



# Jak chytře čistit data

Případová studie řešení  
v České pojišťovně

Data Quality at a Glance  
20.4.2010

[vladimir.kyjonka@cze.sas.com](mailto:vladimir.kyjonka@cze.sas.com)

**THE  
POWER  
TO KNOW®**

# O co šlo

- Česká pojišťovna
  - Hodně historických dat
  - Data v historickém stavu
  - Data v omezené datové struktuře
- V rámci předchozí studie identifikován masivní zdroj znečištění klientských dat
  - Nejstarší systém s daty o životním pojištění - APO
    - Historická platforma – mainframe, data z děrných štítků
    - Historická data
- Úkol: eliminovat zdroj nečištění

# Situace

- V ČP existuje centrální databáze klientů
  - Data z různých systémů ČP
  - Konsolidovaná a vyčištěná
- Nedostatečná kvalita dat ze systému životního pojištění
  - Znehodnocuje možnost využití centrální DB klientů
  - Standardní používané postupy jsou na data krátké
    - Znečištění je nestandardní a velké
    - Paušální opravy jsou málo spolehlivé

# Požadavky (1)

- Dát do pořádku *velmi* znečištěná data
- Aplikovat *flexibilně* nestandardní požadavky postupy, pravidla pro čištění dat
  - Vysoce customisovaná pravidla „šitá na míru“
- „Špeky“
  - Úmyslně zkrácená data
  - Vynechávání přehlásek (př. “MÜLLER“ → “M LLER“)
  - Nestandardní zápisy (Novák Dan a Mar)
  - Jiné údaje v polích (RČ → IČO)

## Požadavky (2)

- **Zajistit vysokou spolehlivost oprav**
  - Opravy promítány do primárních systémů!
- Zohlednit business význam dat, oprav a jejich dopadů
  - Rozsáhlá diskuse co se může migrovat zpátky
  - Každá oprava robustně označena a sledována
- Křížové opravy
  - Využití vazeb – mezi daty opravenými s různou spolehlivostí (pojistná smlkova ↔ pojistná událost, skupiny záznamů klientů a adres)

# Výběr řešení

- Využití specialisované technologie
  - DataFlux: Otevřenost, flexibilita, transparentnost, osvojitelnost, ...
- Služby konsultantů SAS
  - Společný realizační team ČP + SAS
    - Přístup k SASu k řešení
- Výběrové řízení – Proof of Concept
  - Zadavatel si nejdříve ověřil schopnosti týmu i nástroje
  - Přesvědčivé výsledky
- **Ekonomický faktor**

# Ekonomický rozměr nabídky

- Kromě standardních parametrů (cena práce, SW...)
- Orientace na nákladovou efektivitu procesu čištění
- Vstupní data:
  - 24M záznamů
  - Z toho 15000 organizací (0,03%) , zbytek fyzické osoby (lidi)
- 2 sady (přibližně stejně drahých) alogoritmů a pravidel pro 2 různé entity
- Možné priority
  - Náklady vs. účinnost vs. efektivita
- Návrh: Vyčistit 99,93 % dat za téměř poloviční náklady
  - Nabídnuto variantní řešení v různých kombinací priorit

# Vlastní řešení - základní charakteristiky

- 24M záznamů vstupních dat k čištění
  - Pojistné smlouvy pojistné události
  - Velmi názká úroveň datové kvality
- Navíc 14M záznamů „referenčních dat“
  - Data o klientech z jiných systémů
  - Mají nebo mohou mít lepší kvalitu
- Společný realizační team ČP + SAS
  - SAS: 2 techničtí konsultanti DQ, 1 DQ analytik (part time), PM
  - ČP: Až 10 členů teamu: DQ manager, analytici, tech. specialisté, metodici
- Realisace projektu
  - XII.2009 - III.2010
  - Kompletně na pracovišti ČP

# Hlavní úkoly

- Identifikace dat osob a organizací
- Vyčištění a verifikace adres vůči adresnímu registru (UIR-ADR)
- Opravy jmen (parsing, diakritika, zkratky...)
- Kontroly RČ, IČ
- Standardisace hodnot (jména, příjmení, tituly, adresy ...)
- Detekce pohlaví, opravy „mužatek“
  - př.: „Bartošová Václav“
- Křížové opravy
  - V rámci APO
  - Mimo APO – „referenční data“
- Speciality ...

# Speciality

- Standardisace nestandardně hodně nestandardních dat
- Speciální požadavky ze strany zadavatele
  - Nestandardní algoritmy oprav atd.
- Specifické metody konsolidace dat
- Orientace na vysokou spolehlivost
  - Detekce nejednoznačnosti
  - Ne opravy „za každou cenu“
  - Trackování a indikace (skoring) na základě závažnosti *business významu* dat a dopadů jejich oprav

# Výsledky

- Opravy u více než 90% záznamů
  - Alespoň 1 oprava na záznam
- Výstupy – vysoká kvalita
- Odhaleny dříve neviditelné business problémy, př.:
  - Klienti s neplatným typem pojištění
  - Vhyba v RČ – nesprávná indikace věkové skupiny – nesprávná indexace (pro digitální archiv), sazba ...
- Data „zpětně zašpiněna“ (odborně)
  - Ale propojena na vyčištěná data pro budoucí potřebu
- Nástroj i metodika vlastnictvím ČP pro další využití
- Projekt ukončen úspěšným testováním
- Deklarována vysoká spokojenost zákazníka

# Hovoří DQ Manager České pojišťovny

*Štěpán Čábelka oceňuje zejména:*

- Flexibilní vývoj customisovaných postupů na míru
  - Identifikace organizací, „mužatky“, dešifrování zkratk, velmi nestandardní zápisy, „odborné zašpinění“, ...
- Pečlivost a důslednost
  - Nasazení teamu, maximální pokrytí pravidly
- Spojení chytrých technologií a know-how
  - úplné řešení – efektivní pokrytí výjimek

*„To by s žádným jiným nástrojem nebylo možno udělat...“*

# Shrnutí - závěry

- Ekonomický rozměr
- Nestandardní postup – nadstandardní přínos
- Princip spolehlivosti
- Flexibilita: nástroj, team, metodika





# Jak chytře čistit data

Případová studie řešení  
v České pojišťovně

Data Quality at a Glance  
20.4.2010

[vladimir.kyjonka@cze.sas.com](mailto:vladimir.kyjonka@cze.sas.com)

**THE  
POWER  
TO KNOW®**