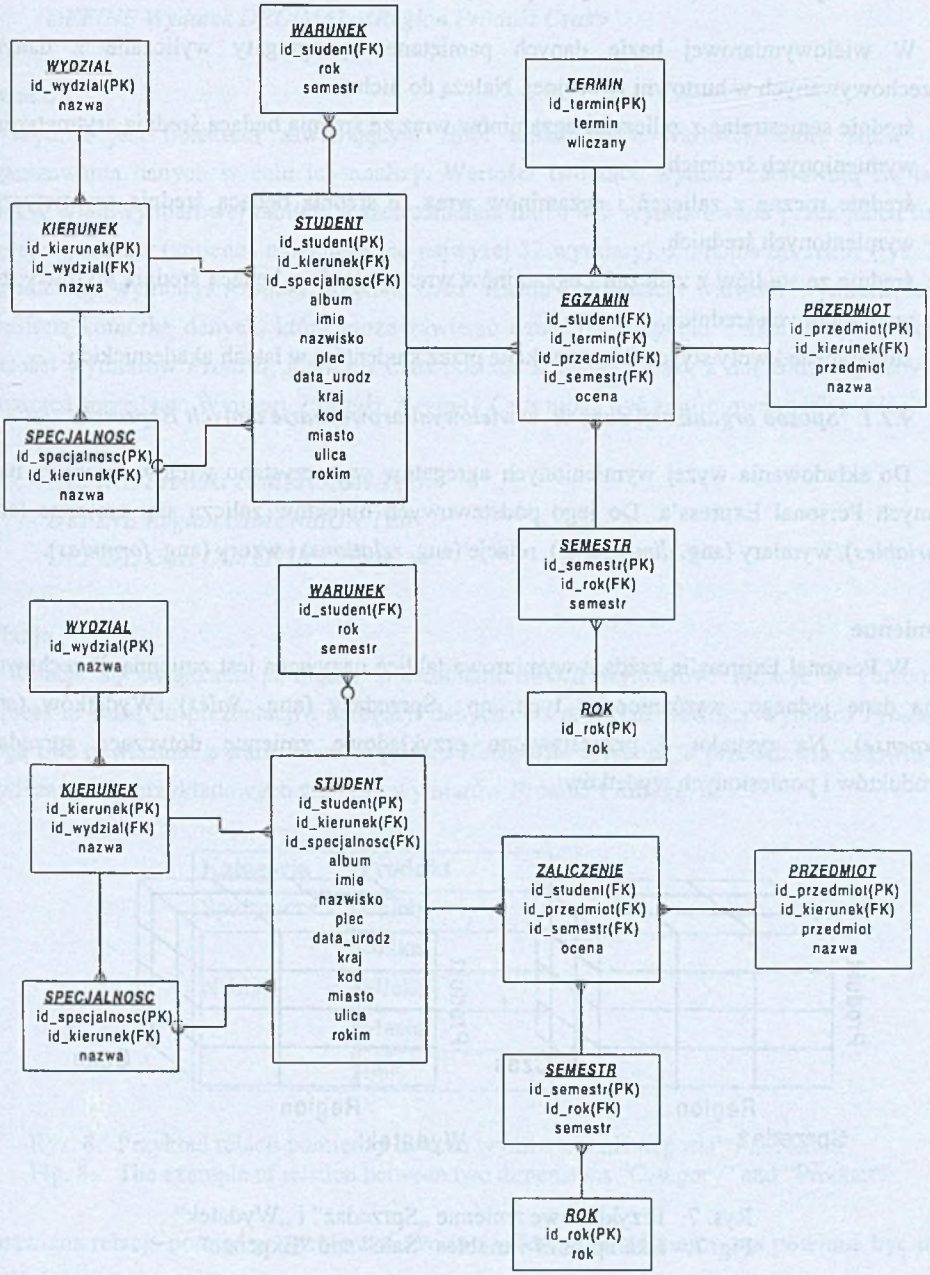
	Wydział	Kierunek	Specjalność	Student	Rok	Semestr	Przedmiot	Termin egz.	Rodzaj styp.
Oceny z egzaminów	√	√	√	√	√	√	√	√	
Oceny z zaliczeń	√	√	√	√	√	√	√		
Kwoty stypendiów	√	√	√	√	√				√
Liczba studentów na kierunku	√	√			√	√			
Liczba studentów na specjalności	√	√	√		√	√			

Rys. 5. Macierz zależności między wymiarami i hurtowniami dziedzinowymi
Fig. 5. The matrix of relations between dimensions and data marts

Biorąc pod uwagę zależności istniejące między wymiarami na końcowy schemat hurtowni wybrano postać płatka śniegu. W ten sposób uzyskuje się strukturę jawnie ukazującą rzeczywistą, hierarchiczną strukturę wymiarów:

- wymiar studenta: *Wydział – Kierunek – Student*,
  Wydział – Kierunek – Specjalność – Student,
- wymiar semestru: *Rok – Semestr*,
- wymiar przedmiotu: *Kierunek – Przedmiot*.

Wybrany fragment schematu globalnej hurtowni danych powstałej po normalizacji i połączeniu wymiarów przedstawia Rysunek 6. Symbolami PK, FK wyróżniono odpowiednio klucze główne oraz klucze obce tabel.



Rys. 6. Fragment globalnego schematu hurtowni danych
Fig. 6. The part of the global data warehouse schema

4.2. Projekt wielowymiarowej bazy danych

W wielowymiarowej bazie danych pamiętane są agregaty wyliczane z danych przechowywanych w hurtowni relacyjnej. Należą do nich:

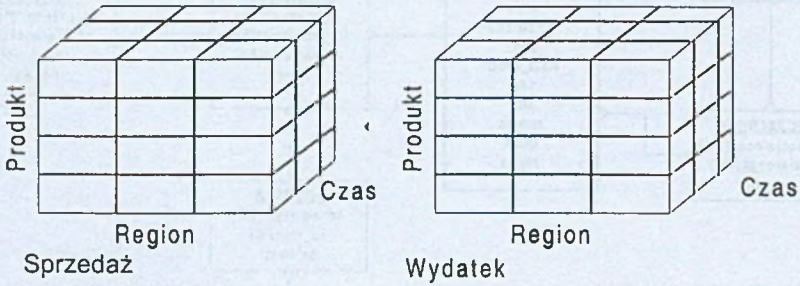
1) średnie semestralne z zaliczeń i egzaminów wraz ze średnią będącą średnią arytmetyczną wymienionych średnich,
2) średnie roczne z zaliczeń i egzaminów wraz ze średnią będącą średnią arytmetyczną wymienionych średnich,
3) średnie ze studiów z zaliczeń i egzaminów wraz ze średnią będącą średnią arytmetyczną wymienionych średnich,
4) sumaryczne kwoty stypendiów uzyskane przez studentów w latach akademickich.

4.2.1. Sposób organizacji danych w wielowymiarowej bazie danych Express'a

Do składowania wyżej wymienionych agregatów wykorzystano wielowymiarową bazę danych Personal Express'a. Do jego podstawowych obiektów zalicza się: zmienne (ang. *variables*), wymiary (ang. *dimensions*), relacje (ang. *relations*) i wzory (ang. *formulas*).

Zmienne

W Personal Express'ie każda n-wymiarowa tablica nazywana jest zmienną. Przechowuje ona dane jednego, wyróżnionego typu, np.: Sprzedaży (ang. *Sales*), Wydatków (ang. *Expense*). Na rysunku 7 przedstawiono przykładowe zmienne dotyczące sprzedaży produktów i poniesionych wydatków.



Rys. 7. Przykładowe zmienne „Sprzedaż" i „Wydatek"
Fig. 7. Examples of variables "Sale" and "Expence"

Aby zdefiniować zmienną, należy podać jej nazwę, typ i wymiary. Definicje zmiennych *Sprzedaż* i *Wydatek* będą następujące:

DEFINE Sprzedaż DECIMAL <Region Produkt Czas>
DEFINE Wydatek DECIMAL <Region Produkt Czas>

Wymiary

Wymiar jest obiektem zawierającym zbiór unikatowych wartości, które służą do organizowania danych w celu ich analizy. Wartości tworzące wymiar zachowują się jak indeksy wielowymiarowej tablicy. Każda zmienna może być wymiarowana przez jeden lub więcej wymiarów (zmienna może mieć co najwyżej 32 wymiary). Zmienna *Sprzedaż* (rys. 7) posiada trzy wymiary: *Produkt*, *Region*, *Czas*. Każda kombinacja wartości wymiaru daje określoną komórkę danych, która może zawierać dane. Na przykład unikalna kombinacja wartości wymiarów *Produkt*, *Region* i *Czas* posiada korespondującą z nią komórkę danych dotyczącą sprzedaży. Wymiary *Produkt*, *Region* i *Czas* mogą być zdefiniowane jak poniżej:

DEFINE Produkt DIMENSION Text
DEFINE Region DIMENSION Text
DEFINE Czas DIMENSION Text

Relacje

Relacje są związkami pomiędzy wartościami dwóch wymiarów. Relacje w Personal Express'ie służą do prezentacji i agregacji danych. Na przykład wartości wymiaru *Produkt* mogą być powiązane z wartościami wymiaru *Kategoria*. Rysunek 8 przedstawia omawiane powiązanie dla przykładowych wartości wymiarów *Produkt* i *Kategoria*.

Kategoria	Produkt
Spożywcze	Chleb
	Bułka
Nabiał	Mleko
	Masło
	Jaja

Rys. 8. Przykład relacji pomiędzy dwoma wymiarami „Kategoria" i „Produkt"
Fig. 8. The example of relation between two dimensions "Category" and "Product"

Omawiana relacja pomiędzy wymiarami *Produkt* i *Kategoria* zdefiniowana powinna być jak poniżej:

DEFINE Kategoria.Produkt RELATION Kategoria <Produkt>

Wzory

Używając Personal Express'a nie jesteśmy ograniczeni tylko do prezentacji danych pamiętanych w zmiennych. Tworząc wzory, można uzyskać nowe dane. Na przykład zysk ze sprzedaży produktów można uzyskać odejmując od wartości zmiennej *Sprzedaż* wartości zmiennej *Wydatek* we wzorze zdefiniowanym jak poniżej:

DEFINE Zysk FORMULA Sprzedaż – Wydatek

4.2.2. Definicja wielowymiarowej bazy danych

Średnie semestralne z zaliczeń i egzaminów wymiarowane są przez wymiary *STUDENT, SEMESTR*. Każdy student przypisany jest do jakiegoś kierunku i wydziału, więc niezbędne są także wymiary *KIERUNEK* i *WYDZIAŁ*. Ponieważ pomiędzy wymiarami studenta, kierunku i wydziału, jak również w przypadku wymiarów *SEMESTR* i *ROK* istnieje zależność, w wielowymiarowej bazie danych powstaną dwie hierarchie:

- studenta, składająca się z wymiaru studenta, kierunku i wydziału,
- semestru, składająca się wymiaru semestru i roku akademickiego.

Wartość będącą średnią arytmetyczną średnich semestralnych z zaliczeń i egzaminów można uzyskać definiując wzór wymiarowany tak samo jak średnie. Opisane obiekty w Personal Express'ie definiuje się następująco:

- definicje wymiarów: *STUDENT, KIERUNEK, WYDZIAL, SEMESTR, ROK*:

  DEFINE Student DIMENSION Text
  DEFINE Kierunek DIMENSION Text
  DEFINE Wydzial DIMENSION Text
  DEFINE Semestr DIMENSION Text
  DEFINE Rok DIMENSION Text
- definicja relacji łączących wymiary *STUDENT, KIERUNEK, WYDZIAL*:

  DEFINE K.S RELATION Kierunek <Student>
  DEFINE W.K RELATION Wydzial <Kierunek>
- definicja relacji łączących wymiary *Semestr* i *Rok*:

  DEFINE R.S RELATION Rok <Semestr>
- definicje zmiennych: *SEMZAL* i *SEMEGZ*, oznaczające kolejno średnią semestralną z zaliczeń i egzaminów:

  DEFINE Semzal DECIMAL <Semestr Student>
  DEFINE Semegz DECIMAL <Semestr Student>
- wzór *SEMZALEGZ*, będący średnią arytmetyczną zmiennych *SEMZAL* i *SEMEGZ*:

  DEFINE Semzalegz FORMULA (Semzal+Semegz)/2

Średnie roczne z zaliczeń i egzaminów wymiarowane są przez wymiar *STUDENT, ROK*. Zmienne mogą współdzielić wymiary, dlatego ponowna definicja wymiarów nie jest

potrzebna. Wartość będącą średnią arytmetyczną średnich rocznych z zaliczeń i egzaminów można uzyskać definiując wzór wymiarowany tak samo jak powyższe średnie. Definicja opisanych obiektów w Personal Express'ie będzie następująca:

- definicje zmiennych *ROKZAL* i *ROKEGZ* oznaczają odpowiednio średnią roczną z zaliczeń i egzaminów:

  DEFINE Rokzal DECIMAL <Rok Student>

  DEFINE Rokegz DECIMAL <Rok Student>
- wzór *ROKZALEGZ*, będąca średnią arytmetyczną zmiennych *ROKZAL* i *ROKEGZ*:

  DEFINE Rokzalegz FORMULA (Rokzal+Rokegz)/2

Średnie ze studiów z zaliczeń i egzaminów wymiarowane są tylko przez wymiar *STUDENT*. Wartość będącą średnią arytmetyczną średnich ze studiów z zaliczeń i egzaminów uzyskano poprzez stworzenie wzoru. Wymienione obiekty w Personal Express'ie definiuje się następująco:

- definicje zmiennych: *STUZAL* i *STUEGZ* oznaczają odpowiednio średnią ze studiów z zaliczeń i egzaminów:

  DEFINE Stuzal DECIMAL <Student>

  DEFINE Stuegz DECIMAL <Student>
- wzór *STUZALEGZ*, będący średnią arytmetyczną zmiennych *STUZAL* i *STUEGZ*:

  DEFINE Stuzalegz FORMULA (Stuzal+Stuegz)/2

Kwoty stypendiów uzyskane w latach akademickich wymiarowane są przez wymiar *STUDENT, RODZAJ i ROK*. Wymiar *RODZAJ* reprezentuje rodzaje stypendiów. Dodatkowo zdefiniowano wzór do przeliczania wartości kwoty podanej w „starych" złotych na PLN. Definicja opisanych obiektów w Personal Express'ie jest następująca:

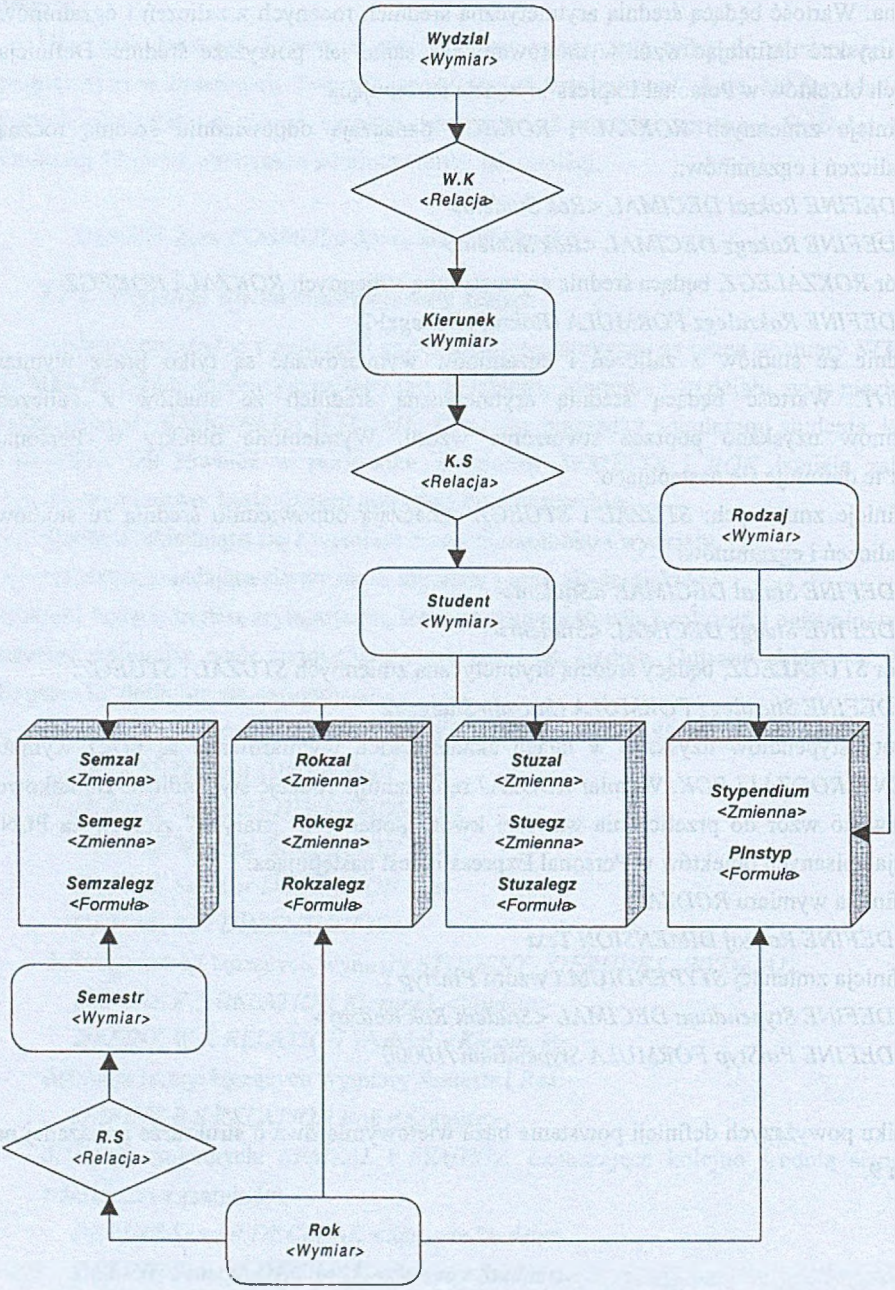
- definicja wymiaru *RODZAJ*:

  DEFINE Rodzaj DIMENSION Text
- definicja zmiennej *STYPENDIUM* i wzoru *Plnstyp* :

  DEFINE Stypendium DECIMAL <Student Rok Rodzaj>

  DEFINE PlnStyp FORMULA Stypendium /10000

W wyniku powyższych definicji powstanie baza wielowymiarowa o strukturze pokazanej na rysunku 9.



Rys. 9. Model bazy wielowymiarowej
Fig. 9. The model of the multidimensional database

5. Metody pozyskiwania danych do relacyjnej hurtowni danych

Ekstrakcja danych jest procesem przekształcania i przenoszenia informacji z baz danych transakcyjnych do hurtowni. Do wykonania tego etapu tworzenia hurtowni danych przystępuje się po zdefiniowaniu wymagań stawianych hurtowni danych, rozpoznaniu struktury systemów transakcyjnych oraz zdefiniowaniu architektury hurtowni. Zazwyczaj proces ekstrakcji składa się z dwóch etapów: etapu wstępnego ładowania danych do hurtowni (tzw. danych historycznych) oraz etapu identyfikowania i ładowania danych przyrostowych. Poniżej przedstawiony został ogólny opis realizacji procesu ekstrakcji danych, opis przekształceń danych źródłowych na dane hurtowni relacyjnej, szczegółowe omówienie algorytmów realizujących zadania procesu ekstrakcji oraz wybrane problemy powstałe w wyniku jego realizacji.

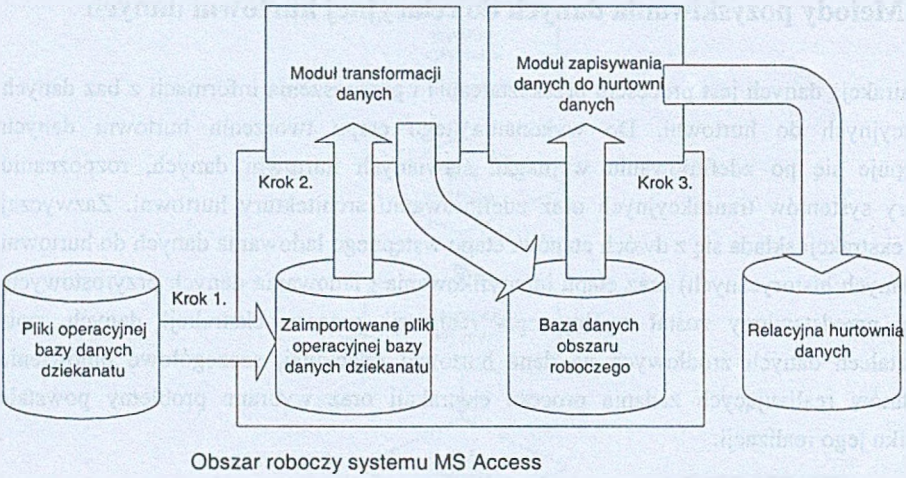
5.1. Ogólny opis realizacji procesu ekstrakcji danych

Ekstrakcję danych przeprowadzono posługując się koncepcją obszaru pośredniego (roboczego) pomiędzy transakcyjnymi systemami źródłowymi a bazą hurtowni danych. Jako reprezentację obszaru pośredniego wybrano przestrzeń roboczą programu Microsoft Access 97, do której można importować pliki w formacie *.dbf, jak również tworzyć własne struktury. Na zawartości plików w obszarze roboczym można wykonywać operacje z wykorzystaniem języka VBA (Visual Basic for Applications). W przedstawianej metodzie pozyskiwania danych do hurtowni można wyróżnić następujące kroki (rys. 10):

Krok 1. Pobieranie danych z operacyjnej bazy danych dziekanatu do obszaru roboczego Access'a za pomocą funkcji importu plików w formacie *.dbf.

Krok 2. Przekazywanie pobranych danych do modułu transformacji danych i zapisywanie przetworzonych danych w bazie danych obszaru roboczego Access'a.

Krok 3. Zapisywanie danych z bazy danych obszaru roboczego do hurtowni danych.



Rys. 10. Realizacja ekstrakcji danych z wykorzystaniem obszaru roboczego systemu MS Access
Fig. 10. The data extraction with usage of workspace of MS Access

5.2. Przetwarzanie danych w module transformacji danych

Proces przekształcania danych pochodzących z systemu obsługi bieżącej obejmuje następujące kroki:

1. Przekształcenia odzwierciedlające odpowiednie pola tabel systemu w dziekanacie na pola tabel relacyjnej hurtowni danych (z uwzględnieniem pól wyliczanych).
2. Nadanie unikalnych indeksów.
3. Konwersja typów danych.
4. Dopracowanie szczegółów.

Dla każdej tabeli wymiaru i faktów istnieje odpowiednia procedura przygotowująca zarówno dane historyczne, jak i przyrostowe oraz ładująca te dane do bazy danych obszaru roboczego. Schemat bazy danych obszaru roboczego przypomina schemat relacyjnej hurtowni danych. Do takich samych tabel, jak w hurtowni relacyjnej, dodane zostały pola pomocnicze przydatne w trakcie przetwarzania danych. Nazwy tabel bazy danych obszaru pośredniego różnią się od nazw tabel operacyjnej bazy danych lub wymiarów i faktów hurtowni przyrostkiem *_MAP*.

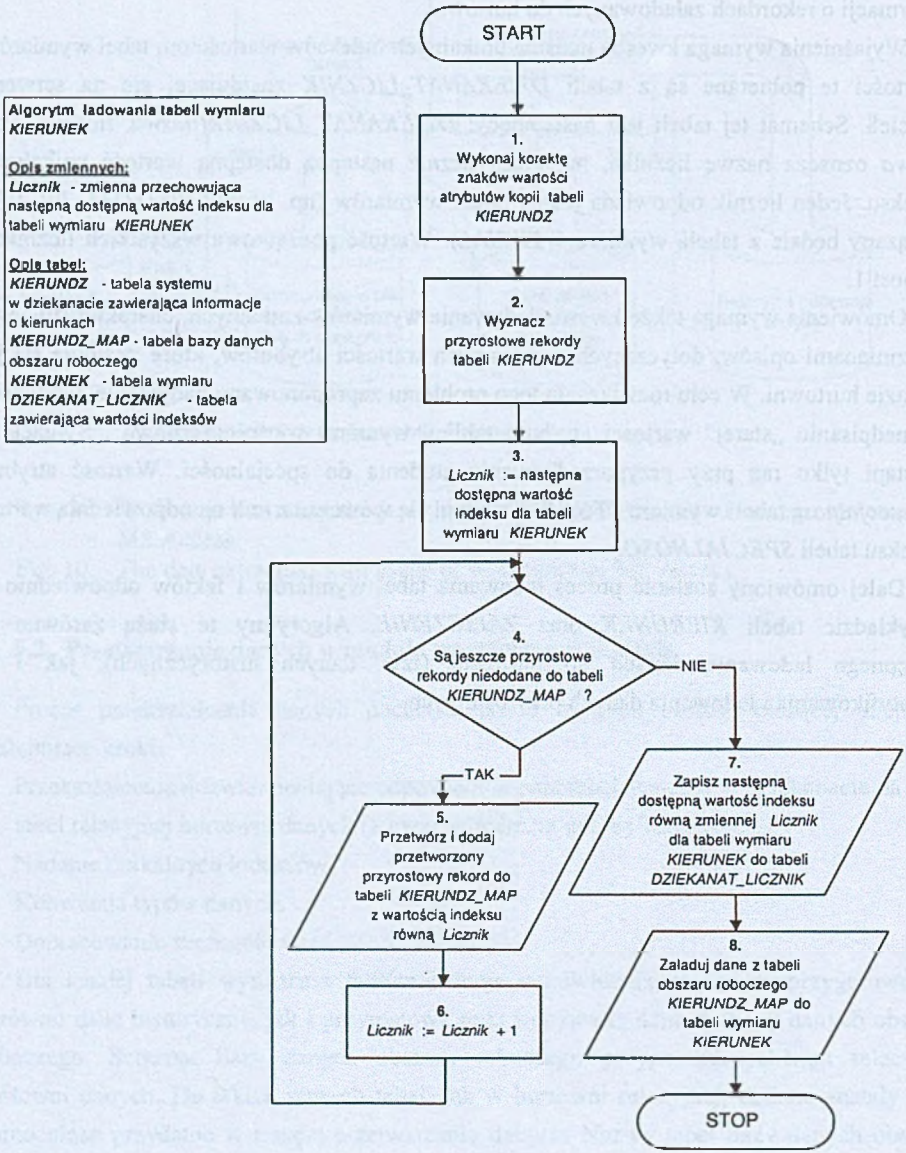
Kolejnym etapem jest ładowanie danych z obszaru roboczego do hurtowni. Podobnie jak w przypadku przetwarzania, zaproponowane zostały odpowiednie procedury, których zadaniem jest przepisanie danych z tabel obszaru roboczego Access'a do tabel hurtowni. W tym procesie każda z omawianych procedur zapisuje dodatkowo wartości indeksów

rekordów tabel z przyrostkiem *_MAP* do tabel z przyrostkiem *_ORA*, w celu utrzymywania informacji o rekordach załadowanych do hurtowni.

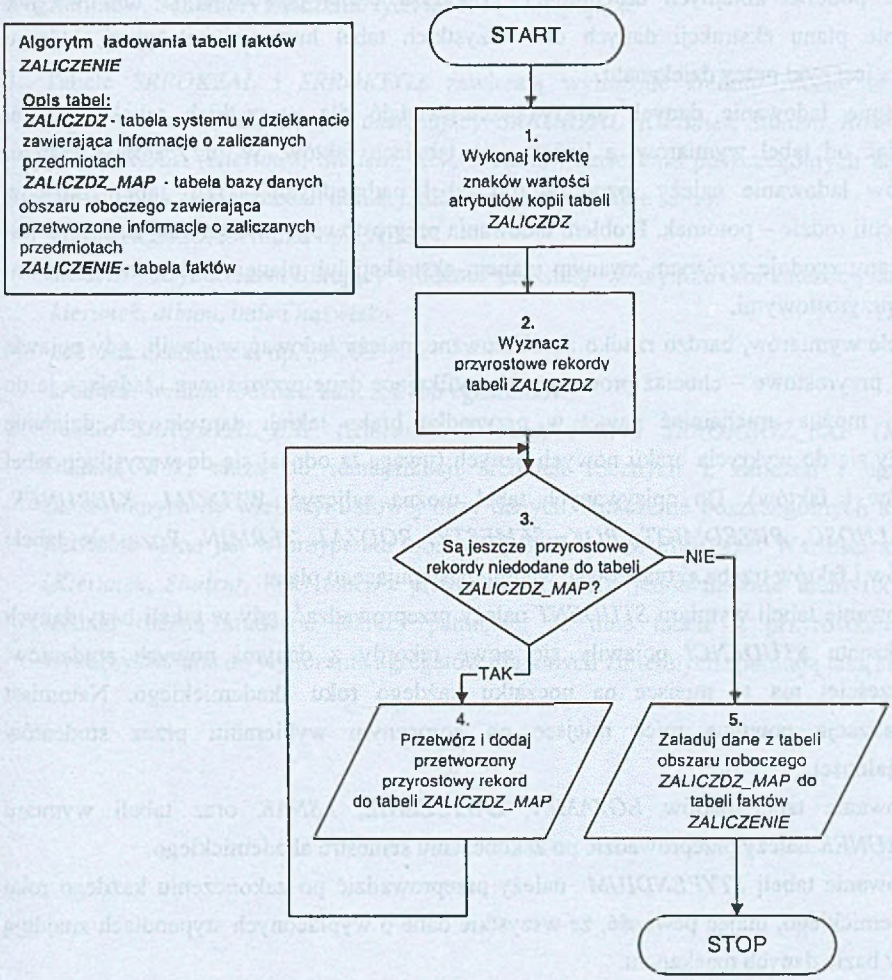
Wyjaśnienia wymaga kwestia nadania unikalnych indeksów wartościom tabel wymiarów. Wartości te pobierane są z tabeli *DZIEKANAT_LICZNIK* znajdującej się na serwerze Oracle8. Schemat tej tabeli jest następujący: *DZIEKANAT_LICZNIK(nazwa, licznik)*. Pole *nazwa* oznacza nazwę licznika, natomiast *licznik* następną dostępną wartość unikalnego indeksu. Jeden licznik odpowiada jednej tabeli wymiarów (np. licznik *WYDZIAL_LICZNIK* związany będzie z tabelą wymiaru *WYDZIAL*). Wartość początkowa wszystkich liczników wynosi 1.

Omówienia wymaga także kwestia ładowania wymiarów zmiennych, charakteryzujących się zmianami opisów, dotyczących tych samych wartości atrybutów, które znajdują się już w bazie hurtowni. W celu rozwiązania tego problemu zaproponowano nadpisanie, polegające na nadpisaniu „starej" wartości atrybutu tablicy wymiaru wartością „nową". Sytuacja ta wystąpi tylko raz przy przyporządkowaniu studenta do specjalności. Wartość atrybutu *id_specjalnosc* tabeli wymiaru *STUDENT* zmieni się wówczas z *null* na odpowiednią wartość indeksu tabeli *SPECJALNOSC*.

Dalej omówiony zostanie proces ładowania tabel wymiarów i faktów odpowiednio na przykładzie tabeli *KIERUNEK* oraz *ZALICZENIE*. Algorytmy te służą zarówno do wstępnego ładowania danych do hurtowni (tzw. danych historycznych), jak i do identyfikowania i ładowania danych przyrostowych.



Rys. 11. Algorytm ładowania statycznej tabeli wymiaru KIERUNEK
Fig. 11. The algorithm of loading the static dimension table COURSE



Rys. 12. Algorytm ładowania tabeli faktów ZALICZENIE
Fig. 12. The algorithm of loading the fact table CREDIT

5.3. Plan procesu ekstrakcji danych

Przedstawione powyżej koncepcje algorytmów ładowania tabel hurtowni relacyjnej spełniają dwie funkcje: funkcję wstępnego ładowania danych do hurtowni oraz funkcję identyfikowania i ładowania danych przyrostowych. Na przykładzie algorytmu ładowania tabeli wymiaru *KIERUNEK* widać, że potrafi on pozyskać dane historyczne podczas

pierwszego uruchomienia, a następnie identyfikować dane przyrostowe i ładować je do hurtowni podczas kolejnych uruchomień. Powyższa cecha powoduje, iż możliwe jest stworzenie planu ekstrakcji danych dla wszystkich tabel hurtowni relacyjnej, którego podstawą jest cykl pracy dziekanatu.

Wstępne ładowanie danych należy przeprowadzić dla wszystkich tabel hurtowni zaczynając od tabel wymiarów, a kończąc na tabelach faktów. W przypadku hierarchii wymiarów ładowanie należy rozpocząć od tabel nadrzędnych – tzn. tabel rodziców w hierarchii rodzic – potomek. Problem ładowania przyrostowego jest bardziej złożony i jest realizowany zgodnie z planem zwanym planem ekstrakcji lub planem ładowania hurtowni danymi przyrostowymi.

Tabele wymiarów, bardzo rzadko aktualizowane, należy ładować w chwili, gdy pojawią się dane przyrostowe – chociaż procedury identyfikujące dane przyrostowe i ładujące je do hurtowni można uruchamiać nawet w przypadku braku takich danych; ich działanie ograniczy się do wykrycia braku nowych danych (uwaga ta odnosi się do wszystkich tabel wymiarów i faktów). Do opisywanych tabel można zaliczyć: *WYDZIAL*, *KIERUNEK*, *SPECJALNOSC*, *PRZEDMIOT*, *ROK*, *SEMESTR*, *RODZAJ*, *TERMIN*. Pozostałe tabele wymiarów i faktów trzeba aktualizować według następującego planu:

1. Ładowanie tabeli wymiaru *STUDENT* należy przeprowadzać, gdy w tabeli bazy danych dziekanatu *STUDENCI* pojawiły się nowe rekordy z danymi nowych studentów. Najczęściej ma to miejsce na początku każdego roku akademickiego. Natomiast aktualizacja powinna mieć miejsce po corocznym wybieraniu przez studentów specjalności.
2. Ładowanie tabel faktów *EGZAMIN*, *ZALICZENIE*, *LSNAK* oraz tabeli wymiaru *WARUNEK* należy przeprowadzić po zakończeniu semestru akademickiego.
3. Ładowanie tabeli *STYPENDIUM* należy przeprowadzić po zakończeniu każdego roku akademickiego, mając pewność, że wszystkie dane o wypłaconych stypendiach znajdują się w bazie danych dziekanatu.

6. Metody pozyskiwania danych do wielowymiarowej bazy danych

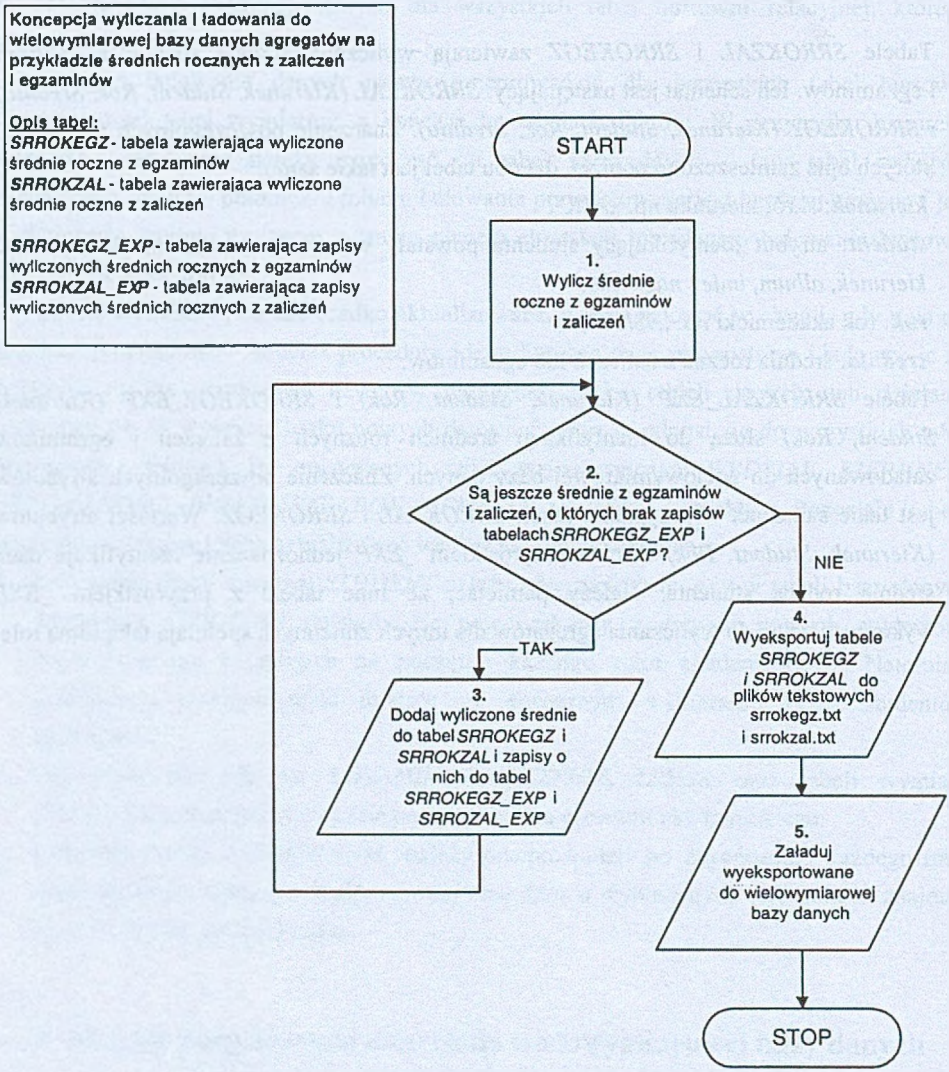
Do wyliczania agregatów wykorzystano dane hurtowni relacyjnej, procedury zaimplementowane w języku VBA oraz obszar roboczy systemu Access. Podobnie jak poprzednio omówiony zostanie tylko jeden reprezentatywny przypadek wyliczania agregatów dla zmiennych *ROKZAL* i *ROKEGZ*.

Aby uzyskać elastyczny sposób wyliczania średnich, utworzone zostały dodatkowe tabele w obszarze roboczym Access'a. Dla omawianego przykładu będą to: *SRROKZAL*,

SRROKZAL_EXP dla średnich z zaliczeń i *SRROKEGZ*, *SRROKEGZ_EXP* dla średnich z egzaminów. Schemat i znaczenie tych tabel są następujące:

1. Tabele *SRROKZAL* i *SRROKEGZ* zawierają wyliczone średnie roczne z zaliczeń i egzaminów. Ich schemat jest następujący: *SRROKZAL (Kierunek, Student, Rok, Srednia)* i *SRROKEGZ (Kierunek, Student, Rok, Srednia)*. Znaczenie poszczególnych atrybutów, których opis zamieszczono poniżej, dla obu tabel jest takie samo:
   - ***kierunek***: skrót kierunku np. „AIR",
   - ***student***: atrybut identyfikujący studenta powstały w wyniku konkatenacji atrybutów *kierunek*, *album*, *imie* i *nazwisko*,
   - ***rok***: rok akademicki np. „98/99",
   - ***srednia***: średnia roczna z zaliczeń lub egzaminów.
2. Tabele *SRROKZAL_EXP (Kierunek, Student, Rok)* i *SRROKEGZ_EXP (Kierunek, Student, Rok)* służą do identyfikacji średnich rocznych z zaliczeń i egzaminów załadowanych do wielowymiarowej bazy danych. Znaczenie poszczególnych atrybutów jest takie samo jak w przypadku tabeli *SRROKZAL* i *SRROKEGZ*. Wartości atrybutów (*Kierunek*, *Student*, *Rok)* tabel z przyrostkiem *_EXP* jednoznacznie identyfikują daną średnią roczną studenta. Należy pamiętać, że inne tabele z przyrostkiem *_EXP* wykorzystywane do wyliczania agregatów dla innych zmiennych spełniają taką samą rolę.

Algorytm wyliczania średnich i ładowania ich do wielowymiarowej hurtowni danych przedstawiony jest na rys.13.



Rys. 13. Algorytm wyliczania i ładowania agregatów do wielowymiarowej bazy danych
Fig. 13. The algorithm of aggregates calculation and loading into multidimensional database

Plan ładowania danych do bazy wielowymiarowej

Podobnie jak w przypadku ładowania danych do relacyjnej hurtowni danych należy przygotować plan wyliczania agregatów i zapisywania ich w bazie wielowymiarowej. Agregaty należy wyliczać po załadowaniu danych z systemu w dziekanacie do relacyjnej

hurtowni. Kolejność ta musi być przestrzegana, gdyż agregaty wyliczane są na podstawie danych hurtowni relacyjnej.

Procedury wyliczające agregaty zostały zaprojektowane tak, aby wyliczać tylko agregaty przyrostowe. Cecha ta powoduje, że podczas pierwszego ładowania baza wielowymiarowa wypełniana jest danymi historycznymi, a podczas kolejnych ładowań danymi przyrostowymi. Plan ładowania zmiennych powinien być następujący:

1. Zmienne *Semzal* i *Semegz* powinny być wypełniane po zakończeniu semestru akademickiego, o ile wszystkie oceny z zaliczeń i egzaminów zostały wprowadzone do relacyjnej hurtowni danych.
2. Zmienne *Rokzal*, *Rokegz*, *Stuzal*, *Stuegz* oraz *Stypendium* powinny być wypełniane po zakończeniu roku akademickiego, o ile wszystkie oceny z zaliczeń i egzaminów, roczne kwoty stypendiów zostały wprowadzone do relacyjnej hurtowni danych

7. Podsumowanie

W pracy przedstawiony został proces projektowania i implementacji hurtowni danych dziekanatu w oparciu o narzędzia firmy Oracle. Ponieważ po zakończeniu każdego semestru do hurtowni ładowane są duże ilości danych szczegółowych dotyczących zaliczeń i egzaminów każdego studenta, na miejsce przechowywania szczegółowych danych historycznych wybrano relacyjną hurtownię zbudowaną na serwerze Oracle8, cechującym się bardzo dobrymi parametrami przetwarzania takiej ilości informacji. Natomiast agregaty, wyliczane z danych przechowywanych w hurtowni relacyjnej, są zapamiętywane w wielowymiarowej bazie danych, która charakteryzuje się szybkim dostępem do danych umieszczonych w jej strukturach.

Połączenie tych dwóch hurtowni pozwoliło stworzyć bardzo efektywny system przeglądania i analizy przechowywanych informacji.

LITERATURA

1. Kimball R., Reeves L., Ross M., Thornthwaite W.: The Data Warehouse Lifecycle Toolkit. J. Willey, 1998.
2. Kimball R.: Data warehouse Toolkit. John Wiley & Sons, Inc, 1996.
3. Widom J.: Researches Problems in Data Warehousing. Proceed. of 4th Int. Conference of Information and Knowledge Management, Nov. 1995.
4. Austin D.: Poznaj Oracle8. MIKOM, Warszawa, luty 1999.
5. Sanna P. i inni: Visual Basic® dla Aplikacji 5 w zastosowaniach LT&P, Warszawa 1998.

6. Codd E.F., Codd S.B., Salley C.T.: Providing OLAP (On-Line Analytical Processing) to User-Analyst.
7. Inmon W. H.: Building the Data Warehouse. John Wiley & Sons, Inc, 1996.
8. Frączek J., Gorawski M., Kozielski S.: Modelowanie struktur wielowymiarowych w hurtowniach danych. Archiwum Informatyki Teoretycznej i Stosowanej, vol. 3, 2000.
9. Gorawski M., Frączek J.: Data Warehouse: Modelowanie danych. Software 2.0 7/99.
10. Gorawski M., Frączek J.: Data Warehouse: Analiza porównawcza MOLAP i ROLAP. Software 2.0 8/99.
11. Gorawski M., Koziatek A.: Data Warehouse: Ekstrakcja danych. Software 2.0 9/99.
12. Kiciński J.: Wielowymiarowość narzędzi OLAP. Software 2.0 5/99.
13. Gorawski M., Koziatek A.: Data Warehouse: Analiza porównawcza środowisk. Software 2.0 10/99.
14. http://www.oracle.com.pl/dokumenty/olapexprservtoenterprise.pdf.
15. http://www.oracle.com.pl/dokumenty/olapexprobjds.pdf.
16. http://www.oracle.com.pl/dokumenty/olapexpranalds.pdf.

Recenzent: Dr hab. inż. Stanisław Wołek, prof. Pol. Rzeszowskiej

Wpłynęło do Redakcji 23 lutego 2001 r.

Abstract

Process of designing and building of the data warehouse is presented in the paper. An information system in dean's office is a source of the data the warehouse. This system contains student exam results, scholarships etc. The date warehouse has two levels: ROLAP – containing detailed data and MOLAP – containing aggregated data. The basic level data warehouse (ROLAP) is a relational database. It has snowflake structure. The facts in the data warehouse are as follows: exam results, sums of scholarships, numbers of students. The dimensions are as follows: time, course, student and university structure. The upper level data warehouse (MOLAP) is the multidimensional database. It contains average values of exams' results and sums of scholarships. Procedures of data extraction from transaction database and data loading to ROLAP and MOLAP data warehouse are also presented in the paper. Project was realized using Oracle8 and Oracle Express tools.