



## NASI SPONSORZY I PARTNERZY



CREDIT SUISSE



Passion for Technology



SQL EXPERT.pl



SPECIALISTA IT PROGRAMISTA DESIGNER  
**GEEKCLUB.pl**  
CODEGURU.pl

SPECIALISTA IT PROGRAMISTA DESIGNER  
**GEEKCLUB.pl**  
wss.pl





## **Jak rozwiązać typowe problemy z zasilaniem hurtowni danych**



# Marcin Szeliga

- Ponad 15 lat doświadczenia z SQL Server
- Trener i konsultant
- Autor książek i artykułów
- Architekt systemów OLTP i DW/BI
- SQL Microsoft Most Valuable Professional od 2006
- Założyciel SQLExpert.pl
  - [marcin@sqlexpert.pl](mailto:marcin@sqlexpert.pl)
  - <http://blog.sqlexpert.pl/>
  - <https://www.facebook.com/SQLExpertpl>



SQL EXPERT.pl



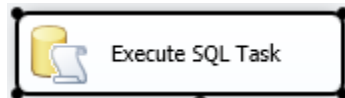
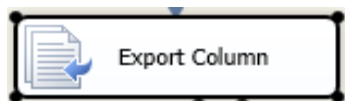
# Agenda

- Problem 1 – **Eksport danych binarnych**
  - Transformacja Export Column vs FileTable
- Problem 2 – **Eliminowanie duplikatów**
  - Transformacja Sort vs klauzula DISTINCT
  - Transformacja Lookup
  - Logika rozmyta (transformacje Fuzzy Lookup i Fuzzy Grouping)
- Problem 3 – **Zarządzanie zmieniającymi się danymi**
  - Czym są wolno zmieniające się wymiary?
  - Transformacja CSD Wizard vs instrukcja MERGE
- Problem 4 – **Efektywna synchronizacja danych**
  - Mechanizm CDC i jego komponenty SSIS



# Eksport danych binarnych

- Jak udostępnić użytkownikom systemów typu Self-Service BI przechowywane w bazach dane binarne (np. grafiki)?
- Można je wyeksportować:
  - Za pomocą transformacji Export Column
  - Zapisując je do tabeli typu FileTable



- Transformacja Export Column
  - Nie wymaga zmiany fizycznej struktury bazy
  - Jest szybsza

# Demo

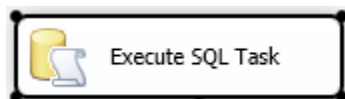
- Eksport danych binarnych



# Eliminowanie duplikatów

## Transformacja Sort vs klauzula DISTINCT

- Importowane z zewnętrznych źródeł dane często zawierają duplikaty
- Można je wyeliminować:
  - Za pomocą klauzuli DISTINCT
  - Za pomocą transformacji Sort z ustawioną opcją usuwania duplikatów



- Tak czy inaczej, dane zostaną posortowane:
  - Albo po stronie serwera SQL Server
  - Albo na komputerze na którym uruchomiony został pakiet SSIS

# Demo

- Eliminowanie duplikatów (transformacja Sort vs klauzula DISTINCT)

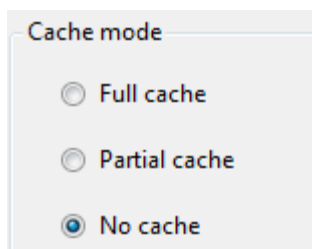
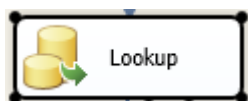




# Eliminowanie duplikatów

## Transformacja Lookup

- Część importowanych z zewnętrznych źródeł danych będzie już w hurtowni
- Wybrać nowe wiersze można:
  - Za pomocą transformacji Lookup
    - W tym wypadku interesować nas będą jedynie niedopasowane wiersze
    - Kluczowe dla wydajności jest włączenie buforowania



- Buforowanie:
  - Może być pełne lub częściowe
  - Wymaga pamięci po stronie komputera, na którym uruchomiony został pakiet

# Demo

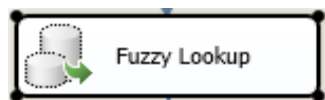
- Eliminowanie duplikatów (transformacja Lookup)



# Eliminowanie duplikatów

## Logika rozmyta

- Importowane z zewnętrznych źródeł dane mogą być błędne
- Dopasować do siebie podobne rekordy można za pomocą transformacji:
  - Fuzzy Grouping
  - Fuzzy Lookup
- Obie dostępne są tylko w edycji Enterprise



- Najpierw dane należy pogrupować, w drugiej kolejności porównać z zapisanymi w hurtowni

# Demo

- Eliminowanie duplikatów (logika rozmyta)























# Zarządzanie zmieniającymi się danymi











## Czym są wolno zmieniające się wymiary?

- Koncepcja wolno zmieniających się wymiarów została zaproponowana przez Ralph Kimballa
- Jej celem jest właściwe opisanie zdarzeń historycznych
  - Jeżeli zmieni się jakiś atrybut danych biznesowych, fakt ten jest zapisywany w tabeli wymiaru
- Występują cztery podstawowe typy SCD:
  - Typ 0: Zmiana jest niedozwolona
  - Typ 1: Historia zmian nie jest przechowywana
  - Typ 2: Historia zmian jest przechowywana
  - Typ 3: Przechowywana jest tylko aktualna i poprzednia wartość atrybutu

# SCD typu 1

DimEmployee	
	EmployeeKey
	ParentEmployeeKey
	EmployeeNationalIDAlternateKey
	ParentEmployeeNationalIDAlternateKey
	SalesTerritoryKey
	FirstName
	LastName
	MiddleName
	NameStyle
	...

BEFORE	
	295
	290
	954276278
	982310417
	8
	Rachel
	Valdez
	B
	0
	...













AFTER	
	295
	290
	954276278
	982310417
	8
	Rachel
	Valdez-Smythe
	B
	0
	...













# SCD typu 2














DimEmployee	
EmployeeKey	
ParentEmployeeKey	
EmployeeNationalIDAlternateKey	
ParentEmployeeNationalIDAlternateKey	
SalesTerritoryKey	
FirstName	
LastName	
MiddleName	
NameStyle	
...	
StartDate	
EndDate	

BEFORE		AFTER	
296		296	298
294		294	294
758596752		758596752	758596752
481044938		481044938	481044938
9		9	10
Lynn		Lynn	Lynn
Tsoflias		Tsoflias	Tsoflias
N		N	N
0		0	0
...		...	...
2003-07-01		2003-07-01	2004-10-09
		2004-10-09	

# SCD typu 3

DimEmployee	
	EmployeeKey
	ParentEmployeeKey
	EmployeeNationalIDAlternateKey
	ParentEmployeeNationalIDAlternateKey
	SalesTerritoryKey
	FirstName
	LastName
	MiddleName
	NameStyle
	...
	PreviousSalesTerritoryKey
	SalesTerritoryKeyEffectiveDate

BEFORE	
	296
	294
	758596752
	481044938
	9
	Lynn
	Tsoflias
	N
	0
	...
	...
	...

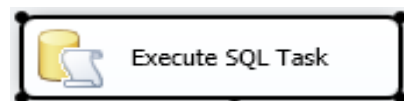
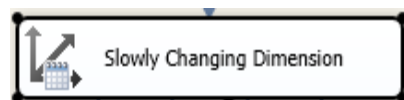
AFTER	
	296
	294
	758596752
	481044938
	10
	Lynn
	Tsoflias
	N
	0
	...
	...
	9
	2004-10-09



# Zarządzanie zmieniającymi się danymi

## Transformacja CSD Wizard vs instrukcja MERGE

- Zmiany typu 1 wymagają aktualizacji rekordu
- Zmiany typu 2 wymagają aktualizacji oryginalnego i wstawienia nowego rekordu
- W obu przypadkach konieczne jest porównanie rekordów



- Transformacja SCD wykonuje wszystkie te operacje wiersz po wierszu
- Instrukcja MERGE operuje na zbiorach
  - Tabele są odczytywane tylko raz

# Demo

- Zarządzanie zmieniającymi się danymi  
(transformacja CSD Wizard vs instrukcja MERGE)



# Efektywna synchronizacja danych

## Mechanizm CDC

- CDC należy włączyć po stronie źródła danych:
  - Na poziomie bazy (sys.sp\_cdc\_enable\_db)
  - Dla poszczególnych tabel (sys.sp\_cdc\_enable\_table)
- Po włączeniu zmiany będą automatycznie rejestrowane
  - Działanie podobne do replikacji transakcyjnej
  - Możliwość wyboru trybu śledzenia:
    - All
    - All with Old Values
    - Net
    - Net with Update Mask
    - Net with Merge
- Dostępny w wersjach 2008 i późniejszych
- Wymaga edycji Enterprise



# Efektywna synchronizacja danych

## Mechanizm CDC i jego komponenty SSIS

- W wersji 2012 dostępne są trzy komponenty:



CDC Control Task



CDC Source



CDC Splitter

- Stosowany przed i po synchronizacji danych do zarządzania informacjami opisującymi jej bieżący stan
- Odczytuje ze źródła dane zmienione po ostatniej synchronizacji
- Dzieli wiersze na podstawie typu oparci, która je zmodyfikowała

# Demo

- Efektywna synchronizacja danych  
(Mechanizm CDC i jego komponenty SSIS)



# Podsumowanie

- Zaprojektowanie i zaimplementowanie procesów ETL średnio zajmuje 2/3 całego czasu trwania projektu DW/BI
- Coraz częściej systemy BI muszą być dostępne 24 godziny na dobę
- Coraz mniej mamy czasu na zasilenie hurtowni danymi
- **Błędy i pomyłki popełnione na tym etapie są wyjątkowo kosztowne**
- Projektując procesy ETL warto od razu uwzględnić wydajność i skalowalność rozwiązania
- I nie zapomnieć o jego elastyczności i łatwości zarządzania
- **Wybór właściwych narzędzi (SQL Server lub odpowiedniej transformacji SSIS) pozwoli uniknąć kosztownych pomyłek**



## NASI SPONSORZY I PARTNERZY



Passion for Technology



SQL EXPERT.pl



Organizacja: Polskie Stowarzyszenie Użytkowników SQL Server - PLSSUG

