

ZASTOSOWANIE STRUKTUR RELACYJNO-OBIEKTOWYCH DO PRZECHOWYWANIA MASOWYCH DANYCH POMIAROWYCH

Tomasz Traczyk
ttraczyk@ia.pw.edu.pl
Karol Stanisławek
karol.stanislawek@cern.ch



Politechnika Warszawska
Wydział Elektroniki i Technik Informacyjnych
Instytut Automatyki i Informatyki Stosowanej



Zastosowanie struktur relacyjno-objektowych do przechowywania masowych danych pomiarowych

- ❖ Wprowadzenie: baza DCDB – założenia i architektura
- ❖ Dane pomiarowe w DCDB
- ❖ Założenia badań zastosowania struktur relacyjno-objektowych
- ❖ Wyniki badań
- ❖ Podsumowanie



Zastosowanie struktur relacyjno-objektowych... KKNTPD'05

2

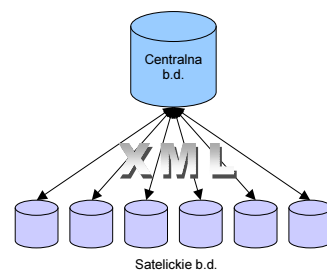
Baza DCDB

Eksperyment ALICE (A Large Ion Collider Experiment)

- Miejsce
 - Europejskie Centrum Badań Jądrowych CERN (Genewa)
- Cel
 - badanie właściwości materii w warunkach zderzeń jonów o wielkich energiach
- Aparatura
 - nowy akcelerator LHC (Large Hadron Collider)
 - ♦ 4 wielkie detektory
 - detektor ALICE
 - ♦ zespół subdetektorów
 - ♦ miliony komponentów!

Baza DCDB (Detector Construction Database)

- Gromadzi dane o
 - budowie detektora
 - jego częściach składowych
- Cechy bazy
 - rozproszona (kilkadziesiąt ośrodków)
 - heterogeniczna (Oracle + PostgreSQL)
 - o generycznej strukturze



Zastosowanie struktur relacyjno-objektowych... KKNTPD'05

3

Użycie DCDB

Bazy satelickie

- Ładowane przez oprogramowanie pomiarowe
- Niekiedy dane edytowane „ręcznie”
- Używane w czasie produkcji komponentów
 - docelowo do likwidacji

Baza centralna

- Ładowana wsadowo
 - transfery z baz satelickich
- Odpytwana interaktywnie
 - zaawansowane wyszukiwanie
 - analizy danych
- Bardzo rzadko dane edytowane „ręcznie”

Aplikacje

- Składniki
 - edytor metadanych *Dictionary Wizard*
 - oprogramowanie pomiarowe (LabView + LabServer)
 - aplikacja główna *RABBIT*
 - system transferu danych
- Technologia
 - dostęp przez WWW
 - JSP + Struts
 - analizy danych: pakiet *Root* (CERN)
 - transfery danych w XML (obecnie wdrażany protokół SOAP)



Zastosowanie struktur relacyjno-objektowych... KKNTPD'05

4

Wprowadzenie: baza DCDB – założenia i architektura

Dane pomiarowe w DCDB

Założenia badań zastosowania struktur relacyjno-objektowych

Wyniki badań

Podsumowanie

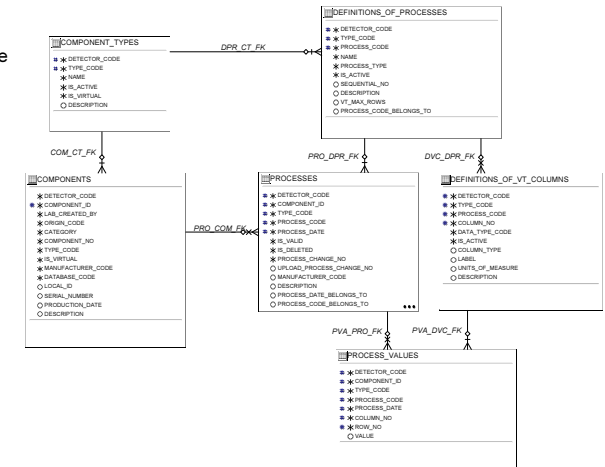


Dane pomiarowe w DCDB

Dane pomiarowe w DCDB

- Cele
 - rejestrują wyniki testów komponentów detektora
 - mają zasadnicze znaczenie dla poprawności interpretacji przyszłych wyników
- Postać
 - wirtualna tablica
 - kolumny odpowiadają torom pomiarowym
 - wiersze odpowiadają punktom pomiarowym

Generyczna realizacja relacyjna



Dane pomiarowe w DCDB, c.d.

Przechowywanie danych pomiarowych

- Pierwsze podejście: naiwna struktura r-o
 - całe tablice wyników pomiarów
 - trudne operowanie na bazie
 - nienadzwyczajna wydajność (zwłaszcza w PostgreSQL)
- Obecnie: struktura czysto relacyjna
 - każdy punkt pomiarowy w osobnym wierszu tabeli
 - łatwe operowanie na bazie
 - bardzo duże zużycie przestrzeni dyskowej (zwłaszcza w PostgreSQL)
- Przyszłość: zoptymalizowana struktura r-o
 - badania na Oracle wykonane
 - badania na PostgreSQL w toku

Trudności

- Bardzo duża ilość danych
 - w czasie badań
 - ♦ > 1 mln testów
 - ♦ > 200 mln punktów pomiarowych
 - spodziewany przyrost
 - ♦ ok. 10 razy
- Wielkie rozmiary indeksów

Użyte rozwiązania dla VLDB (Oracle)

- Partycjonowanie tabel
- Tabele zorganizowane indeksowo (*index organized tables*)
 - z kompresją indeksu



Wprowadzenie: baza DCDB – założenia i architektura

Dane pomiarowe w DCDB

Założenia badań zastosowania struktur relacyjno-objektowych

Wyniki badań

Podsumowanie



Założenia badań

Cele do osiągnięcia

- Mniejsze zużycie pamięci dyskowej
 - mniejsze koszty
 - mniejsze trudności w administrowaniu
- Krótsze czasy ładowania danych
 - nadążanie za napływem danych
 - uniknięcie niepokoju użytkowników
- Dobra wydajność typowych zapytań
 - możliwość efektywnej pracy fizyków

Badane struktury (Oracle)

- Relacyjna – komórkami
- Nested table* – kolumnami
- Nested table* – wierszami
 - » badanie porzucono
- Nested table* – komórkami
- Varying array* – kolumnami
- Varying array* – wierszami
- Varying array* w *varying array* – kolumnami
- Nested table* w *nested table* – kolumnami
- Varying array* w *nested table* – kolumnami

Badane operacje

- Ładowanie danych
- Wyszukiwanie wszystkich danych komponentu
- Tworzenie histogramu wyników pomiarów
- Uśrednianie wartości pomiaru
- Wyszukiwanie kolumn
- Modifikacja danych

❖ Wprowadzenie: baza DCDB – założenia i architektura

❖ Dane pomiarowe w DCDB

❖ Założenia badań zastosowania struktur relacyjno-objektowych

❖ Wyniki badań

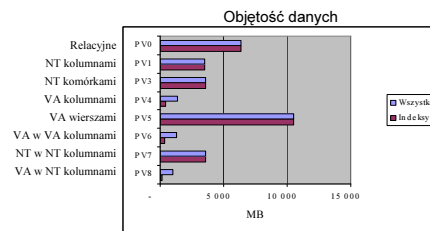
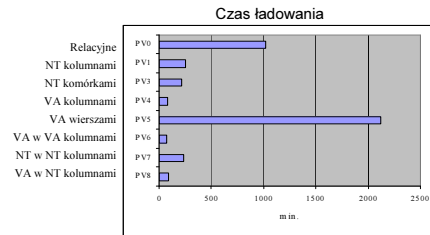
❖ Podsumowanie



Wyniki badań

Ładowanie danych

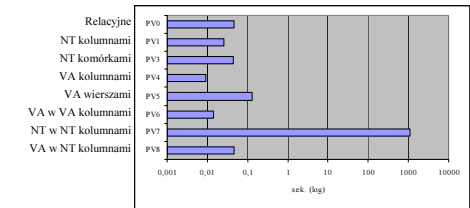
- Ładowanie danych do tabel bazy centralnej
- Kryteria
 - czas ładowania
 - objętość danych
- Wyniki
 - najlepsze: *varying arrays* kolumnami
 - najgorsze: reprezentacje wierszami
 - odrzucone: *nested table* wierszami



Wyniki badań, c.d.

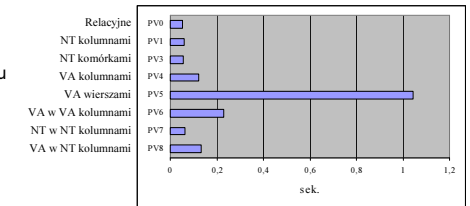
Wyszukiwanie całego testu

- Zapytanie zwracające wszystkie wyniki
 - jednego testu
 - dla jednego komponentu
- Kryterium
 - czas wykonania
- Wyniki
 - najlepsze: *varying array* kolumnami
 - najgorsze: *nested table* w *nested table*



Histogramowanie

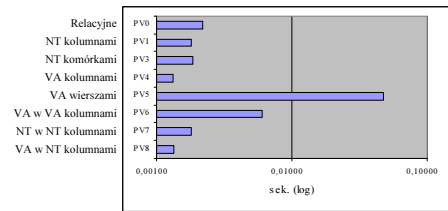
- Zapytanie liczące histogram
 - jednego parametru
 - dla wszystkich komponentów jednego typu
- Wyniki
 - najlepsze: relacyjne i *nested table*
 - najgorsze: wierszami



Wyniki badań, c.d.

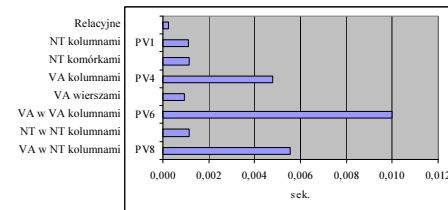
Tworzenie wykresu

- Zapytanie wyszukujące wyniki pomiaru
 - jednego toru pomiarowego (kolumny)
 - dla jednego testu
 - dla jednego komponentu
- Wyniki
 - najlepsze: *varying array* kolumnami
 - najgorsze: wierszami



Modyfikacja danych

- Zmiana pojedynczej „komórki” danych
- Wyniki
 - najlepsze: relacyjne i *nested table*
 - najgorsze: *varying array* w *varying array*



- ❖ Wprowadzenie: baza DCDB – założenia i architektura
- ❖ Dane pomiarowe w DCDB
- ❖ Założenia badań zastosowania struktur relacyjno-objektowych
- ❖ Wyniki badań
- ❖ Podsumowanie



Podsumowanie

Wyniki badań

- Stosując odpowiednio dobrane struktury r-o można
 - znacznie zredukować czas ładowania i objętość danych pomiarowych DCDB
 - zachowując dobrą wydajność typowych zapytań
- Wyniki badań nie dają jednoznacznej odpowiedzi co do wyboru struktury
 - nie ma lidera we wszystkich „konkurencjach”
 - trzeba jeszcze wziąć pod uwagę
 - dodatkowe czynniki techniczne
 - wyniki badań dla PostgreSQL
 - wskazanie na reprezentację *varying array* kolumnami

Dalsze prace

- Dalsze badania na Oracle
 - wpływ innych czynników: partycjonowanie, indeksowanie
- Dokończenie badań na PostgreSQL
- Wybór struktury i konwersja danych na obu platformach
- Dostosowanie aplikacji do struktur r-o i testy zmienionej aplikacji
- Ostateczna akceptacja

Zastosowanie struktur relacyjno-objektowych
do przechowywania
masowych danych pomiarowych

