

Mieczysław L. Owoc, Maciej Pondel

ADMINISTROWANIE HURTOWNIAMI DANYCH W ŚRODOWISKU PRZETWARZANIA GRIDOWEGO

1. Wstęp

Hurtownie danych, traktowane jako bardziej zaawansowane formy baz danych, w znacznym stopniu wzbogacają współczesne technologie informatyczne. Ich istotna rola wynika z coraz większego znaczenia funkcji analitycznych wykorzystywanych przy podejmowaniu decyzji, a zarazem z możliwości prawie nieograniczonego korzystania z zewnętrznych źródeł informacji. Administrowanie zasobami hurtowni, pomimo bardzo rozwiniętych narzędzi bazodanowych, ciągle jest zadaniem złożonym – szczególnie wobec coraz większych wolumenów przechowywanych danych, a także ze względu na wielowymiarowość dokonywanych analiz.

Dużą innowacyjność rozwiązań reprezentują serwery baz danych działające zgodnie z założeniami przetwarzania gridowego. W takim środowisku system zarządzania bazą danych powinien m.in. wykorzystywać rozwiązania z poprzednich okresów (czyli uczyć się) i w konsekwencji w sposób zautomatyzowany reagować na występujące w jego funkcjonowaniu nieprawidłowości. Celem artykułu jest zbadanie nowych możliwości administrowania hurtowniami danych w takim środowisku.

Kolejne części artykułu dotyczą: istoty samego administrowania hurtownią danych (AHD), podstawowych zasad i celów przetwarzania gridowego oraz użyteczności tego nowego środowiska w realizacji wybranych funkcji administrowania AHD (przykłady). W artykule odwołano się do platformy serwera Oracle 10g jako jednego z wiodących rozwiązań w zakresie przetwarzania gridowego.

2. Kluczowe funkcje administrowania hurtownią danych

Ze względu na coraz bardziej rozwinięte środowisko współczesnych serwerów baz danych rola administratora baz danych (ABD) zyskuje na znaczeniu i wymaga odpowiedniego wsparcia ze strony narzędzi programowych. Celem pracy ABD jest

zapewnienie ciągłej i bezkonfliktowej eksploatacji systemu wykorzystującego technologie baz danych. Technologie te odwołują się do zaawansowanych baz danych działających w środowisku rozproszonym, wykorzystujących multimedia itd., co jeszcze bardziej komplikuje proces eksploatacji systemów informatycznych.

Do klasycznych funkcji ABD należą: instalowanie bazy danych oraz konfiguracja parametrów, definiowanie schematów baz danych, zapewnienie bezpieczeństwa i poufności baz danych, monitorowanie i strojenie bazy danych czy wreszcie obsługa procesów dotyczących wymiany danych z otoczeniem (por. [Theriault, Carmichael, Viscuzi 2003; Owoc, Sachański 2005]). Należy podkreślić, że poszczególne funkcje mogą być wspomagane przez specjalizowane pakiety wiodących serwerów baz danych, a także być realizowane za pośrednictwem odpowiednich poleceń języka SQL (por. [Pondel 2004]).

Cele administrowania hurtownią danych są zbieżne z celami administrowania każdym systemem informatycznym, zwłaszcza bazą danych, a ponadto obejmują czynności specyficzne dla AHD. W tej części szcharakteryzujemy dwa etapy, częściowo odmienne od wcześniej wymienionych.

Pierwszym z etapów jest **administrowanie ładowaniem danych**. Kompetencje potrzebne do prowadzenia tego etapu wykraczają poza standardowe funkcje administratora bazy danych. W niektórych obszarach zbieżne są one z kompetencjami projektanta hurtowni. Bardzo często etap ten jest realizowany za pomocą procedur uruchamianych automatycznie. W jego skład wchodzi:

- Wybieranie danych operacyjnych z systemów źródłowych – powinny one trafiać do hurtowni w sposób automatyczny, niemniej jednak konieczny jest tutaj stały nadzór, a czasami może być konieczna „ręczna” konfiguracja połączeń pomiędzy systemami. Dane mogą pochodzić bezpośrednio z baz danych systemów transakcyjnych, z systemów typu *middleware*, zapewniających integrację danych, lub z plikowych źródeł danych, które mogą być przesyłane za pomocą np. protokołu ftp, http czy poprzez sieć lokalną. Funkcją administratora jest zapewnienie procedurom dostępu do danych przez odświeżanie połączeń (zmiana haseł, ponowna konfiguracja itp.) czy kopiowanie plików do odpowiednich lokalizacji.
- Ładowanie danych do tabel tymczasowych – za pomocą procedur stworzonych na poziomie implementacji hurtowni danych.
- Ocenianie stopnia poprawności danych – na etapie tworzenia hurtowni powinny powstać mechanizmy badające spójność i poprawność danych dostarczonych do hurtowni. W razie błędów danych administrator powinien ustalić przyczyny wystąpienia niespójności i wyeliminować je.
- Przetwarzanie danych, zwane czyszczeniem danych – uruchamianie procedur napisanych specjalnie dla wybranej hurtowni, uwzględniających jej specyfikę,
- Ładowanie danych do odpowiednich tabel – przenoszenie danych do docelowych tabel hurtowni danych.

Administrator powinien przede wszystkim kontrolować poprawność wszystkich procesów zachodzących w tym etapie. Powinien też odpowiednio dobierać częstotliwość oraz czas uruchamiania konkretnych procesów, tak aby zapewnić równowagę pomiędzy aktualnością systemu – pokazywaniem przez hurtownię możliwie jak najświeższych danych – a minimalizacją czasu trwania procesu ładowania, podczas którego możliwości korzystania z hurtowni są mocno ograniczone.

Drugi etap jest związany z dbaniem o właściwą wydajność systemu – czyli należy do typowych prac **administratora bazy danych**, jednak ze względu na specyfikę hurtowni – przyjrzyjmy się tym zadaniom bardziej szczegółowo. W skład tego etapu wchodzi:

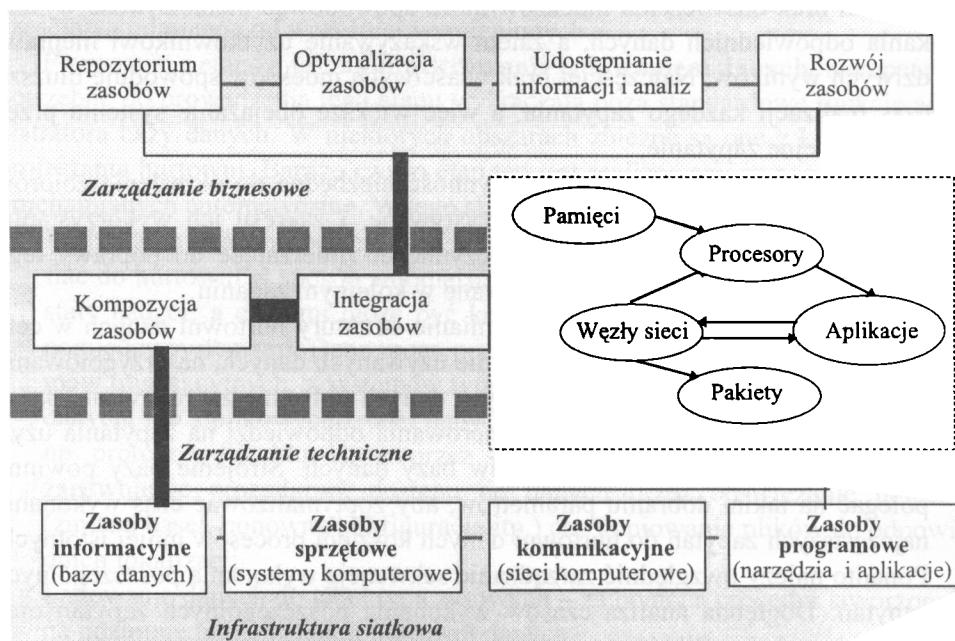
- Pielęgnacja indeksów i synonimów danych operacyjnych – indeksy powinny być regularnie odświeżane po każdym załadowaniu nowych danych. Indeksy wspomagają wyszukiwanie danych, a co za tym idzie, skracają czas realizacji zapytań do hurtowni danych. W niektórych systemach baz danych przy części zapytań brak odświeżenia indeksów może spowodować niemożliwość wyszukania odpowiednich danych, a zatem wskazywanie użytkownikowi nieprawdziwych wyników. Najczęściej brak właściwych indeksów spowoduje dłuższy czas realizacji każdego zapytania, a więc większe obciążanie systemu przez każde kolejne zapytanie,
- Monitorowanie zapytań – jest to czynność niezbędna do określenia zbiorów danych, które są najczęściej wykorzystywane, i zakresu ich wykorzystania. Efektem monitorowania mogą być czynności zmierzające do poprawy tego niepożądanego stanu, scharakteryzowane w kolejnym zadaniu.
- Korekta bazy, polegająca m.in. na zmianie struktury hurtowni danych w celu optymalizacji zapytań i archiwizacji nie używanych danych, na przygotowaniu wycinka najczęściej wykorzystywanych danych w formie zagregowanej, która będzie stanowiła lepszą bazę do generowania odpowiedzi na zapytania użytkownika, czy na zmianie parametrów bazy danych. Strojenie bazy powinno polegać na takim dobraniu parametrów, aby zoptymalizować czas wykonania najczęstszych zapytań do hurtowni danych kosztem procesów mniej istotnych. Ponadto należy uwzględnić zarządzanie sekwencją wykonania poszczególnych zapytań. Dogłębna analiza czasów wykonania poszczególnych zapytań oraz powstających konfliktów pomaga w ustaleniu priorytetów oraz kolejkwania wykonania poszczególnych procesów por. (por. [Szlapin 2002]).

Prace administratora hurtowni danych w drugim etapie można podsumować jako ciągłe dostosowywanie wykorzystania dostępnych zasobów – mocy obliczeniowej systemu oraz pamięci dyskowej – do jak najbardziej efektywnej pracy całego systemu, w szczególności do optymalizacji czasów wykonania zapytań do hurtowni.

3. Istota i możliwości przetwarzania gridowego

W ramach ciągłych poszukiwań rozwiązań technologicznych, grupujących coraz większy zakres funkcjonalny systemów komputerowych oraz optymalizację wykorzystywanych zasobów, istotny krok do przodu stanowi tzw. przetwarzanie gridowe, będące przedmiotem wielu projektów o znaczeniu lokalnym i globalnym (zob. np. [Grid 2006]). Jego istota sprowadza się do orientacji na realizowane usługi, wdrażanie otwartych standardów, współpracę i wirtualizację. Architektura przetwarzania siatkowego (*grid computing*) coraz częściej traktowana jest jako przełom w technologiach informatycznych (por. [Owoc 2004; Owoc, Wolasiński]).

W koncepcji tego typu przetwarzania, zapoczątkowanego jeszcze w latach sześćdziesiątych, kluczową rolę odgrywa grid, tzn. siatka, jako globalne połączenie różnych i rozproszonych zasobów, które ze sobą współpracują i którymi można skutecznie zarządzać. Ogólną ideę takiej infrastruktury przedstawiono na rys. 1.



Rys. 1. Architektura informacyjna wykorzystująca infrastrukturę siatkową

Źródło: opracowanie własne na podstawie [Owoc 2004; Kelly 2004].

W dolnej części rysunku znajduje się infrastruktura siatkowa podstawowa dla całego przetwarzania, obejmująca struktury informacyjne, sprzętowo-programowe oraz umożliwiające wzajemną wymianę informacji. Stanowią one odpowiednio dobrane zasoby, zespolone w jeden spójny układ, będący przedmiotem zarządzania technicznego. W odniesieniu do hurtowni danych układ ten charakteryzuje się

dużym stopniem zróżnicowania w ramach każdego z wyodrębnionych zasobów, począwszy od heterogenicznych baz danych (zasoby informacyjne) aż po odmienne środowiska sprzętowo-programowe.

Część środkowa rysunku obejmuje procesy kluczowe dla środowiska siatkowego (w rozumieniu również hurtowni danych), związane z ciągłym utrzymywaniem takiego układu. Sprowadzono je do zadań konfiguracyjnych (określonych jako kompozycja zasobów) oraz łączących je w spójną całość (integracja zasobów). Po prawej stronie mamy zaznaczone komponenty istotne dla tworzonych w ten sposób siatek (począwszy od różnych rodzajów stosowanych pamięci aż po aplikacje i pakiety).

Część górna rysunku odnosi się do merytorycznych zadań kojarzonych z biznesem. Zasadniczą rolę w takim układzie spełnia swoiste repozytorium pozwalające na dość klarowne dysponowanie poszczególnymi zasobami. Umożliwia ono dokonywanie ciągłych korekt działającego układu poprzez odpowiednie procedury optymalizacyjne. W konsekwencji zapewnione zostają warunki sprawnego informowania i wspomagania decyzji (np. poprzez przygotowanie odpowiednich analiz z zachowaniem klasycznych własności BD: poufności, niezawodności czy współbieżnego dostępu do zasobów). Zgodnie z założeniami, układ taki powinien dysponować skutecznymi mechanizmami adaptacyjnymi do różnych warunków działania – co symbolicznie ujęto jako „rozwój zasobów”.

Przedstawiona koncepcja przetwarzania siatkowego jest zbieżna z zadaniami przedstawionymi wcześniej w obszarze funkcji administratora HD. Takie cele jak optymalizacja i integracja zasobów stanowią naturalne czynności wykonywane w ramach budowanej hurtowni danych. Znajduje to odzwierciedlenie w formie zarówno dedykowanych funkcji pakietów służących do tworzenia samej hurtowni danych, np. Oracle Warehouse Builder, jak i całej architektury siatkowej dostępnej w środowisku zaawansowanych serwerów BD.

4. Przykłady realizacji zadań administrowania hurtowniami danych wspomaganego technologią gridową na przykładzie Oracle

Jak pokazano w części poprzedniej, jednym z najtrudniejszych wyzwań stojących przed administratorem hurtowni danych jest zapewnienie sprawnej obsługi zadań składających się na proces zasilania HD, a także utrzymywania zasobów tej hurtowni w takiej formie, aby były zgodne z wyznaczonymi celami użytkowymi.

Podstawowym narzędziem administratora bazy danych Oracle jest Oracle Enterprise Manager (OEM). Jest to zintegrowana konsola pozwalająca na zarządzanie najważniejszymi produktami Oracle działającymi w przedsiębiorstwie. Dostarczana wraz z bazą danych Oracle 10g wersja OEM Grid Control pozwala na zarządzanie m.in. bazą danych, serwerami aplikacji, a także instancjami klastra aplikacyjnego oraz Gridu. OEM powala także na stałe monitorowanie wielorakich wskaźników i parametrów. Umożliwia ponadto definiowanie domyślnych akcji,

które będą wywoływane automatycznie przy przekroczeniu założonych (krytycznych) wartości [Masewicz 2005].

Charakterystykę środowiska programowego wspomagającego realizację wybranych zadań AHD rozpoczniemy od funkcji umożliwiającej wgląd w zawartość „agregatów” danych, składających się na wielowymiarowe przechowywanie struktur informacyjnych hurtowni danych (kostek, wymiarów czy hierarchii). Rysunek 2 prezentuje przykłady odwzorowań elementów HD.

Cube Name **COST_CUBE**
 Display Name **Cost Analysis**
 Schema **SH**
 Description **Unit Cost. Unit Price <TIMES PRODUCTS CHANNELS PROMOTIONS>**
 Fact Type **TABLE**
 Fact Schema **SH**
 Fact Table **COST**
 Status **Valid**

Dimensions

Dimension Alias	Dimension Owner	Dimension Object	Dimension Dependency
CHANNELS_DIM	SH	CHANNELS_DIM	
PRODUCTS_DIM	SH	PRODUCTS_DIM	
PROMOTIONS_DIM	SH	PROMOTIONS_DIM	
TIMES_DIM	SH	TIMES_DIM	

Measures

Measure Name	Display Name	Source Column	Data Type
UNIT_COST	Cost	UNIT_COST	NUMBER
UNIT_PRICE	Price	UNIT_PRICE	NUMBER

Aggregation

Measure Name	Dimension Name	Operator	Argument Column
UNIT_COST	TIMES_DIM	SUM	
UNIT_PRICE	TIMES_DIM	SUM	

Rys. 2. Definicje elementów strukturalnych hurtowni danych odwołujących się do tabeli faktów obrazujących analizę kosztową

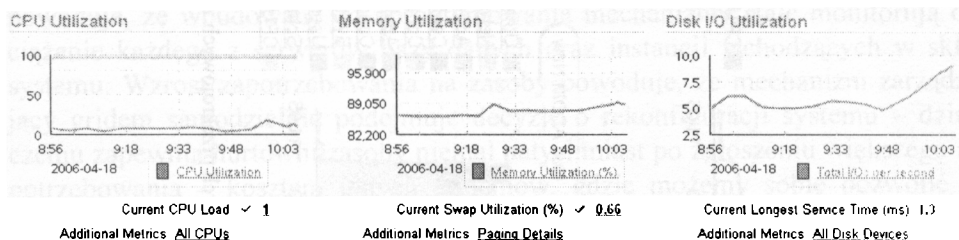
Źródło: opracowanie własne z wykorzystaniem Oracle Enterprise Manager.

Kolejne zadanie to **monitorowanie różnych zasobów**, istotnych dla funkcjonowania HD. OEM pozwala monitorować (co przedstawiono na rys. 3) przede wszystkim:

- wykorzystanie mocy obliczeniowej,
- wykorzystanie pamięci operacyjnej,
- wykorzystanie pamięci dyskowej,
- występowanie błędów przy pracy z bazą danych.

Hurtownia danych w większości przypadków nie jest systemem, od którego nie wymaga się podwyższonej niezawodności. Przerwy w jej działaniu (choćby zwią-

zane z ładowaniem danych) są rzeczą naturalną i nie wymaga się od administratora, aby zapewnił użytkownikom aplikacji działającej ze stuprocentową niezawodnością, tak jak to ma miejsce w przypadku systemów transakcyjnych.



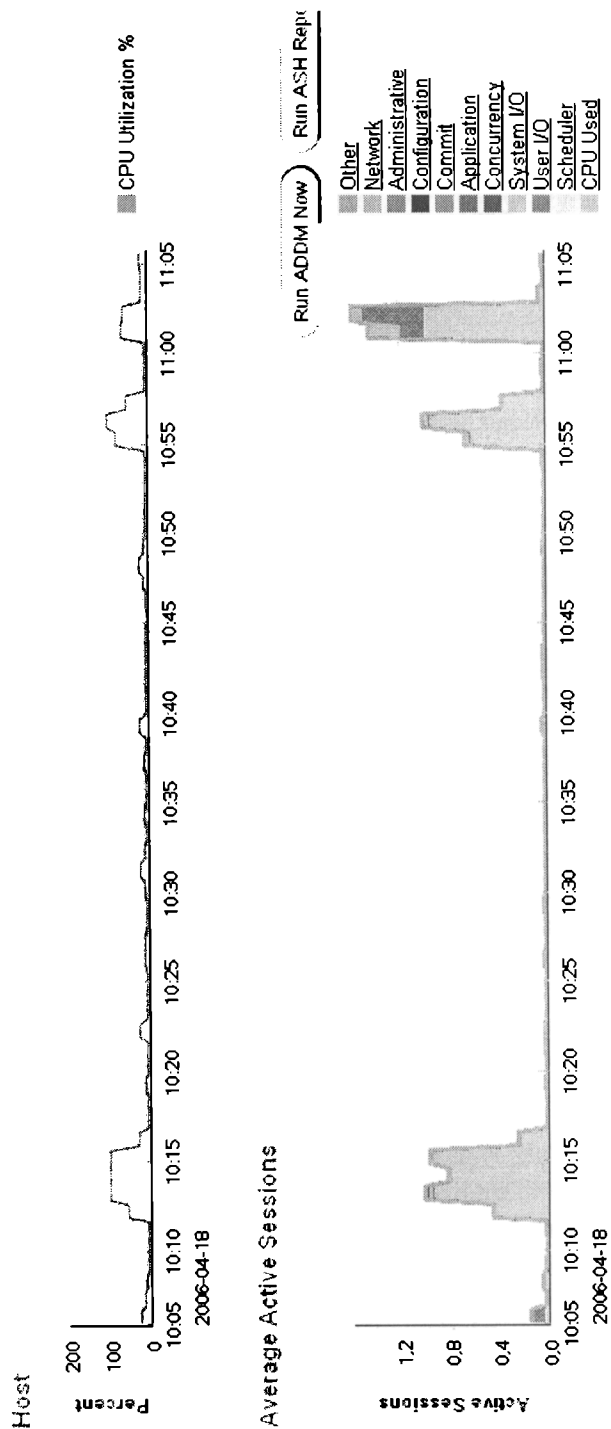
Rys. 3. Monitorowanie obciążenia poszczególnych zasobów systemu

Źródło: opracowanie własne z wykorzystaniem Oracle Enterprise Manager.

Również inaczej niż w systemach transakcyjnych wygląda sprawa z obciążeniem serwera bazy danych przez użytkowników hurtowni. Zapytania do niej kierowane mają zazwyczaj charakter przekrojowy, co powoduje, że do wygenerowania odpowiedzi na zapytanie użytkownika może być użyta ogromna liczba rekordów znajdujących się w hurtowni. Zapytania pojedynczego użytkownika mogą spowodować wysokie obciążenie zasobów serwera bazy danych przez długi czas. Powszechnie są zapytania, na które odpowiedzi generowane są w czasie przekraczającym kilkadziesiąt minut. Wzrost zapotrzebowania na zasoby spowodowane jednym zapytaniem można zaobserwować na rys. 4.

Inną cechą funkcjonowania hurtowni danych są dynamiczne okresy wzrostu jej obciążenia. Nie jest ona systemem wykorzystywanym bez przerwy. Z reguły ma określoną grupę użytkowników (zarząd, analitycy, pracownicy controllingu), co zmniejsza częstość jej wykorzystania. Ich potrzeby informacyjne wzrastają zawsze przy zamykaniu okresów rozliczeniowych (miesiąca, kwartału czy roku finansowego). Nie sposób jednak opisać okresów wzrostu wykorzystania hurtowni danych za pomocą prostych reguł. Wprowadzenie nowego produktu na rynek, intensywna promocja czy rozpoczęcie współpracy z nowym partnerem handlowym to momenty, po których przychodzi okres analizy ich wpływu na wyniki firmy. Można więc stwierdzić, że cechą charakterystyczną hurtowni danych jest duże okresowe zapotrzebowanie na zasoby sprzętowe serwerów baz danych – przy czym dokładne określenie owych okresów może być dla administratora systemu bardzo trudne.

W sytuacji gdy system odwołuje się do kilku instancji bazy danych, administrator, korzystając z własnego doświadczenia, może dokonywać „ręcznego” przełączania instancji bazy danych pomiędzy serwerami tak, aby w przewidywanym „gorącym” okresie zapewnić użytkownikom hurtowni możliwie największą moc obliczeniową oraz pamięć. Problemem jednak jest sytuacja, w której wzrost obciążenia wymyka się standardowym regułom. Przewidywanie wzrostu potrzeb informacyjnych użytkowników może wychodzić poza kompetencje administratora



Rys. 4. Zmiana poziomu wykorzystania zasobów spowodowana wykorzystaniem przez system pojedynczych przekrojowych zapytań do hurtowni danych

Źródło: opracowanie własne z wykorzystaniem Oracle Enterprise Manager.

systemu informatycznego, gdyż wiąże się z zaistniałymi okolicznościami biznesowymi. Innym problemem jest to, iż zwiększone zapotrzebowanie może mieć charakter krótkotrwały – administrator może nie zdążyć zareagować.

Zarządzanie obciążeniem za pomocą architektury przetwarzania siatkowego powoduje, że wbudowane do oprogramowania mechanizmy stale monitorują obciążenie każdego z serwerów bazy danych oraz instancji wchodzących w skład systemu. Wzrost zapotrzebowania na zasoby powoduje, że mechanizm zarządzający gridem samodzielnie podejmuje decyzje o rekonfiguracji systemu – dzięki czemu zapewnia hurtowni zasoby niemal natychmiast po zgłoszeniu większego zapotrzebowania – kosztem innych systemów, gdzie możemy sobie pozwolić na chwilowy spadek wydajności.

5. Wnioski

Przedstawiona charakterystyka rozwiązań dotyczących implementacji wybranych funkcji ABD w środowisku serwera Oracle 10g potwierdza dużą innowacyjność narzędzi wspomagających AHD. W ramach przedstawionych funkcji środowisko narzędziowe zapewnia automatyzację wielu zadań związanych z monitorowaniem i diagnostyką hurtowni danych zgodnie z założeniami przetwarzania siatkowego. Dotyczą one m.in. monitorowania obciążenia poszczególnych zasobów czy też szczegółowej analizy skutków wykonywania złożonego zapytania do bazy. Podobne zastosowanie mamy w przypadku konieczności administrowania procesu zasilania danymi samej hurtowni. W dalszych badaniach należałoby ocenić realizację innych funkcji AHD.

Literatura

- Foster I. i in., *The Physiology of the Grid: An Open Grid Services Architecture for Distributed Systems Integration*, Open Grid Service Infrastructure WG, Global Grid Forum, 22 czerwca 2002.
- Foster I., Kesselman C., *The Grid: Blueprint for a New Computing Infrastructure*, Morgan-Kaufmann 1999.
- Foster I., Kesselman C., Tuecke S., *The Anatomy of the Grid: Enabling Scalable Virtual Organizations*, „International J. Supercomputer Applications” 2001, nr 15(3).
- IDC White Paper, *Oracle 10g: Putting Grids to Work*, 2004.
- Grid Computing: <http://www.grid.org/home.htm> (dostępny – kwiecień 2006)
- Joseph J., Ernest M., Fellenstein C., *Evolution of Grid Computing Architecture and Grid Adoption Models*, „IBM Systems Journal” 2004, vol. 43, nr 4.
- Kelly D.A., *Core to Your Business*, „Oracle Magazine” maj/czerwiec 2004, s. 26-35.
- Masewicz M., *RAC, GRID, Data Guard – technologie zwiększające wydajność i niezawodność bazy danych Oracle 10g*, XI Seminarium PLOUG, Warszawa, czerwiec 2005 - http://www.ploug.org.pl/seminarium/seminarium_XI/pliki/RAC_GRID_DataGuard.pdf (dostępny – kwiecień 2006).
- Oracle Technical Article, *Manageability: What's In It for the DBA*, 2005.

- Owoc M.L., *Przetwarzanie siatkowe na przykładzie Oracle 10g*, [w:] *Nowoczesne technologie informacyjne w zarządzaniu*, red. E. Niedzielska, H. Dudycz, M. Dyczkowski, Prace Naukowe Akademii Ekonomicznej we Wrocławiu nr 1044, AE, Wrocław 2004.
- Owoc M.L., Sachańbiński J., *Wybrane funkcje administrowania zasobami informatycznymi w przetwarzaniu siatkowym* [w:] *Nowoczesne technologie informacyjne w zarządzaniu*, red. E. Niedzielska, H. Dudycz, M. Dyczkowski, Prace Naukowe Akademii Ekonomicznej we Wrocławiu nr 1081, AE, Wrocław 2005.
- Owoc M.L., Walasiński T., *Przetwarzanie siatkowe jako przełom w technologiach informatycznych*, Zeszyty Naukowe Uniwersytetu Szczecińskiego, Studia Informatica nr 20 (w druku).
- Pondel M., *Systemy analityczne*, [w:] *Generowanie wiedzy dla przedsiębiorstwa – metody i techniki*, red. M. Nycz, AE, Wrocław 2004.
- Szlapin G., *Hurtownie danych*, [w:] *Wstęp do informatyki gospodarczej*, red. A. Rokicka-Broniatowska, SGH, Warszawa 2002.
- Therhault M., Carmichael R., Viscusi J., *Oracle9i. Administrowanie bazami danych od podstaw*, Helion, Gliwice 2003.

DATAWAREHOUSE ADMINISTRATION USING GRID COMPUTING

Summary

Announced progress of data processing effectiveness in *grid technology* environment is strictly tied to warehouse administrator's (WHA) functions. The main goal of the paper is to point out key WHA functions can be supported by grid computing and analysis of two of them: monitoring and tuning apart of presentation data warehouse contents. Practical aspects of the problem are presented using Oracle 10g server.

Dr inż. Mieczysław L. Owoc jest kierownikiem Katedry Systemów Sztucznej Inteligencji, Akademii Ekonomicznej Wrocław; zajmuje się nowoczesnymi technologiami informacyjnymi, ze szczególnym uwzględnieniem inteligentnych narzędzi wspomagających infrastrukturę zarządzania. e-mail: mieczyslaw.owoc@ae.wroc.pl.

Mgr inż. Maciej Pondel jest asystentem w Katedrze Systemów Sztucznej Inteligencji i doktorantem Akademii Ekonomicznej we Wrocławiu, przygotowuje rozprawę dotyczącą pozyskiwania informacji i wiedzy na potrzeby inteligentnego przedsiębiorstwa, e-mail: maciej.pondel@ae.wroc.pl.