

# Ügyfél- és címadatok feldolgozása Talenddel

2012.október 4.

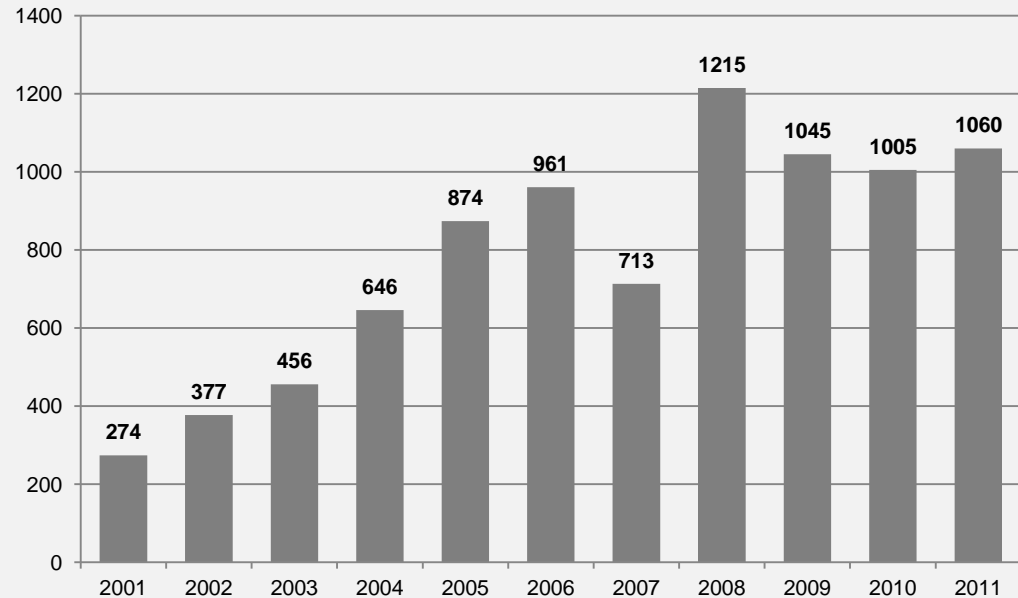
Dr. Miskolczi Mátyás, Kiss György

go the extra mile

# A Stratisről 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:

**ORACLE**<sup>®</sup>

**SAS**<sup>®</sup>

**Microsoft**<sup>®</sup>

**talend**<sup>\*</sup>  
\*open data solutions

## Alkalmazás-szolgáltatások

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

## 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 adatszervezés  
Előre elkészített alkalmazásrészek  
Jobban kommunikálható technikai folyamatok

## Az adatkezelés folyamata Talenddel

# Adatállomásozítás 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 cím adatok

## Cím adatok megfelelő átstrukturálása

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

## Google cím adatok 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)
- Folyamatszemplélet: 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 előben az adatokat!

# Adatállomásozgatás és adattisztítás

The screenshot displays the Stratis ETL Designer interface. The main workspace shows a data flow diagram for the job 'Job Stratis\_demo\_stg 0.1.r227'. The diagram consists of several components and their connections:

- Inputs:** "STRSST\_PARTNER\_1" (25504 rows in 3.32s, 7691.19 rows/s) and "VW\_IRANYITOSZAMOK" (18618 rows in 1.7s, 10977.59 rows/s).
- Transformations:** "row1 (Main)", "row2 (Main)", "row3 (Lóókup, Filter)", "tFilterRow\_7", "tMap\_1", "Insert1 (Main)", "tAggregateRow\_1", "row4 (Main)", "tMap\_4", "Truncate\_insert (Main)".
- Outputs:** "STRSST\_PARTNER\_1\_CORR" (25504 rows in 1.01s, 25201.58 rows/s).
- Other Jobs:** "STRSST\_PARTNER\_2" (261 rows in 0.25s, 1031.62 rows/s) connects to "tMap\_3" and "Truncate\_Insert2 (Main)", which outputs to "STRSST\_PARTNER\_2\_CORR" (255 rows in 0.41s, 617.43 rows/s). "STRSST\_PARTNER\_3" (437 rows in 0.24s, 1843.88 rows/s) connects to "tMap\_5" and "Truncate\_insert3 (Main)", which outputs to "STRSST\_PÁRTNER\_3\_CORR" (370 rows in 0.41s, 904.65 rows/s).

The bottom panel shows the 'Job Stratis\_demo\_stg' execution log:

```
[statistics] connecting to socket on port 3518
[statistics] connected
[statistics] disconnected
Job Stratis_demo_stg ended at 16:50 28/09/2012. [exit code=0]
```

The interface also includes a 'Repository' pane on the left with 'Business Models', 'Job Designs', and 'Metadata' sections. The 'Out' pane on the bottom left lists components like 'tAggregateRow\_1', 'tFilterRow\_7', 'tMap\_1', 'tMap\_3', 'tMap\_4', 'tMap\_5', and 'tOracleInput' components. The 'Palette' on the right contains various components like 'filter', 'tFileList', 'tAssert', and 'tAdvancedFileOutputX...'. The 'Execution' pane on the bottom right shows a table with 'Name' and 'Value' columns.



# Egységes adatstruktúra kialakítása

The screenshot displays the Talend Studio interface for a data integration job. The main workspace shows a job design with three source components on the left: "STRSST\_PARTNER\_1\_CORR", "VW\_PARTNER\_2\_CORR", and "STRSST\_PARTNER\_3\_CORR". These are connected to a central "tUnite\_1" component, which is then connected to a target component "STRSODS\_PARTNER".

Performance statistics for the source components are shown:

- "STRSST\_PARTNER\_1\_CORR": 25504 rows in 40.12s, 635.74 rows/s, row1 (Merge order:1)
- "VW\_PARTNER\_2\_CORR": 258 rows in 40.12s, 6.43 rows/s, row2 (Merge order:3)
- "STRSST\_PARTNER\_3\_CORR": 370 rows in 40.12s, 9.22 rows/s, row3 (Merge order:2)

The "tUnite\_1" component outputs 26132 rows in 40.96s at a rate of 637.97 rows/s, labeled as row4 (Main).

The bottom panel shows the execution log for "Job Stratis\_demo\_ods":

```
[statistics] disconnected
Job Stratis_demo_ods ended at 17:21 28/09/2012. [exit code=0]
```

The interface also includes a left-hand navigation pane with categories like Business Models, Job Designs, and Metadata. A right-hand palette shows various components like tAssert, tFileList, and tUnite. The bottom status bar shows the current job and its execution status.

# Deduplikáció

The screenshot displays the Stratis ETL Designer interface for a job named "Job Stratis\_demo\_dwh 0.1.r276". The main workspace shows a data flow diagram with the following components and metrics:

- STRSODS\_PARTNER**: 140 rows in 0.81s, 171.99 rows/s (row1 (Main))
- tMatchGroup\_1**: 140 rows in 0.1s, 1443.3 rows/s (row2 (Main))
- tFilterRow\_10**: 139 rows in 0.18s, 759.56 rows/s (row3 (Filter order:1))
- tMap\_1**: 139 rows in 0.4s, 346.63 rows/s (Truncate\_insert1 (Main))
- STRSPARTNER**: Target of the main flow.
- tMap\_3**: 1 rows in 0.1s, 10.31 rows/s (row4 (Reject order:2))
- Truncate\_insert2 (Main)**: 1 rows in 0.18s, 5.46 rows/s
- STRSPARTNER\_DUPLICATES**: Target of the reject flow.

The interface includes a left sidebar with a project tree, a bottom status bar with execution logs, and a right sidebar with a component palette.

**Execution Log:**

```
[statistics] connecting to socket on port 3923
[statistics] connected
[statistics] disconnected
Job Stratis_demo_dwh ended at 16:40 28/09/2012. [exit code=0]
```

# Adatgazdagítás: geokódok és Google cím adatok

The screenshot displays a data integration tool interface with the following components:

- Repository:** Shows a tree view of the project structure, including folders like CDC Foundation, Queries, and Table schemas, and tables like STR\_PARTNER and STR\_PARTNER\_DUPLICATES.
- Designer:** Shows a workflow diagram with the following steps:
  - "STRSPARTNER"**: 1 rows in 0.88s, 1.14 rows/s, row2 (Main)
  - tMap\_1**: 1 rows in 0.88s, 1.13 rows/s
  - AddressGeoCode (Main)**: 1 rows in 0.88s, 1.13 rows/s
  - tGoogleAddressRow\_1**: 1 rows in 1.89s, 0.53 rows/s, row1 (Main)
  - PARTNER\_GEOCODE**
- Job Stratis\_demo\_geocode:** Shows the execution log:

```
Starting job Stratis_demo_geocode at 15:39 28/09/2012.
[statistics] connecting to socket on port 3641
[statistics] connected
[statistics] disconnected
Job Stratis_demo_geocode ended at 15:39 28/09/2012. [exit code=0]
```
- Palette:** Lists various integration components like Big Data, Business Intelligence, Cloud, etc.

Köszönjük a figyelmet!