Extracting Useful Data from Free-form Text files

Michael Wong
TELUS

Vancouver SAS Business Analytics Forum
November 2, 2011

Problem

- Telephone Directory Data from our legacy mainframe was supplied as text files
- Some of the columns in these files contain information that was free-formatted
- Needed to understand what type of information can be gained from the text files

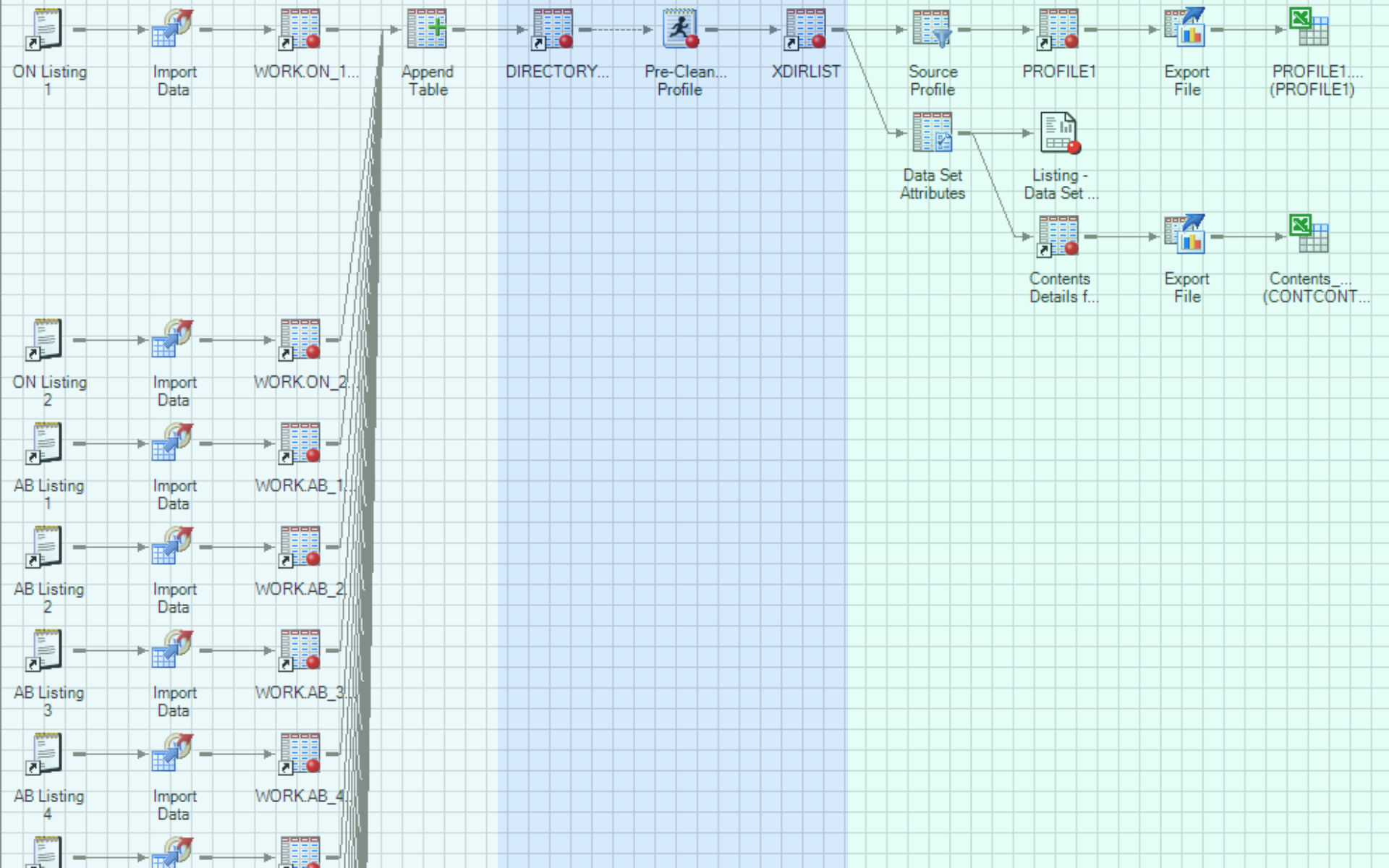
Solution

- Import text files into SAS data set
- Standardize Content
- Quantify and Qualify Data

Import text files into SAS data set

Standardize Content

Quantify and Qualify



Standardize Content

```
*************************************************************;
***************Pre-Cleansed Profile SAS Code*************;
*************************************************************;
DATA sastabls.xdirlist (RENAME = (Column1 = ServiceProvide
                                  Column2 = SPID
                                  Column3 = ListingType
                                  Column4 = TelephoneNumber
                                  Column5 = CustomerName1
                                  Column6 = CustomerName2
                                  Column7 = AddrNoStreet
                                  Column8 = AddrLocality
                                  Column9 = AddrPostalCode
                                  Column10 = Count));
SET sastabls.directory_listing_201106;
LENGTH  TN_Format       $015.
                CustName1
                CustName2
                Addr1Street
                Addr2Loc
                Addr3PC         $010.
                Area_code       $003
                TN_Prov         $002;

Area_code = SUBSTR(Column4,1,3);
IF Area_code IN ('250','604','778') THEN TN_Prov = 'BC';
ELSE IF Area_code IN ('403','587','780') THEN TN_Prov = 'AB';
ELSE IF Area_code IN ('226','289','416','519','613','647','705','807','905')
        THEN TN_Prov = 'ON';
ELSE TN_Prov = 'ER';

Column10 = 1 ;
TN_Format = LEFT(Column4);
TN_Format = TRANWRD(TN_Format,'0','x');
TN_Format = TRANWRD(TN_Format,'1','x');
TN_Format = TRANWRD(TN_Format,'2','x');
TN_Format = TRANWRD(TN_Format,'3','x');
TN_Format = TRANWRD(TN_Format,'4','x');
TN_Format = TRANWRD(TN_Format,'5','x');
TN_Format = TRANWRD(TN_Format,'6','x');
TN_Format = TRANWRD(TN_Format,'7','x');
TN_Format = TRANWRD(TN_Format,'8','x');
TN_Format = TRANWRD(TN_Format,'9','x');
IF Column5 NE ' ' THEN CustName1 = 'NOTBLANK';
ELSE CustName1 = 'BLANK';
IF Column6 NE ' ' THEN CustName2 = 'NOTBLANK';
ELSE CustName2 = 'BLANK';
IF Column7 NE ' ' THEN Addr1Street = 'NOTBLANK';
ELSE Addr1Street = 'BLANK';
IF Column8 NE ' ' THEN Addr2Loc = 'NOTBLANK';
ELSE Addr2Loc = 'BLANK';
Addr3PC = UPCASE(Column9);
Addr3PC = LEFT(COMPRESS(Addr3PC,COMPRESS(Addr3PC,'ABCDEFGHIJKLMNOPQRSTUVWXYZ0123456789')));
IF SUBSTR(Addr3PC,1,1) IN ('A','B','C','D','E','F','G','H','I','J','K','L','M','N','O','P','Q','R','S','T','U','V','W','X','Y','Z')
        THEN SUBSTR(Addr3PC,1,1) = 'X';
IF SUBSTR(Addr3PC,2,1) IN ('1','2','3','4','5','6','7','8','9','0') THEN SUBSTR(Addr3PC,2,1) = 'X';
IF SUBSTR(Addr3PC,3,1) IN ('A','B','C','D','E','F','G','H','I','J','K','L','M','N','O','P','Q','R','S','T','U','V','W','X','Y','Z')
        THEN SUBSTR(Addr3PC,3,1) = 'X';
IF SUBSTR(Addr3PC,4,1) IN ('1','2','3','4','5','6','7','8','9','0') THEN SUBSTR(Addr3PC,4,1) = 'X';
IF SUBSTR(Addr3PC,5,1) IN ('A','B','C','D','E','F','G','H','I','J','K','L','M','N','O','P','Q','R','S','T','U','V','W','X','Y','Z')
         THEN SUBSTR(Addr3PC,5,1) = 'X';
IF SUBSTR(Addr3PC,6,1) IN ('1','2','3','4','5','6','7','8','9','0') THEN SUBSTR(Addr3PC,6,1) = 'X';

RUN;
```

Standardize Content

- Parsed the raw text files visually into identifiable columns in a single SAS data set

- Identified Free-formatted columns

- Created additional columns for descriptive & analytic purposes

Columns

Name	Type	Length	Format	Informat	Label
ServiceProvide	Character	15	$15.0	$15.0	Column1
SPID	Character	7	$7.0	$7.0	Column2
ListingType	Character	3	$3.0	$3.0	Column3
TelephoneNumber	Character	15	$15.0	$15.0	Column4
CustomerName1	Character	30	$30.0	$30.0	Column5
CustomerName2	Character	313	$313.0	$313.0	Column6
AddrNoStreet	Character	96	$96.0	$96.0	Column7
AddrLocality	Character	35	$35.0	$35.0	Column8
AddrPostalCode	Character	6	$6.0	$6.0	Column9
TN_DIR	Character	10	$10.0	$10.0	10-Digit version of TelephoneNumber
AREA_CODE	Character	3	$3.0	$3.0	Area Code of TelephoneNumber
TN_PROV	Character	2	$2.0	$2.0	Province of TelephoneNumber

Quantify & Qualify: Column Content Formats

Format Type			Listing Type			Grand Total
			Bus	Gov	Res	
TN_Format: xxx-xxx-xxxx			390771	2310	3109734	3502815
All records have standard 10-digit format						
	CustName1	CustName2	Bus	Gov	Res	Grand Total
	BLANK	BLANK			16	16
	BLANK	NOTBLANK	3	26	5	34
	NOTBLANK	BLANK	9844	91	142	10077
	NOTBLANK	NOTBLANK	380924	2193	3109571	3492688
			97.5%	94.9%	100.0%	99.7%
Over 90% of the records have content in the Customer Name fields						
Addr1Street	Addr2Loc	Addr3PC	Bus	Gov	Res	Grand Total
BLANK	BLANK	1X1			1	1
BLANK	BLANK	XXXXXX	568		21191	21759
NOTBLANK	NOTBLANK	XXXXXX	307327	590	1811488	2119405
			78.8%	25.5%	58.9%	61.1%
NOTBLANK	NOTBLANK	0X2X6X			1	1
NOTBLANK	NOTBLANK	2TXXXX			1	1
NOTBLANK	NOTBLANK	2X0			1	1
NOTBLANK	NOTBLANK	X			2	2
NOTBLANK	NOTBLANK	XSXXXX			1	1
NOTBLANK	NOTBLANK	XX5XXX			1	1
NOTBLANK	NOTBLANK	XX8A6			2	2
NOTBLANK	NOTBLANK	XXXK1W			1	1
NOTBLANK	NOTBLANK	XXXN1A			1	1
NOTBLANK	NOTBLANK	XXXN3S			1	1
NOTBLANK	NOTBLANK	XXXR1J	1			1
NOTBLANK	NOTBLANK	XXXR2H			1	1
NOTBLANK	NOTBLANK	XXXX8X			2	2
NOTBLANK	NOTBLANK	XXXXXO			1	1
NOTBLANK	NOTBLANK	XXXXXZ			1	1
NOTBLANK	NOTBLANK	BLANK	27360	1720	188734	217814
BLANK	BLANK	BLANK	55515		1088303	1143818
			85.8%	100.0%	65.0%	67.3%
65% or more of the records have content in the Address fields with 60% having the standard 6-character Postal Code format						
Grand Total			**390771**	**2310**	**3109734**	**3502815**

Quantify & Qualify: Directory Customer Name Inconsistencies

CBU_CID_COMP1	CBU_CID_COMP2	Listing Type	CustomerName1	CustomerName2
		B	Autosense	
		B	Evancic	Perrault Robertson Ltd bankrptcy trustees
SISTERS OF MERCIFUL JESUS		B		Of Merciful Jesus Sister
CURVES FOR WOMEN		B	Curves	
GREEK ORTHODOX COMMUNITY CHURCH		B	Greek	Orthodox Church Of St Demetrois
PUSCH	CULTURES UNITED LTD	B	Pusch	
ROYAL LEPAGE	612287 BC LTD	B	Royal	LePage Coronation West Realty
FOREST PRACTICES BOARD		G		C Province of
FOREST PRACTICES BOARD		G	Fraser Health Authority	
PLACE DES ARTS		G	Coquitlam	City of
PR GEO HOSPITAL MAINTENANCE DIAL UP	NORTHERN HEALTH	G	Northern Health: Omineca	
PR GEO HOSPITAL MAINTENANCE DIAL UP	NORTHERN HEALTH	G	Northern	Health Authority
		R	Lenkewich	J
PANDA TANK & VAC TRUCK SERVICES INC		R	Swanberg	
PLACE DES ARTS		R	MacDonell	D
	0 / A PLEASE			

Large amount of space between start and end words

Business customer name split

TELUS Customer Name different from Directory Customer Name

Residential have Last Name first then First Name last

Quantify & Qualify: Addr1Street

Addr1Street
1 Main St
1-1425 Main
9900 Carleton
9900 Carleton St
Bag 2
Bsmt 908 17 Av SW
Ctge 3247 Lefeuvre
Elv1 160 W 3rd
Frnt 619 Lakeshore Dr
Rear 45865 Hocking
Rm 304 11808 St Albert Tr
Rm308 8180 Macleod Tr SE
Sturgeon Industrial Park
Sturgeon Industrial Pk
Trlr 8710 Horton Rd SW
Unit A5 10160 152 St
Upb 202 Centre St SE
Upr 95 Queens
Uprlv 8127 Fraser Av
Upstrs 24 Northmount Dr NW
Utlty 675 W Hastings
Village Square Mall NE
Vlge At Pigeon Lake
W Old N Thompson Hwy
W2 430 Stewart
West On 132 Av
Wetaskiwin- Direct Line (No Charge) Dial
acty-1330 Pinetree Way
bsmt 10230 95 St
dwnstrs 137 Banff Av
front 2018 20 Av
ft Pemberton
lwr 770 Bernard
pent2-1061 Fort
rear 18341 Fraser Hwy

Some select examples that highlight the variability due to the free form text nature of the raw data source

- Non-Canada Post Standard Format for Street Number, Name and spelling
- Using abbreviations

Quantify & Qualify: Addr2Loc – City Name

Raw Source (Addr2Loc)	Upper Cased	Canada Post Format	Total
Agassiz	AGASSIZ	AGASSIZ	1
AGASSIZ-HARRISON	AGASSIZ-HARRISON	Not Matched	21
Agasz	AGASZ	Not Matched	106
Alb	ALB	Not Matched	1
ALBERTA BEACH	ALBERTA BEACH	ALBERTA BEACH	24
Ald	ALD	Not Matched	422
Aldrsyde	ALDRSYDE	Not Matched	1
Black Creek	BLACK CREEK	BLACK CREEK	3
Black Diamond	BLACK DIAMOND	BLACK DIAMOND	2
BLACKFALDS	BLACKFALDS	BLACKFALDS	61
Blckflds	BLCKFLDS	Not Matched	1
Blk Dmd	BLK DMD	Not Matched	2
Blk Dmnd	BLK DMND	Not Matched	45
Rd Deer	RD DEER	Not Matched	6
Rd Deer Cnty	RD DEER CNTY	Not Matched	6
RED DEER	RED DEER	RED DEER	2502
RED DEER 2	RED DEER 2	Not Matched	5
Red Deer County	RED DEER COUNTY	RED DEER COUNTY	1
Vancouver	VANCOUVER	VANCOUVER	1
VCR	VCR	Not Matched	17123
VIC	VIC	Not Matched	2526
Victoria	VICTORIA	VICTORIA	1
York	YORK	YORK	2
YORK REGION	YORK REGION	Not Matched	2
ZAMA	ZAMA	Not Matched	8
Not Matched to Canada Post Format			52974
Grand Total			157257

- 33.7% of theAddr2Loc were not matched to the Canada Post Format
- Some select examples that highlight the variability due to the free form text nature of the raw data source i.e. Non-Canada Post Standard for municipality name and spelling

Quantify & Qualify: CustomerName2

CustomerName2			
Imaging Productions Inc		fax line	
Canada		fax line	
Wood Centre		fax line	
Glass Western Ltd		fax line	
fax line			
Derek		fax line	Dr
David Constituency Office	fax line		MLA
Foundation		fax line	
& Company		fax	
Buck Oilfield Services Ltd		faxline	
s Equestrian Inc		fax line	Sister

- The Directory Listing Customer Name fields combined is 343 characters long
- the CustomerName2 field not only contains the second part of the customer name but also additional information
- Above is an example where the listing is designated as a Fax line in the Directory Listing
- This was discovered while visually investigating the source data
- A quick query searching for the string ‘fax’ gained 1602 records
- Further Text Mining would be needed to fully uncover additional information

Summary of Useful Data

- All of the supplied Telephone Number were in a format that required minimal adjusting
- Approximately 10% of the records have blank customer Name fields
- 65% or more of the records have content in the Address fields
- 40% of the records do not have the standard 6-character Postal Code format
- The Customer Name and Addresses will require extensive data manipulation and verification
- A visual perusal of the raw data is useful and may indicate an opportunity for text mining

Post-Analysis Decisions

- The supplied Telephone Number is flagged in the Business Marketing Data Mart as the primary contact phone number for marketing purposes
- The supplied Telephone Number is used to match against other company data sources to provide additional customer metrics for the Business Marketing Data Mart