



THE  
POWER  
TO KNOW.

# Forum Technologii SAS 2007

8 maja 2007 Hotel Marriott

SAS® Scalable  
Performance Data Server  
celną odpowiedzią na  
potrzeby użytkowników  
hurtowni

---

Sławomir Bokiniec, SAS Polska

# Agenda

- SAS SPD Server w skrócie

Podstawowe potrzeby użytkowników hurtowni:

- Efektywny ETL mieszczący się w okienku czasowym
- Wydajne zapytania
- Bezpieczeństwo
- Uniwersalność
- Niskie koszty wdrożenia i administracji

# SAS Scalable Performance Data Server

## SPD Server w skrócie

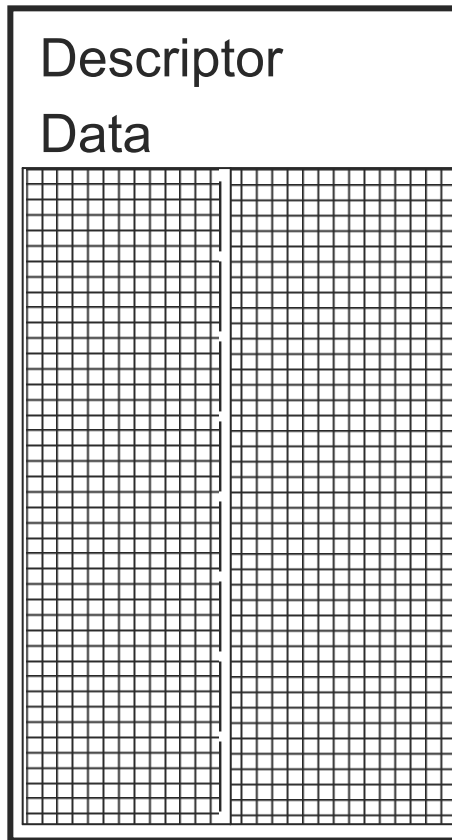
- Skalowalne partycjonowane I/O
  - Skuteczne wykorzystanie dysków dla osiągnięcia przepustowości
- Wielowątkowe przetwarzanie zapytań
  - Skuteczne wykorzystanie i indeksów prostych i złożonych
- Efektywne ładowanie i utrzymywanie dowolnie dużych zbiorów
- Równoległe przetwarzanie warunków where
  - Skuteczne wykorzystanie zarówno prostych jak i złożonych indeksów
- Brak kosztów przetwarzania transakcji
  - Użycie dostępnych procesorów do przetworzenia zapytania
- Kompatybilność z SAS-em
- Implicit and explicit pass through SQL

# Potrzebny efektywny ETL

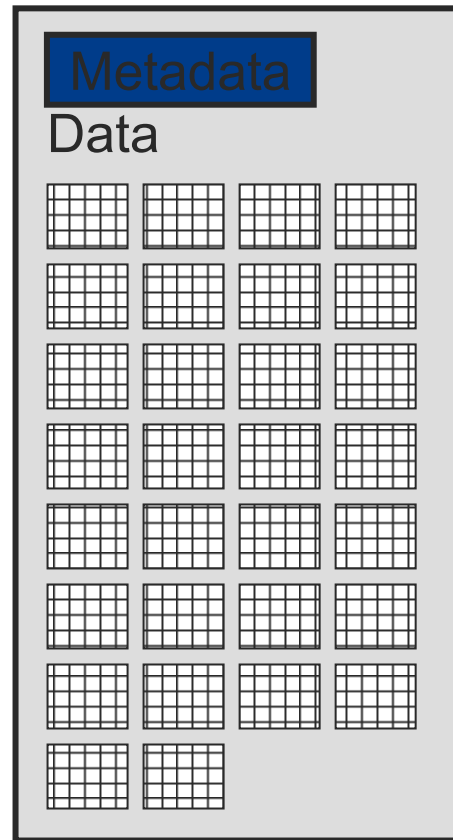
- „ET” – power of SAS
- „L” - efektywne ładowanie
- Różne sposoby (do)ładowania
  - Append
  - Append z podmianą rekordów po kluczu unikalnym
- Równoległe (wielowątkowe) modyfikacje indeksów
- Klastry – struktury zawierające wiele tablic SPDS-owych
  - Index MinMax
- Integracja z narzędziami SAS 9: DI Studio i SMC
- Dynamic locking

# Ewolucja struktury tablicy

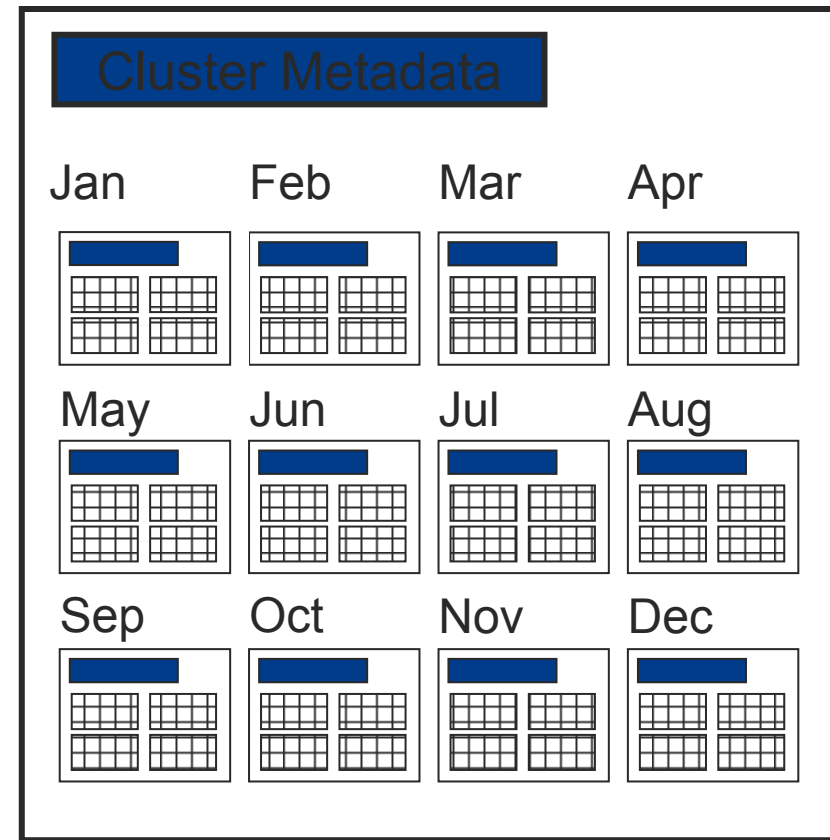
## Traditional SAS Table



## SPD Server Table



## Cluster Table



# Klastry - korzyści

- Przyspieszony ETL
- Dowolność zrównoleglenia ładowania/modyfikowania elementów klastra
- Uproszczenie zarządzania dużymi zbiorami
  - sekundowy czas tworzenia klastra (trochę dłuższy przy indeksach unikalnych)
- Dodatkowa warstwa do optymalizacji
  - Wykorzystanie indeksów MinMax
- Brak ograniczeń (vs widok)
- Uproszczony i przyspieszony backup
- I inne

# ETL – bijemy rekordy wydajnościowe

Załadowanie 1,27 TB data martu

- 3 lata transakcji międzynarodowej firmy handlowej
  - 100 mln klientów, 3 lata zakupów; 6,8 mld rekordów
  - 1549 sklepów w tym online, 400000 produktów
- Załadowanie struktury gwiazdy przygotowanej dla potrzeb użytkowników biznesowych
- Zastosowano DI Studio i SPD Server (dynamic cluster)
  - Transformacje: Lookup, join, File reader, Loop, User- written code i Table loader
- Wykonane zadania:
  - Odczyt plików tekstowych, walidacja danych i sprawdzanie spójności, wyliczenie 2 dodatkowych kolumn w tablicy faktów, type II slowly changing dimenions, poindeksowanie wszystkich tablic, załadowanie
- Sprzęt
  - Sun E25k, Solaris 10, 24x 2,95 GHz US-IV+ CPU
  - 20 Sun StorageTek ST6140 macierzy, QFS 4,5
- Rekordowy czas 2h:26s
- Dalsze szczegóły - white paper

# Potrzebne wydajne zapytania

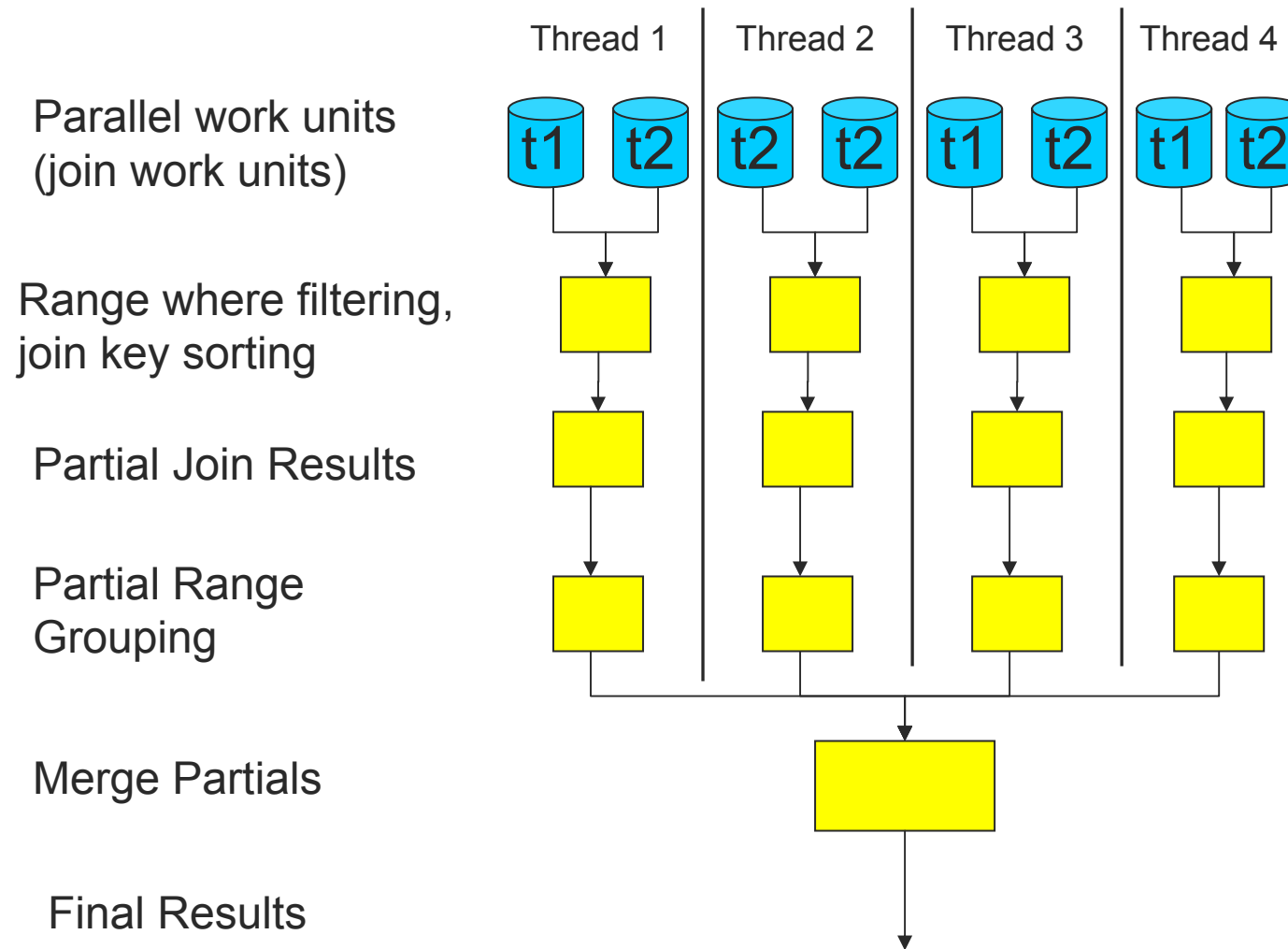
Optymalizacje przygotowane dla wydajnego przetwarzania dowolnie dużych zbiorów:

- Agregacje
  - Parallel Group By
  - PGB pogrupowany po formatach użytkownika
- Optymalizacje dla warunków where
  - Where costing
- Łączenie zbiorów
  - Parallel join
  - Star join
- Index scan

## Wydajne zapytania - fundamenty

- Wysoce wydajne funkcje I/O
  - Testy I/O wykonywane przy pomocy SPDS-a dają obraz fizycznej wydajności I/O
- Uniwersalne indeksy
  - Część bitmapowa i B-tree
  - Skuteczne przy warunkach where, joinach
  - Oszczędne dla dysku

# Parallel Join z PGB – prostota i skuteczność



# Potrzebne bezpieczeństwo

- Szyfrowanie połączenia
- ACL-e oparte na standardzie RACF
- Uprawnienia od biblioteki do kolumny
- Audytowanie dostępu
- Oraz stabilność środowiska
  - Każdy użytkownik obsługiwany przez dedykowany proxy
  - Małe obciążenie RAM-u

# Potrzebna uniwersalność

- Różne sposoby wykorzystania
  - Klienci generujący SQL
  - W przetwarzaniach batchowych
  - W zapytaniach ad-hoc
- Możliwość udostępniania bibliotek przez serwer metadanych
- Otwartość – udostępnianie danych poza SAS-a
  - JDBC
  - ODBC
  - i inne

# Wymagane niskie koszty wdrożenia i administracji

- Dowolność wyboru platformy sprzętowej
- Architektura SMP – uniwersalność wykorzystania wszystkich zasobów
- Efektywność wykorzystania sprzętu
- Łatwość wdrożenia: 4GL, DI Studio, „full SAS power behind”
- Minimalny udział administratorów

# Przykład kosztów administracyjnych

- TIM
  - Jeden z największych operatorów GSM na świecie
  - 48 milionów klientów w Europie i Ameryce Południowej
  - 3 TB danych
  - Największa tablica ~100G
  - ~300GB miesięczny przyrost danych
  - Miesięcznie modyfikowane od 10 to 99GB
  - Compaq Tru64 Unix & EMC8730 via 8 FC-SW
- Potrzeby administracyjne
  - 1 osoba – 2 dni w miesiącu

## Nowości w 4.4

- Widoki zmaterializowane
- Monitorowanie w SMC – nowe możliwości
  - Dodatkowo do
    - Zarządzania użytkownikami, ACL-ami i serwerami
  - Monitorowanie otwartych połączeń i tablic
  - Usuwanie połączeń
- Wsparcie dla LDAP-a
- SPDS for Solaris on Opteron

# Nowość – Widoki zmaterializowane

- Wynik działania widoku przechowywany w ukrytej tablicy
- Dynamiczne odświeżanie widoku
  - przy dostępie do widoku
  - gdy tablica była zmodyfikowana
  - Po MVREFRESHTIME= (domyślnie 30 s.)
- Dostęp tylko przez pass-through
  - EXECUTE (Create Materialized View <viewname> as Select ...) BY [sasspds | alias];

# Podsumowanie

- Spełniamy Państwa potrzeby w obszarze hurtowni danych i Business Intelligence
- SPD Server komponentem pakietu Intelligent Storage odpowiadającym za przetwarzanie dowolnie dużych zbiorów
- Gdy czas ładowania nie mieści się w okienku czasowym, gdy zapytania trwają zbyt długo, gdy koszty utrzymania środowiska hurtowni są zbyt duże, ... SAS proponuje
  - Scalable Performnace Data Server v4
- Jesteśmy firmą „customer driven”