profinit.
NEW FRONTIER GROUP

IT řešení, které přináší ovoce

NF SK | NEW FRONTIER SLOVAKIA

Dátové sklady
Dátová kvalita

Miroslav Dávid

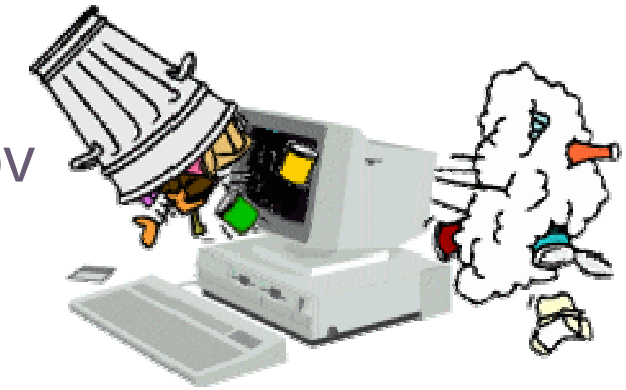
1.10.2010

Pokročilé databázové technológie, FIIT STU

Obsah

Úvod

- V dátach sú **chyby** (a vždy budú)
- Chyby spôsobujú ekonomické straty:
  1. Nesprávne fungovanie business procesov
  2. GARBAGE IN/GARBAGE OUT
     = správa a spracovanie chybných a nadbytočných dát



- Nárast zložitosti systémov → nárast dopadu nekvalitných dát

- Akceptovateľná dátová kvalita
  - Postačujúca pre riadenie business procesov
  - Zvýšenie už neprináša business hodnotu adekvátnu investícii

Dátová kvalita a integrácia

- **Systémová integrácia**
  = prepojenie častí informačných systémov a ztotožnenie IT s business procesmi (DWH, MDM, EAI, ...)
  - Účel: vyššia efektivita, prehľadnosť a pružnosť

- Chyby sa automaticky propagujú z jednej časti do ďalších
  - Nesplnený účel integrácie
  - DWH: chyby sa šíria ETL procesmi

"**70%** projektov dátových skladov má za následok spustenie dodatočného projektu na riadenie kvality dát s celým zmenovým riadením, ktoré je nutné kvôli dodatočnosti takéhoto snaženia" (Gartner)

Dôsledky nekvalitných dát

Data Quality Iceberg

Viditeľné ihneď

Viditeľné neskôr

Šesť dimenzí dátovej kvality

Úplnosť	Uvedené sú všetky potrebné/požadované dáta.
Syntax a formát	Syntax a formát dát sú korektné – zodpovedajú definícii.
Konzistencia	Jednotlivé dátové položky sú jednotne pomenované a chápané.
Presnosť	Dáta poskytujú presný a aktuálny obraz reality.
Jednoznačnosť	Dáta neobsahujú duplicity.
Integrita	Štruktúra dát, vzťahy mezi jednotlivými zdrojovými dátami (databázovými tabuľkami) a atribúty sa konzistentne udržujú a rozvíjajú.

Presnosť nie je vždy možné určiť iba podľa dátovej množiny, ale je potrebné porovnanie s dôveryhodným referenčným zdrojom.

Príklad

NEW FRONTIER SLOVAKIA

CUST. ID	CUSTOMER NAME	BUSINESS TYPE	CUSTOMER ADDRESS	CUSTOMER CITY	CUST. STATE	CUST. ZIP	CUSTOMER PHONE	CUSTOMER FAX	LAST IN-VOICE DATE
90	WENDY WOLF	BUSINESS	3345 GATEW AY DR	INDIANAPOLIS	IN	46224	3172913421		30.8.2001
109	NANCY BUNKER	PERSON	APT A 4556 WATERWAY	BROAD RIPPLE	IN	47950	3174262323		10.2.2004
12	MARYS GIFT SHOP	PERSON	435 MAIN ST	DANVILLE	IL	47978	3178567221	3178523434	1.2.2003
21	MARGANS CANDIES	BUSINESS	5657 W TENTH ST	INDIANAPOLIS	IN	46234	3172714398		14.9.1999
221	JANE DUNNE	PERSON	2337 S SHELBY ST	INDIANAPOLIS	IN	47834	3175634402		21.3.2003
232	LESLIE GLEASON	PERSON	798 HARDRAW AY DR	INDIANAPOLIS	IN	47856	317545690		15.6.2004
287	DAVID G LACEY	PERSON	9880 ROCKVILLE RD	INDIANAPOLIS	IN	46244	3172719991	3172719992	19.3.2002
288	JOSEPH LEDDIN	PERSON	567 US 31	WHITELAND	IN	49980	3178879023		4.12.2000
333	JASONS AND DALL	BUSINESS	INDIANAPOLIS	9 FIR CROVE	IN	46222	3172978886	3172978887	
345	ANGELO DOBKO	PERSON	42 CAMUS AVENUE	LEBANON	IN	49967	7658970090		
43	SCHYELERS NOVELTIES	BUSINESS	17 MAPLE ST	LEBANON	IN	48990	3174346758		25.5.1999
121	MRS JANE DUNNE	PERSON	2337 SHELBY STREET	INDIANAPOLIS	IN	47834	3175634402		30.7.2003
432	SCOTTYS MARKET	BUSINESS	43 LONDON ROAD	BROWNSBURG	IN	45687	3178529835	3178529836	8.9.2003
560	ANDYS CANDIES	BUSINESS	9 CARLISLE PARK	NASHVILLE	IN	48756	8123239871		14.9.2004
590	EDWARD LEDESMA	PERSON	409 SHADELAND AVENUE	INDIANAPOLIS	IN	43278	3175456768		19.2.2004
610	RAGANS HOBBIES	BUSINESS	451 GREEN	PLAINFIELD	IN	46818	3178393441	3178399090	10.2.2004
1021	ANNA LEDESMA	PERSON	409 SHADELAND AV	NDIANAPOLIS	IN	43278	3175456768		15.3.2004

úplnosť syntax a formát konzistencia jednoznačnosť integrita

Zdroje nekvality dát

- Chyby pri zadávaní dát
  - Vznikajú nepresnosťou alebo nevedomosťou front-end používateľov
  - Možno minimalizovať prednastavenými štandardizovanými hodnotami alebo formátom

Ján Nov8k

Gaštanová 10	Bratislava – Staré mesto	811 02
Gaštanová 10	Bratislava – Staré mesto, 811 02	

12-02-2008
18.10.2008

- Nedokonalosť business procesov
- Nedokonalosť technologických postupov

Riadenie kvality dát

- Poskytnutie infraštruktúry na transformáciu dát do akceptovateľnej kvality
- Iteratívne sa opakujúci proces

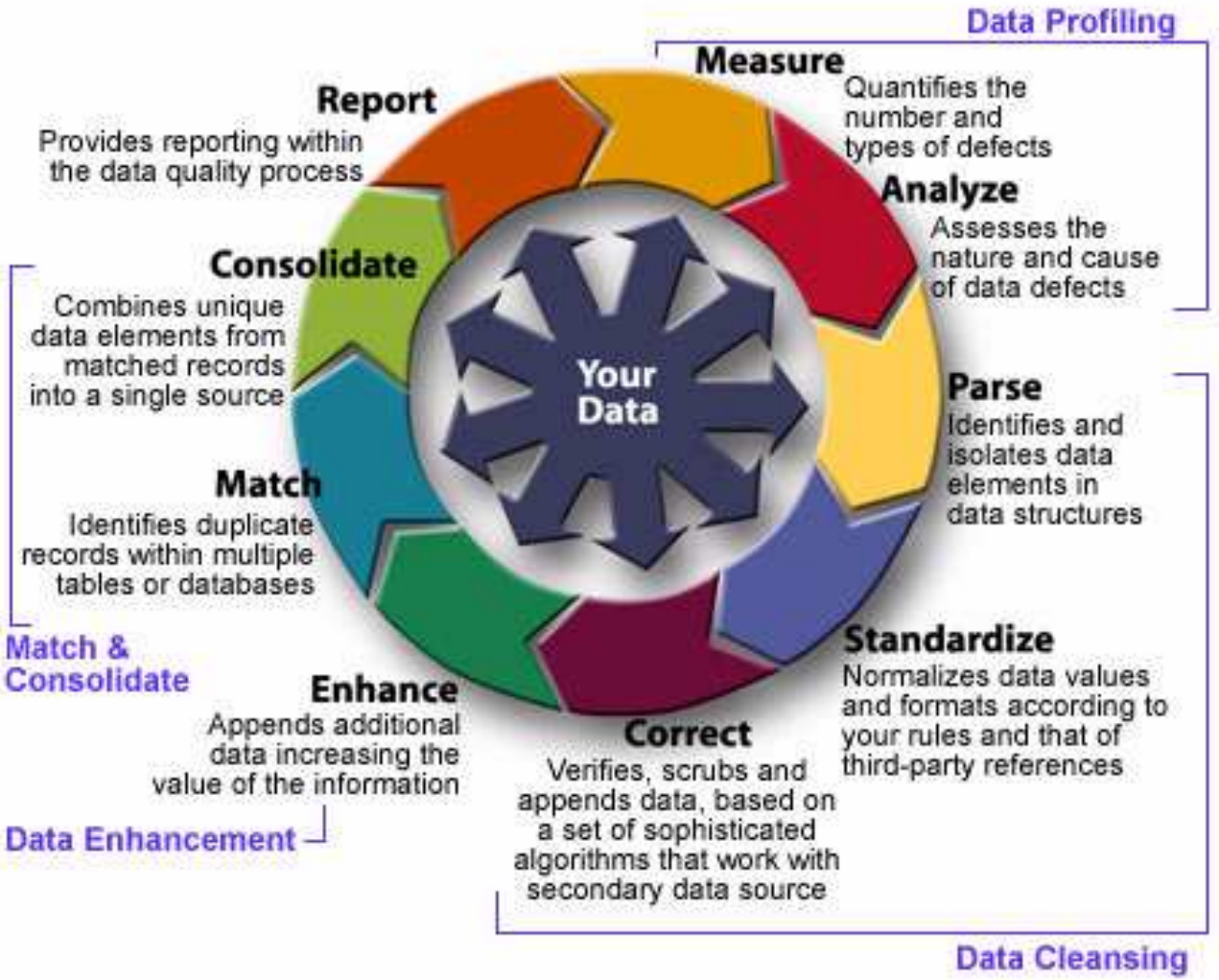
Zložky:

1. Metodika riadenia kvality dát
   - Množina procesov, ktoré merajú kvalitu dôležitých dát a zvyšujú ju na akceptovateľnú
   - Zaručuje, že business procesy a aplikácie závislé na dátach dávajú očakávané výsledky
2. Softwarový nástroj
   - Analýza, čistenie, kontrola

Kroky zvyšovania dátovej kvality

Konsolidácia
- presnosť
- jednoznačnosť
- integrita

Porovnávanie
- analýza pre konsolidáciu



Profilácia dát
- analýza pre čistenie

Čistenie dát
- úplnosť
- syntax a formát
- konzistencia

Profilácia dát

Business pravidlá, metriky pre dátovú kvalitu atribútov

Vytvorenie plánu profilácie
= pravidlá pre identifikáciu dátových chýb na úrovni tabuliek a atribútov

- Analýza frekvencie distribúcie hodnôt
- Analýza formátu hodnôt
- Analýza chýbajúcich hodnôt
- Detekcia hraničných hodnôt
- ...

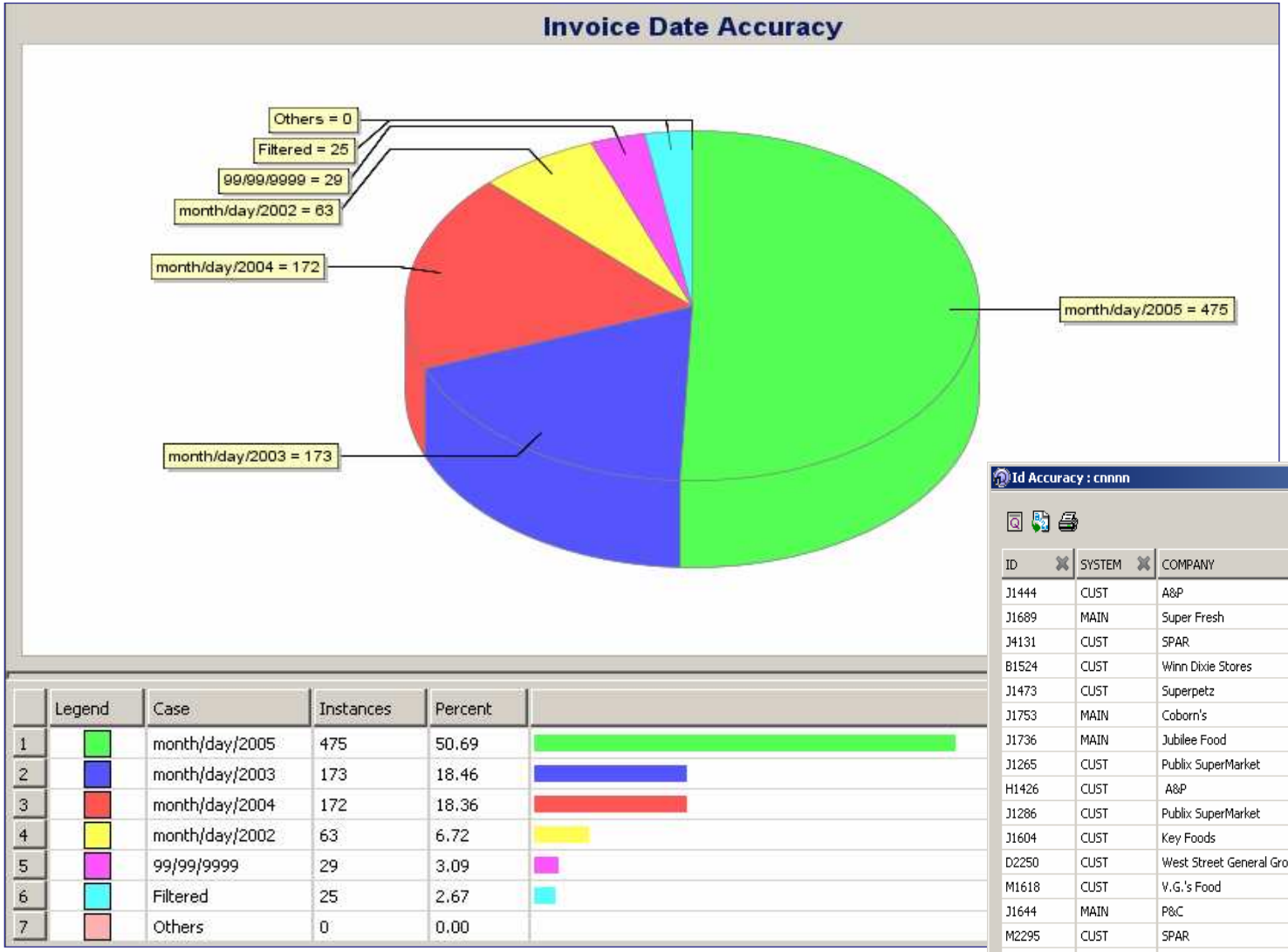
Nízko kvalitné dáta

Spustenie plánu profilácie

- Podrobné reporty kvality
- Súhrnné reporty kvality

Extraho-vané dáta

Profilácia dát - výstup



	Legend	Case	Instances	Percent
1		month/day/2005	475	50.69
2		month/day/2003	173	18.46
3		month/day/2004	172	18.36
4		month/day/2002	63	6.72
5		99/99/9999	29	3.09
6		Filtered	25	2.67
7		Others	0	0.00

Atribút Invoice Date:

- 50,69% month/day/2005
- 18,46% month/day/2003
- 18,36% month/day/2004
- 3,09% 99/99/9999
- 2,67%
- 0,00% iné hodnoty

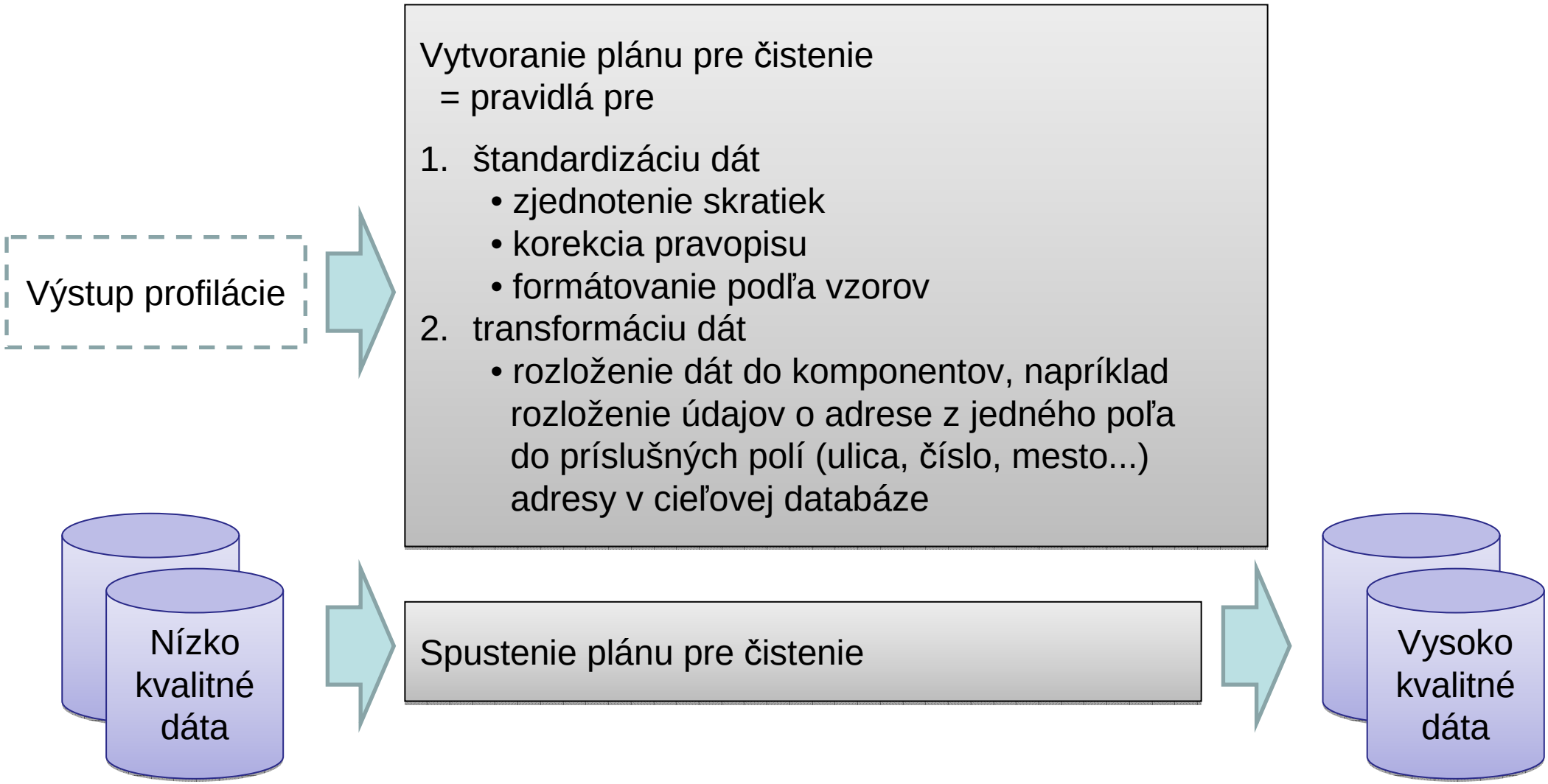
Id Accuracy : cnnnn

ID	SYSTEM	COMPANY	ADDR1	ADDR2	ADDR3	ADDR4
J1444	CUST	A&P	110	Washington St	Morristown	7960
J1689	MAIN	Super Fresh	300	S Best Ave	Walnutport	18088
J4131	CUST	SPAR	51	Old Springville Rd	Penrith	CA10 1JG
B1524	CUST	Winn Dixie Stores	7851	Palm River Road	TAMPA	33619
J1473	CUST	Superpetz	302	S Us Highway 11 And 15	Selinsgrove	17870
J1753	MAIN	Coborn's	2150	Dakota Ave S	Huron	57350
J1736	MAIN	Jubilee Food	1530	East St	Idaho Falls	83401
J1265	CUST	Publix SuperMarket	4851	Whitesburg Dr S Ste B	Huntsville	35802-162
H1426	CUST	A&P	614	Clinton St	Hoboken	7030
J1286	CUST	Publix SuperMarket	1020	Us Highway 27 S	Avon Park	33825 510
J1604	CUST	Key Foods	2722	E Tremont Ave	Bronx	10466
D2250	CUST	West Street General Grocers	185	West St	Sheffield	S1 4EW
M1618	CUST	V.G.'s Food	8503	Davison Rd	Davison	48423
J1644	MAIN	P&C	61	Market Lane	Bradford	5033
M2295	CUST	SPAR	6	Patrax	LA NUCIA	3530
M1337	CUST	Shop 'n Save	3411	Nameoki Commons	Granite City	62040
M2251	CUST	Wharfedale Value & Convenience	458	Southcoates Lane	Hull	HU9 3UA

Load complete

Records: 24 Page 1 of 1

Čistenie dát



Porovnávanie

Business pravidlá, metriky pre dátovú kvalitu atribútov

Vytvorenie plánu pre porovnávanie

- Porovnanie dát v rámci jednej množiny dát → identifikácia duplikátov a súvisiacich záznamov
- Porovnanie dát v rôznych množinách dát → spájanie množín dát, pripojenie nových dát

Nízko kvalitné dáta

Spustenie plánu porovnávania

- Podrobné reporty kvality
- Súhrnné reporty kvality

Extraho-vané dáta

Porovnávanie dát - výstup

ID	ID_1	COMPANY_1	ADDR1_1	ADDR2_1	ADDR3_1	ADDR4_1	COUNTRY_CODE_1	Edit Distance	Best Match
Cluster 1									
92	1345	THE CORNER STORES	971	MID STR S NUTFIELD, REDHILL		RH1 4JH	GB	0.80971	779
779	2224	THE CORNER STORES	102	MID STR S NUTFIELD	REDHILL	RH1 4JH	GB	0.80971	92
Cluster 2									
518	1891	KWIKSAVE	B8	GEORGES SHOPPING CNTR	GRAVESEND	DA11 0TA	GB	0.952021	890
890	5239	KWIKSAVE	B8	GORGE SHOPPING CNTR	GRAVESEND	DA11 0TA	GB	0.952021	518
Cluster 3									
28	1278	PUBLIX SUPERMARKET	951	N STATE RD 434	ALTAMONTE SPRINGS	32714-7026	US	0.898187	29
29	1279	PUBLIX SUPERMARKET	851	S STATE RD 434	ALTAMONTE SPRINGS	32714-4811	US	0.898187	28
Cluster 4									
822	2270	ALCAMPO	114	SANTA COLOMBA	BARCELONA	8030	ES	1	908
856	2322	ALCAMPO		SANTA COLOMA 114	BARCELONA	8030	ES	0.942057	907
907	8876	ALCAMPO		SANTA COLOMBA, 114	BARCELONA	8030	ES	0.942057	856
908	9221	ALCAMPO	114	SANTA COLOMBA	BARCELONA	8030	ES	1	822

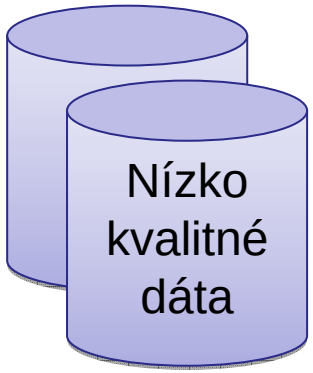
- Vytvorenie zoskupení (clusters) s podobnými záznamami
- Podobnosť je definovaná napr. pomocou "edit distance"

Konsolidácia

Výstup porovnávania

Vytvorenie konsolidačného plánu
= pravidlá pre

- spojenie duplicitných záznamov
- vytvorenie prepojenia súvisiacich dát
- pridanie dát z referenčného zdroja
- náhradenie nepresných dát referenčnými dátami



Spustenie konsolidačného plánu

Na záver

- Nekvalitné dáta stoja americké firmy ročne **600 miliárd** dolárov
- Na základe auditu jedna európská firma objavila, že nevystavila faktúru na **4%** objednávok – čo predstavovalo **80 miliónov** dolárov (DM Review)
- V roku 1992 sa v ČR vrátilo **96 000** daňových preplatkov späť z dôvodu nedoručiteľnej adresy
- Nesprávne uvedené ceny v databázi obchodných reťazcov stoja ročne amerických zákazníkov **2,5 miliard** dolárov na preplatkoch

Úloha

- Vymyslieť kroky pre zlepšenie kvality tabuľky CUSTOMER
  1. profilácia, porovnávanie (analýza) - akým spôsobom identifikovať uvedené chyby?
  2. čistenie, konsolidácia – akým spôsobom upraviť dáta, aby boli bez chýb?

Príklad

NEW FRONTIER SLOVAKIA

CUST. ID	CUSTOMER NAME	BUSINESS TYPE	CUSTOMER ADDRESS	CUSTOMER CITY	CUST. STATE	CUST. ZIP	CUSTOMER PHONE	CUSTOMER FAX	LAST IN-VOICE DATE
90	WENDY WOLF	BUSINESS	3345 GATEW AY DR	INDIANAPOLIS	IN	46224	3172913421		30.8.2001
109	NANCY BUNKER	PERSON	APT A 4556 WATERWAY	BROAD RIPPLE	IN	47950	3174262323		10.2.2004
12	MARYS GIFT SHOP	PERSON	435 MAIN ST	DANVILLE	IL	47978	3178567221	3178523434	1.2.2003
21	MARGANS CANDIES	BUSINESS	5657 W TENTH ST	INDIANAPOLIS	IN	46234	3172714398		14.9.1999
221	JANE DUNNE	PERSON	2337 S SHELBY ST	INDIANAPOLIS	IN	47834	3175634402		21.3.2003
232	LESLIE GLEASON	PERSON	798 HARDRAW AY DR	INDIANAPOLIS	IN	47856	317545690		15.6.2004
287	DAVID G LACEY	PERSON	9880 ROCKVILLE RD	INDIANAPOLIS	IN	46244	3172719991	3172719992	19.3.2002
288	JOSEPH LEDDIN	PERSON	567 US 31	WHITELAND	IN	49980	3178879023		4.12.2000
333	JASONS AND DALL	BUSINESS	INDIANAPOLIS	9 FIR CROVE	IN	46222	3172978886	3172978887	
345	ANGELO DOBKO	PERSON	42 CAMUS AVENUE	LEBANON	IN	49967	7658970090		
43	SCHYELERS NOVELTIES	BUSINESS	17 MAPLE ST	LEBANON	IN	48990	3174346758		25.5.1999
121	MRS JANE DUNNE	PERSON	2337 SHELBY STREET	INDIANAPOLIS	IN	47834	3175634402		30.7.2003
432	SCOTTYS MARKET	BUSINESS	43 LONDON ROAD	BROWNSBURG	IN	45687	3178529835	3178529836	8.9.2003
560	ANDYS CANDIES	BUSINESS	9 CARLISLE PARK	NASHVILLE	IN	48756	8123239871		14.9.2004
590	EDWARD LEDESMA	PERSON	409 SHADELAND AVENUE	INDIANAPOLIS	IN	43278	3175456768		19.2.2004
610	RAGANS HOBBIES	BUSINESS	451 GREEN	PLAINFIELD	IN	46818	3178393441	3178399090	10.2.2004
1021	ANNA LEDESMA	PERSON	409 SHADELAND AV	NDIANAPOLIS	IN	43278	3175456768		15.3.2004

úplnosť syntax a formát konzistencia jednoznačnosť integrita

Diskusia

NEW FRONTIER SLOVAKIA