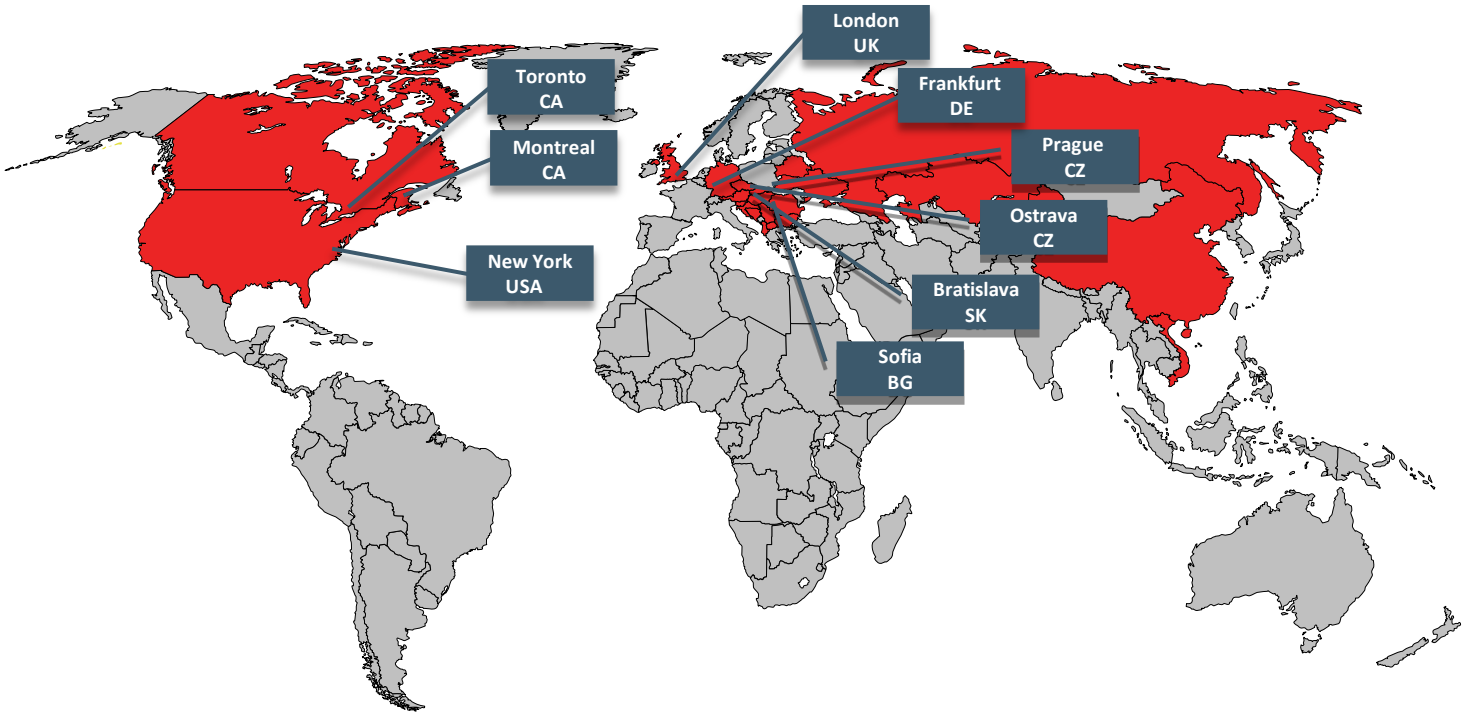
Adastra Bulgaria and Data Quality

July Karchev
DWH Technical Lead

Georgi Pamukov, Galina Guneva
DWH Consultants

October 12, 2011
CAREER DAYS 2011, IT, TELECOMMUNICATIONS, BPOSofia

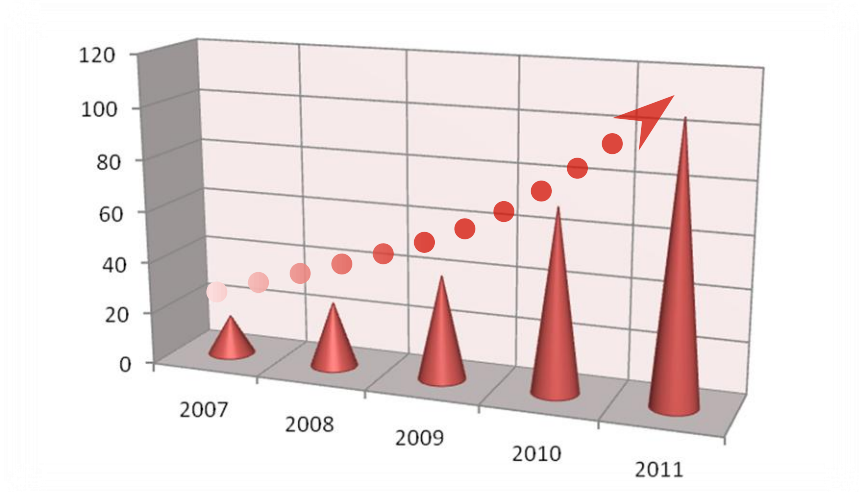
Adastra Group



- Established in 1994, headquarter in Toronto, Canada
- Offices in **7** countries
- Over **420** successful projects in 24 countries
- Over **800** consultants
- 32 international quality awards

Adastra Bulgaria

- Business Information Management Leader
- 100+ IT professional consultants
- International and local projects for world-leading companies



- Sofia office
  - 29th Panayot Volov street
- Varna office
  - 2 Dunav street

Selected Partners

Information management

ADASTRA

Data Quality – the problem

- Everyone is affected by poor data quality
  - Directly – 2 or 3 identical mailings from the same sales organization in the same week
  - In less direct ways – 20 minute wait on hold for a customer service department
  - Dangerously – deliberate identity theft
- Impact on business and government agencies
  - Ineffective operations- data quality problems cost companies about 10% of their revenue

Bad Data – You'll Cry

- Low quality data costs billions

  Gartner has reported
  - 25% of critical data within large businesses is somehow inaccurate or incomplete
  - and that 50% of implementations fail due to a lack of attention to data quality issues
  - data quality problems cost U.S. businesses over **US$600** billion a year

What is Data Quality

- *Data quality is something everyone desires*

- Everybody has a different idea of what data quality means
  - Each organization has own rules on the quality of data
  - Each company department has own data quality expectations

- Definition of data quality
  - “fitness to serve its purpose” - to what extent the data are appropriate for their purpose
  - In practice this means identifying assets of data quality objectives associated with any dataset and then measuring that datasets conformance to these objectives

The many faces of Data Quality

- **Accuracy -** does data reflect real world state
  - company identifier exists in the official register of companies
  - street, street number, city, ZIP, country code found in an authorized/official etalon of addresses
- **Completeness -** extent to which the expected attributes of data are provided
  - Empty address fields, blank characters, “_”, “-”
- **Consistency -** does pieces of info match together
  - Трабант 2105
- **Timeliness/Relevance -** is the data up-to-date
  - Old pricing info
- **Validity -** is the data in a reasonable range/format
  - 2090

Accuracy – data quality gone wrong

	src_name	src_gender	src_birth_date	src_sin	src_card	src_address
1	Dr. John Smith	M	12/16/1978	000000000	88682239496	14618 110 Ave Surrey V3R2A9
2	Smith W. John	MALE	16.12.1978	095-242-434	266805807984498	Surrey 14618 110 Ave
3	John William Smith		781216	SIN095242434		25 Linden Str Toronto M4X 1V5
4	Dr. J.W. Smith	M	11/16/78	095242433	4334338874158398	
5	John Smith		16.11.1978	095252433	43176816175728	
6	Smith John		16.11.1978		43176816175728	
7	John Smiht		16.11.1978	95252433	NULL	
8	Jane Watson		1982	420347213	974997714973	
9	Watson Jane	FEMALE	5.1.1982	420-347-213		8500 Leslei street Toronto L3T 7M8
10	Jane Smith	F	1982-05-01	SIN420347213	720847891758473	
11	J. Smith			420-347-213		
12	Janette Smith	F	B. 2015-01-05	SIN	54590972345186400	

Social Insurance Number not corresponding to authorized/official etalons

Inconsistent pattern of values in the field

There is no such numbers on this streets

	src_street	src_city	src_province	src_zip
1	8500 Leslie	Toronto	Ontario	L3T7M8
2	128 Yonge St	Tronto	ON	M5C1XX
3	25 Linden Str	TORONNTO	SOFIA OKRAG	2227
4	8500 Leslie street	Marham	Onntario	L37
5	8500 Leslei street	Toronto	Ontsrio	L3T 7M8
6	8500 Leslei st	Totonto	on	L3T 7M8
7	128 St John St	St John	NB	
8	20 St John Ave	Oromocto	NB	
9	5867 Eagle Island	Vancouver	Britisch Columbia	V7W1V5

Accuracy – data quality gone wrong (*cont.*)

Duplicated data – this is several different records for the same person

	src_name	src_gender	src_birth_date	src_sin	src_card	src_address
1	Dr. John Smith	M	12/16/1978	000000000	88682239496	14618 110 Ave Surrey V3R2A9
2	Smith W. John	MALE	16.12.1978	095-242-434	266805807984498	Surrey 14618 110 Ave
3	John William Smith		781216	SIN095242434		25 Linden Str Toronto M4X 1V5
4	Dr. J.W. Smith	M	11/16/78	095242433	4334338874158390	
5	John Smith		16.11.1978	095252433	431768161757282	8500 Leslie L3T 7M8 Toronto
6	Smith John		16.11.1978		431768161757282	8500 Leslie street Marham
7	John Smiht		16.11.1978	95252433	NULL	
8	Jane Watson		1982	420347213	974997714973	600-8500 Bugatica str. Toronto L3T 7M8
9	Watson Jane	FEMALE	5.1.1982	420-347-213		8500 Leslei street Toronto L3T 7M8
10	Jane Smith	F	1982-05-01	SIN420347213	720847891758473	
11	J. Smith			420-347-213		
12	Janette Smith	F	B. 2015-01-05	SIN	54590972345186400	

Completeness – data quality gone wrong

	src_name	src_gender	src_birth_date	src_sin	src_card	src_address
1	Dr. John Smith	M	12/16/1978	000000000	88682239496	14618 110 Ave Surrey V3R2A9
2	Smith W. John	MALE	16.12.1978	095-242-434	266805807984498	Surrey 14618 110 Ave
3	John William Smith			SIN095242434		25 Linden Str Toronto M4X 1V5
4	Dr. J.W. Smith	M		095242433	4334338874158390	
5	John Smith			095252433	431768161757282	8500 Leslie L3T 7M8 Toronto
6	Smith John		16.11.1978		431768161757282	8500 Leslie street Marham
7	John Smiht		16.11.1978	95252433	NULL	
8	Jane Watson		1982	420347213	974997714973	600-8500 Bugatica str. Toronto L3T 7M8
9	Watson Jane	FEMALE	5.1.1982	420-347-213		8500 Leslei street Toronto L3T 7M8
10	Jane Smith	F	1982-05-01	SIN420347213	720847891758473	
11	J. Smith			420-347-213		
12	Janette Smith	F	B. 2015-01-05	SIN	54590972345186400	

Missing date & month

Missing/blank Values

'-' values in 'src_province' field

	src_street	src_city	src_province	src_zip
22	618 REVTELL SYND NW	EDMONTON	AB	T6R2H9
23	283 PARKB SIEEET S	HAMILTON	ON	L8P3G5
24	2046 SIRWOCCO DVIRE SW	CALGARY	-	T3H2M8
25	15 MCDGILL STRETE S	SMITHS FALLS	-	K7A3M4
26	229 SPRUCEE STRET S	TIMMINS	-	P4N2M8
27	4508 BEEDIE	BURNABY	-	V5J5L2
28	1460 ESTHER-BLONDIN	QUEBEC	QC	G1Y3N7
29	295 FATHER TOBIN	BRAMPTON	ON	L6R0N2
30	273 RIVER	SUNNY CORNER	NB	E9E1C9

Consistency – data quality gone wrong

	src_street	src_city	src_province	src_zip
1	8500 Leslie	Toronto	Ontario	L3T7M8
2	128 Yonge St	Tronto	ON	M5C1XX
3	25 Linden Str	TORONNTO	Britisch Columbia	2227
4	8500 Leslie street	Marham	Onntario	L37
5	8500 Leslei street	Toronto	Ontsrio	L3T 7M8
6	8500 Leslei st	Totonto	on	L3T 7M8
7	128 St John St	St John	NB	
8	20 St John Ave	Oromocto	NB	
9	5867 Eagle Island	Vancouver	Britisch Columbia	V7W1V5

Validity – data quality gone wrong

	src_name	src_gender	src_birth_date	src_sin	src_card	src_address
1	Dr. John Smith	M	12/16/1978	000000000	88682239496	14618 110 Ave Surrey V3R2A9
2	Smith W. John	MALE	16.12.1978	095-242-434	266805807984498	Surrey 14618 110 Ave
3	John William Smith		781216	SIN095242434		25 Linden Str Toronto M4X 1V5
4	Dr. J.W. Smith	M	11/16/78	095242433	4334338874158390	
5	John Smith		16.11.1978	095252433	431768161757282	8500 Leslie L3T 7M8 Toronto
6	Smith John		16.11.1978		431768161757282	8500 Leslie street Marham
7	John Smiht		16.11.1978	95252433	NULL	
8	Jane Watson		1982	420347213	974997714973	600-8500 Bugatica str. Toronto L3T 7M8
9	Watson Jane	FEMALE	5.1.1982	420-347-213		8500 Leslei street Toronto L3T 7M8
10	Jane Smith	F	1982-05-01	SIN420347213	720847891758473	
11	J. Smith			420-347-213		
12	Janette Smith	F	B. 2015-01-05	SIN	54590972345186400	

Impossible date of birth – value out of range

Not valid card numbers (too short)

Data Quality Market

- Growing need for data quality analysts
  - Employment of data quality analysts was projected to increase 20% from 2008 to 2018, according to the U.S. Bureau of Labor Statistics (BLS).
- According to a Gartner report, the BI and analytics market is expected to reach $10.8 billion in 2011.

Data Quality – anchors

- Data Quality Experts and Consultants
  - Business Analyst
  - Data Analyst
  - Data Steward
- Methodologies
- Software Tools

Case Study – DQA Project

- The Company

  Company with 2 million customers , 3 million accounts. Enterprise information system handles customer registration and servicing, billing, invoicing and payments.

- The Problem

  Company reported difficulties with target marketing campaigns and inability to send invoices due to incorrect or missing address information.

  Corporate reports have showed differences in sold products versus billed amounts.

- The Challenge

  Identify data with quality issues and implement solution for DQ improvement.

  Evaluate the amount of impacted address information and resolve as much as possible of the affected information.

Project environment: DQ Analyzer [PROFILE & ANALYZE]

ataccama DQ Analyzer

Data Quality Assessment

Various data analyses to reveal basic and hidden DQ issues

Unmatched performance

Analyze millions of DB or CSV records in minutes using an easy-to-use wizard

Regular Expressions

Validate format/structure of the data, extract partial information from unstructured text

Rule-based Engine

Context-based validation using customizable rules

- **Unique data profiling tool** to reveal basic and hidden data quality issues
- **Easy to work with** (tutorials and illustrative samples)
- **Completely FREE for commercial use**
- www.ataccama.com

Project environment:
Ataccama DQC [CLEANSE & MATCH]

ataccama
Data Quality Center

Data Cleansing & Enrichment

Improve quality of individual records, enrich the data using external data sources.

Match & Merge

Correctly match related records, create representative “golden records”.

DQ Firewall

Prevent poor quality data from entering the systems by leveraging DQC as the validation procedure.

DQ Reporting & Monitoring

Set up and run DQ reports periodically to monitor quality of your data.

- **Essential tool for complex DQM** designed to evaluate, monitor and manage quality of data
  - Assessment **[Data Profiling]**
  - Prevention **[DQ Firewall]**
  - Measurement **[DQ Reporting]**
  - Control **[DQ Monitoring]**
- Bundled with
  - Vertical-specific and country-specific sets of **business rules**
  - Localized **dictionaries** and **knowledge bases**
- **Flexible** and platform-independent
- Scalable and **high-performance** oriented (incremental batch and online mode)

Activities in a Data Quality Project

1. Profiling

   basic analysis - metadata discovery and definition

2. Parse

   extract individual elements and store in correct fields

3. Cleanse and Standardize

   remove non-relevant information and “noise” from the content of the data

   reach uniform structure and enrich fro etalons

4. Match and Merge

   identify and consolidate records that refer to the same business

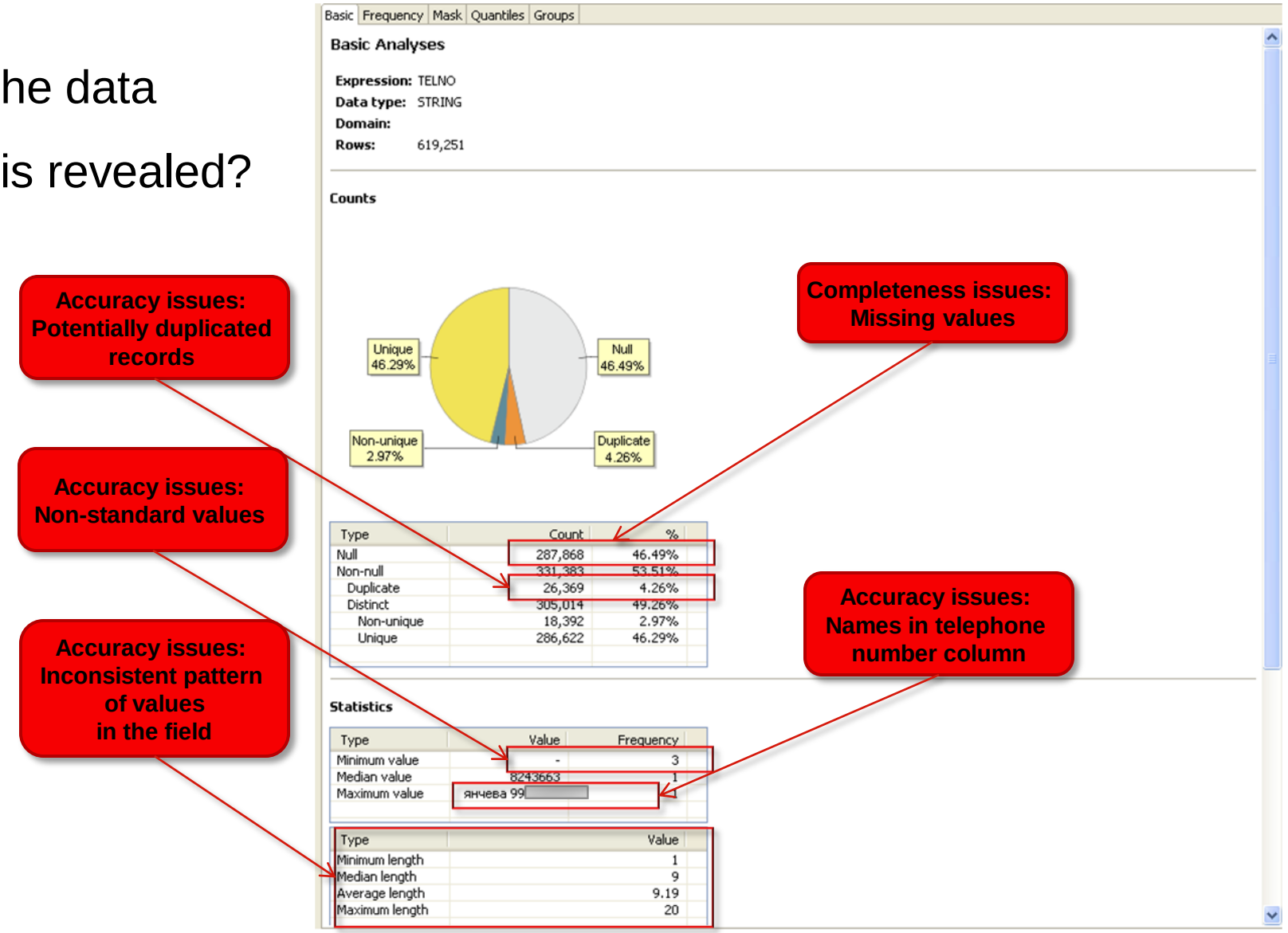
   object(customer for example)

5. Enrich

   adding useful, but optional, information to existing data or complete data

Project activities – 1. Profiling (*Basic*)

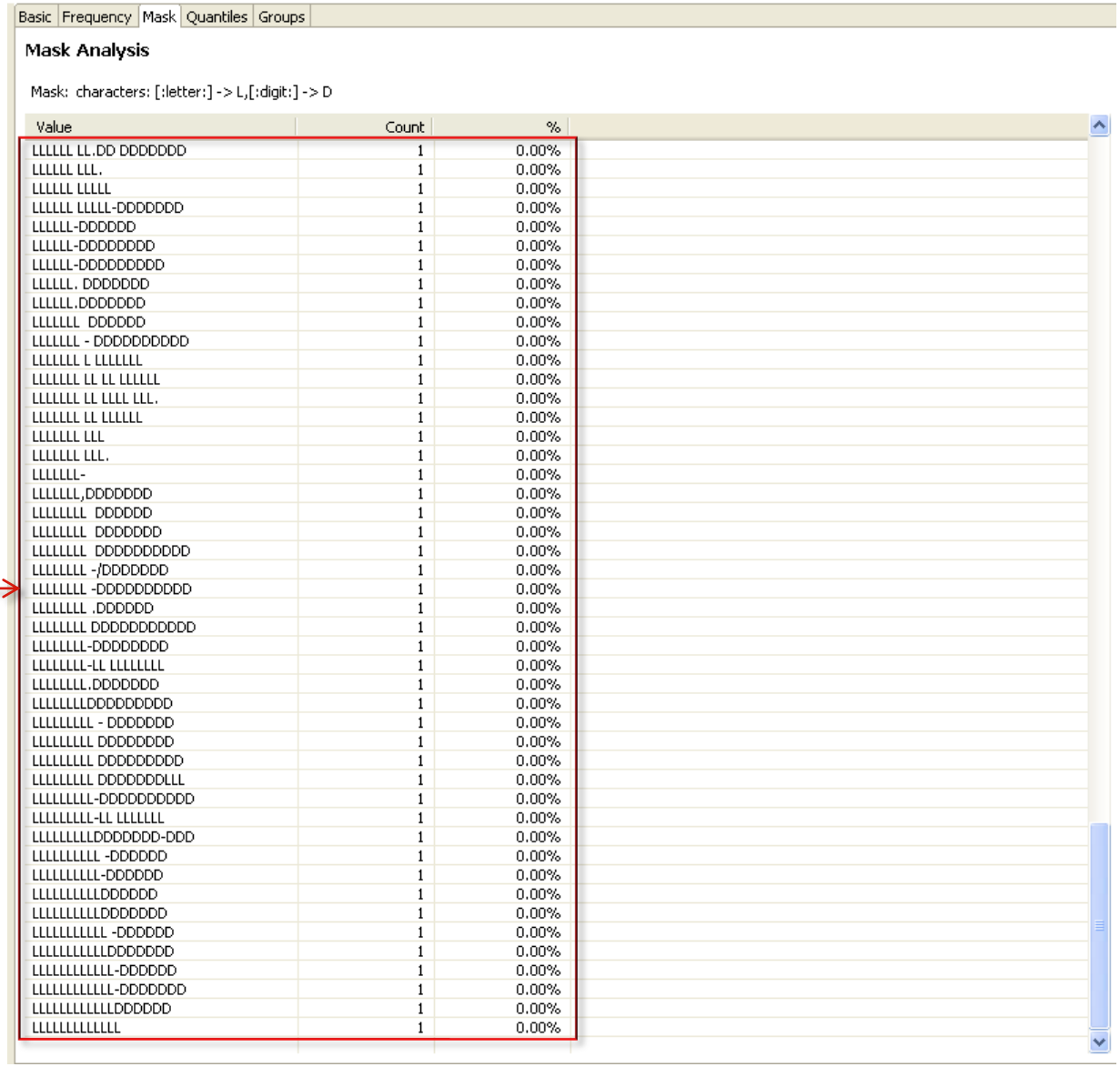
- What the data analysis revealed?



Project activities - 1. Profiling (*Mask*)

- What the data analysis revealed?

Accuracy issues:
Inconsistent pattern of values in the telephone number field
L – represents letter
D – represents digit



Value	Count	%
LLLLLL LL.DD DDDDDDD	1	0.00%
LLLLLL LLL.	1	0.00%
LLLLLL LLLLL	1	0.00%
LLLLLL LLLLL-DDDDDDD	1	0.00%
LLLLLL-DDDDDD	1	0.00%
LLLLLL-DDDDDDDD	1	0.00%
LLLLLL-DDDDDDDDD	1	0.00%
LLLLLL. DDDDDDD	1	0.00%
LLLLLL.DDDDDDD	1	0.00%
LLLLLL DDDDDD	1	0.00%
LLLLLL - DDDDDDDDD	1	0.00%
LLLLLL L LLLLLL	1	0.00%
LLLLLL LL LL LLLLL	1	0.00%
LLLLLL LL LLLL LLL.	1	0.00%
LLLLLL LL LLLLL	1	0.00%
LLLLLL LLL	1	0.00%
LLLLLL LLL.	1	0.00%
LLLLLL-	1	0.00%
LLLLLL,DDDDDDD	1	0.00%
LLLLLLL DDDDDD	1	0.00%
LLLLLLL DDDDDDD	1	0.00%
LLLLLLL DDDDDDDDD	1	0.00%
LLLLLLL -/DDDDDDD	1	0.00%
LLLLLLL -DDDDDDDDD	1	0.00%
LLLLLLL .DDDDDD	1	0.00%
LLLLLLL DDDDDDDDDD	1	0.00%
LLLLLLL-DDDDDDDD	1	0.00%
LLLLLLL-LL LLLLLLL	1	0.00%
LLLLLLL.DDDDDDD	1	0.00%
LLLLLLLDDDDDDDDD	1	0.00%
LLLLLLLL - DDDDDDD	1	0.00%
LLLLLLLL DDDDDDDD	1	0.00%
LLLLLLLL DDDDDDDDD	1	0.00%
LLLLLLLL DDDDDDDLLL	1	0.00%
LLLLLLLL-DDDDDDDDD	1	0.00%
LLLLLLLL-LL LLLLLLL	1	0.00%
LLLLLLLLDDDDDDD-DDD	1	0.00%
LLLLLLLLL -DDDDDD	1	0.00%
LLLLLLLLL-DDDDDD	1	0.00%
LLLLLLLLLDDDDDD	1	0.00%
LLLLLLLLLDDDDDDD	1	0.00%
LLLLLLLLLL -DDDDDD	1	0.00%
LLLLLLLLLLDDDDDDD	1	0.00%
LLLLLLLLLL-DDDDDD	1	0.00%
LLLLLLLLLL-DDDDDDD	1	0.00%
LLLLLLLLLLDDDDDD	1	0.00%
LLLLLLLLLLL	1	0.00%

Project activities - 2. Parsing

- Goal:
  - When different types of data are in a single field, extraction of individual elements and storing in correct fields are needed;
  - This will also allow performing of cleansing, standardizing and enrichment of the data.

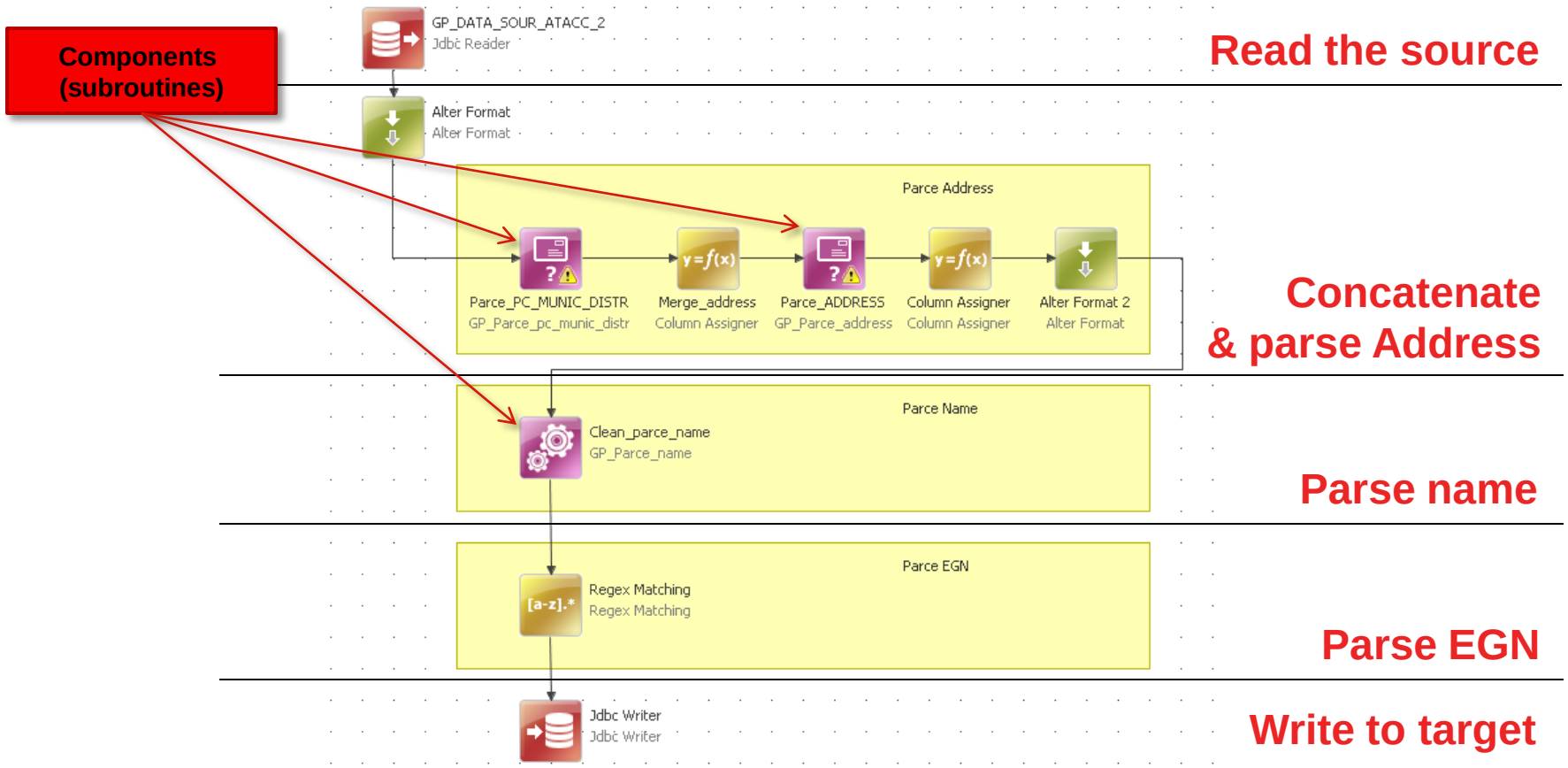
1	1000 СРЕДЕЦ КВ.ЦЕНТЪР	УЛ.ХРИСТО БЕЛЧЕВ/ГАВРИЛ ГЕНОВ/	Стр:21
2	1000 СРЕДЕЦ КВ.ЦЕНТЪР	УЛ.ИВАН ДЕНКОГЛУ Стр:44	<null>
3	1000 СРЕДЕЦ КВ.ЦЕНТЪР	УЛ.ХРИСТО БЕЛЧЕВ/ГАВРИЛ ГЕНОВ/	Стр:18
4	1000 СРЕДЕЦ КВ.ЦЕНТЪР	УЛ.АНГЕЛ КЪНЧЕВ №2	<null>
5	1000 СРЕДЕЦ КВ.ЦЕНТЪР	УЛ.ГРАФ ИГНАТИЕВ Стр:12	<null>
6	1000 СРЕДЕЦ КВ.ЦЕНТЪР	УЛ.ГРАФ ИГНАТИЕВ Стр:14	<null>
7	1000 СРЕДЕЦ КВ.ЦЕНТЪР	УЛ.ГРАФ ИГНАТИЕВ Бл:16	<null>
8	1000 СРЕДЕЦ КВ.ЦЕНТЪР	УЛ.ГРАФ ИГНАТИЕВ Стр:16 Ет:1	<null>
9	1000 СРЕДЕЦ КВ.ЦЕНТЪР	УЛ.ГРАФ ИГНАТИЕВ Стр:16 Ет:2	<null>
10	1000 СРЕДЕЦ КВ.ЦЕНТЪР	УЛ.ГРАФ ИГНАТИЕВ № 18	<null>

PARSING
Splitting of the individual elements and storing in correct fields

	POSTAL_CODE	MUNICIPLITY	TYPE	DISTRICT	STREET_TYPE	STREET	STREET_NO	BLOCK_NO	BUILDING_NO	ENTRANCE	FLOOR	APARTMENT_NO
1	1000	СРЕДЕЦ	КВ.	ЦЕНТЪР	УЛ	ХРИСТО БЕЛЧЕВ/ГАВРИЛ ГЕНОВ	<null>	<null>	21	<null>	<null>	<null>
2	1000	СРЕДЕЦ	КВ.	ЦЕНТЪР	УЛ	ИВАН ДЕНКОГЛУ	<null>	<null>	44	<null>	<null>	<null>
3	1000	СРЕДЕЦ	КВ.	ЦЕНТЪР	УЛ	ХРИСТО БЕЛЧЕВ/ГАВРИЛ ГЕНОВ	<null>	<null>	18	<null>	<null>	<null>
4	1000	СРЕДЕЦ	КВ.	ЦЕНТЪР	УЛ	АНГЕЛ КЪНЧЕВ	2	<null>	<null>	<null>	<null>	<null>
5	1000	СРЕДЕЦ	КВ.	ЦЕНТЪР	УЛ	ГРАФ ИГНАТИЕВ	<null>	<null>	12	<null>	<null>	<null>
6	1000	СРЕДЕЦ	КВ.	ЦЕНТЪР	УЛ	ГРАФ ИГНАТИЕВ	<null>	<null>	14	<null>	<null>	<null>
7	1000	СРЕДЕЦ	КВ.	ЦЕНТЪР	УЛ	ГРАФ ИГНАТИЕВ	<null>	16	<null>	<null>	<null>	<null>
8	1000	СРЕДЕЦ	КВ.	ЦЕНТЪР	УЛ	ГРАФ ИГНАТИЕВ	<null>	<null>	16	<null>	1	<null>
9	1000	СРЕДЕЦ	КВ.	ЦЕНТЪР	УЛ	ГРАФ ИГНАТИЕВ	<null>	<null>	16	<null>	2	<null>
10	1000	СРЕДЕЦ	КВ.	ЦЕНТЪР	УЛ	ГРАФ ИГНАТИЕВ	18	<null>	<null>	<null>	<null>	<null>

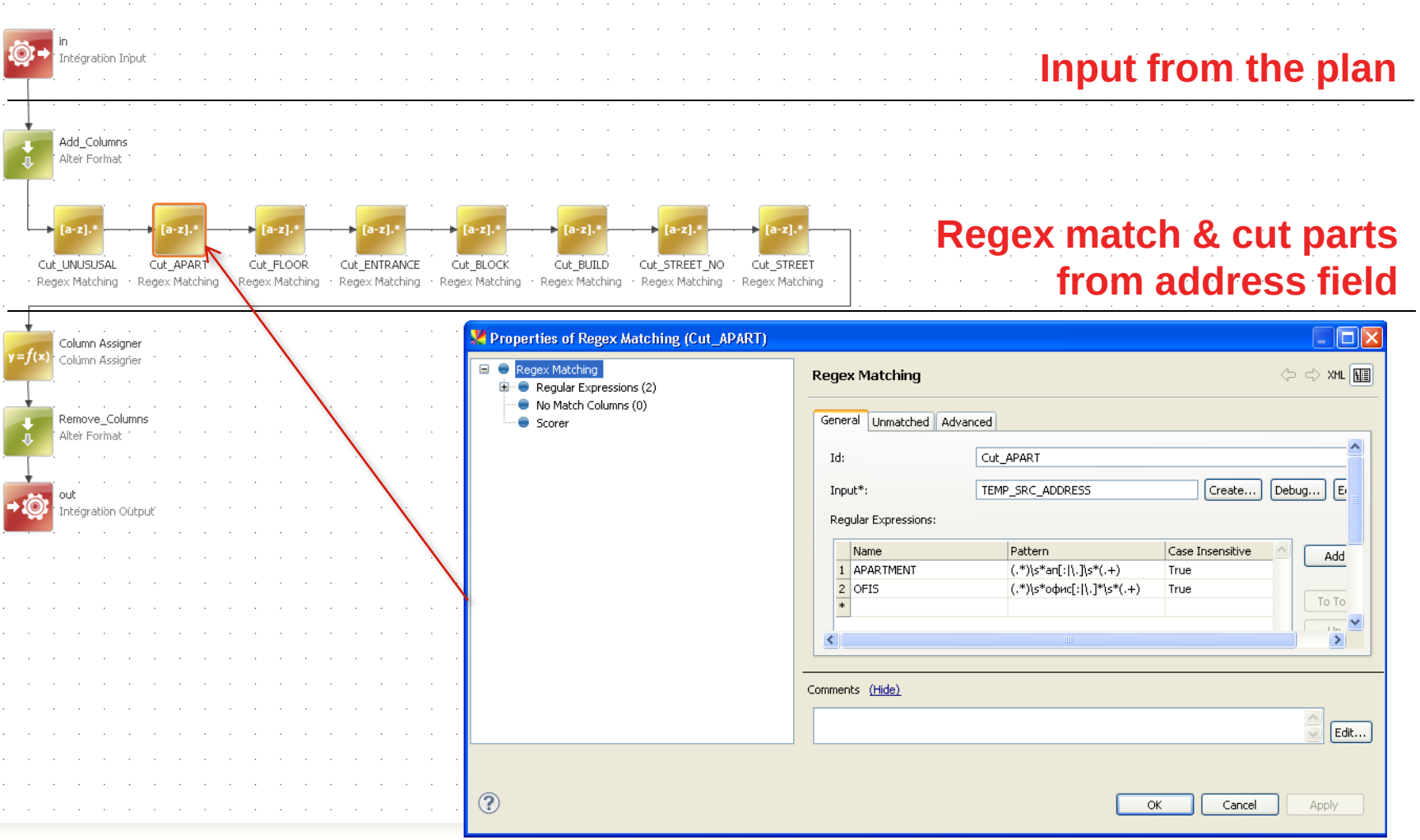
Project activities - 2. Parsing (*Plan*)

- Parsing plan – Ataccama power in practice



Project activities - 2. Parsing (*Plan components*)

- Parsing plan – components



Project activities – 3. Cleanse & Standardize

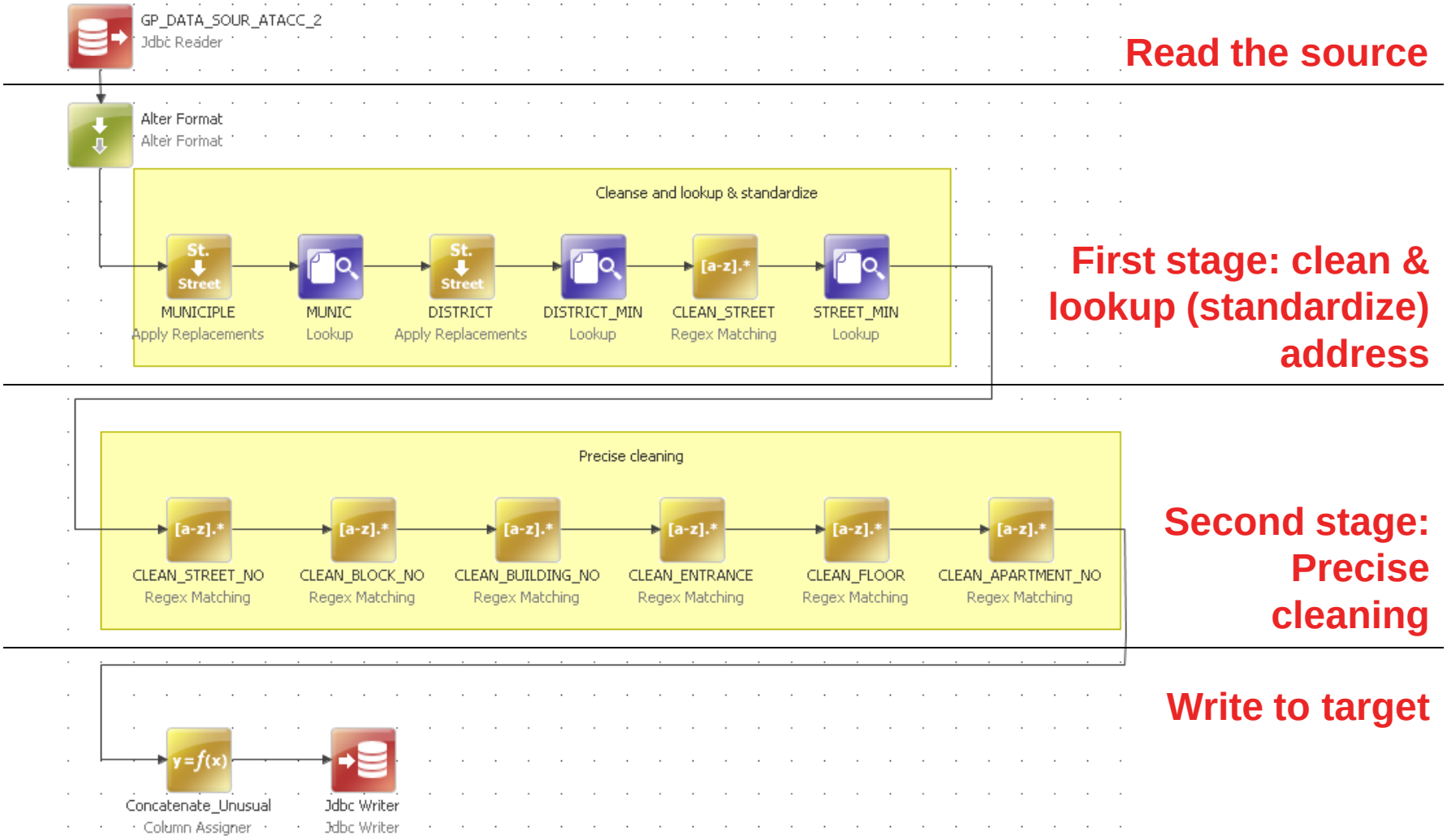
- Goal:
  - Clearing different formats;
  - Standardization through approximately lookups to clean up spelling errors.

СРДЕЦ	1	0.00%
СРЕДЕЦ	1	0.00%
СРЕДЕД	1	0.00%
СРЕДЕЦ.	1	0.00%
СРЕДЕЦИ	1	0.00%
СРЕДИКА	1	0.00%
СРЕЗЕЦ	1	0.00%
СРЕЦЕЦ	1	0.00%
СТДЕНТСКА	1	0.00%
СТИДЕНТСКА	1	0.00%
СТИУДЕНТСКА	1	0.00%
СТКУДЕНТСКА	1	0.00%
СТУДЕНТСКА	1	0.00%
СТУДЕН	1	0.00%
СТУДЕНКСА	1	0.00%
СТУДЕНТКА	1	0.00%
СТУДЕНТСК	1	0.00%
СТУДЕНТСКИА	1	0.00%
СТУДЕНТСКО	1	0.00%
СТУДНТСКА	1	0.00%
СТУДУНТСКА	1	0.00%
СТЪДЕНТСКА	1	0.00%
СУДЕНТСКА	1	0.00%

NULL	9,667	1.56%
КРАСНО СЕЛО	48,789	7.88%
ЛЮЛИН	48,533	7.84%
МЛАДОСТ	46,706	7.54%
ТРИАДИЦА	40,732	6.58%
СЛАТИНА	33,632	5.43%
ЛОЗЕНЕЦ	32,454	5.24%
ВИТОША	31,671	5.11%
НАДЕЖДА	29,973	4.84%
ПОДУЯНЕ	27,832	4.49%
ИСКЪР	27,595	4.46%
ОВЧА КУПЕЛ	24,311	3.93%
СРЕДЕЦ	23,272	3.76%
ВЪЗРАЖДАНЕ	23,169	3.74%
ОБОРИЩЕ	22,035	3.56%
СЕРДИКА	21,392	3.45%
КРАСНА ПОЛЯНА	20,266	3.27%
ВРЪБНИЦА	18,295	2.95%
ИЗГРЕВ	17,985	2.90%
ИЛИНДЕН	16,536	2.67%
СТУДЕНТСКА	14,034	2.27%
КРЕМИКОВЦИ	12,537	2.02%
НОВИ ИСКЪР	10,985	1.77%
ПАНЧАРЕВО	9,875	1.59%

Project activities – 3. Cleanse & Standardize (*Plan*)

- Cleanse plan – Ataccama power in practice



Project activities – 4. Match and Merge

- Goal – achieve a single view of customers
- Prerequisite – defined rules for matching (matching keys)

Cleansed data					
First	Last	G	SIN	Birth Date	Address
John	Smith	M		1978-12-16	V3R 2A9;BC;Surrey;14618 110 Avenue
John	Smith	M	095242434	1978-12-16	V3R 2A9;BC;Surrey;14618 110 Avenue
John	Smith	M	095242434		M4X 1V5;ON;Toronto;25 Linden Street
	Smith	M		1978-11-16	
John	Smith	M	095252433	1978-11-16	L3T 7M8;ON;Markham;8500 Leslie Str.
John	Smith	M		1978-11-16	L3T 7M8;ON;Markham;8500 Leslie Str.
John	Smiht		095252433	1978-11-16	
Jane	Watson	F	420347213		L3T 7M8;ON;Markham;8500 Leslie Str.
Jane	Watson	F	420347213	1982-01-01	L3T 7M8;ON;Markham;8500 Leslie Str.
Jane	Smith	F	420347213	1982-01-05	
J.	Smith		420347213		

Project activities – 4. Match and Merge *(cont.)*

Cleansed data					
First	Last	G	SIN	Birth Date	Address
John	Smith	M		1978-12-16	V3R 2A9;BC;Surrey;14618 110 Avenue
John	Smith	M	095242434	1978-12-16	V3R 2A9;BC;Surrey;14618 110 Avenue
John	Smith	M	095242434		M4X 1V5;ON;Toronto;25 Linden Street

Golden record					
First	Last	G	SIN	Birth Date	Address
John	**Smith**	**M**	**095242434**	**1978-12-16**	**M4X 1V5;ON;Toronto;25 Linden Street**

V3R 2A9;BC;Surrey;14618 110 Avenue

The most frequent address

The newest permanent address

Project activities – 5. Enrich

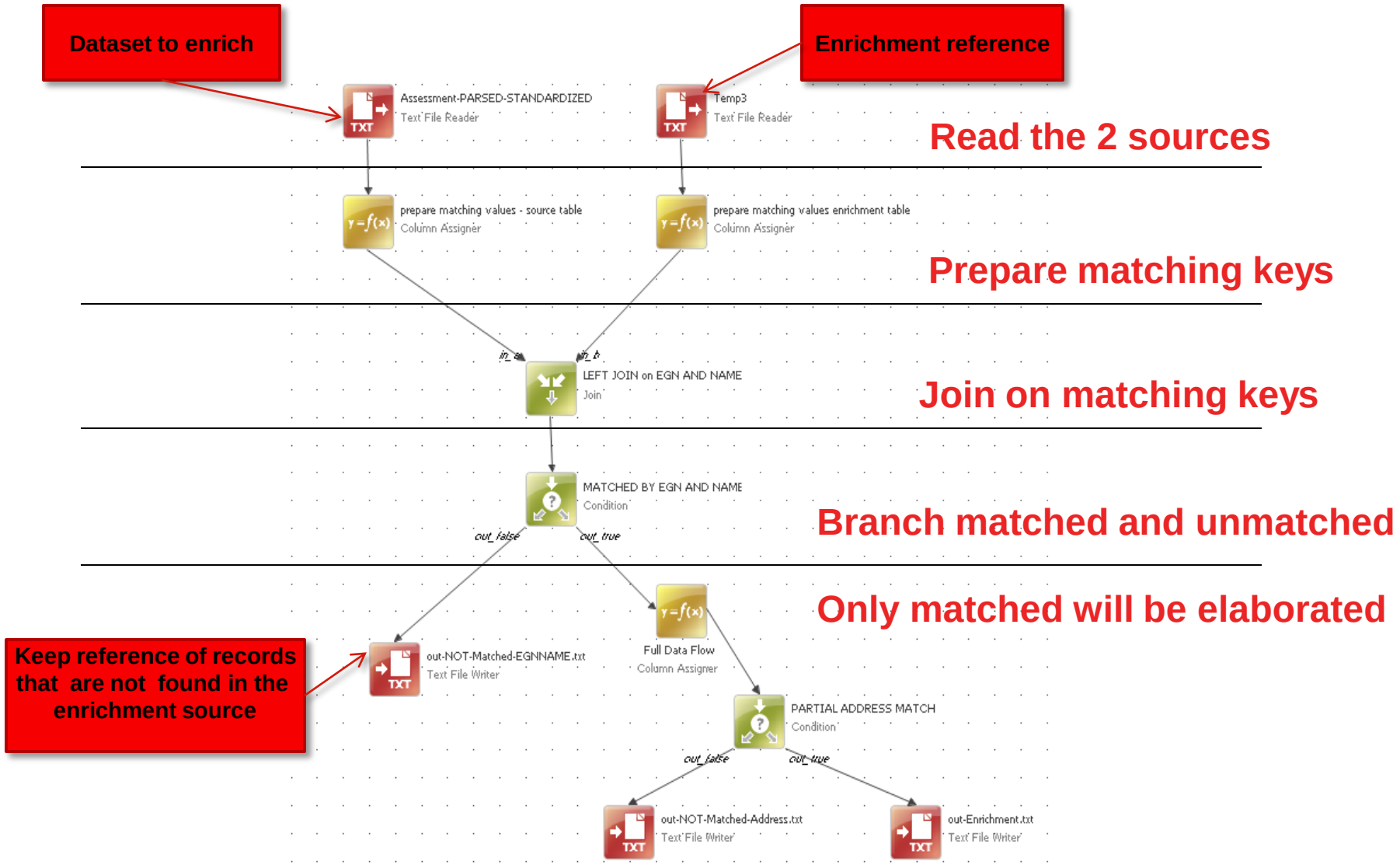
Enriched ZIP

Missing ZIP

	pur_zip	pur_district	std_district	titles_out_quarter	pur_quarter	enr_POSTCODE
1		СТУДЕНТСКА	СТУДЕНТСКИ	КВ.	ДЪРВЕНИЦА	1756
2	1618	КРАСНО СЕЛО	КРАСНО СЕЛО	ЖК.	КРАСНО СЕЛ	1618
3	1229	НАДЕЖДА	НАДЕЖДА	ЖК.	НАДЕЖДА III	1229
4		ЛЮЛИН	ЛЮЛИН	ЖК.	ЛЮЛИН-IX	1373
5	1281	НОВИ ИСКЪР	НОВИ ИСКЪР		С СЛАВОВЦИ	1281
6	1000	ЛОЗЕНЕЦ	ЛОЗЕНЕЦ	КВ.	ЛОЗЕНЕЦ	1164
7	1504	ОБОРИЩЕ	ОБОРИЩЕ	КВ.	ДОКТОРСКИ ПАМЕТНИК	1504
8	1172	ИЗГРЕВ	ИЗГРЕВ	ЖК.	ДИАНАБАД	1504
9	1309	ИЛИНДЕН	ИЛИНДЕН	ЖК.	СВЕТА ТРОИЦА	1309
10	1680	ТРИАДИЦА	ТРИАДИЦА	ЖК.	БОКАР	1404
11	1680	ТРИАДИЦА	ТРИАДИЦА	ЖК.	БОКАР	1404
12	1505	СЛАТИНА	СЛАТИНА	КВ.	РЕДУТА	1505
13	1231	НАДЕЖДА	НАДЕЖДА	ЖК.	СВОБОДА	1309
14		ЛОЗЕНЕЦ	ЛОЗЕНЕЦ	КВ.	ЛОЗЕНЕЦ	1407
15		ЛОЗЕНЕЦ	ЛОЗЕНЕЦ	КВ.	ЛОЗЕНЕЦ	1407
16	1220	НАДЕЖДА	НАДЕЖДА	ЖК.	НАДЕЖДА II	1220
17		ОВЧА КУПЕЛ	ОВЧА КУПЕЛ	КВ.	ОВЧА КУПЕЛ 1	1618
18	1618	ОВЧА КУПЕЛ	ОВЧА КУПЕЛ	КВ.	ОВЧА КУПЕЛ	1618
19	1618	ОВЧА КУПЕЛ	ОВЧА КУПЕЛ	КВ.	ОВЧА КУПЕЛ	1618
20	1592	ИСКЪР	ИСКЪР	ЖК.	ДРУЖБА-I	1592
21	1379	ВЪЗРАЖДАНЕ	ВЪЗРАЖДАНЕ	ЖК.	СЕРДИКА	1379
22	1505	ОБОРИЩЕ	ОБОРИЩЕ	КВ.	ПОДУЯНЕ-ЦЕНТЪР	1320
23		ИСКЪР	ИСКЪР	ЖК.	ДРУЖБА-I	1330
24	1618	КРАСНО СЕЛО	КРАСНО СЕЛО	ЖК.	БЪКСТОН	1618
25	1618	КРАСНО СЕЛО	КРАСНО СЕЛО	ЖК.	БЪКСТОН	1618
26		ИЛИНДЕН	ИЛИНДЕН	КВ.	ЗАХАРНА ФАБРИКА	1618
27		ВИТОША	ВИТОША	ЖК.	БЪКСТОН	1618
28		КРАСНО СЕЛО	КРАСНО СЕЛО	ЖК.	БОРОВО	1680

Project activities – 5. Enrich (Plan)

- Goal - elaborate with additional information from reference sources



Project Delivery

Good Data = Good Business

- Cleansed Data
  - Corrected typing errors, removed dummy characters, etc.
- Merged Data
  - No duplications of records
- Elaborated Data
  - All addresses are completed with ZIP codes
- Refined Data Quality Rules
  - Implementing rules to observe data quality
- Data Quality Report
  - Current Data Quality status and quantified DQ issues
  - Next steps of DQ improvement

Key Takeaways

- Good company
  - Adastra is a solid world company and gives a career opportunity for motivated young people without experience
- Good perspectives
  - Data Quality is an increasing niche and perspective market
- Good skills
  - Gain both business knowledge and technical skills with best-of-breed technologies

Thank You

ADASTRA Bulgaria

29 Panayot Volov Str.
1527 Sofia, Bulgaria
Tel: +359 2 960 00 30

2 Dunav Str.
9000 Varna, Bulgaria
Tel: +359 2 960 29 95

www.bg.adastragrp.com
infobg@adastragrp.com
jobsbg@adastragrp.com

ADASTRA GROUP North America
8500 Leslie Str., Suite 600
Markham, Ontario
CANADA L3T 7M8
Tel: +1 905 881 7946
info@adastragrp.com