3. Vežbe

Projektovanje i Implementacija ETL procesa. Materijalizovani pogledi

Izvođači laboratorijskih vežbi

- Marko Knežević (kancelarija TMD 9b)
- Nikola Obrenović

Termin konsultacija

- Marko Knežević: petak 15:00 TMD 9b
  marko.knezevic(AT)uns.ac.rs
- Nikola Obrenović
  nikob(AT)uns.ac.rs

ETL osnove

- Extract – Transform – Load
- Predstavlja proces koji migrira podatke iz izvornih sistema/baza/fajlova u DW bazu podataka
- Mora da uskladi iste podatke iz različitih sistema po tipu, ograničenjima i semantici
- MS SQL Server Integration Services (SSIS)
- MS Business Intelligence Development Studio

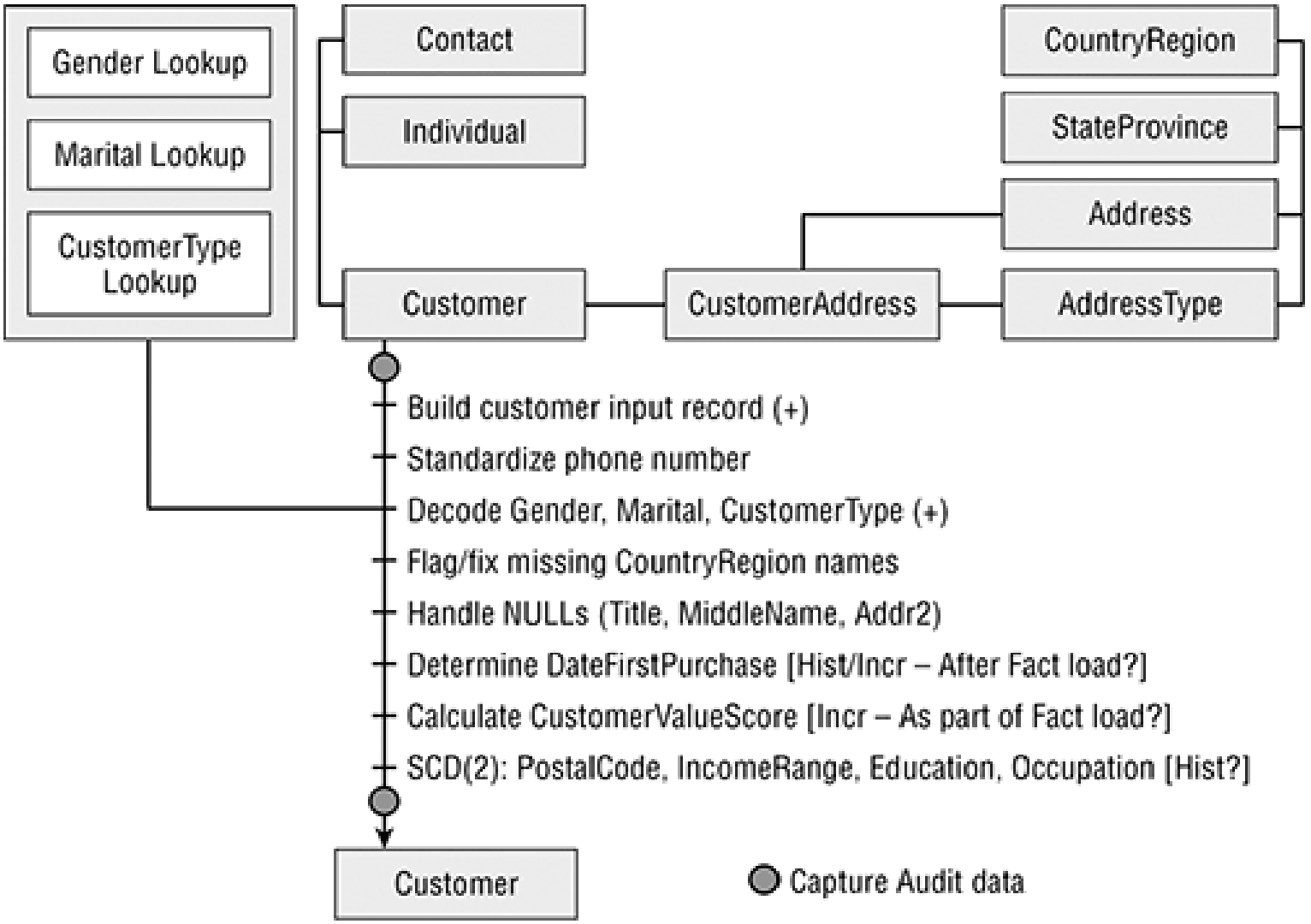
Razvoj ETL procesa

- Odrediti tehnologije koje će se koristiti za ekstrakciju
- Definisati dokaz da je migracija uspešna
- Definisati mesto gde se ispravljaju greške – poželjno je da to bude u izvornom sistemu
- Arhivirati izvorne podatke - Da li potrebno?
- Definisati ETL plan

ETL plan

- ETL plan - specifikacija transformacija za svaki odredišni DW atribut:
  - koji su ulazni podaci
  - funkcija koja transformiše ulazne podatke u DW atribut
- Sprovesti profilisanje podataka
- Sprovesti analizu nedostajućih vrednosti i odrediti kako da se one nadoknade
- Sarađivati sa vlasnicima podataka
- Definisati mehanizme za inkrementalni ETL proces u okviru izvora podataka

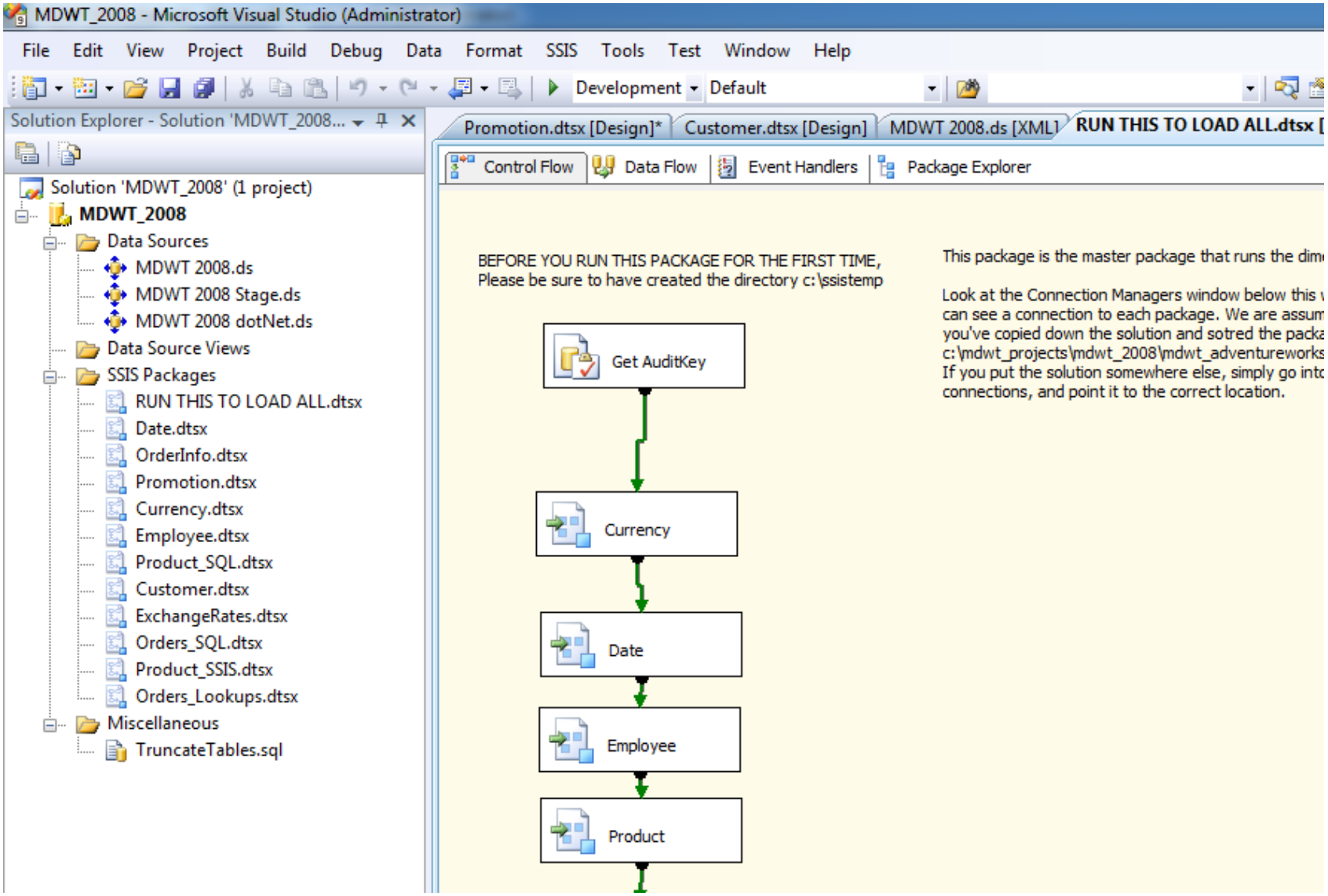
ETL Plan


Gender Lookup
Marital Lookup
CustomerType Lookup
Contact
Individual
Customer
CountryRegion
StateProvince
Address
CustomerAddress
AddressType
Build customer input record (+)
Standardize phone number
Decode Gender, Marital, CustomerType (+)
Flag/fix missing CountryRegion names
Handle NULLs (Title, MiddleName, Addr2)
Determine DateFirstPurchase [Hist/Incr – After Fact load?]
Calculate CustomerValueScore [Incr – As part of Fact load?]
SCD(2): PostalCode, IncomeRange, Education, Occupation [Hist?]
Customer
Capture Audit data

ELT proces

- ETL proces predstavlja skup paketa koji se izvršavaju po definisanom redosledu i/ili u paraleli
- ETL paket ≈ C# Procedura
- Postoji glavni ETL paket koji upravlja izvršavanjem svih ostalih
- Preporuka: jedan paket puni jednu tabelu u DW ili staging bazi
- Odvojeni ETL procesi, jedan za prvi load i jedan za naredne inkrementalne

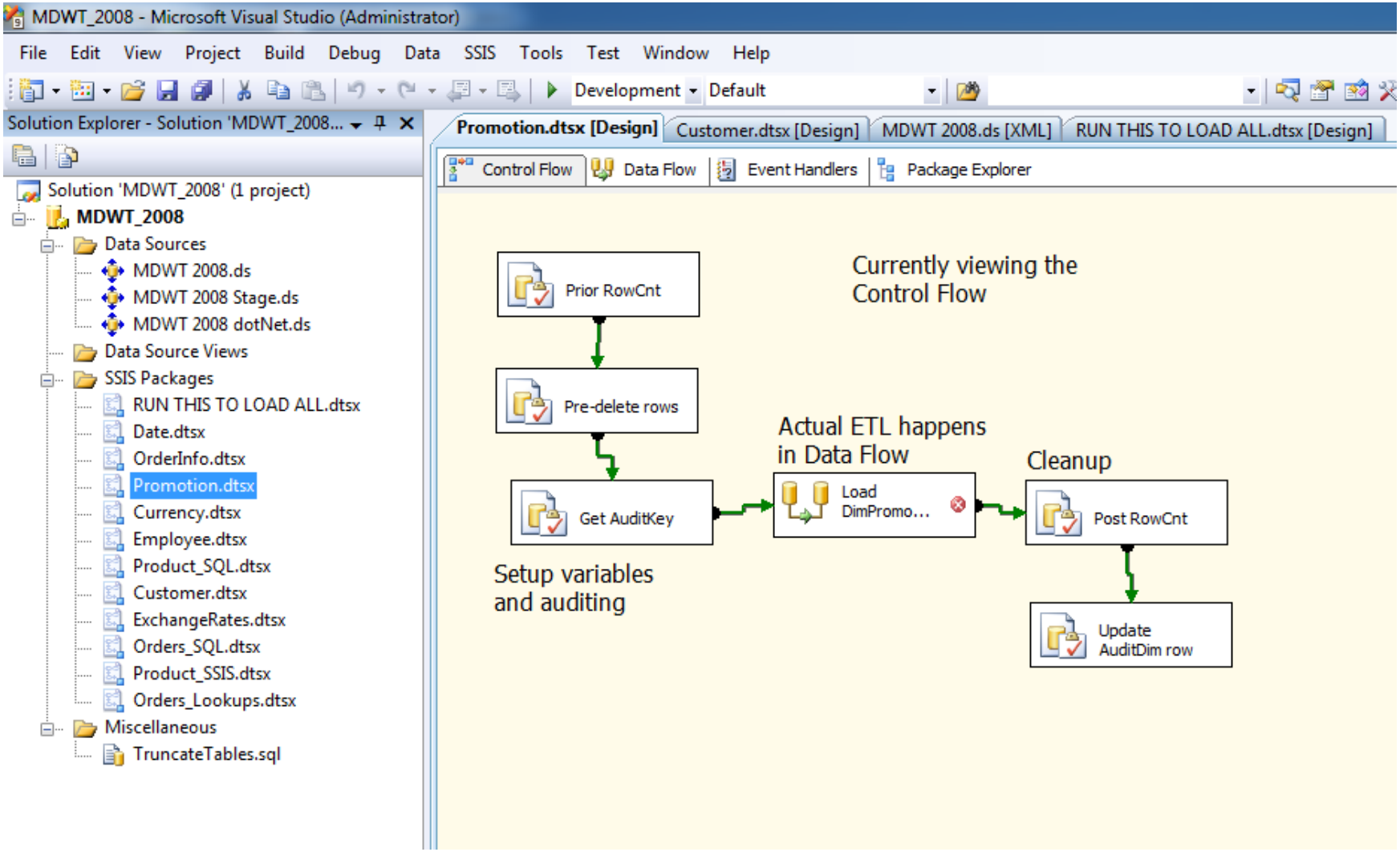
Glavni ETL paket



ETL primeri

- Primere kopirati na lokaciju C:\MDWT_Projects
- Napraviti folder C:\SSIStemp

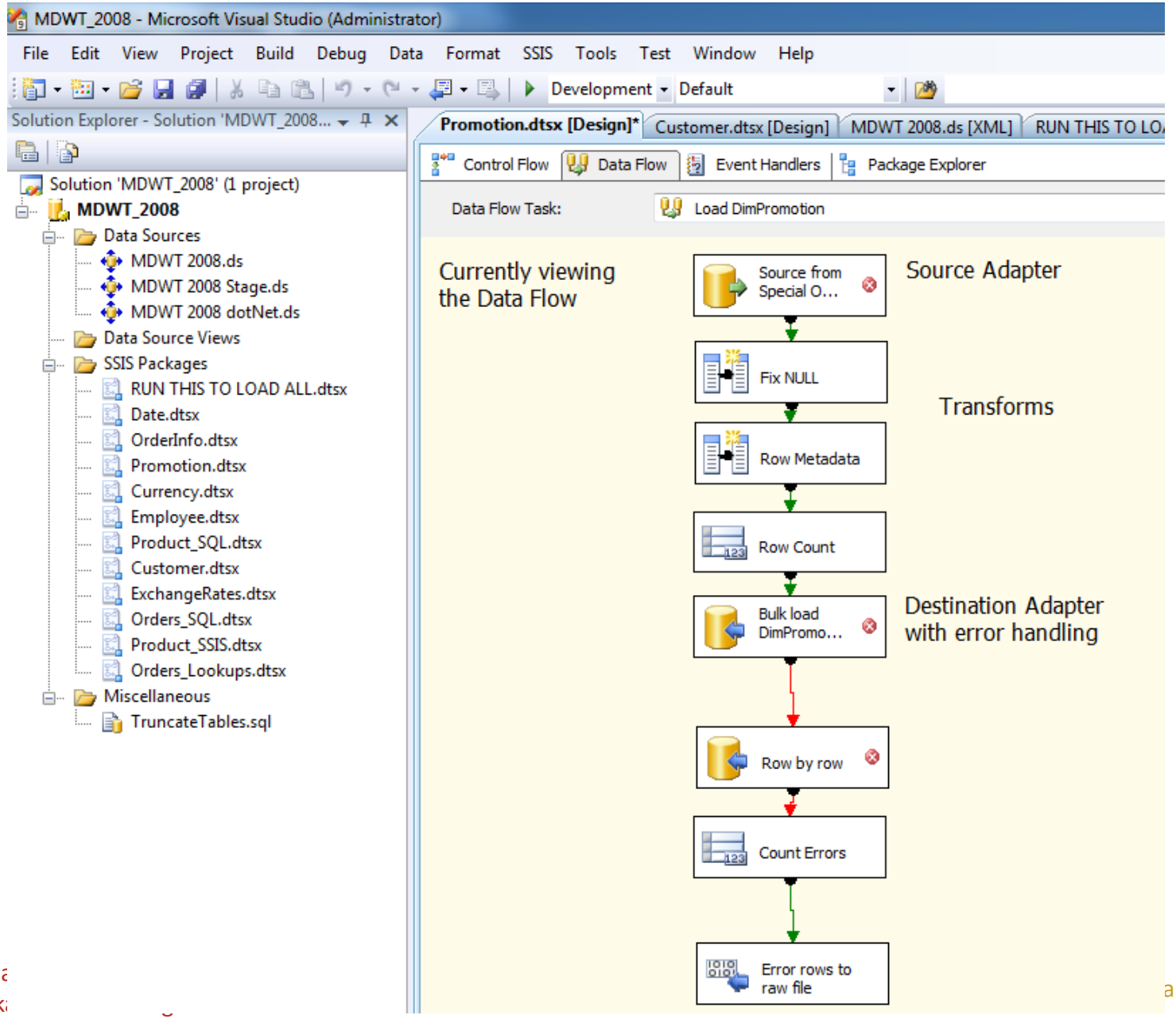
ELT Paket – Control Flow



Control flow

- postoji samo jedan control flow po paketu
- sekvenca taskova
- najkorišćeniji tipovi taskova:
  - Execute SQL
  - Execute package
  - Data flow
  - Script
- Data flow task sadrži logiku transformacije podataka

Data Flow


MDWT_2008 - Microsoft Visual Studio (Administrator)
File Edit View Project Build Debug Data Format SSIS Tools Test Window Help
Development
Default
Solution Explorer - Solution 'MDWT_2008...
Solution 'MDWT_2008' (1 project)
MDWT_2008
Data Sources
MDWT 2008.ds
MDWT 2008 Stage.ds
MDWT 2008 dotNet.ds
Data Source Views
SSIS Packages
RUN THIS TO LOAD ALL.dtsx
Date.dtsx
OrderInfo.dtsx
Promotion.dtsx
Currency.dtsx
Employee.dtsx
Product_SQL.dtsx
Customer.dtsx
ExchangeRates.dtsx
Orders_SQL.dtsx
Product_SSIS.dtsx
Orders_Lookups.dtsx
Miscellaneous
TruncateTables.sql
Promotion.dtsx [Design]*
Customer.dtsx [Design]
MDWT 2008.ds [XML]
Control Flow
Data Flow
Event Handlers
Package Explorer
Data Flow Task:
Load DimPromotion
Currently viewing the Data Flow
Source from Special O...
Source Adapter
Fix NULL
Row Metadata
Transforms
Row Count
Bulk load DimPromo...
Destination Adapter with error handling
Row by row
Count Errors
Error rows to raw file

Data Flow

- Jedan ili više izvora podataka - izvor podataka je prethodni control flow task
- Niz transformacija
- Jedan ili više odredišta podataka – naredni control flow taskovi

Error Flow

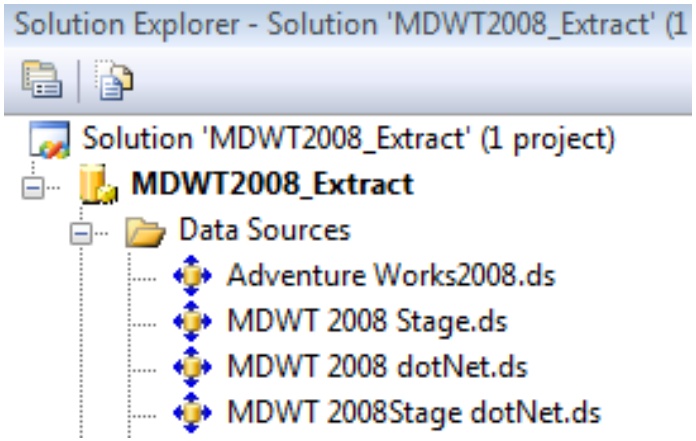
- Alternativni data flow
- Izvor ili transformacija definiše koji slogovi završavaju u error flow
- Primeri:
  - Neuspela konverzija podataka
  - Podaci nezadovoljavaju specificirano ograničenje
- Greška u jednom redu:
  - može da se zaustavi ceo proces
  - da se preskoči red
  - da se ispravi greška

Control i Data Flow taskovi

- MSDN: http://msdn.microsoft.com/en-us/library/ms139892.aspx

Extract Data

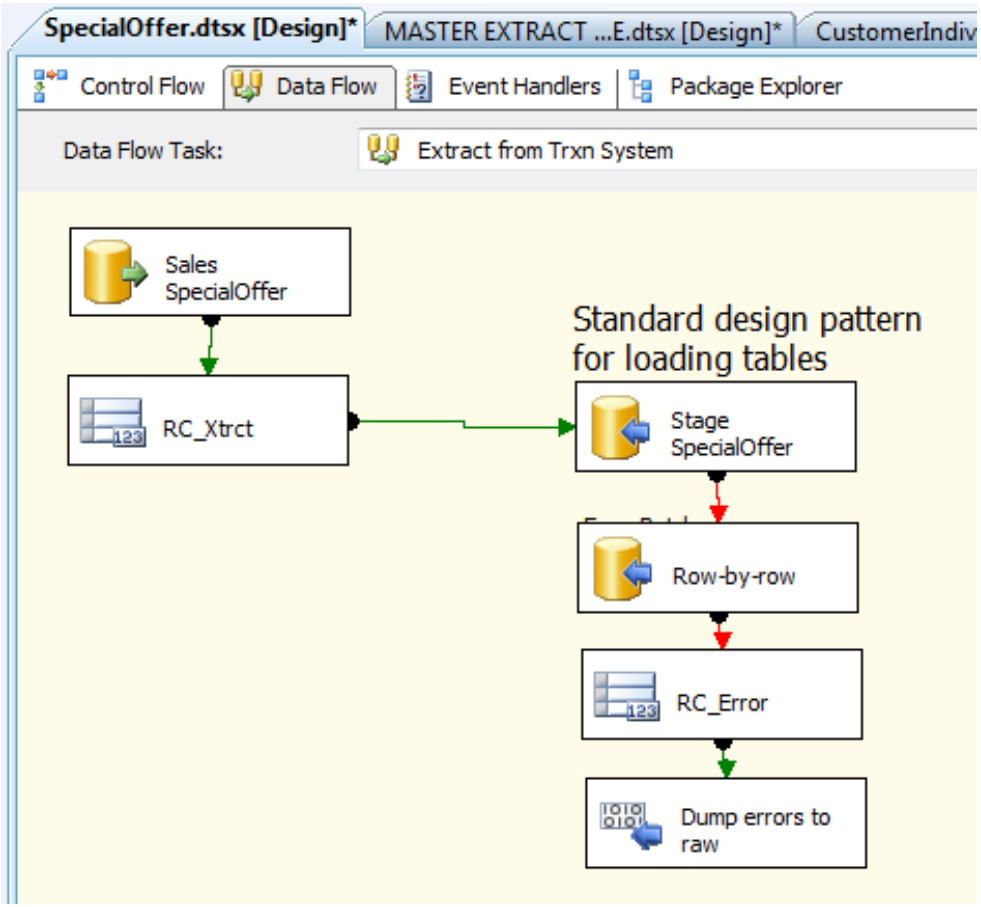
- definisati izvor podataka
- SSIS podržava OLE DB i .Net providere (MS ili od nekog drugog proizvođača)



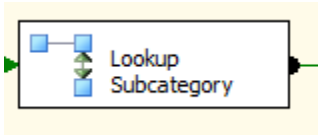
- ako izvor podataka nema provider, najlakše je prvo eksportovati podatke u csv/txt fajlove
- podaci se ektrakuju neizmenjeni u staging bazu ili fajlove
- minimizira se opterećenje izvornog sistema

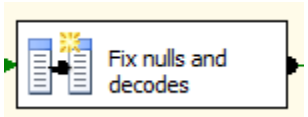
Loading Data

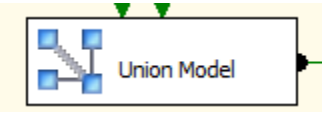
- 2 načina:
  - red po red: sporo
  - bulk: svi redovi odjednom, brzo



Cleaning Data

- sastavni deo data flow taska
- 2 načina čišćenja podataka na nivou kolone:
  - kroz SQL upit za load podataka (Product_SQL.dtsx)
  - upotrebom data flow taskova (Product_SSIS.dtsx):
    - Lookup transformacija 

    - Derived Column transformacija 

    - Union transformacija 

- omogućava da se otkriju problemi u izvorima podataka

Conforming Data

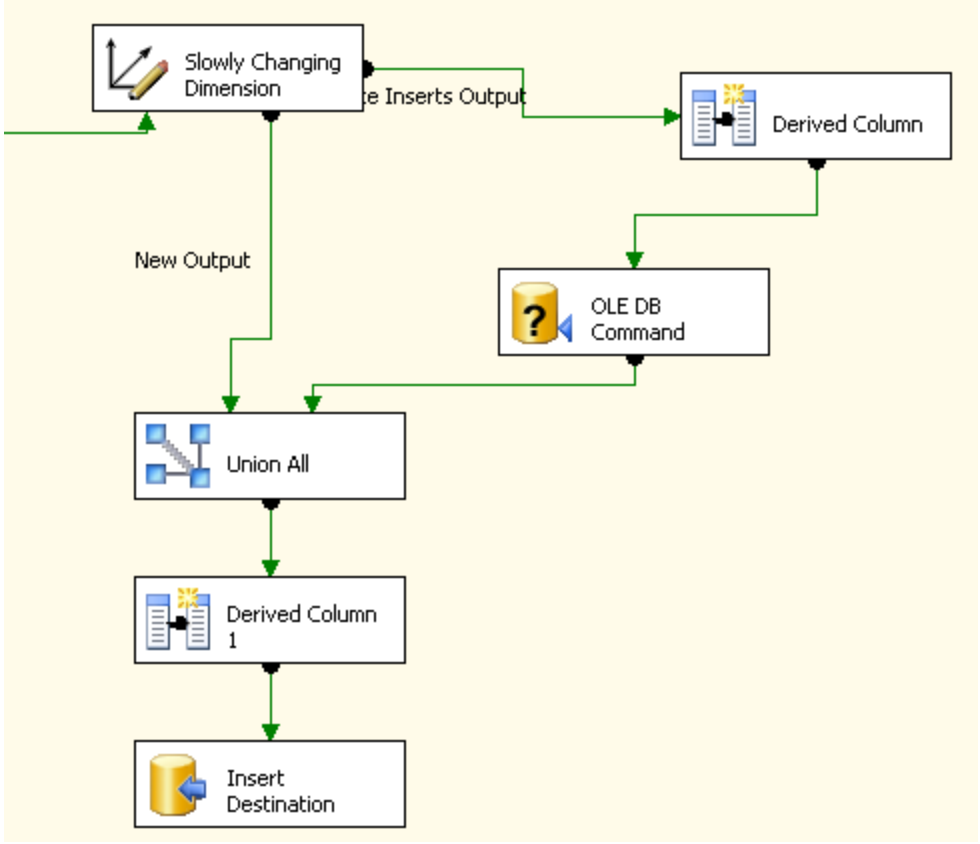
- Usklađivanje podataka između različitih sistema:
  - Detekcija duplikata: Fuzzy Lookup i Fuzzy Grouping transformacije
  - Konsolidacija podataka po tipu, ograničenjima i formatu

Load dimenzija

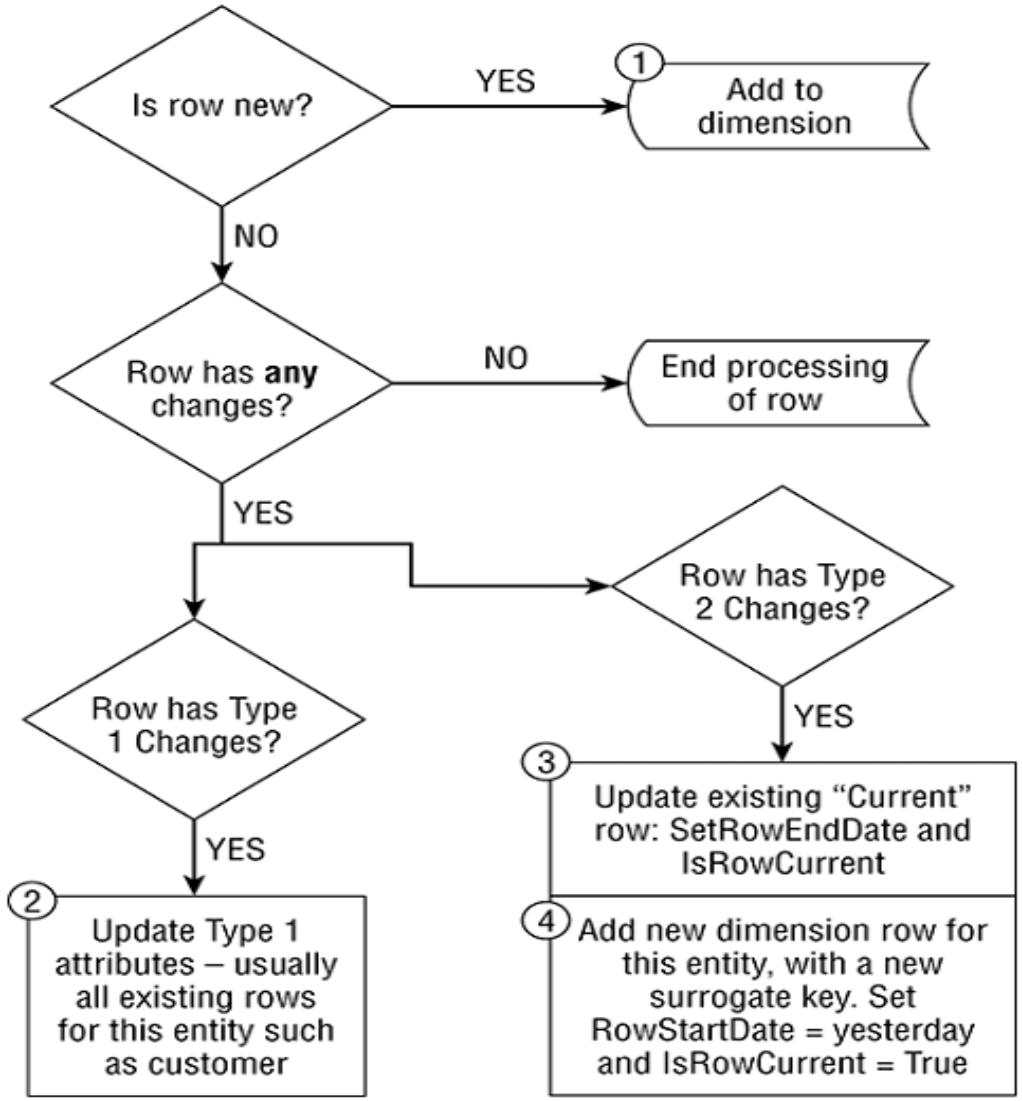
- Implementirati surogatni ključ IDENTITY kolonom
- Hijerarhije se prave spajanjem tabela iz izvora ili se prave u Excelu (npr. vremenska dimenzija)
- Junk dimenzije se prave spajanjem tabela ili iz Excela/skripte
- Inkrementalni load bez SPD:
  - Za male dimenzije – truncate i ponovno punjenje
  - update (OLE DB transformacija)
  - insert u posebnu tabelu pa izvršiti MERGE naredbu

Sporo promenjive dimenzije

- Dodati kolone RowStartDate, RowEndDate (i RowIsCurrent)
- Upotreba wizzarda Slowly Changing Dimension



Sporo promenjive dimenzije



Load činjenice

- Odvojeni ETL paketi, jedan za prvi (bulk) load i jedan za naredne inkrementalne
- (Bulk) insert ili update naredbe nad tabelom činjenica
- Tri tipa činjenice:
  - transakcija
  - presek (snapshot)
  - akumulirana

Load činjenice

- Transaction grain činjenica
  - Bulk load: insert
  - Inkrementalni load: update (OLE DB transformacija) ili insert u posebnu tabelu pa izvršiti MERGE naredbu
- Snapshot činjenica
  - red se ažurira svakodnevno i zamrzne se na kraju perioda – zahteva dve ETL transformacije
  - Inkrementalni load: update (OLE DB transformacija) ili insert u posebnu tabelu pa izvršiti MERGE naredbu
- Akumulirane činjenice
  - Ažuriranje jednog istorijskog podatka može da izazove izmenu na velikom broju redova - uglavnom je brže ponovo load-ovati (delete+insert) poslednjih n meseci

Load činjenice

- Strani surogatni ključevi se dodeljuju Lookup operacijom tokom data flow-a
- Ako lookup ne uspe (odluka zavisi od zahteva):
  - preskočiti red činjenice
  - prebaciti nevalidne redove u posebnu tabelu
  - napraviti placeholder za dimenziju
  - prekinuti obradu
- Primer: Orders_Lookups.dtsx
- Iako su Lookup operacije navedene serijski, SSIS ih automatski paralelizuje

Load činjenice

- Alternativni pristup dodele stranih surogatnih ključeva: SQL upiti sa spojevima tabela
- Koristi se OUTER JOIN da bi se pokupile sve činjenice
- Primer: Orders_SQL.dtsx

Audit dimenzija

- sadrži red za svaki put kada je SSIS paket
- sadrži informacije ko je i kada je pokrenuo paket
- svaka tabela u DW sadrži dva strana ključa ka Audit tabeli:
  - jedan definiše kada je red kreiran i
  - drugi definiše kada je poslednji put ažuriran

Kašnjenje

- Zakasnele dimenzije:
  - napraviti novi red ukoliko je dimenzija tipa 2 i prevezati činjenice
  - StartDate novog reda je timestamp od koga važi nova vrednost dimenzije
  - EndDate novog reda je EndDate stagor reda
- Zakasnele činjenice
  - identifikovati dimenziju u odgovarajuće vremenskom trenutku preko StartDate i EndDate kolona

Error Schema

- Skup tabela koji je namenjen da se loguju greške tokom ETL procesa
- ,,Error'' tabele
  - po strukturi iste sa ciljnim tabelama dimenzione šeme
  - Smeštaju se redovi koji imaju grešku
  - Tabele mogu biti smeštene i u obične fajlove

Materijalizovani pogledi

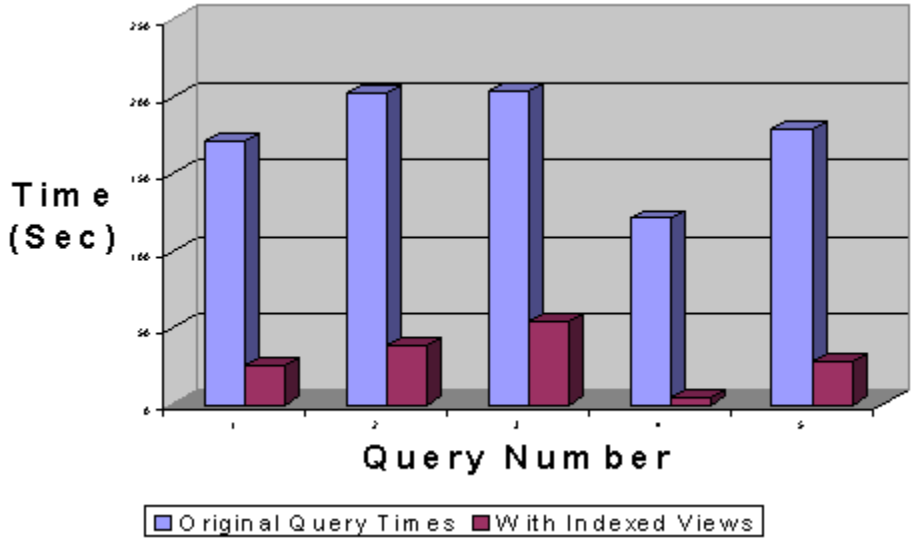
- SQL Server termin - indexed view
- Kreirati unique clustered index nad pogledom:
  `CREATE UNIQUE CLUSTERED INDEX <ind_name> ON <table_name> (<col_name>);`
- Samo kolona `<col_name>` postaje sastavni deo indeksa ali sve kolone pogleda se čuvaju u bazi
- Pogled se automatski ažurira kad se ažuriraju tabele
- Optimizator upita samostalno pronađe i koristi materijalizovane poglede

Kada se kreiraju materijalizovani pogledi?

- Tipovi aplikacija:
  - Data marts
  - Data warehouses
  - OLAP stores and sources
  - Data mining workloads.
- Aplikacije koje imaju sledeće funkcionalnosti:
  - Spojevi i agregacije velikih tabela
  - Često ponavljani kompleksni upiti
  - Ponavljajuće agregacije nad istih skupom kolona
  - Ponavljani spojevi istih tabela nad istim ključevima

Ubrzanje performansi

- MSDN: The represented queries varied in complexity (for example, the number of aggregate calculations, the number of tables used, or the number of predicates) and included large multi-million row tables from a real production environment.



Kada se NE kreiraju materijalizovani pogledi?

- Nad tabelama OLTP sistema gde su česta ažuriranja
- Automatsko održavanje materijalizovanog pogleda postaje preskupo

Materijalizovai pogledi

- Preduslovi:
  - Pogled sme da referencira samo tabele a ne i druge poglede
  - Sve referencirane tabele moraju biti u istoj bazi kao i pogled i imati istog vlasnika
  - Pogled mora biti kreiran sa SCHEMABINDING opcijom, koja povezuje pogled sa šemom kojoj pripadaju tabele
  - Korisničke funkcije koje referencira pogled takođe moraju biti kreirane sa SCHEMABINDING opcijom
  - Tabele i funkcije u definiciji pogleda moraju biti navedene sa nazivom šeme
  - ...