



Matching For Money!

Adventures with SAS Dataflux

Leslie.Urquhart@Bell.ca

Let me take you back.....

The Telephone Way to a Happier Day

Try it today when the dishes are done, beds made, clothes in the washer. You've earned a break.

So relax a little and pick up the telephone. Enjoy a cheerful visit with a friend or loved one.

It's so easy to do, whatever the miles may be. For no one is ever far away by telephone.

It helps to make any day a happier day at both ends of the line.



"It's fun to phone"

BELL TELEPHONE SYSTEM



Data-Flux Capacitor



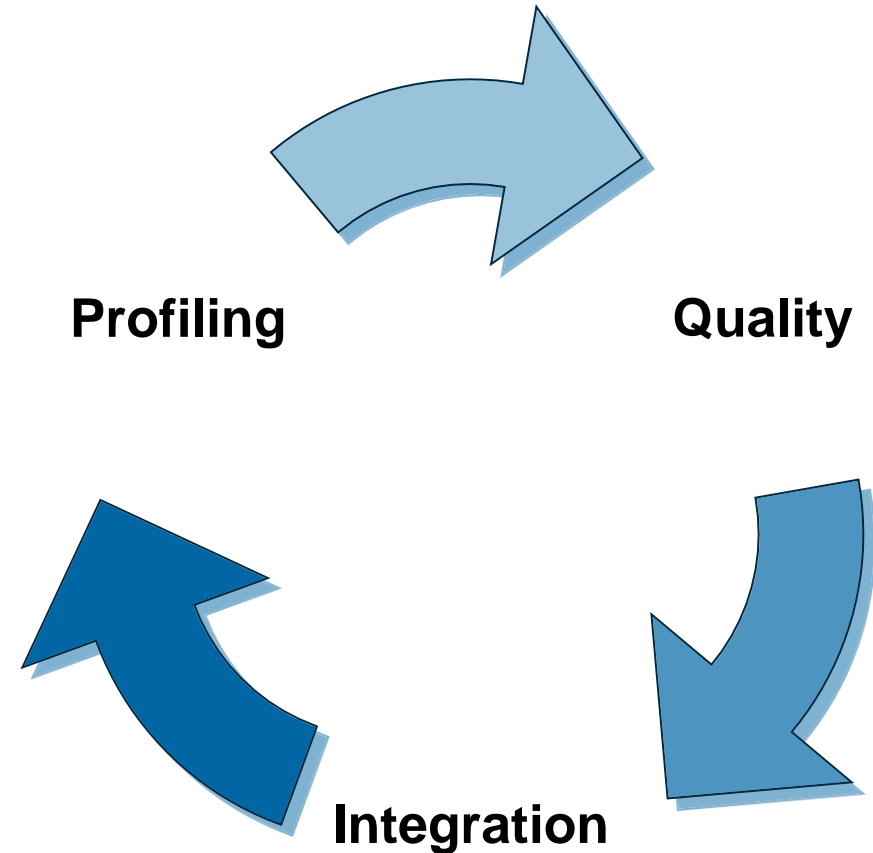


2011

DataFlux Methodology

The three steps that Bell used to make more valuable enterprise data:

1. Data Profiling
2. Data Quality
3. Data Integration





Data Profiling

dfPower Profile (Viewer) - sngl_cust_all

File View Tools Help

Field: CUST_LAST_NM
Defined type: CHARACTER
Defined length: 28 chars

Column Profiling | Frequency Distribution | Pattern Frequency Distribution | Percentiles | Outliers | Notes

Value	Count	Percentage
SMITH	47669	0.45
TREMBLAY	42825	0.40
GAGNON	30191	0.28
ROY	29901	0.28
BROWN	29004	0.27
MARTIN	27913	0.26
GAUTHIER	22367	0.21
WILSON	21756	0.20
BOUCHARD	20722	0.19
MORIN	18874	0.18
LEE	18796	0.18
LAVOIE	17790	0.17
TAYLOR	17554	0.17
LEBLANC	17459	0.16
CAMPBELL	17336	0.16
JOHNSON	17135	0.16
THOMPSON	16931	0.16
JONES	16871	0.16
WILLIAMS	16712	0.16
FORTIN	16342	0.15
PELLETIER	16235	0.15
BERGERON	15198	0.14
MILLER	15058	0.14
WHITE	14585	0.14
SIMARD	14193	0.13
ANDERSON	14121	0.13
GIRARD	14004	0.13
MACDONALD	13997	0.13
BOUCHER	13848	0.13
YOUNG	13043	0.12
SCOTT	12827	0.12
SINGH	12727	0.12
ROBINSON	12566	0.12
POIRIER	12519	0.12
BEAULIEU	12249	0.12
MOORE	12114	0.11
WONG	11902	0.11
CARON	11783	0.11
CÔTÉ	11773	0.11
STEWART	11707	0.11
LAPOINTE	11534	0.11
LEFEBVRE	11461	0.11
LANDRY	11413	0.11
PATEL	11404	0.11

Filter and Sort | Query Builder | Data | Describe | Graph | Analyze | Export | Send To

	OPEN_CUST_FIRST_NM	OPEN_CUST_LAST_NM	OPEN_CITY	OPEN_PROVINCE	Population
12249	MARTIN	TREMBLAY	QUÉBEC	QC	MATCH STR < 7
12250	MARCEL	TREMBLAY	JONQUIÈRE	QC	MATCH STR < 7
12251	MARC-ANDRÉ	TREMBLAY	SAINT-HUBERT	QC	MATCH STR < 7
12252	MARIE	TREMBLAY	CHICOUTIMI	QC	MATCH STR < 7
12253	MICHEL	TREMBLAY	CHICOUTIMI	QC	MATCH STR < 7
12254	MARIO	TREMBLAY	CHICOUTIMI	QC	MATCH STR < 7
12255	MARIE-BLANCHE	TREMBLAY	QUÉBEC	QC	MATCH STR < 7
12256	MICHÈLE	TREMBLAY	MONTRÉAL	QC	MATCH STR < 7
12257	MICHELE	TREMBLAY	QUEBEC	QC	MATCH STR < 7
12258	MONIQUE	TREMBLAY	QUÉBEC	QC	MATCH STR < 7
12259	MARIE-ÈVE	TREMBLAY	JONQUIÈRE	QC	MATCH STR < 7
12260	MICHEL	TREMBLAY	LA MALBAIE	QC	MATCH STR < 7
12261	MARIE-CLAUDE	TREMBLAY	QUÉBEC	QC	MATCH STR < 7
12262	MICHEL	TREMBLAY	CHICOUTIMI	QC	MATCH STR < 7
12263	MONIQUE	TREMBLAY	QUÉBEC	QC	MATCH STR < 7
12264	MARIO	TREMBLAY	QUÉBEC	QC	MATCH STR < 7
12265	MYRIAM	TREMBLAY	LA BAIE	QC	MATCH STR < 7
12266	M	TREMBLAY	CHICOUTIMI	QC	MATCH STR < 7

Ready





Data Quality

- One field with many elements: First Middle and Last Name
- Typos: & instead of 7
- Copy & Paste errors
- Long data overflowed to next field
- Email address in foreign country field
- John & Mary Smith
- Units, Apartments and Suites
- Abbreviations: QUE, QR, Quebec, QC

LESLIE A URQUHART	1-10 GRANDVIEW DR	10 GRANDVIEW DR APT 1	(999) 999-9999
L ANN URQUHART	3128 MAIN STREET	3128 MAIN ST	999 999 9999
URQUHART, LESLIE ANN	17 HIDDEN RANCH CIRCLE	17 HIDDENRANCH CIR	999-999-9999
URQUHART, LESLIE A	420 9 AV SE APP 565	APT 565 420 9 AVE SE	1 (999) 999.9999





Data Integration



- » Parsing
- » Tokens
- » Transformation
- » Standardization

Standardization Properties

Name:

Locale:

Standardization fields

Available:

- CUST_SYS_ID
- BAN
- CUBS_DEBTOR_NO
- CAN
- CUST_FIRST_NM
- CUST_LAST_NM
- CUST_BILL_ADDR
- BILL_ADDR2
- City
- Province

Selected:

Field Name	Definition	Scheme	Output Name
REF_TELNUM	Non-Number Removal		REF_TELNUM_Std
CUST_CONTACT	Non-Number Removal		CUST_CONTACT_TN_Std
CO_USER_CONT	Non-Number Removal		CO_USER_CONTACT_NO_Std
CUSTOMER_SSN	Non-Number Removal		CUSTOMER_SIN_Std
BIRTH_DATE	Date (DMY)		BIRTH_DATE_Std
DRIVR_LICNS_NO	Non-Alphanumeric Rem		DRIVR_LICNS_NO_Std

Preserve null values Add standardization flag field



Data Integration – Match Codes

SAS Enterprise Guide

File Edit View Tasks Program

oth_matches_feb1

Filter and Sort Query Builder Data

	CID1	NAM
1	784041	GYB47\$\$\$\$\$BY
2	784041	GYB47\$\$\$\$\$B&
3	775202	4Y\$\$\$\$\$B7J
4	775202	4Y\$\$\$\$\$B7J
5	766227	M3437\$\$\$\$\$B73
6	766227	M3437\$\$\$\$\$B73
7	757907	B38YB\$\$\$\$\$B7
8	757907	B38YB\$\$\$\$\$B7
9	740751	4**MB4B\$\$\$\$\$4#
10	740751	4**MB4B\$\$\$\$\$4#
11	717824	2B8YB\$\$\$\$\$42Y
12	717824	2B8YB\$\$\$\$\$42Y
13	717710	LB\$\$\$\$\$37B
14	717710	LB\$\$\$\$\$3W
15	696440	B3Y**27\$\$\$\$\$42E
16	696440	B3Y**27\$\$\$\$\$4**
17	687260	3YM\$\$\$\$\$FY
18	687260	3YM\$\$\$\$\$FY
19	684662	**2YW\$\$\$\$\$4**
20	684662	**2YW\$\$\$\$\$4**
21	679014	4**W\$\$\$\$\$&B
22	679014	4**W\$\$\$\$\$&B
23	675040	MMW&\$\$\$\$\$3
24	675040	MMW&\$\$\$\$\$3
25	664723	3**23Y****\$8B
26	664723	3**23Y****\$8**
27	656148	8LY\$\$\$\$\$B73
28	656148	8LY\$\$\$\$\$B73
29	653684	GYB3YB4\$\$\$\$\$42
30	653684	GYB3YB4\$\$\$\$\$42
31	652250	34MB****\$J&
32	652250	34MB****\$3**
33	649464	4MB4\$\$\$\$\$3W
34	649464	4MB4\$\$\$\$\$J2Y
35	636976	GWB\$\$\$\$\$3W
36	636976	GWB\$\$\$\$\$3W
37	629709	YBWB\$\$\$\$\$C4
38	629709	YBWB\$\$\$\$\$C4
39	623408	J2M\$\$\$\$\$1W
40	623408	J2M\$\$\$\$\$N\$\$\$\$\$
41	620810	J24B****\$C@P\$\$\$\$\$
42	620810	J24B****\$C@P\$\$\$\$\$
43	620810	J24B****\$C@P\$\$\$\$\$
44	620810	J24B****\$C@P\$\$\$\$\$
45	620810	J24B****\$C@P\$\$\$\$\$
46	620810	J24B****\$C@P\$\$\$\$\$
47	620810	J24B****\$C@P\$\$\$\$\$
48	620810	J24B****\$C@P\$\$\$\$\$
49	620810	J24B****\$C@P\$\$\$\$\$
50	620810	J24B****\$C@P\$\$\$\$\$
51	620810	J24B****\$C@P\$\$\$\$\$
52	620810	J24B****\$C@P\$\$\$\$\$
53	620810	J24B****\$C@P\$\$\$\$\$
54	620810	J24B****\$C@P\$\$\$\$\$
55	620810	J24B****\$C@P\$\$\$\$\$
56	620810	J24B****\$C@P\$\$\$\$\$
57	620810	J24B****\$C@P\$\$\$\$\$
58	620810	J24B****\$C@P\$\$\$\$\$
59	620810	J24B****\$C@P\$\$\$\$\$
60	620810	J24B****\$C@P\$\$\$\$\$
61	620810	J24B****\$C@P\$\$\$\$\$
62	620810	J24B****\$C@P\$\$\$\$\$
63	620810	J24B****\$C@P\$\$\$\$\$
64	620810	J24B****\$C@P\$\$\$\$\$
65	620810	J24B****\$C@P\$\$\$\$\$
66	620810	J24B****\$C@P\$\$\$\$\$
67	620810	J24B****\$C@P\$\$\$\$\$
68	620810	J24B****\$C@P\$\$\$\$\$
69	620810	J24B****\$C@P\$\$\$\$\$
70	620810	J24B****\$C@P\$\$\$\$\$
71	620810	J24B****\$C@P\$\$\$\$\$
72	620810	J24B****\$C@P\$\$\$\$\$
73	620810	J24B****\$C@P\$\$\$\$\$
74	620810	J24B****\$C@P\$\$\$\$\$
75	620810	J24B****\$C@P\$\$\$\$\$
76	620810	J24B****\$C@P\$\$\$\$\$
77	620810	J24B****\$C@P\$\$\$\$\$
78	620810	J24B****\$C@P\$\$\$\$\$
79	620810	J24B****\$C@P\$\$\$\$\$
80	620810	J24B****\$C@P\$\$\$\$\$
81	620810	J24B****\$C@P\$\$\$\$\$
82	620810	J24B****\$C@P\$\$\$\$\$
83	620810	J24B****\$C@P\$\$\$\$\$
84	620810	J24B****\$C@P\$\$\$\$\$
85	620810	J24B****\$C@P\$\$\$\$\$
86	620810	J24B****\$C@P\$\$\$\$\$
87	620810	J24B****\$C@P\$\$\$\$\$
88	620810	J24B****\$C@P\$\$\$\$\$
89	620810	J24B****\$C@P\$\$\$\$\$
90	620810	J24B****\$C@P\$\$\$\$\$
91	620810	J24B****\$C@P\$\$\$\$\$
92	620810	J24B****\$C@P\$\$\$\$\$
93	620810	J24B****\$C@P\$\$\$\$\$
94	620810	J24B****\$C@P\$\$\$\$\$
95	620810	J24B****\$C@P\$\$\$\$\$
96	620810	J24B****\$C@P\$\$\$\$\$
97	620810	J24B****\$C@P\$\$\$\$\$
98	620810	J24B****\$C@P\$\$\$\$\$
99	620810	J24B****\$C@P\$\$\$\$\$
100	620810	J24B****\$C@P\$\$\$\$\$

Match Codes (Parsed) Properties

Name: Match Codes Address

Output field: CUST_BILL_ADDR_Mc

Locale: English (Canada)

Sensitivity: 95

Definition: Address

Address

City - State/Province - Postal Code

Date (DMY)

Name

Phone

Tokens:

Token Name	Field Name
Street Number	Street Number
Pre-direction	
Street Name	Street Name
Street Type	
Post-direction	
Address Extension	
Address Extension Num	Unit_Number_Strd

Generate null match codes for blank field values

Preserve null values

OK Cancel Help

Ready

No profile selected





Data Integration - Clustering

dfReport Viewer [Report Title: Other Provinces - Match Report]

File View Tools Help

Clustering Properties

Name: Notes...

Output cluster ID field: Options...

Treat blank field values as nulls Compact cluster numbers

Override clustering memory size MB Cluster

Do not cluster field: All clusters

Sort output by cluster number Single-row clusters only

Multi-row clusters only

Conditions

Available Selected fields:

OR
DRIVR_LICNS_NO_Std
CUST_BILL_ADDR_mc
OR
BIRTH_DATE
Name_Initial_mc
POSTAL_CD
OR
Name_Initial_mc
EMAIL_ADDRESS_Mc

Additional Outputs...

OK Cancel Help

CID1	BALANCE	Name_mc_95	refel	name_contactnu	name_sin	sys_id	dl_sin	sin_addr	name_dl
257	142.86	M43\$\$\$\$\$\$FYW8B\$\$\$\$	false	false	false	false	false	false	false
257	40.14	M43\$\$\$\$\$\$FYW8B\$\$\$\$	false	false	false	false	false	false	false
1431	0	BFY\$\$\$\$\$\$~Y\$\$\$\$\$\$	false	false	false	false	false	false	false
1431	154.97	BFY\$\$\$\$\$\$~Y\$\$\$\$\$\$	false	false	false	false	false	false	true
1431	-4.78	BFY\$\$\$\$\$\$~Y\$\$\$\$\$\$	false	false	false	false	false	false	true
3916	637.14	F4W\$\$\$\$\$\$8WF~4 \$\$\$	false	false	false	false	false	false	false
3916	0	F4W\$\$\$\$\$\$8WF~\$\$\$	false	false	false	false	false	false	false
5804	862.85	FB8Y\$\$\$\$\$\$NBB\$\$\$\$	false	false	false	false	false	false	false
5804	255.96	FB8Y\$\$\$\$\$\$NBB\$\$\$\$	false	false	false	false	false	false	false
6139	199.55	48\$\$\$\$\$\$W4\$\$\$\$\$\$	false	false	false	false	false	false	false
6139	637.2	48\$\$\$\$\$\$W4\$\$\$\$\$\$	false	false	false	false	false	false	false
7256	100.88	BW~\$\$\$\$\$\$Y8R\$\$\$\$	false	false	false	false	false	false	false
7256	50.48	BW~\$\$\$\$\$\$Y8R\$\$\$\$	false	false	false	false	false	false	false
10499	139.13	J2W~B\$\$\$\$\$\$JY74\$\$\$\$	false	false	false	false	false	false	false
10499	142.52	J2W~B\$\$\$\$\$\$JY74\$\$\$\$	false	false	false	false	false	false	false
10499	156.43	J2W~B\$\$\$\$\$\$JY74\$\$\$\$	false	false	false	false	false	false	false
10499	139.31	J2W~B\$\$\$\$\$\$JY74\$\$\$\$	false	false	false	false	false	false	false
10499	142.88	J2W~B\$\$\$\$\$\$JY74\$\$\$\$	false	false	false	false	false	false	false
10499	151.51	J2W~B\$\$\$\$\$\$JY74\$\$\$\$	false	false	false	false	false	false	false
10499	144.6	J2W~B\$\$\$\$\$\$JY74\$\$\$\$	false	false	false	false	false	false	false
11191	330.93	8YB~BFY\$\$\$287\$\$\$\$	false	true	false	true	false	true	true
11191	421.33	8YB~BFY\$\$\$287\$\$\$\$	false	true	false	true	false	false	true
12754	1.14	3W7\$\$\$\$\$\$42B\$\$\$\$	false	false	false	false	false	false	false
12754	770.69	3W7\$\$\$\$\$\$42B\$\$\$\$	false	false	false	false	false	false	false
12754	168.5	3W7\$\$\$\$\$\$42B\$\$\$\$	false	false	false	false	false	false	false
12790	39.12	2BMW7\$\$\$\$\$\$8@#F\$\$\$\$	false	false	false	false	false	false	false

Cluster Field Name: CID1
SR Field Name: BAN
Total Clusters: 312



Assigning Match Strengths

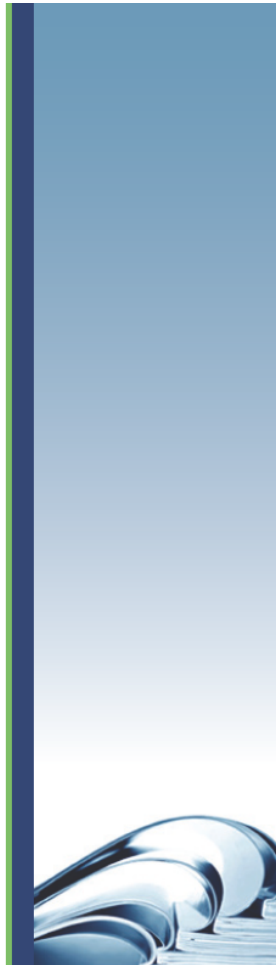
CID1	BALANCE	Name mc 95	Name Initial mc	match strength	name addr city	name city couser	name city refname	name city occup	name city empname	name couser tel	name ref tel	name contact num	name sin	sys id	dl sin	sin addr	name dl	dl addr	dob name pc	name email
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T
257	1	CID1	MATCH_STRENGTH	CUST_LAST_NM	CUST_BILL_ADDR	INITIAL	CUST_FIRST_NM	NAME_MC_95	NAME_INITIAL_MC	EMAIL_ADDRESS	BIRTH_DATE	CUST								
257	2	20705	7	BEAULIEU	250 RUE LAFRANCE	D	D		MJ2BM\$\$\$\$\$8\$\$\$\$\$\$	BEAULIEU.thibert@sympatico.ca										
	3	20705	105	BEAULIEU	617 MOREAU	D	DENIS	MJ2BM\$\$\$\$\$8P4\$\$\$\$\$	MJ2BM\$\$\$\$\$8\$\$\$\$\$\$	BEAULIEUdenis@sympatico.ca	28Mar1953 0:00:00.00									
	4	20705	7	BEAULIEU	2082 10 RUE	D	DENIS	MJ2BM\$\$\$\$\$8P4\$\$\$\$\$	MJ2BM\$\$\$\$\$8\$\$\$\$\$\$	BEAULIEUdenis@sympatico.ca										
431	5	20705	7	BEAULIEU	90 TURCOT	D	DAVID	MJ2BM\$\$\$\$\$8&V_\$\$\$\$\$	MJ2BM\$\$\$\$\$8\$\$\$\$\$\$	BEAULIEU.david@sympatico.ca										
431	6	20705	10	BEAULIEU	67 RUE MARCEL-CHAPUT	D	DIANE	MJ2BM\$\$\$\$\$87&P\$\$\$\$\$	MJ2BM\$\$\$\$\$8\$\$\$\$\$\$	BEAULIEU.diane@sympatico.ca										
431	7	20705	3	BEAULIEU	APP 2 9 RUE CHOLETTE	D	DANIELLE	MJ2BM\$\$\$\$\$8&P7\$\$\$\$\$	MJ2BM\$\$\$\$\$8\$\$\$\$\$\$	daniel.BEAULIEU7@sympatico.ca	15Aug1955 0:00:00.00									
	8	20705	7	BEAULIEUS	617 RUE MOREAU	D	D		MJ2BM\$\$\$\$\$8\$\$\$\$\$\$											
	9	20705	7	BEAULIEU	2700 RUE DES FLORALIES	D	DAVID	MJ2BM\$\$\$\$\$8&V_\$\$\$\$\$	MJ2BM\$\$\$\$\$8\$\$\$\$\$\$	BEAULIEU.david@sympatico.ca										
	10	20705	105	BEAULIEU	617 MOREAU	D	DENIS	MJ2BM\$\$\$\$\$8P4\$\$\$\$\$	MJ2BM\$\$\$\$\$8\$\$\$\$\$\$		28Mar1953 0:00:00.00									
374	11	20705	7	BEAULIEU	8865 AV JOLIOT-CURIE	D	DIANE	MJ2BM\$\$\$\$\$87&P\$\$\$\$\$	MJ2BM\$\$\$\$\$8\$\$\$\$\$\$	BEAULIEUdoiron@sympatico.ca										
374	12	20705	14	BEAULIEU	617 RUE MOREAU	D	DENIS	MJ2BM\$\$\$\$\$8P4\$\$\$\$\$	MJ2BM\$\$\$\$\$8\$\$\$\$\$\$	BEAULIEU.denis@sympatico.ca										
	13																			
916	637.14	F4W\$\$\$\$\$8WF~4 \$\$\$		F4W\$\$\$\$\$8\$\$\$\$\$\$		7	1	0	0	0	0	0	0	0	0	0	0	0	0	0
916	0	F4W\$\$\$\$\$8WF~\$\$\$\$\$		F4W\$\$\$\$\$8\$\$\$\$\$\$		7	1	0	0	0	0	0	0	0	0	0	0	0	0	0
804	862.85	FB8Y\$\$\$\$\$NBB\$\$\$\$\$		FB8Y\$\$\$\$\$N\$\$\$\$\$\$		7	1	0	0	0	0	0	0	0	0	0	0	0	0	0
804	255.96	FB8Y\$\$\$\$\$NBB\$\$\$\$\$		FB8Y\$\$\$\$\$N\$\$\$\$\$\$		7	1	0	0	0	0	0	0	0	0	0	0	0	0	0
5139	199.55	4&\$\$\$\$\$W4\$\$\$\$\$		4&\$\$\$\$\$W\$\$\$\$\$\$		7	1	0	0	0	0	0	0	0	0	0	0	0	0	0
5139	637.2	4&\$\$\$\$\$W4\$\$\$\$\$		4&\$\$\$\$\$W\$\$\$\$\$\$		7	1	0	0	0	0	0	0	0	0	0	0	0	0	0
7256	100.88	BW~\$\$\$\$\$Y&R\$\$\$\$\$		BW~\$\$\$\$\$Y\$\$\$\$\$\$		7	0	0	0	0	0	0	0	0	0	0	0	0	1	0
7256	50.48	BW~\$\$\$\$\$Y&R\$\$\$\$\$		BW~\$\$\$\$\$Y\$\$\$\$\$\$		7	0	0	0	0	0	0	0	0	0	0	0	0	1	0



References

Parsing and Standardization

A DataFlux White Paper
Prepared by David Loshin



DF Job Flow

Questions?



Leslie Urquhart – Specialist Risk Analysis
110 King St W – Floor 2
Hamilton ON L8P 4S6
(905) 977-5763
leslie.urquhart@bell.ca