Hurtownie danych w praktyce

Fakty i mity

Dr inż. Maciej Kiewra

QBICO

Parę słów o mnie...

- 8 lat pracy zawodowej z hurtowniami danych
- Projekty realizowane w kraju i zagranicą
- Certyfikaty Microsoft z Business Intelligence (Microsoft SQL Server 2005 i 2008)
- Od 2007 prowadzę firmę doradczą zajmującą się tworzeniem, audytowaniem, wdrażaniem i utrzymaniem hurtowni danych i wspierającego je oprogramowania
- Przez 3 lata pracownik dydaktyczno – naukowy Politechniki Wrocławskiej (zajęcia z hurtowni danych)

QBICO

Niezbędnym elementem każdej hurtowni danych są kostki OLAP.

- **Fałsz** – kostki OLAP są tylko i wyłącznie pomocniczym sposobem organizacji danych.
- Kostki OLAP mogą być z powodzeniem pominięte np. gdy narzędzie do generowania raportów ad – hoc jest w stanie automatycznie generować zapytania w języku SQL

QBICO

W kostkach OLAP zapytania wykonują się szybciej niż w tabelach relacyjnych.

- **Fałsz** – w kostkach OLAP będą wykonywały się szybciej zapytania wymagające agregacji w locie dużej ilości danych
- W tabelach będą wykonywały się szybciej zapytania krzyżujące dużą ilość danych opisowych bez agregacji (np. pokaż mi wszystkie faktury, wystawione wszystkim moim klientom na wszystkie możliwe usługi)

QBICO

Narzędzia typu „OLAP in memory” pozwolą przygotować i wdrożyć średniej wielkości hurtownię w kilka tygodni

- **Fałsz** – te cudowne narzędzia pozwalają dokonywać efektownych operacji na z góry przygotowanych danych
- Przy wdrażaniu zakłada się, że dane już są w jednym dużym zbiorze np. pliku tekstowym
- Przy dzisiejszych pojemnościach pamięci RAM wczytajmy cały plik do pamięci i „po sprawie”
- W ocenie czasu wdrożenia pomijane są takie czasochłonne operacje jak uzgadnianie danych z tabel źródłowych, definiowanie wspólnej bazy pojęć, czyszczenie danych itp.

QBICO

Nawet jeśli nie wdrażamy kostek OLAP powinniśmy stosować model wielowymiarowy.

- **Prawda** – organizacja danych w postaci tabel faktów i wymiarów jest ogólnie przyjętym sposobem organizacji danych w hurtowniach ze względu na:
  - Fakt, że jest on powszechnie obowiązujący w branży
  - Dużą czytelność dla użytkowników biznesowych
  - Wydajność przy sporządzaniu zapytań
  - Prostotę, elegancję oraz łatwą rozszerzalność

QBICO

Dane w tabelach faktów powinny być maksymalnie zagregowane.

- **Fałsz** – dane w tabelach faktów powinny być przechowywane na poziomie atomowym np. pozycja na fakturze – zamiast tylko zbiorczej sumy
- Hurtownia danych powinny być „przygotowana” do sporządzenia dowolnych zestawień
- Przechowywanie danych zagregowanych znacznie je zubaża

QBICO

Przy kopiowaniu danych z systemów źródłowych tabele docelowe nie powinny posiadać więzów integralności.

- **Prawda** – więzy integralności (np. klucze obce) spowalniają proces kopiowania danych
- Kłóci się to z dobrą praktyką mówiącą o tym, że odczyt z systemów dziedzinowych powinien być tak krótki jak to tylko możliwe (uwaga na blokowania!!!)

QBICO

Wprowadzenie hurtowni danych pozwala poprawić jakość danych w systemach dziedzinowych.

- **Prawda** – nieodzownym elementem rozwoju hurtowni danych jest tzw. profiling danych pozwalający wykryć bardzo dużą liczbę błędów na źródle
- Część awarii po wdrożeniu hurtowni jest powodowana przez błędne dane źródłowe

QBICO

Najlepszymi kandydatami na klucze główne w tabelach wymiarów są klucze główne z tabel dziedzinowych.

- **Fałsz** – co zrobimy gdy do tabeli DIM_PRACOWNIK trzeba będzie kopiować osoby zatrudnione także we właśnie przejętej spółce posługującej się innym systemem kadrowym?
- Odwzorowanie tabela źródłowa – tabela wymiaru nie zawsze jest 1 do 1

QBICO

W hurtowni danych nie ma potrzeby przechowywania identyfikatorów z systemów dziedzinowych

- **Fałsz** – identyfikatory te okazują się bardzo przydatne, gdyż:
  - Znacząco ułatwiają diagnostykę
  - W wymiarach ujednoliconych tzw. *conformed dimension* pozwalają w bardzo szybki sposób ocenić integrację tych samych danych z różnych systemów dziedzinowych

QBICO

Przy kopiowaniu danych z systemów dziedzinowych dobrze jest zapisać dokładną kopię odczytanych danych

- **Prawda** – są to tzw tabele stage'owe, bardzo przydatne, gdyż:
  - Ułatwiają prowadzenie diagnostyki i audytów (bez angażowania źródła)
  - Pozwalają odtworzyć wykonanie procesu ETL w przypadku awarii bez ponownego łączenia się ze źródłem
  - Dane na źródle są ulotne

QBICO

Z systemów źródłowych nie kopiujemy niepełnych rekordów (np. faktury bez numeru NIP klienta)

- **Fałsz** – bardzo rzadko systemy źródłowe posiadają pełne i „czyste” dane
- Brak w hurtowni części faktur sprzedażowych powoduje, że zmniejszamy obroty firmy!!!
- Rozwiązanie: kopiujemy niekompletną fakturę, a jej kontrahenta oznaczamy jako nieznanego

QBICO

Przy tworzeniu raportów nie ma nic złego w krzyżowaniu danych z hurtowni z danymi z systemów dziedzinowych.

- **Fałsz** gdyż:
  - System dziedzinowy może zostać wymieniony
  - Może zmienić się struktura danych
  - Dane z systemu źródłowego mogą być nieujednolicone lub zduplikowane
  - Niebezpieczeństwo blokowań w systemach źródłowych

QBICO

Liczby zmiennoprzecinkowe dobrze reprezentują wartości pieniężne

- **Fałsz** – wartości pieniężne zazwyczaj podawane są do dwóch miejsc po przecinku
- Wartości pieniężne to liczby stałoprzecinkowe!
- Użycie liczb zmiennoprzecinkowych = błąd zaokrąglenia

QBICO

Dziękuję za uwagę!

QBICO