OPEN SOURCE BI FÓRUM 2012

ÜZLETI INTELLIGENCIA A TÖMEGEKNEK

stratis

Ügyfél- és címadatok feldolgozása Talenddel

2012.október 4.

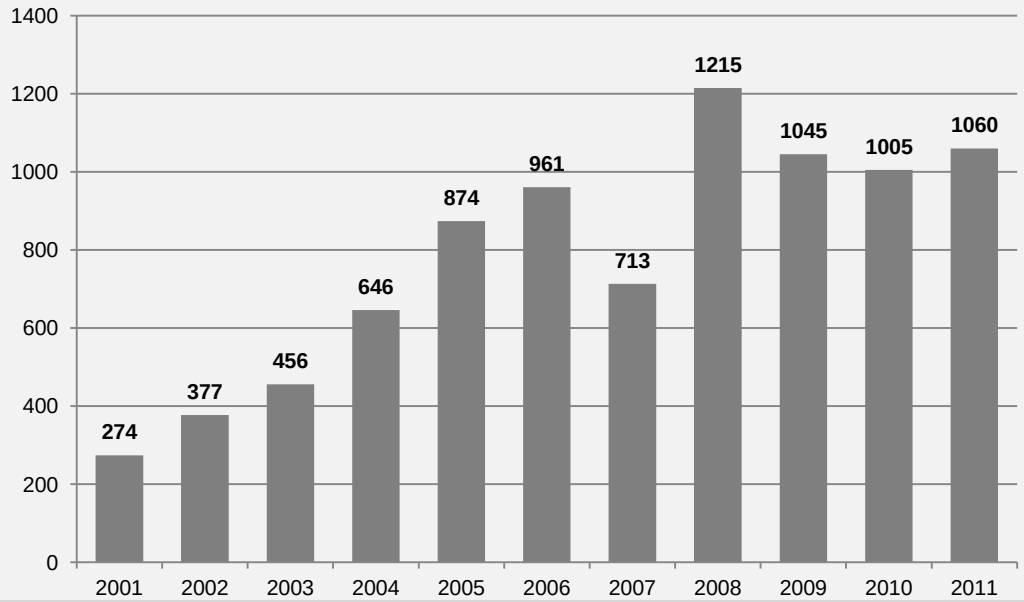
Dr. Miskolczi Mátyás, Kiss György

go the extra mile

A Stratisról röviden

Jellemzők

- Alapítva: 1998
- Tisztán magyar tulajdon
- 50 tanácsadó
- 140 ügyfél
- 500+ projekt



Szolgáltatások

- Üzleti és IT stratégiai tanácsadás
- Folyamatmenedzsment
- Speciális szakértői szolgáltatások

Bővült a Stratis szolgáltatáspalettája

Adatvagyon-gazdálkodás

- Adatintegráció
- Adatminőség-javítás
- Törzsadat-kezelés
- Eszközrendszerek az adatkezelés üzleti folyamatainak szervezéséhez
- Kiemelt szállítói technológiák:

Alkalmazás-szolgáltatások

- Vállalati architektúra (TOGAF)
- Alkalmazás modernizáció
- Cloud integráció
- IT biztonság, kockázatelemzés

Megközelítés

A Stratis hozzáadott értéke

- Komplex megoldás a stratégiai szinttől a technikai megvalósításig
- Tapasztalatok és kompetenciák az üzleti folyamatok és integrációs műveletek terén
- A vállalatvezetés és vezetői információszolgáltatás legjobb gyakorlatának megfelelő eljárások alkalmazása

Mit hoztunk?

- Esettanulmány általános üzleti problémára
- Feladat: adatintegráció és adattisztítás
- Cél: mesteradatok előállítása
- Konkrét technikai megoldás Talend segítségével
- A megoldás alkalmazhatósága: több rendszer ügyféltörzsének / címadatainak integrációja

Alaphelyzet és Cél

Az alap-adatállomány jellemzői

- Ügyféladatok (név, vállalati adatok, kontaktadatok)
- 9 forrásállomány
- 3 séma
- Adatbázis, flatfile vegyesen
- 25-30.000 rekord

A kívánt állapot bemutatása

- Tiszta, megbízható adatok
- Egységes struktúra
- Egy adatállomány (db)
- Szabványos és transzparens feldolgozási folyamat

Szükséges eszköz: Integrált adatkezelési platform

Centralizált metaadat-kezelés

- Könnyen illeszthető a vállalati architektúrába
- Transzparens folyamatok
- Automatikus dokumentációkészítés
- Csoportmunka-támogatás
- Change Data Capture támogatás
- Beépített ütemező

Alacsony fejlesztési költség
Gyors, egyszerű
adatkarbantartási folyamatok

Intuitív grafikus felület

- Intuitív grafikus felület

Gyorsabb, olcsóbb ETL fejlesztések
Könnyű kezelhetőség
Fejlesztői (sql, java) ismeretek nélküli adatműveletek
Előre elkészített alkalmazásrészek
Jobban kommunikálható technikai folyamatok

stratis

Az adatkezelés folyamata Talenddel

go the extra mile

Adatállomásoztatás és adattisztítás

Betöltés

- 3 eltérő struktúrájú adatforrás
- Egyazon egyedhalmazra vonatkozó adatok

Előfeldolgozás

- Az adatok megfelelőségi szabályok szerinti ellenőrzése
- A szabályoktól eltérő adatok tisztítása, vagy a rekord töltésének elvetése
    - Telefonszám
    - Név
    - E-mail cím
    - Cím (településnév validálása/javítása referencia-adatállomány alapján)

Talend Open Studio for Data Quality

Talend Open Studio for Data Integration

Egységes adatstruktúra kialakítása

Parsing

- Egy mezőben tárolt névadatok felbontása elemi szintre (parsing)
- Egy mezőben tárolt címadatok felbontása elemi szintre

Egységes adatállományba való betöltés

- A forrásadatokat részletes bontásban tároló struktúra
- Az eredeti adatok tisztított és kiegészített (gazdagított) verziója
- Csak a megfelelő adattartalmú rekordok betöltése

Talend Open Studio
for Data Quality

Talend Open Studio
for Data Integration

Deduplikáció

Egyezőségvizsgálat

- Egyedi kulcs definiálása
- Egyezőségi kritériumok definiálása (súlyozás, scoring)

Mesteradatok és duplikátumok szétválasztása

- A minőségi kritériumoknak megfelelő adatok áttöltése a mestertáblába
- A duplikátumok elkülönítése
- A duplikátumok mesteradatok közötti „párjának” eltárolása

Talend Open Studio
for Data Quality

Talend Open Studio
for Data Integration

Adatgazdagítás: geokódok és Google címadatok

Címadatok megfelelő átstruktúrálása

- Külön mező a geokódoláshoz szükséges címstruktúra létrehozásához
- tGoogleGeocode éa/vagy tGoogleAddressRow pluginok használata

Google címadatok lekérése

- A Google ingyenes szolgáltatása (napi 2.500 rekordra)
- A Google címadataival pótolhatók az esetleges hiányok, javíthatók a pontatlanságok

Talend Open Studio
for Data Quality

Talend Open Studio
for Data Integration

A tisztított adatállomány jellemzői

Egységes struktúra

- Elemi szintre bontott, mezőszinten azonosított adatelemek
- Leíró metaadatokkal kiegészített rekordok és oszlopok
- Egységes törzsadat-kezelés céljára előkészített állomány

Megbízható adattartalom

- Szabályoknak való megfelelés = 100%
- Az adatok valós tisztasága > 90% (manuális tisztítást nem végeztünk)
- Deduplikált állomány, minden rekord egy példányban szerepel
- Gazdagított adatok külső adatforrásból kiegészítve

Talend Open Studio
for Data Quality

Talend Open Studio
for Data Integration

A fejlesztői munka támogatása

Üzleti felhasználók számára nyújtott előnyök

- Intuitív grafikus felület: átlátható, könnyen megtanulható
- Automatikus kódgenerálás: az egyszerűbb feladatok nem igényelnek fejlesztői ismereteket (pl: SQL, Regex)
- Folyamatszemlélet: az üzleti logika könnyebb implementálása

Fejlesztők számára nyújtott előnyök

- Előre elkészített konnektorok: több mint 500 rendszerhez kész csatoló
- Más alkalmazásokban tárolt logika implementálása: kész eljárások beemelése, meghívása
- Komplex operátorok egyszerű paraméterezéssel testre szabhatók
- A létrehozott eljárások publikálhatók az üzleti felhasználók számára
- Java / Eclipse keretrendszer
- Nyílt java/perl kód generálása
- Automatikus dokumentációkészítés

És végül …

A

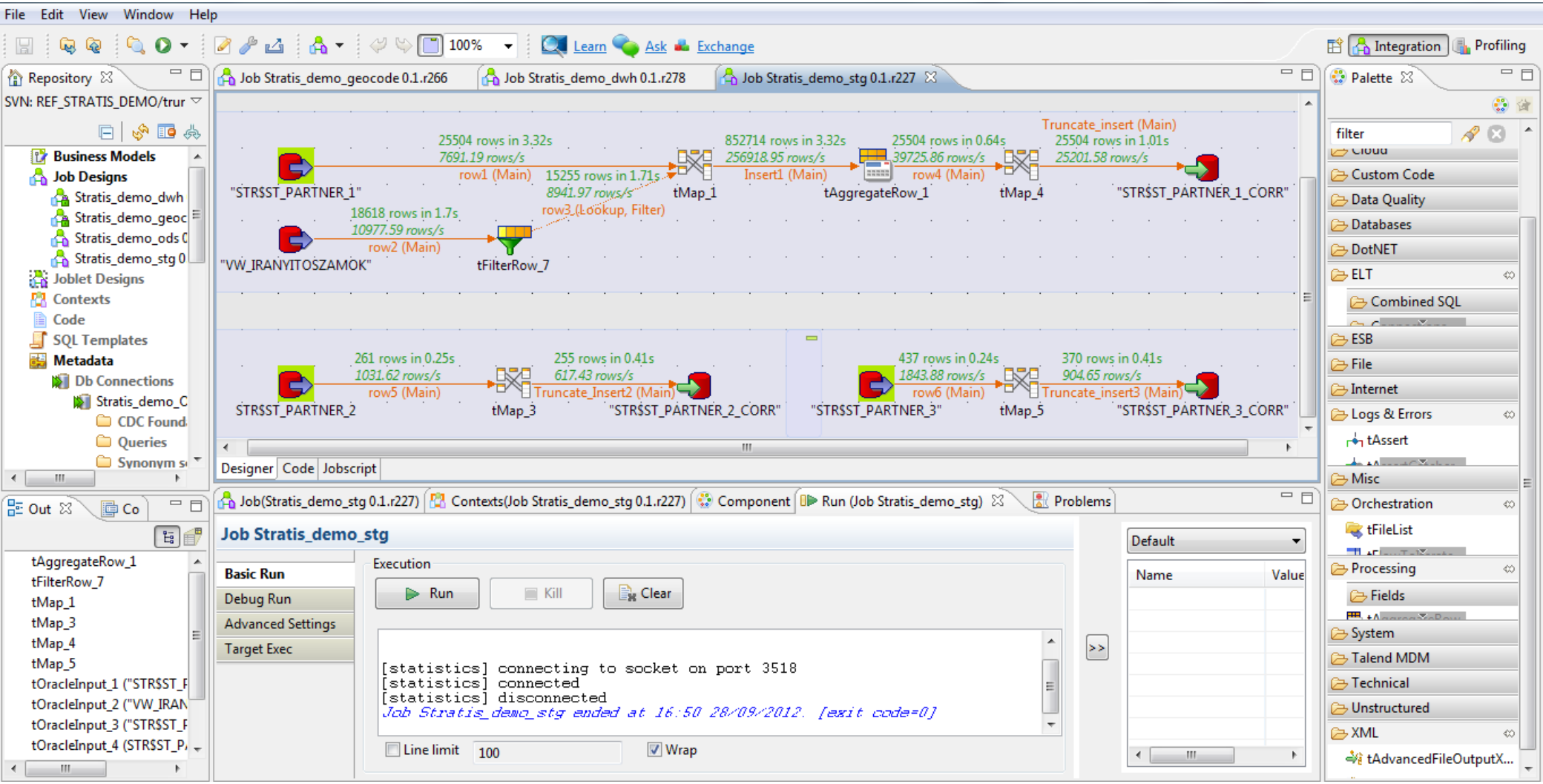
hazai partnere a

stratis

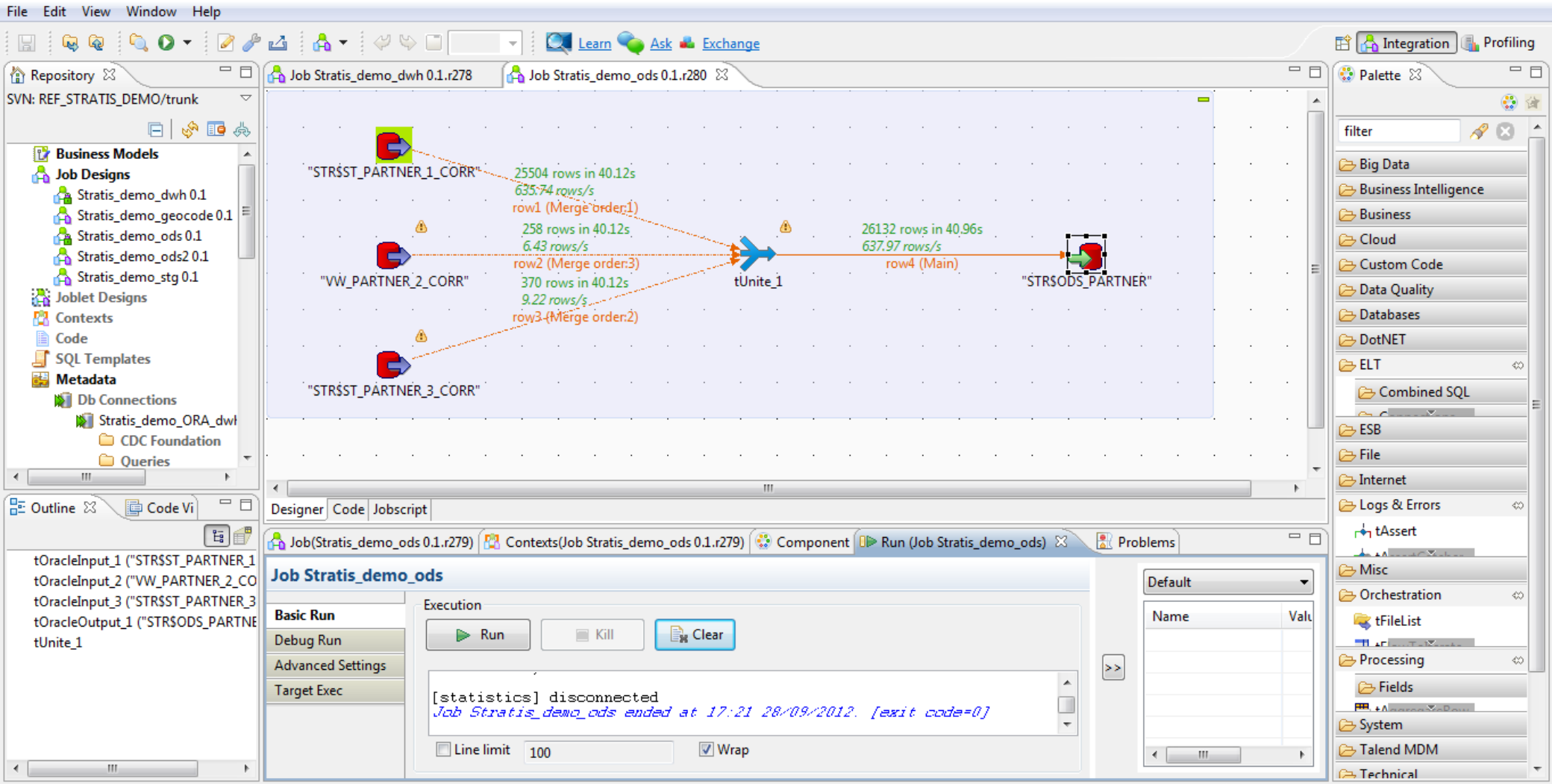
… és most lássuk élőben az adatokat!

go the extra mile

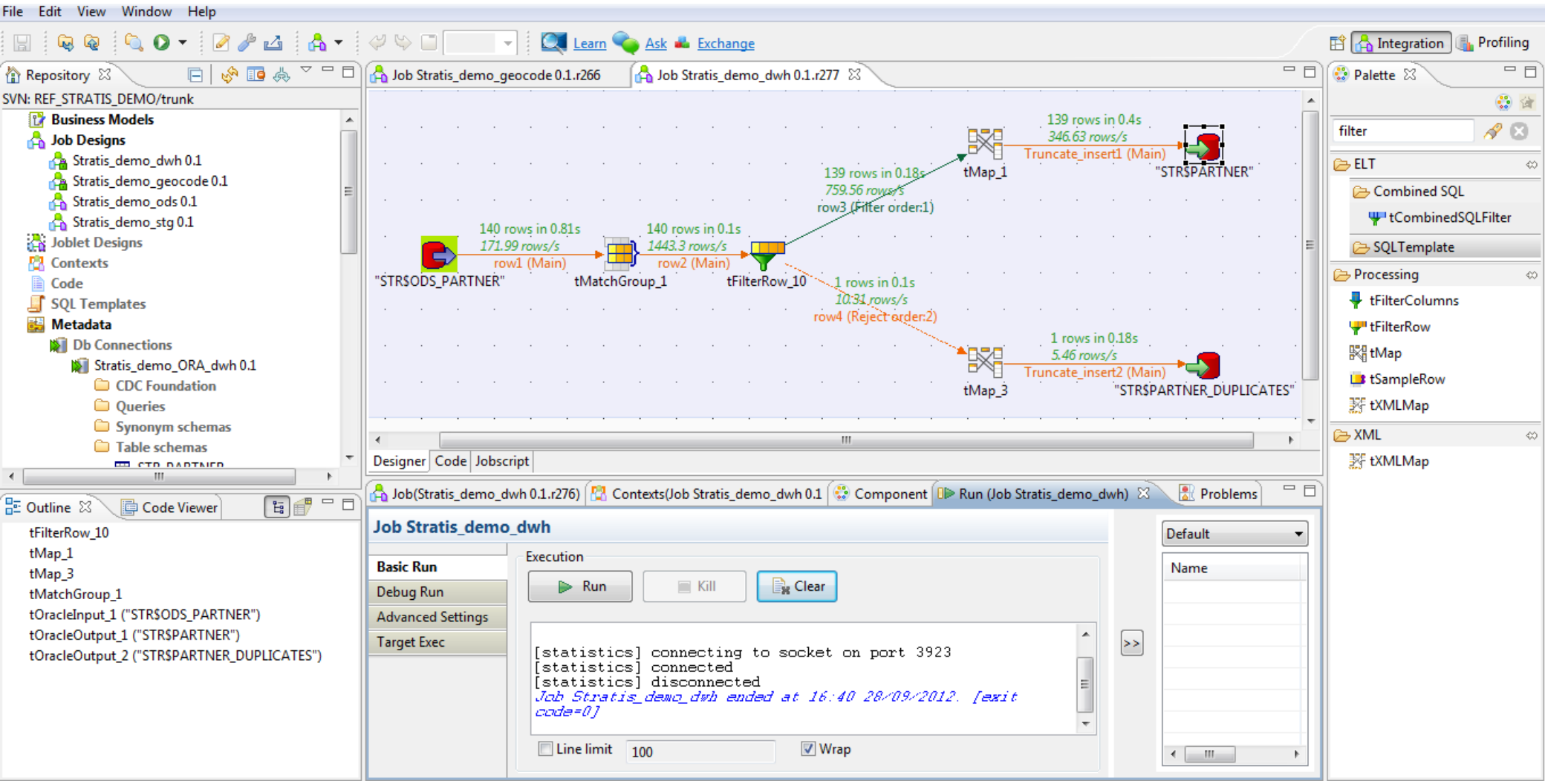
Adatállomásoztatás és adattisztítás



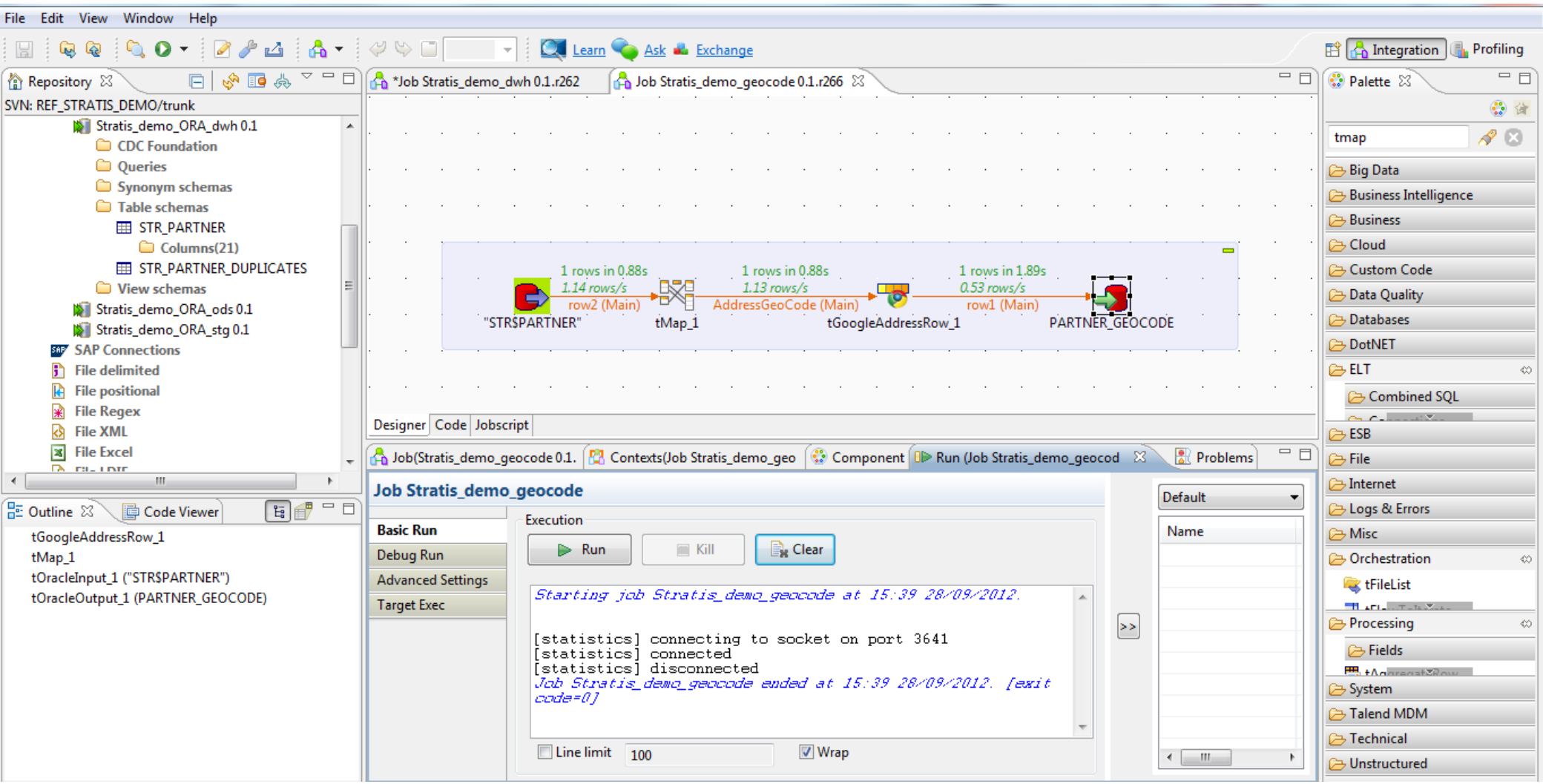
Egységes adatstruktúra kialakítása



Deduplikáció



Adatgazdagítás: geokódok és Google címadatok



stratis

Köszönjük a figyelmet!

go the extra mile