Data Analysis and Integration

Data profiling

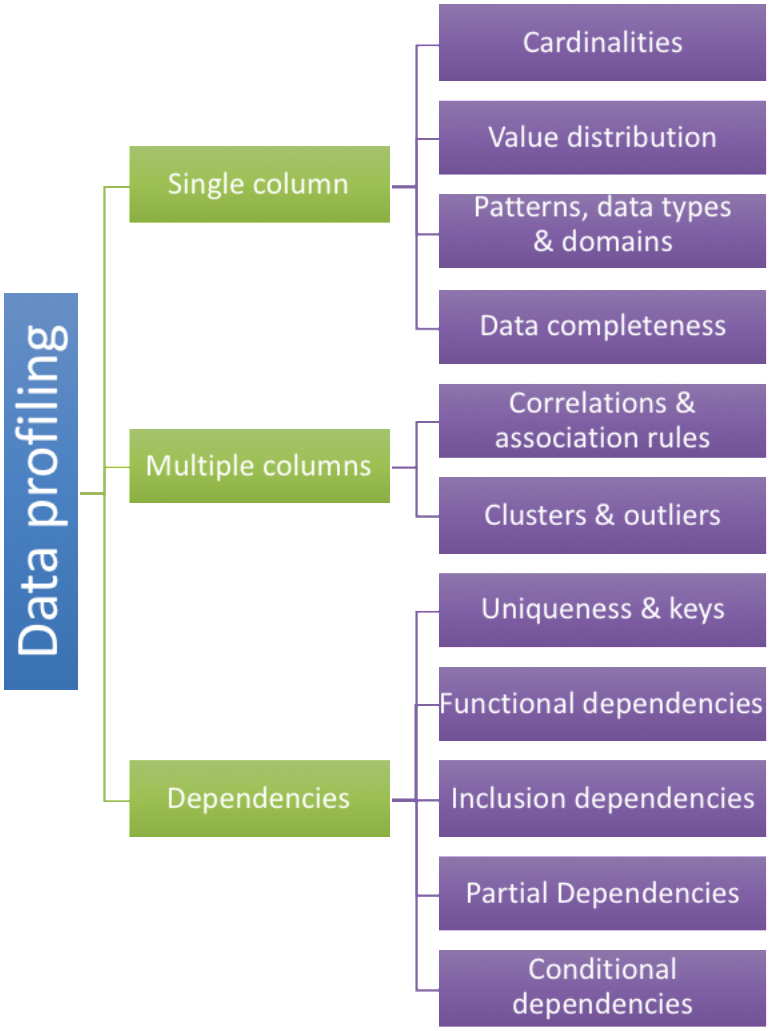
Data profiling

- What is data profiling?
  - analyze the contents of a data source
  - gather statistics about the data contained therein
    - minimum, maximum, average, range, value distribution, etc.
  - identify data quality problems
    - missing or incomplete data, errors, inconsistencies, etc.
  - understand the logic and relationships between data
    - unique values, keys, foreign keys and other constraints

Data profiling

- These tasks (and many more) are data profiling
  - number of rows, number of null values, number of distinct values
  - minimum value, maximum value, minimum length, maximum length
  - single- and multi-column frequency histogram
  - precision of numeric values, length of string values
  - data type discovery
  - uniqueness and constancy
  - single- and multi-column primary key discovery
  - single- and multi-column foreign key discovery
  - value overlap (cross-domain analysis)
  - text field profiling
  - pattern discovery (e.g. phone number patterns)
  - soundex frequency histogram
  - etc.

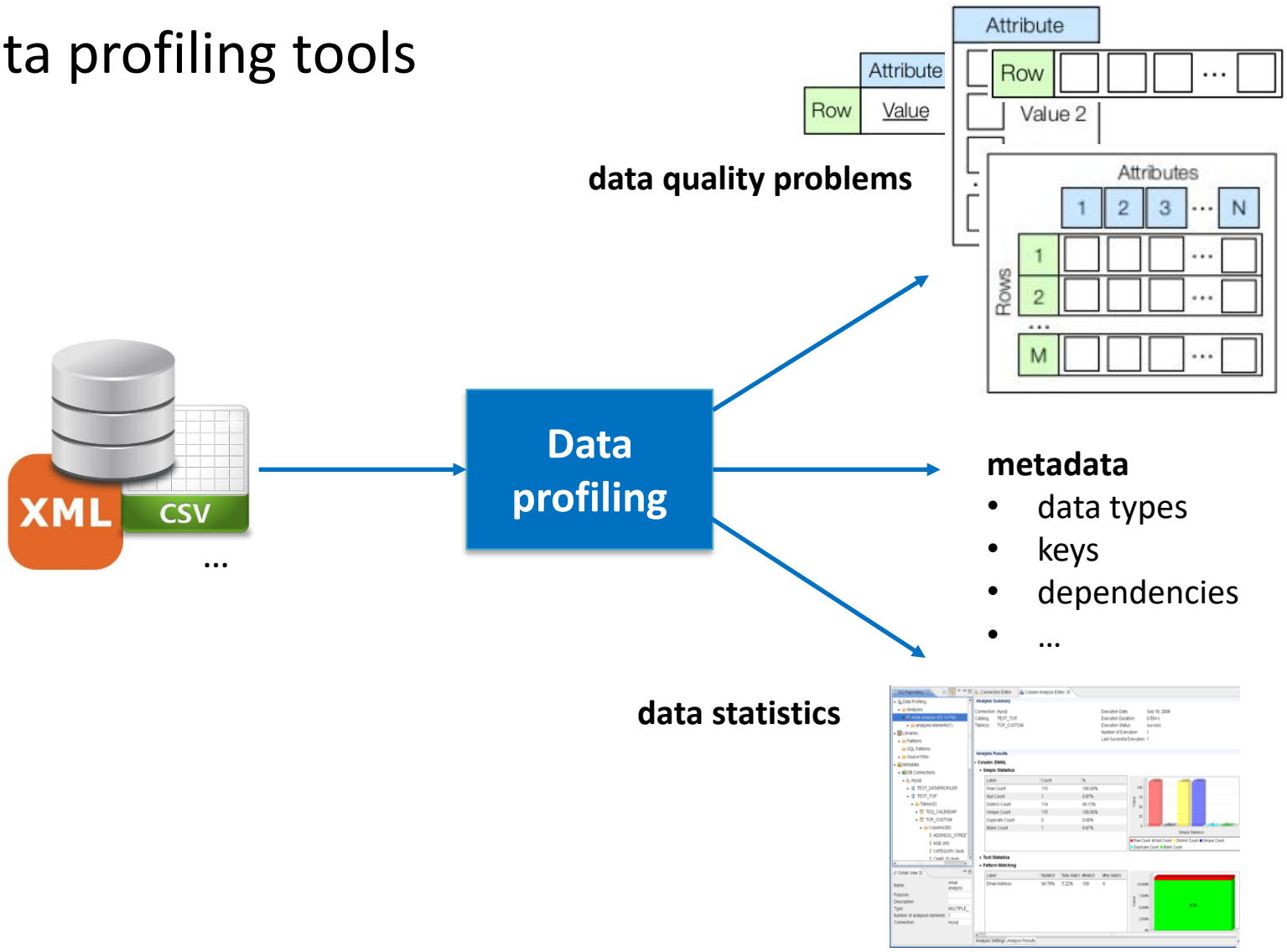
Data profiling



Z. Abedjan, L. Golab, F. Naumann
Profiling relational data: a survey
VLDB Journal, vol. 24, no. 4, 2015

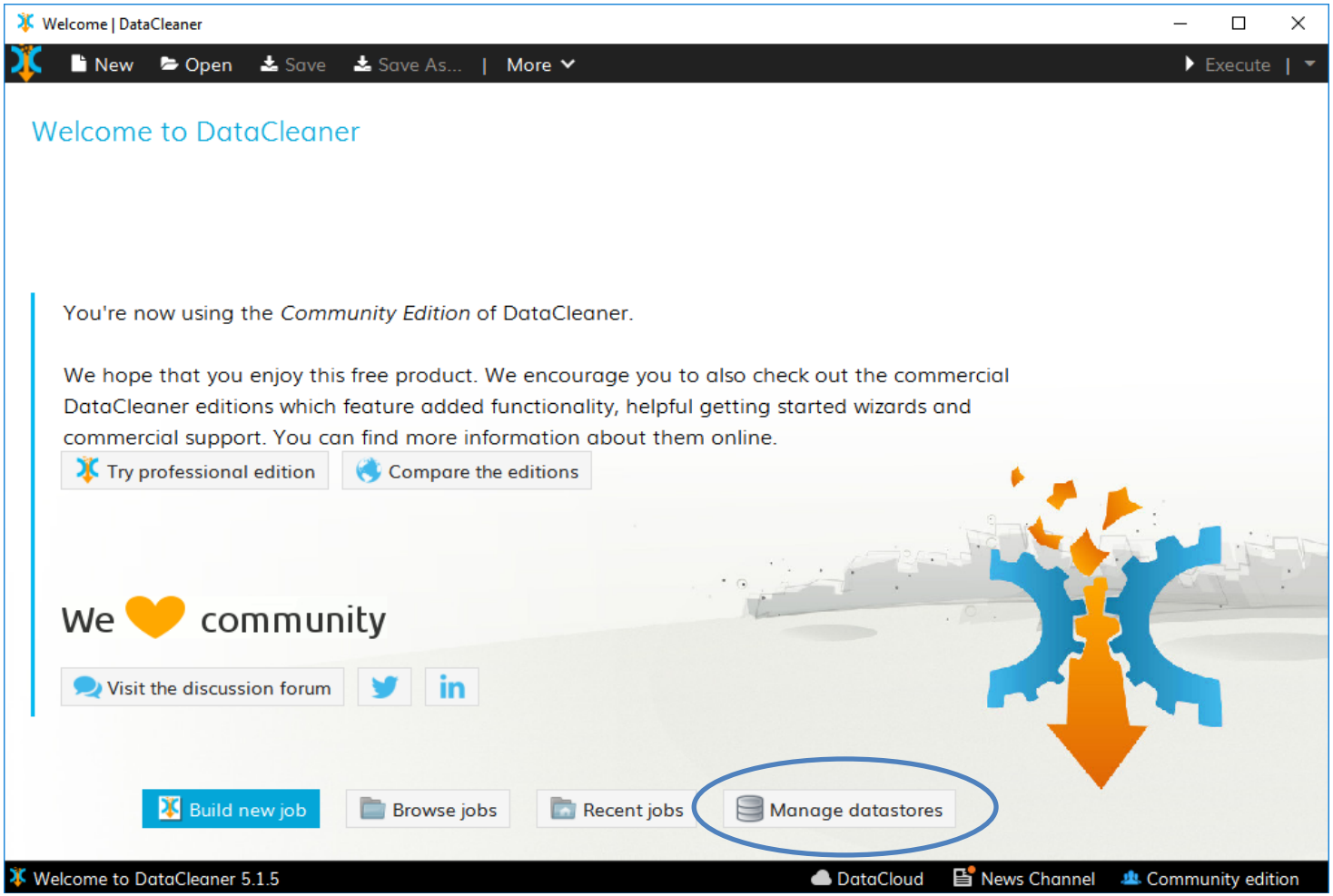
Data profiling

- Data profiling tools



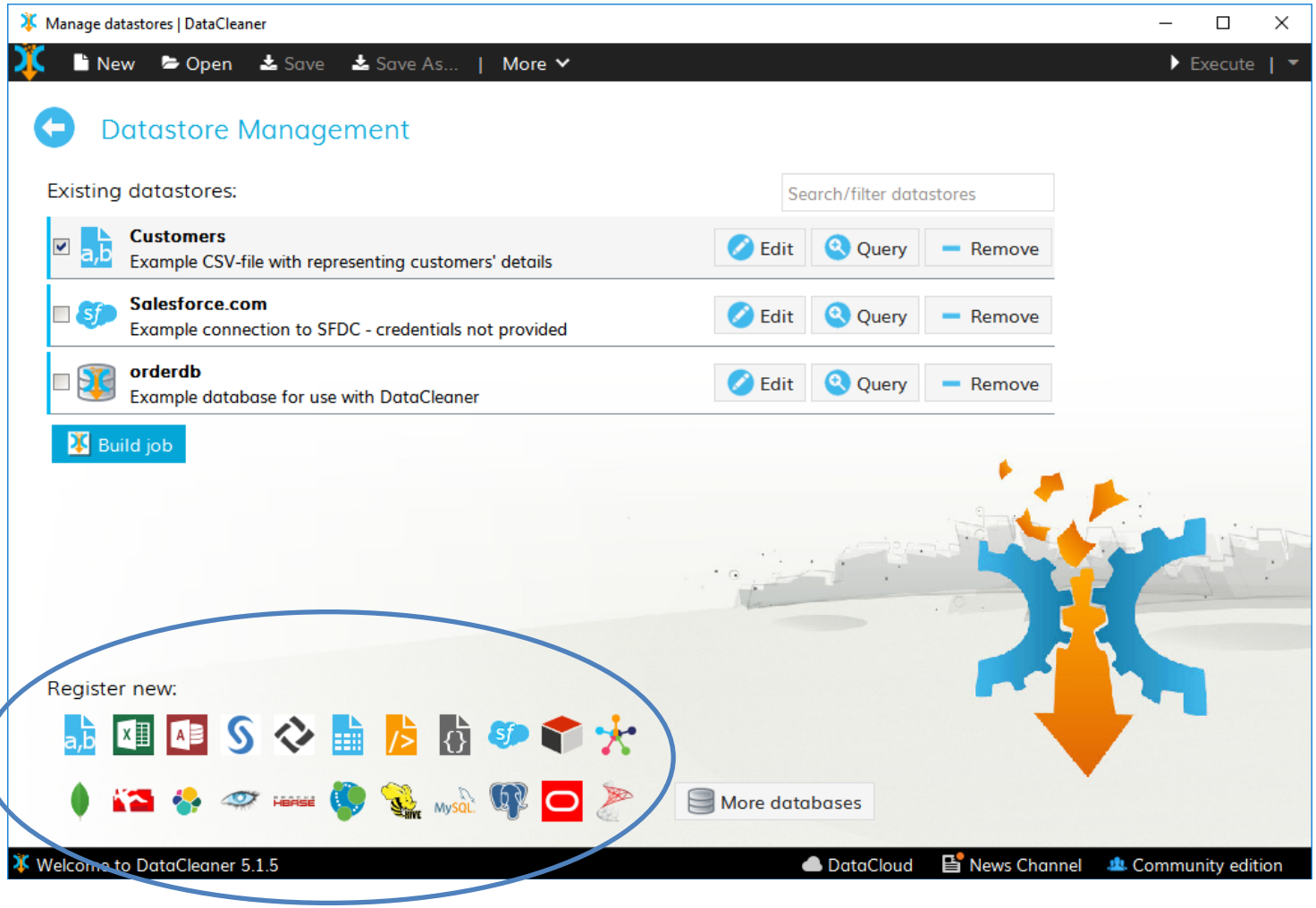
Data profiling

- Data profiling tools



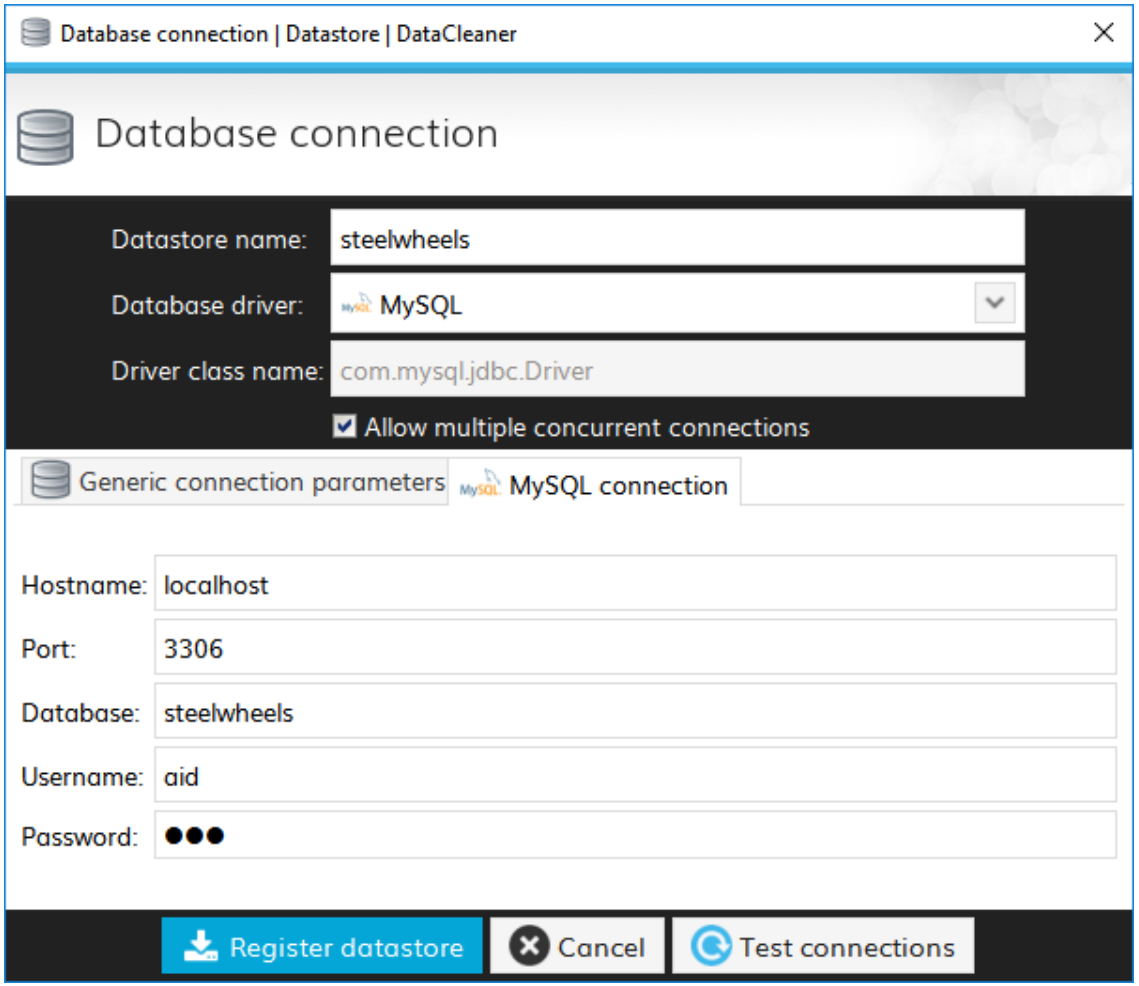
Data profiling

- A datastore can be a file, database, etc.



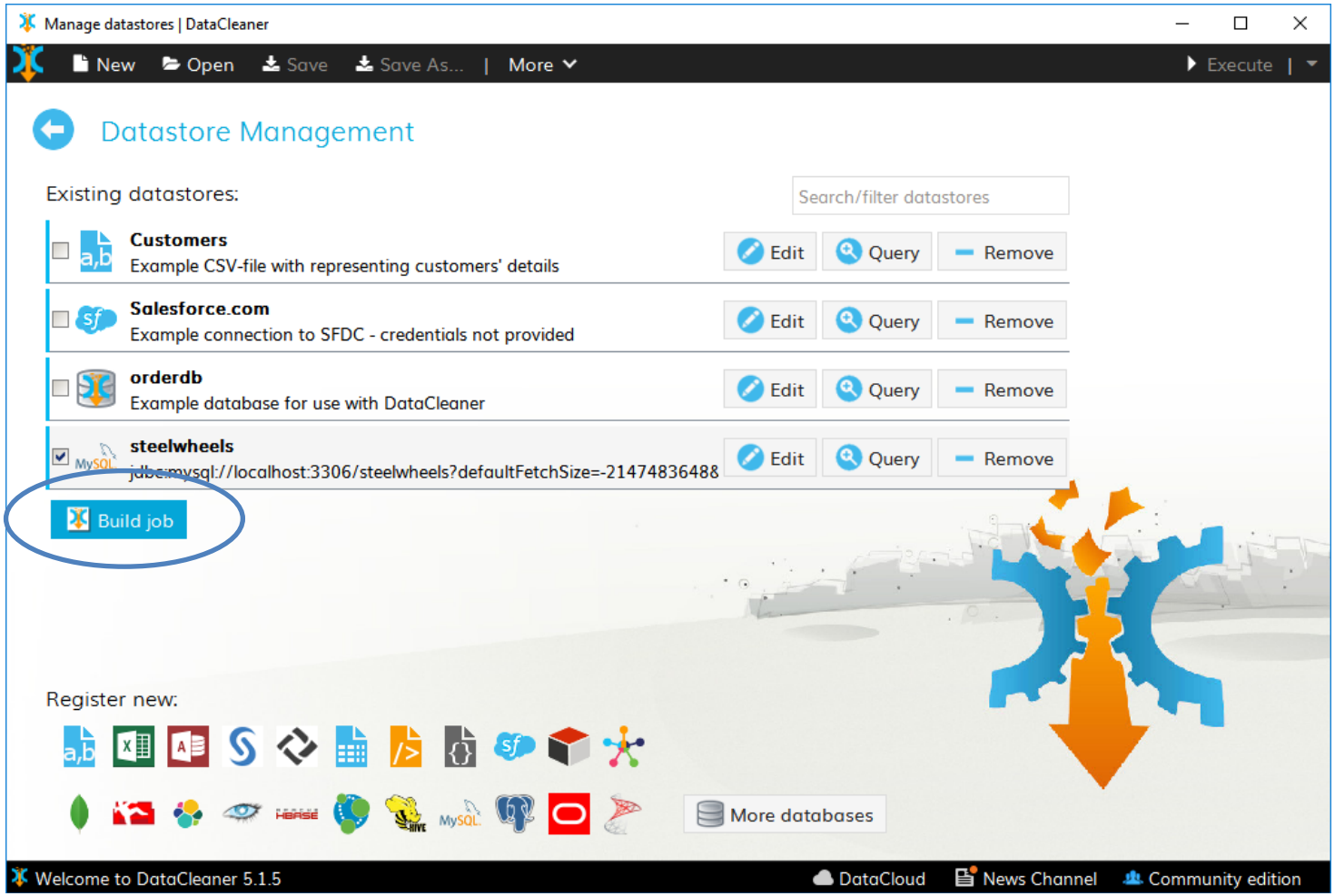
Data profiling

- Creating a new datastore



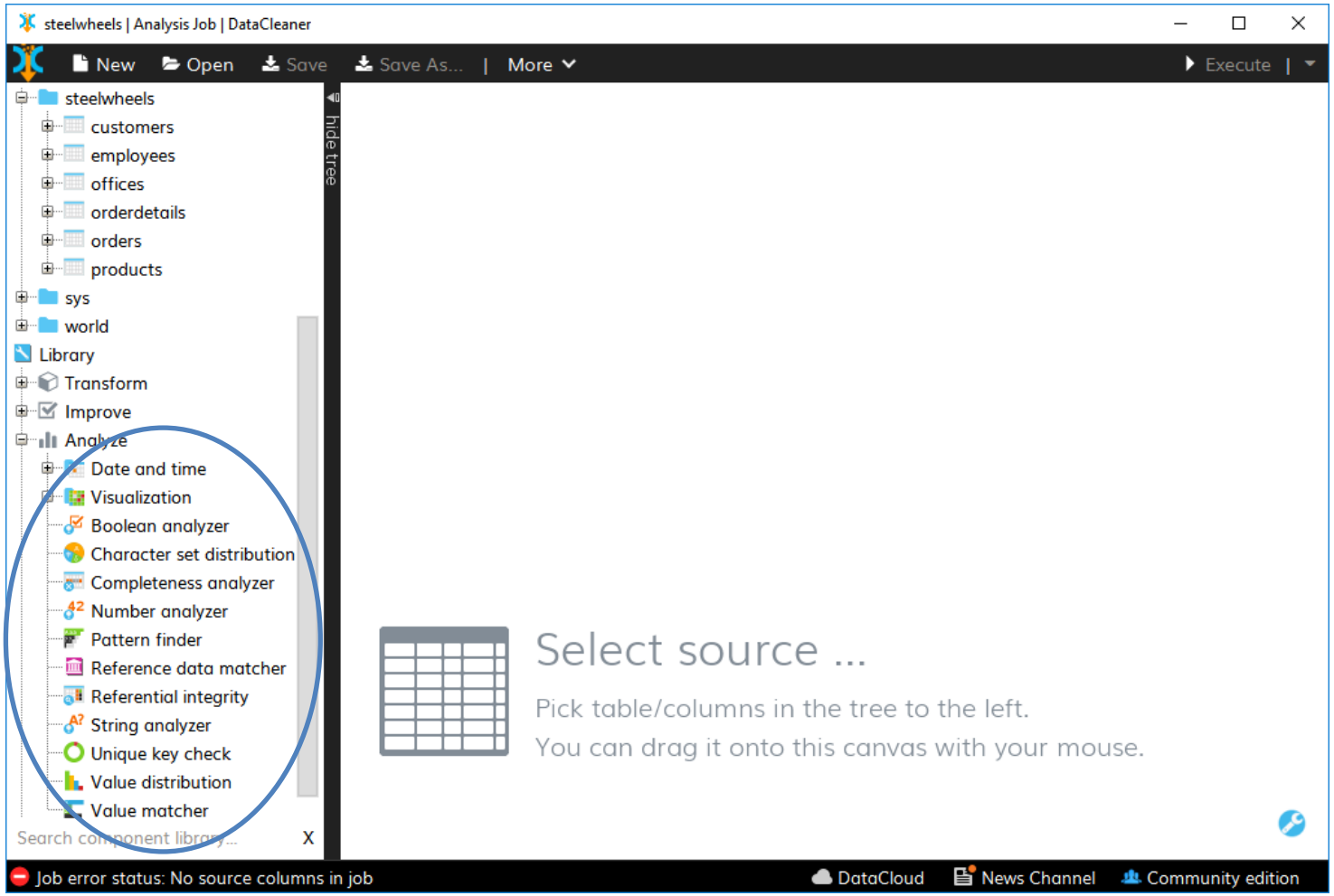
Data profiling

- Creating a new datastore



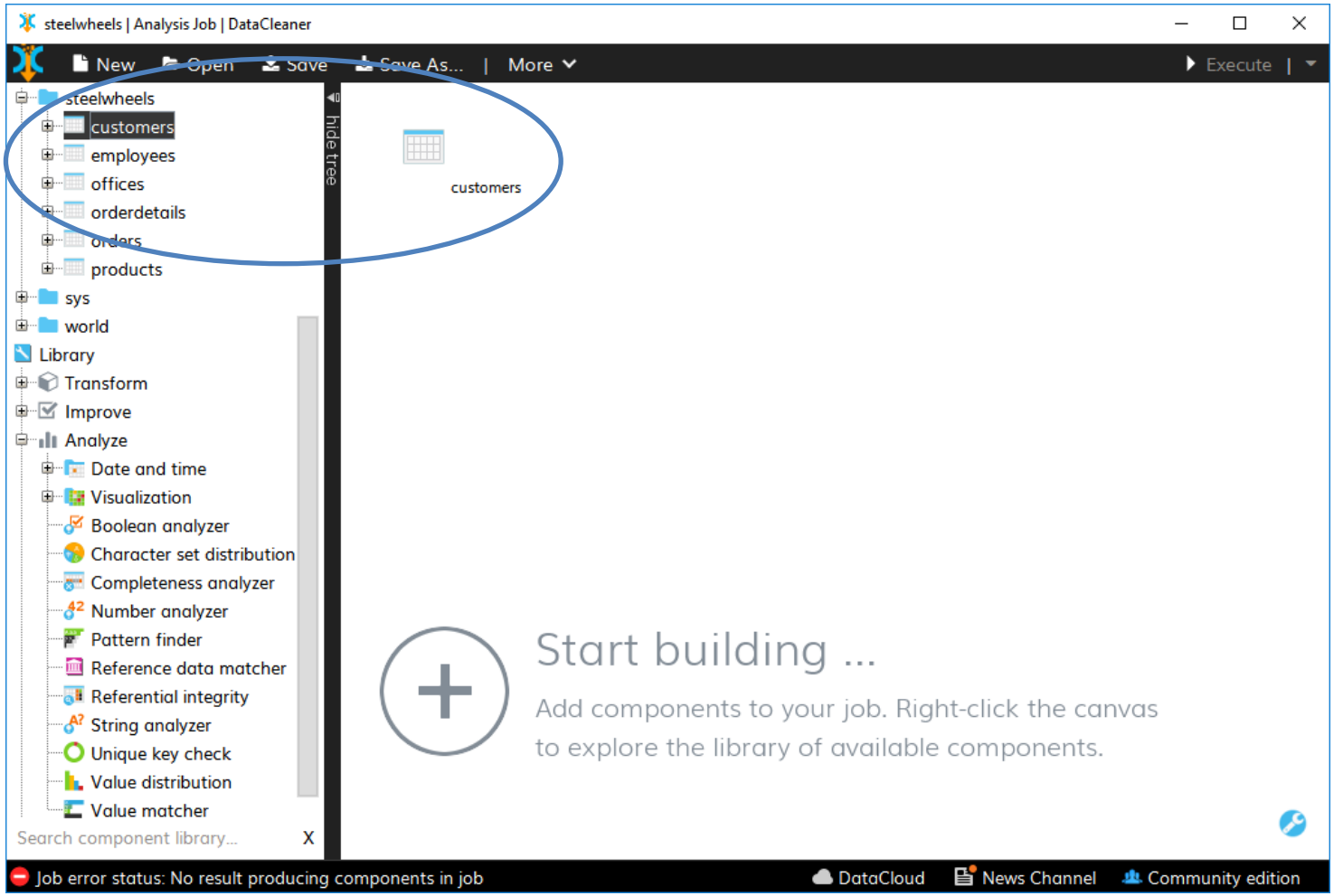
Data profiling

- Several options for data profiling tasks



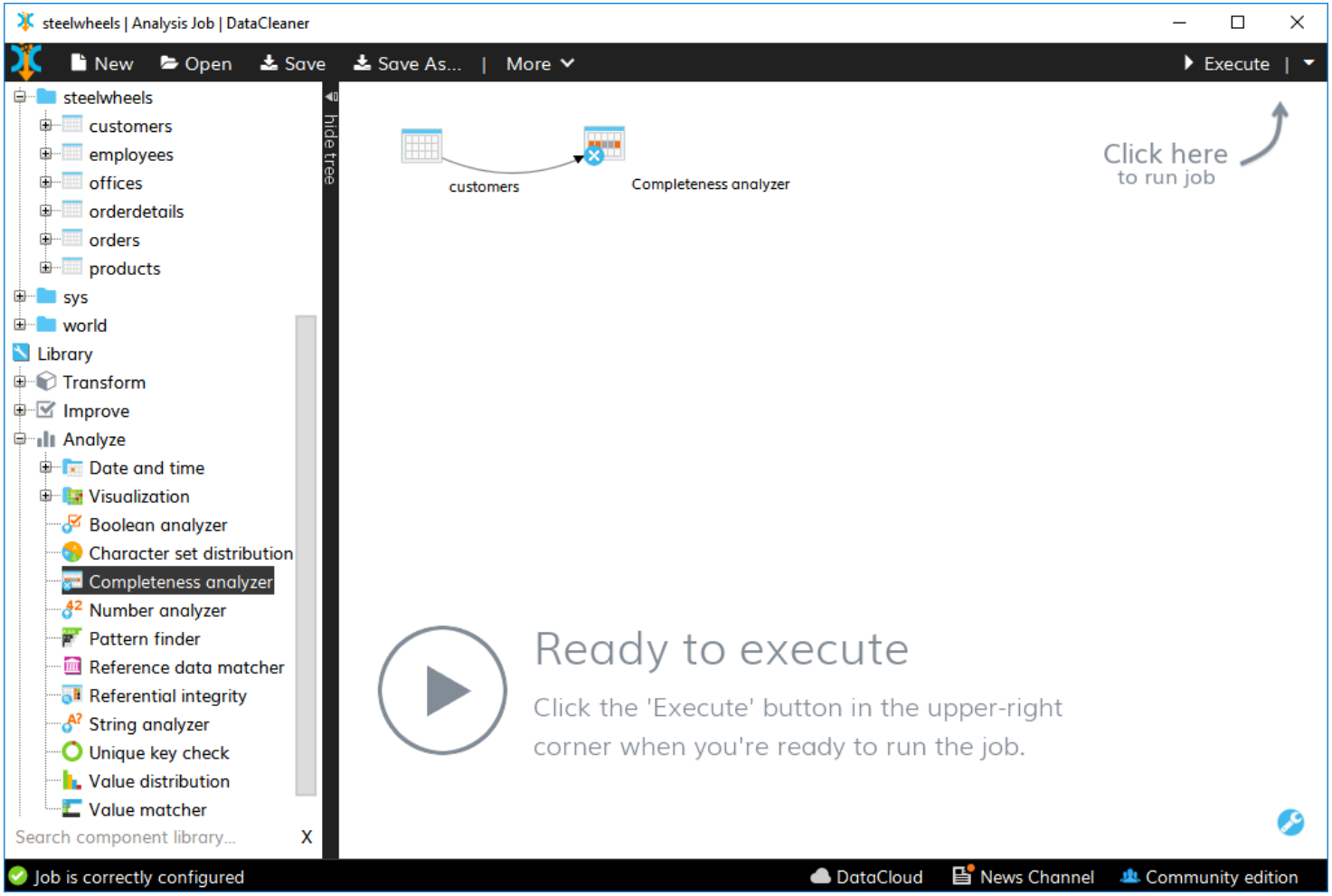
Data profiling

- Selecting the data source



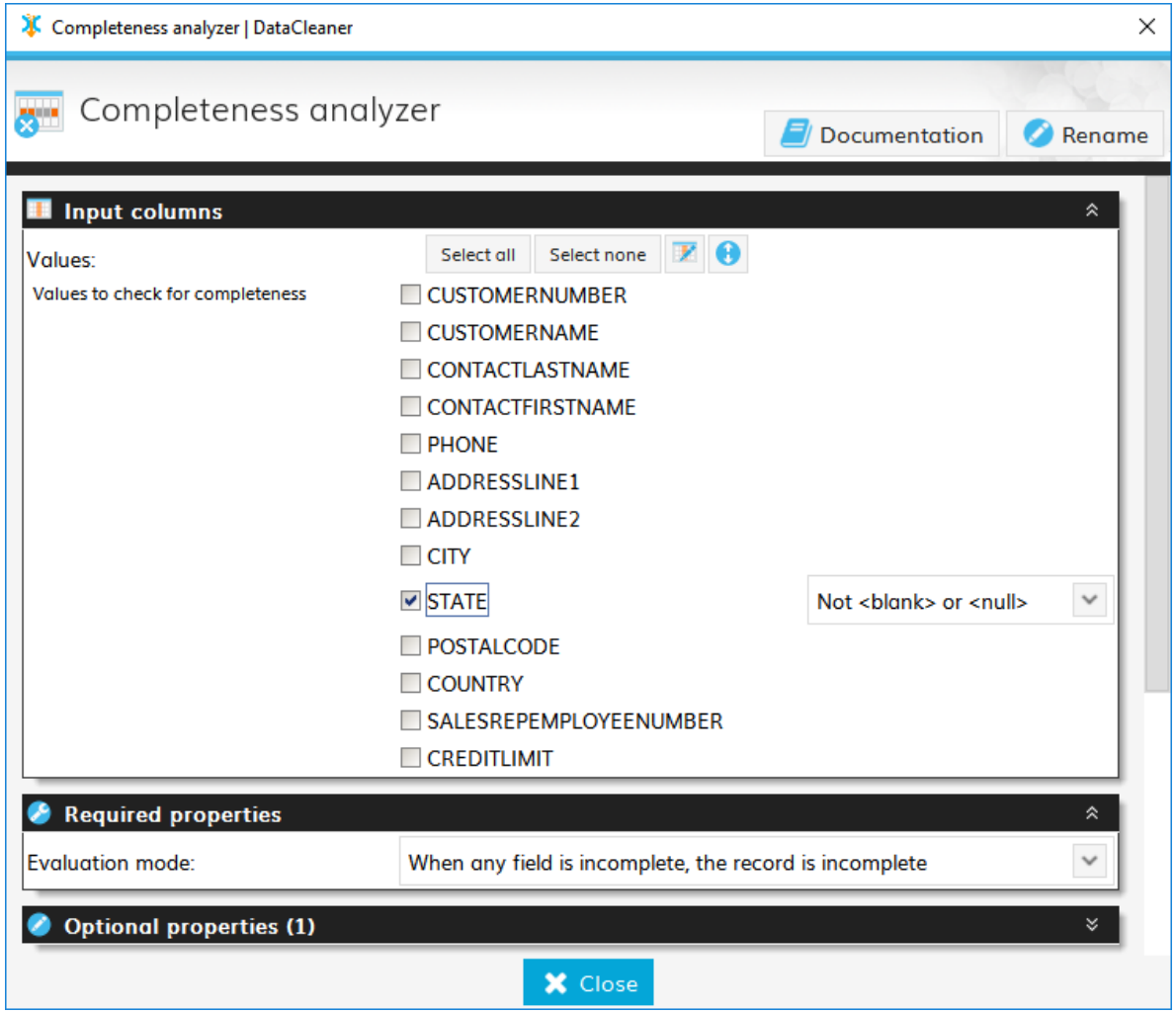
Data profiling

- Completeness analysis



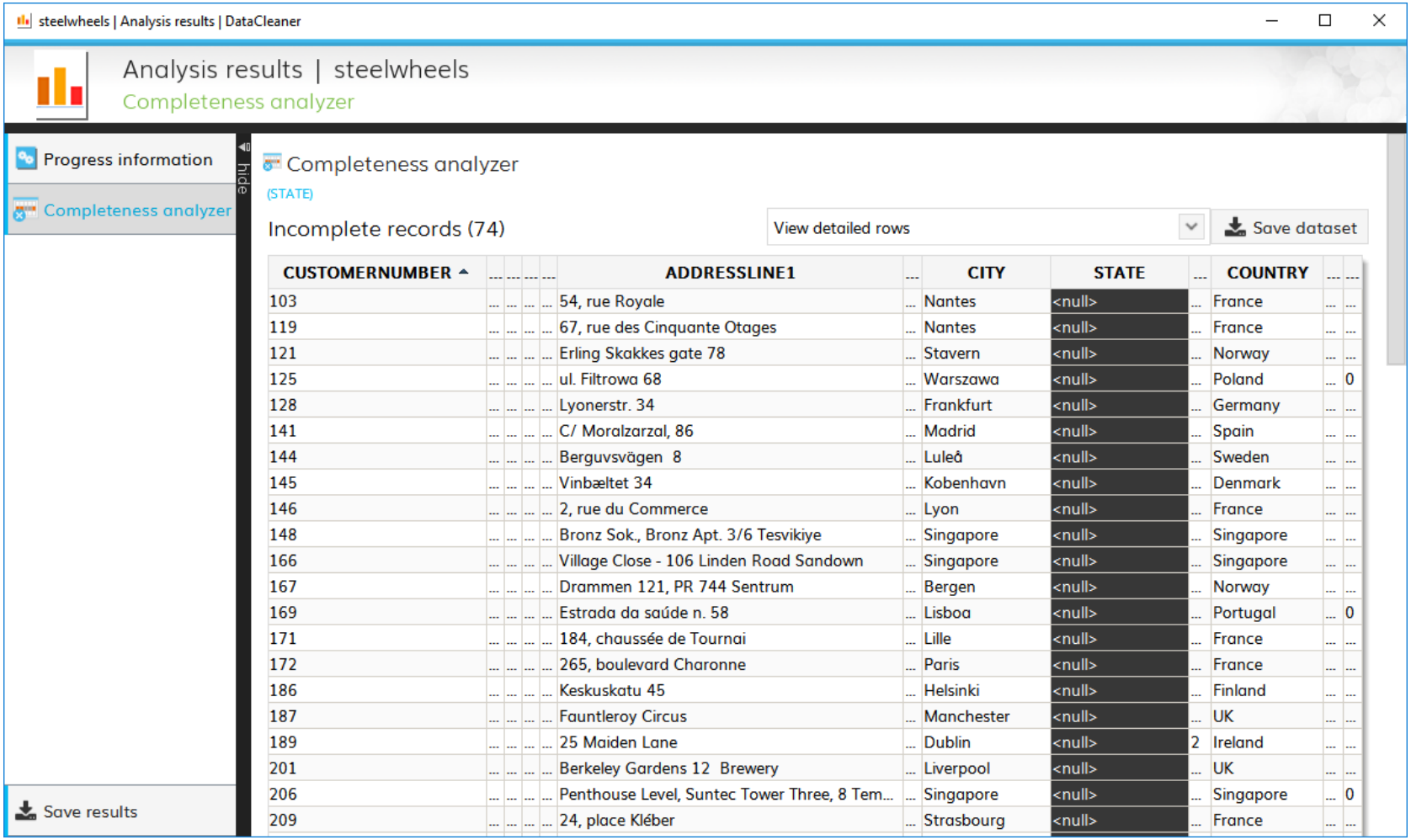
Data profiling

- Completeness analysis



Data profiling

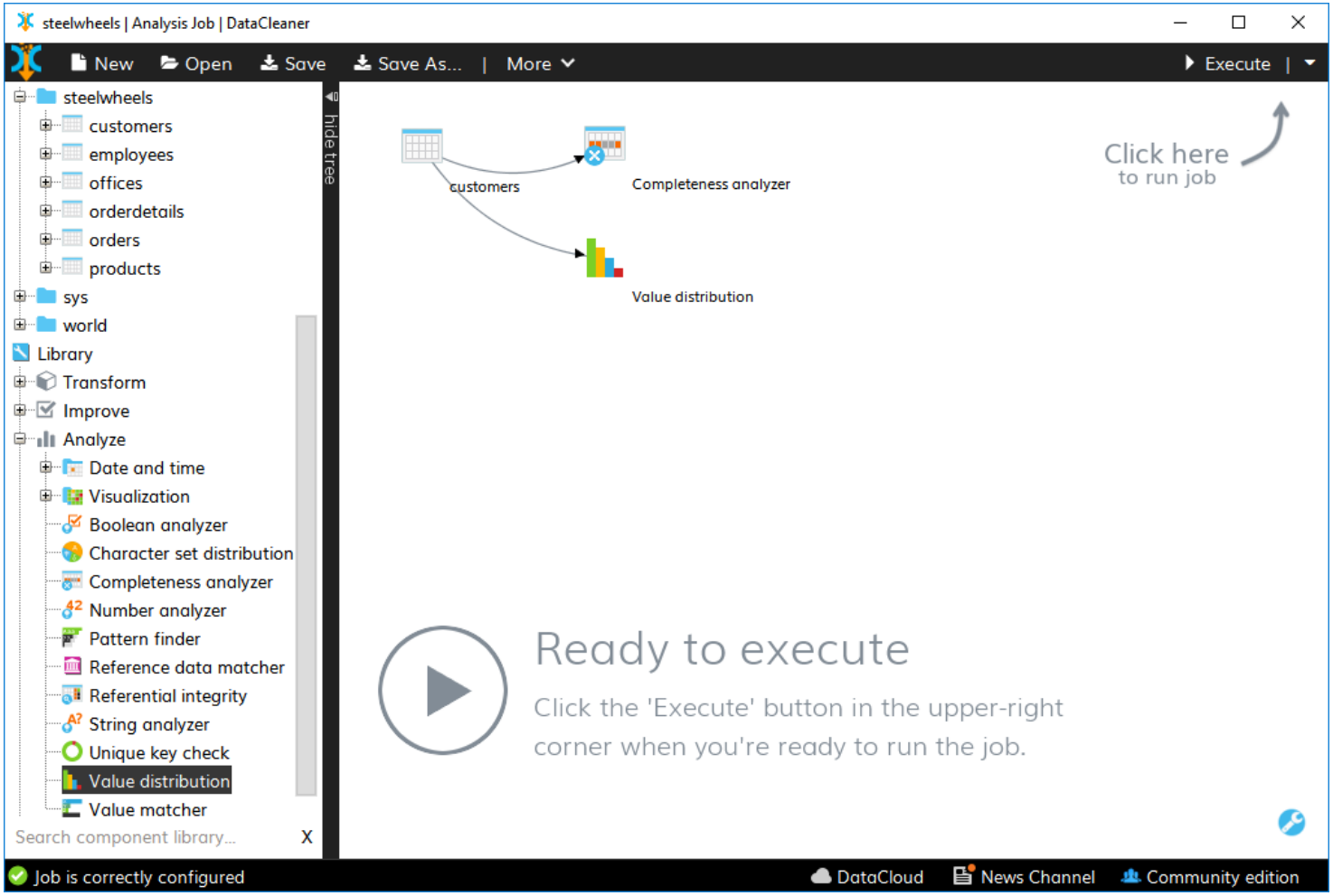
- Completeness analysis



CUSTOMERNUMBER	ADDRESSLINE1	CITY	STATE	COUNTRY
103	54, rue Royale	Nantes	<null>	France
119	67, rue des Cinquante Otages	Nantes	<null>	France
121	Erling Skakkes gate 78	Stavern	<null>	Norway
125	ul. Filtrowa 68	Warszawa	<null>	Poland
128	Lyonerstr. 34	Frankfurt	<null>	Germany
141	C/ Moralzarzal, 86	Madrid	<null>	Spain
144	Berguvsvägen 8	Luleå	<null>	Sweden
145	Vinbæltet 34	Kobenhavn	<null>	Denmark
146	2, rue du Commerce	Lyon	<null>	France
148	Bronz Sok., Bronz Apt. 3/6 Tesvikiye	Singapore	<null>	Singapore
166	Village Close - 106 Linden Road Sandown	Singapore	<null>	Singapore
167	Drammen 121, PR 744 Sentrum	Bergen	<null>	Norway
169	Estrada da saúde n. 58	Lisboa	<null>	Portugal
171	184, chaussée de Tournai	Lille	<null>	France
172	265, boulevard Charonne	Paris	<null>	France
186	Keskuskatu 45	Helsinki	<null>	Finland
187	Fauntleroy Circus	Manchester	<null>	UK
189	25 Maiden Lane	Dublin	<null>	Ireland
201	Berkeley Gardens 12 Brewery	Liverpool	<null>	UK
206	Penthouse Level, Suntec Tower Three, 8 Tem...	Singapore	<null>	Singapore
209	24, place Kléber	Strasbourg	<null>	France

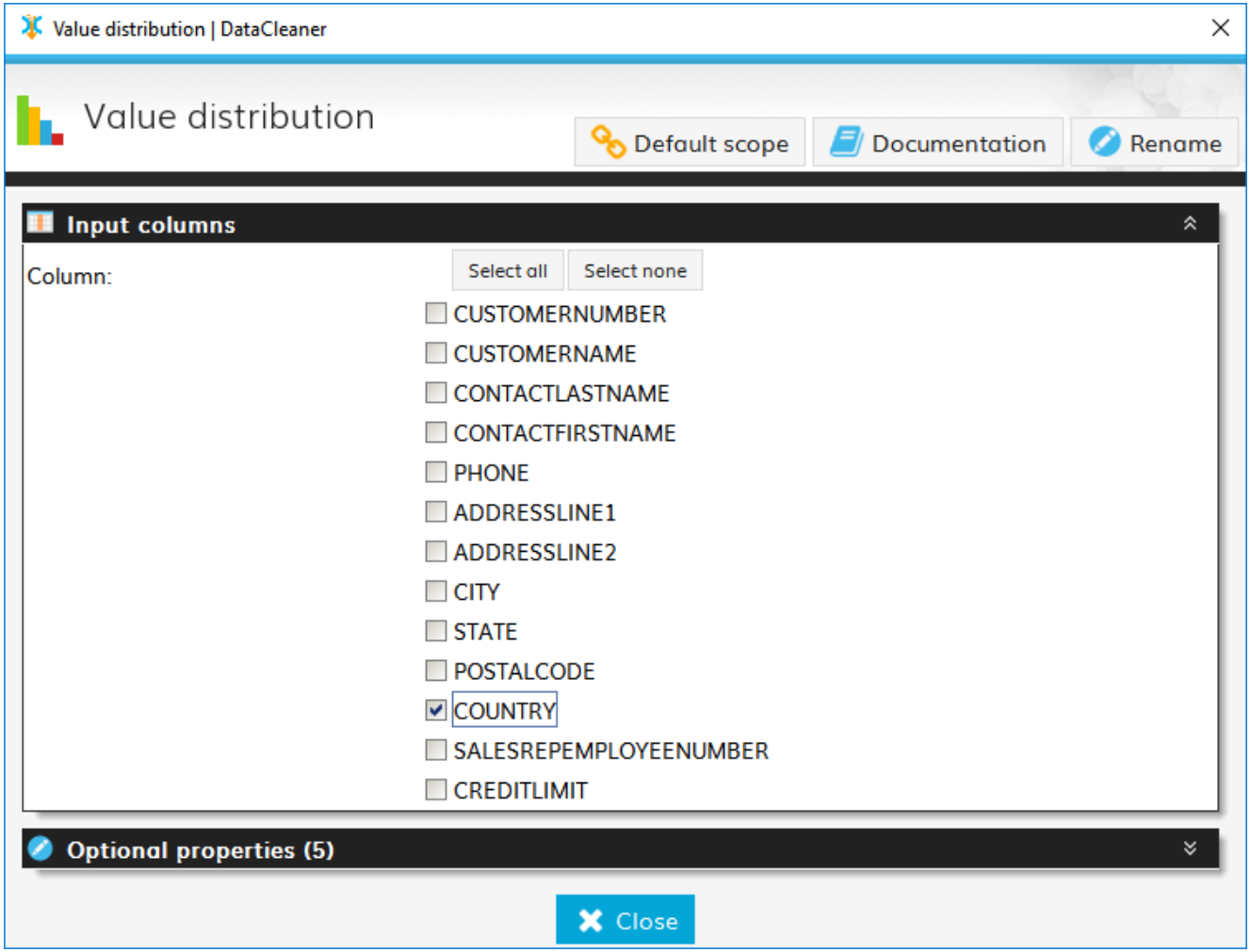
Data profiling

- Value distribution



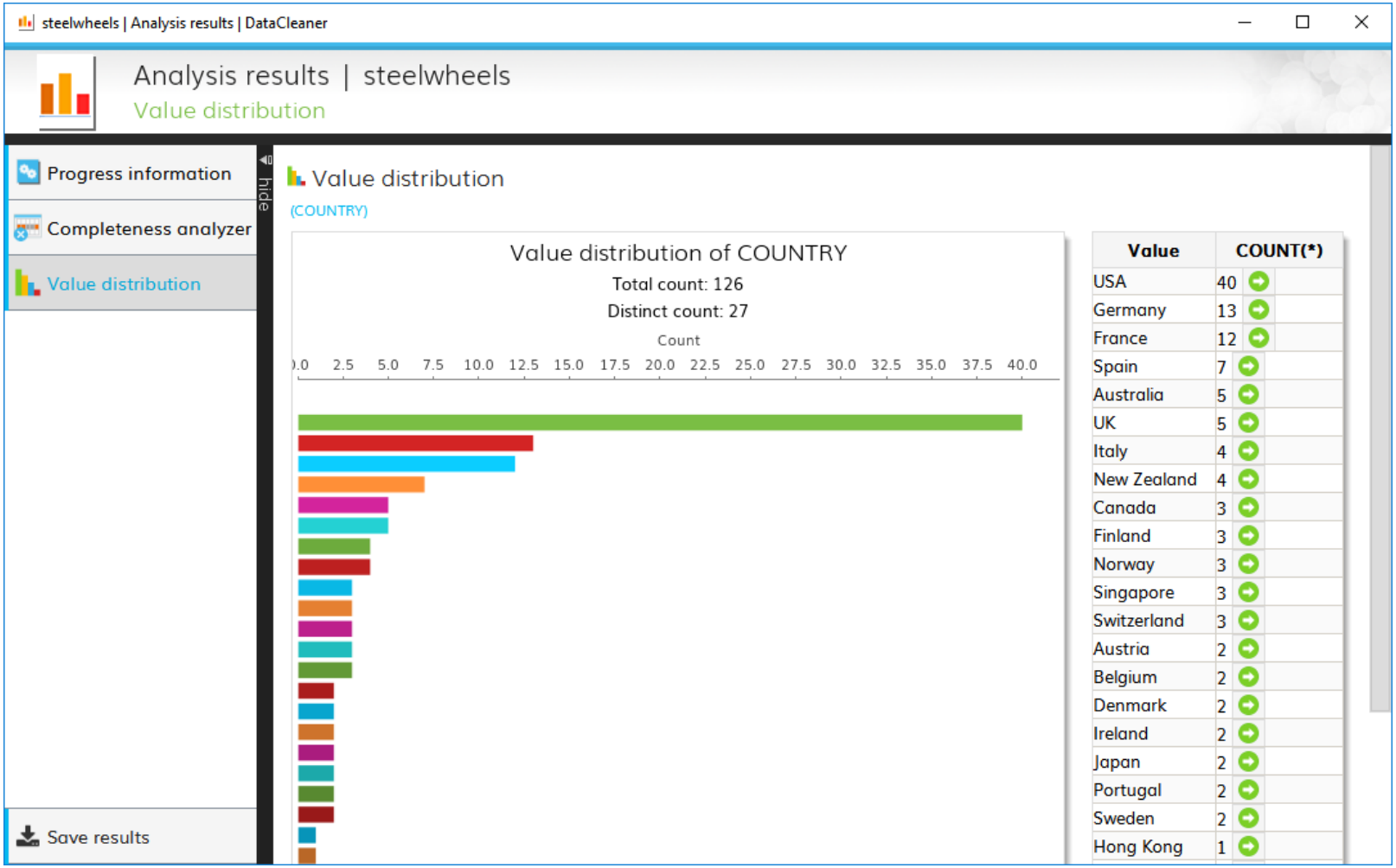
Data profiling

- Value distribution



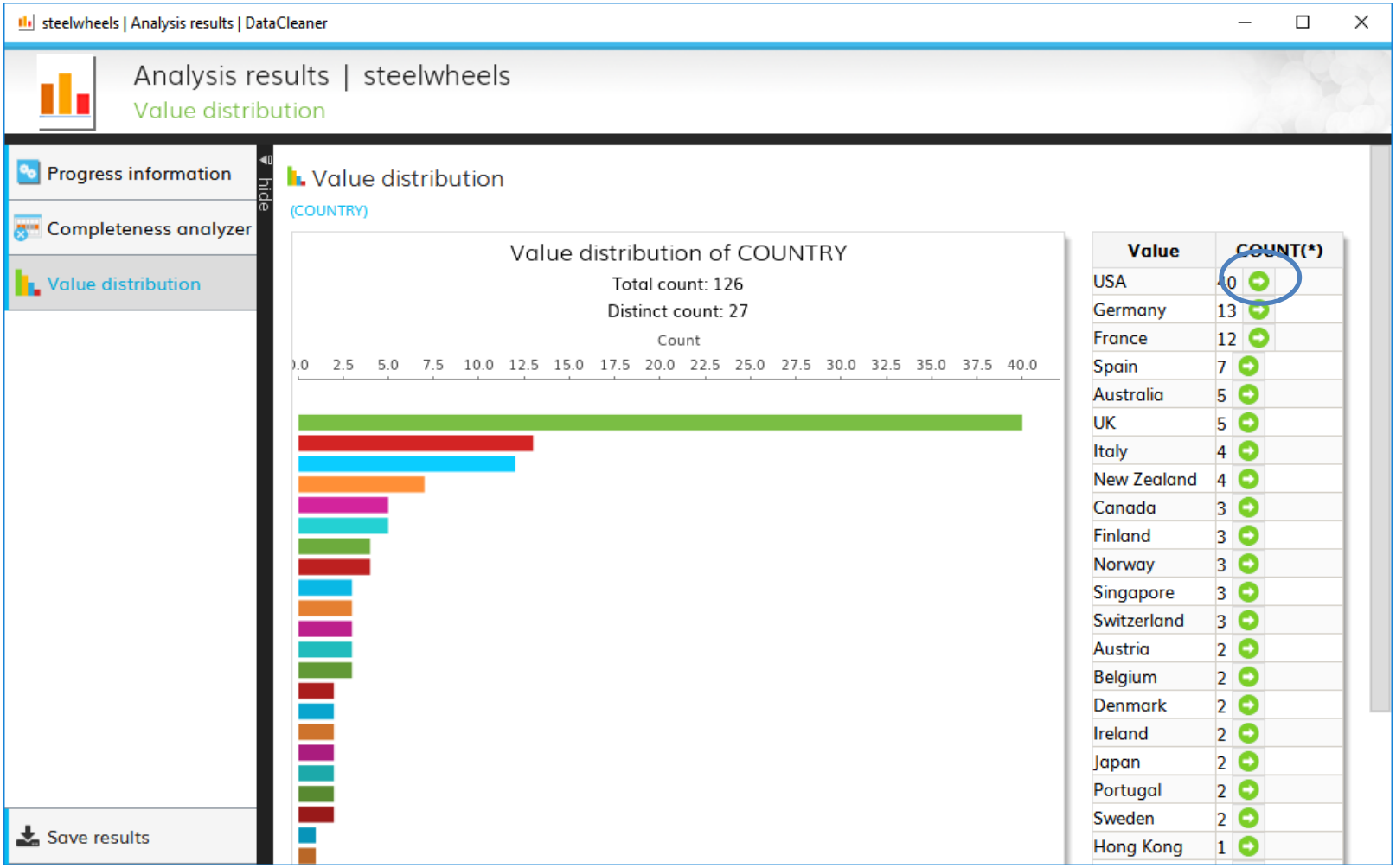
Data profiling

- Value distribution



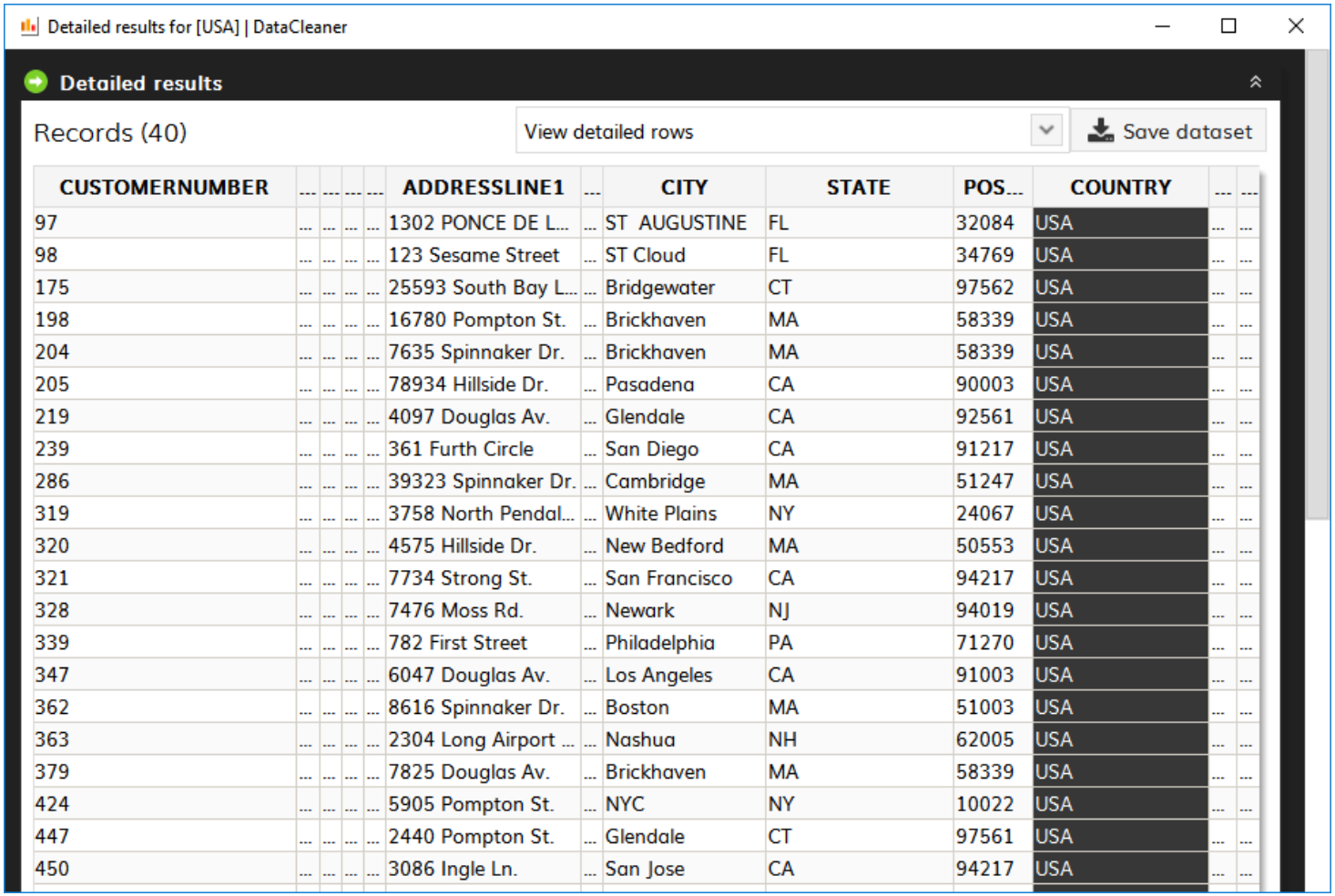
Data profiling

- Value distribution



Data profiling

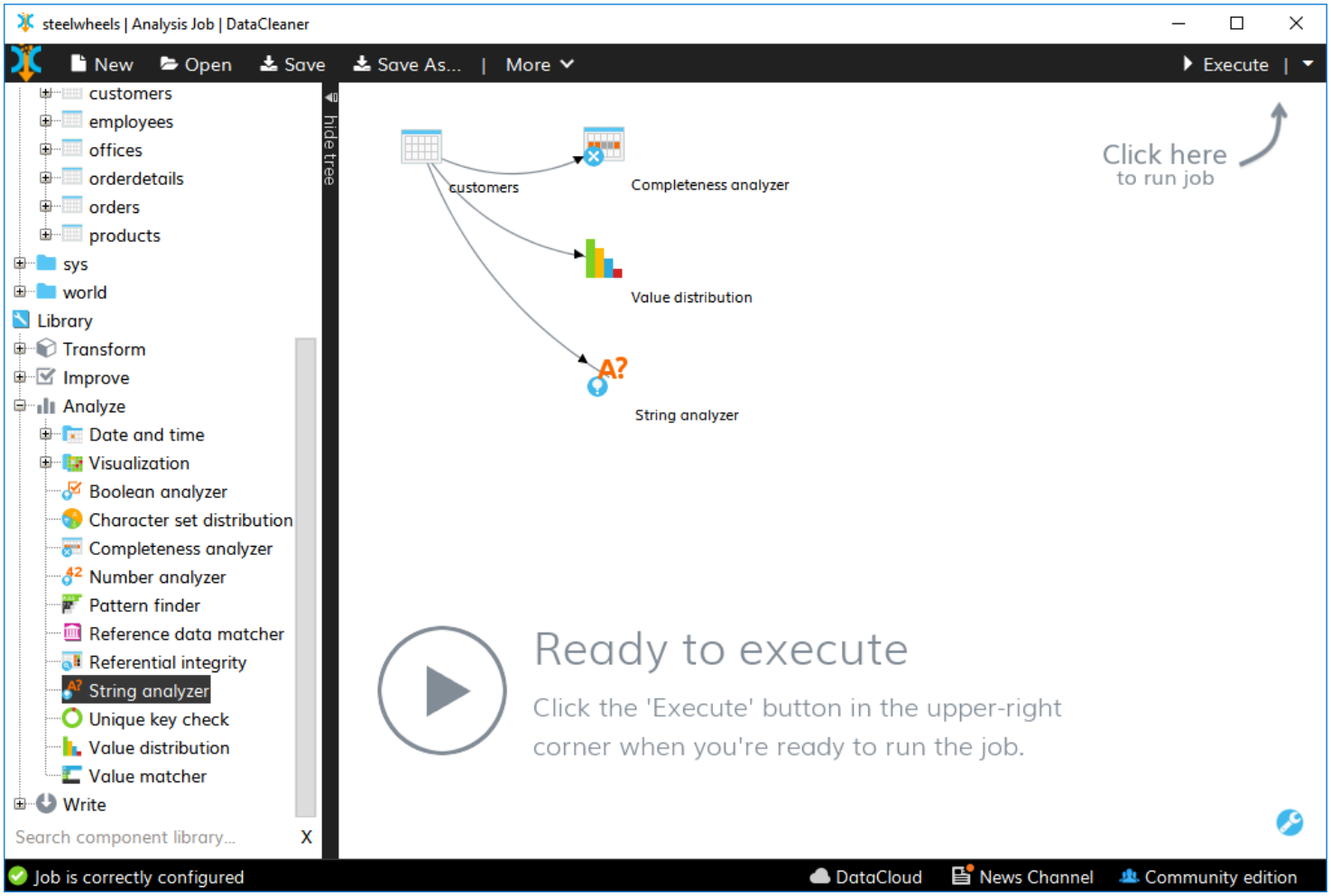
- Value distribution – detailed results



CUSTOMERNUMBER	ADDRESSLINE1	CITY	STATE	POS...	COUNTRY
97	1302 PONCE DE L...	ST AUGUSTINE	FL	32084	USA
98	123 Sesame Street	ST Cloud	FL	34769	USA
175	25593 South Bay L...	Bridgewater	CT	97562	USA
198	16780 Pompton St.	Brickhaven	MA	58339	USA
204	7635 Spinnaker Dr.	Brickhaven	MA	58339	USA
205	78934 Hillside Dr.	Pasadena	CA	90003	USA
219	4097 Douglas Av.	Glendale	CA	92561	USA
239	361 Furth Circle	San Diego	CA	91217	USA
286	39323 Spinnaker Dr.	Cambridge	MA	51247	USA
319	3758 North Pendal...	White Plains	NY	24067	USA
320	4575 Hillside Dr.	New Bedford	MA	50553	USA
321	7734 Strong St.	San Francisco	CA	94217	USA
328	7476 Moss Rd.	Newark	NJ	94019	USA
339	782 First Street	Philadelphia	PA	71270	USA
347	6047 Douglas Av.	Los Angeles	CA	91003	USA
362	8616 Spinnaker Dr.	Boston	MA	51003	USA
363	2304 Long Airport ...	Nashua	NH	62005	USA
379	7825 Douglas Av.	Brickhaven	MA	58339	USA
424	5905 Pompton St.	NYC	NY	10022	USA
447	2440 Pompton St.	Glendale	CT	97561	USA
450	3086 Ingle Ln.	San Jose	CA	94217	USA

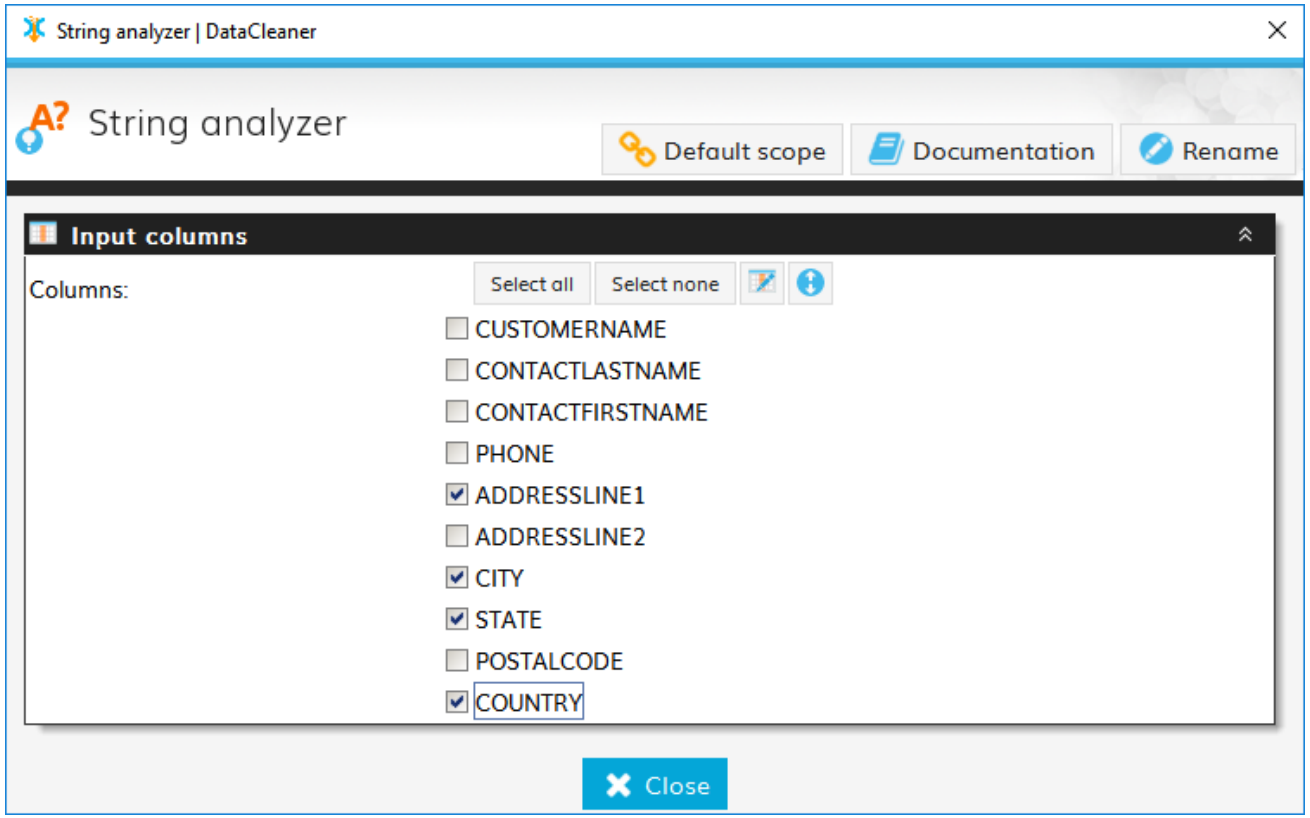
Data profiling

- String analysis



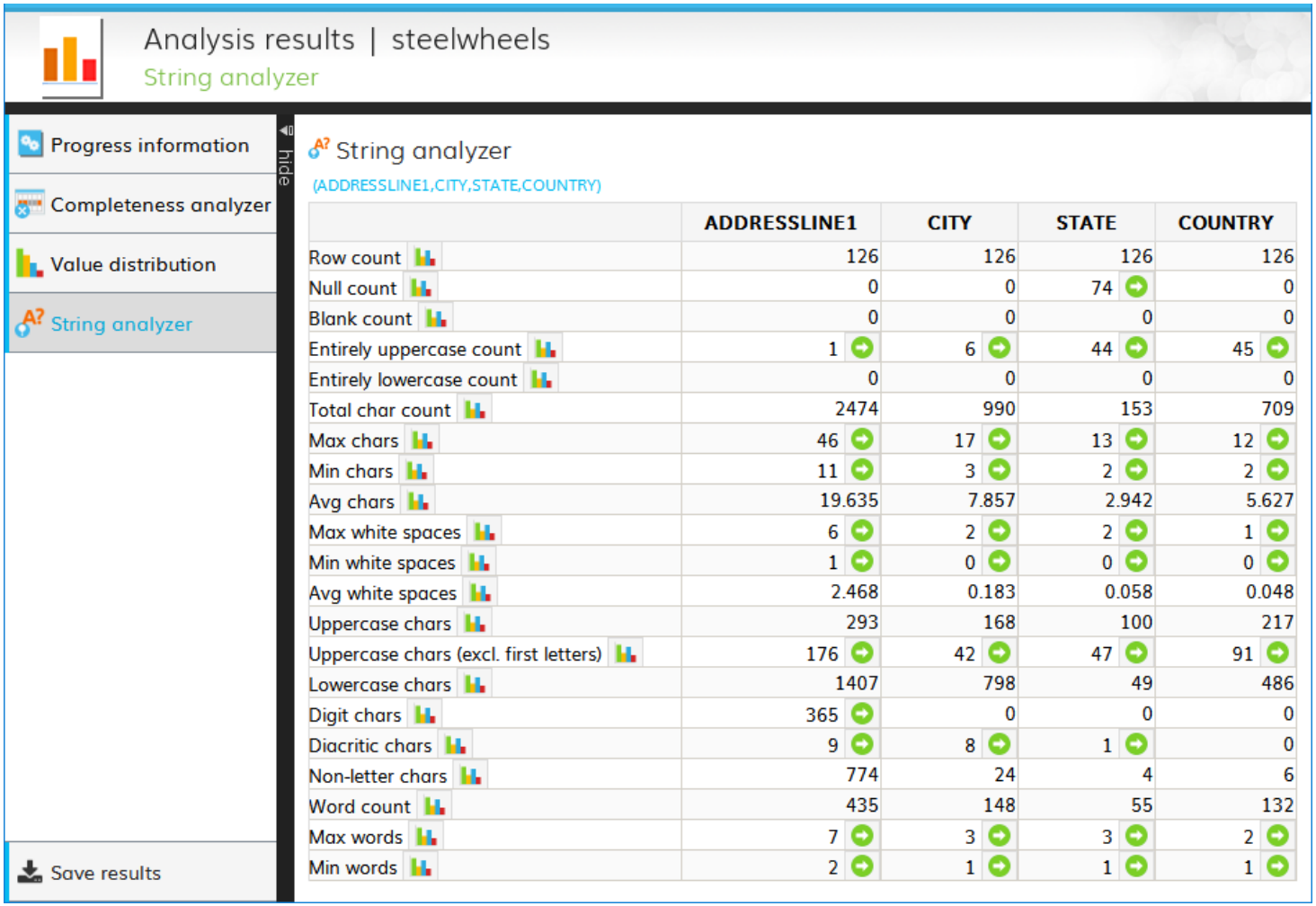
Data profiling

- String analysis



Data profiling

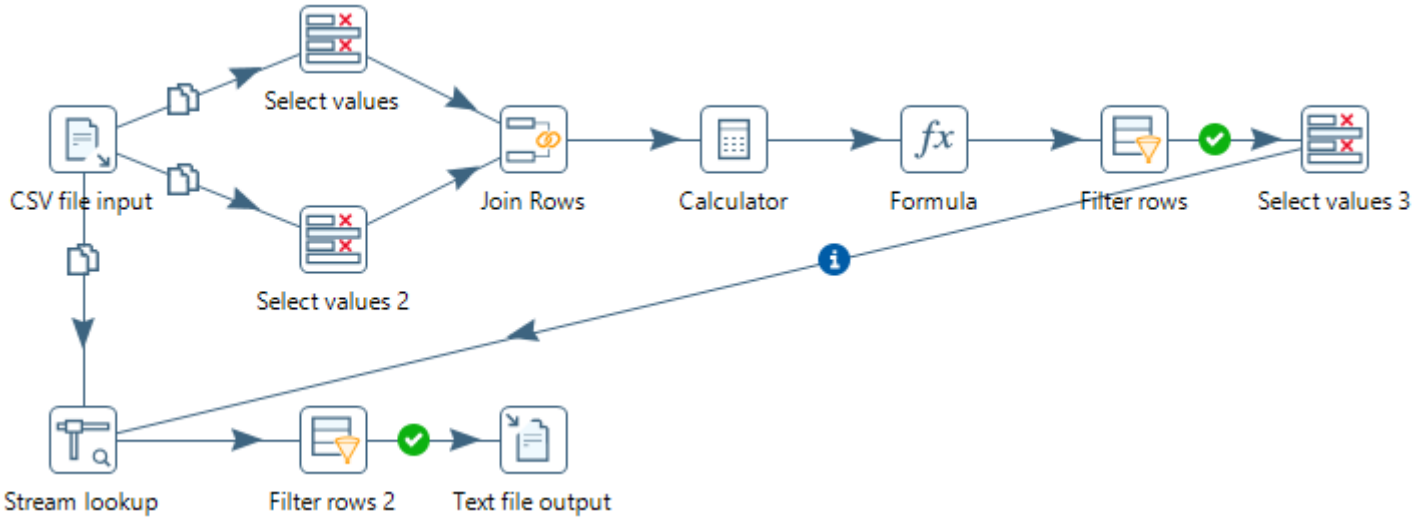
- String analysis



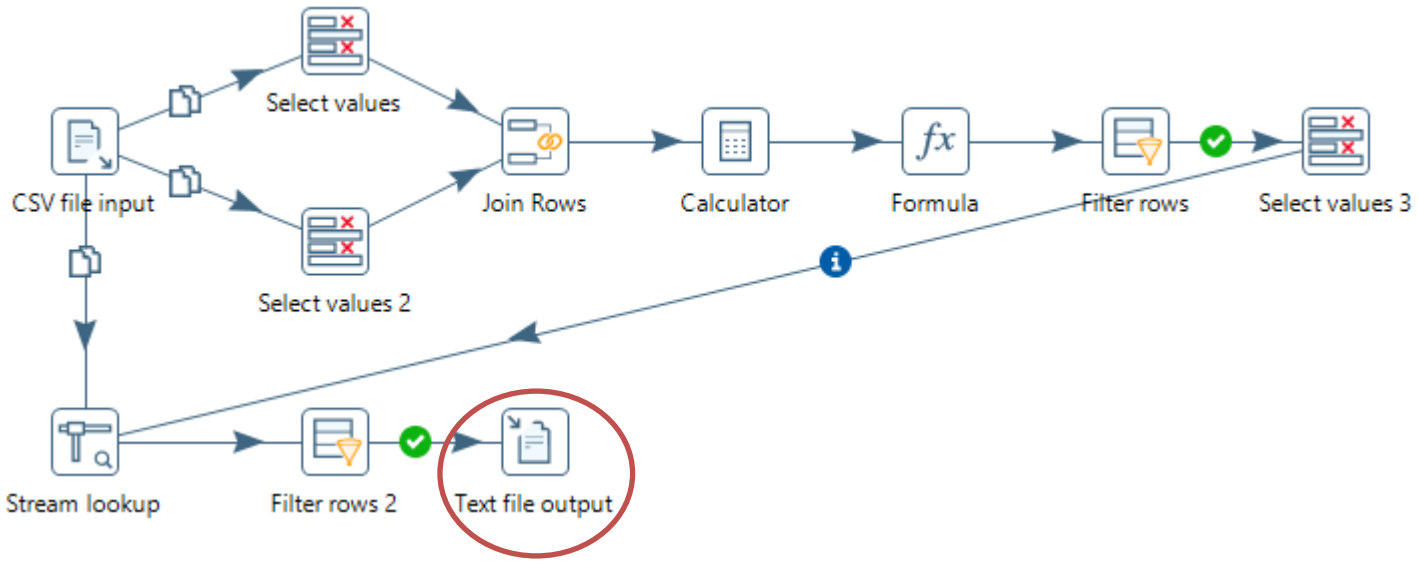
	ADDRESSLINE1	CITY	STATE	COUNTRY
Row count	126	126	126	126
Null count	0	0	74	0
Blank count	0	0	0	0
Entirely uppercase count	1	6	44	45
Entirely lowercase count	0	0	0	0
Total char count	2474	990	153	709
Max chars	46	17	13	12
Min chars	11	3	2	2
Avg chars	19.635	7.857	2.942	5.627
Max white spaces	6	2	2	1
Min white spaces	1	0	0	0
Avg white spaces	2.468	0.183	0.058	0.048
Uppercase chars	293	168	100	217
Uppercase chars (excl. first letters)	176	42	47	91
Lowercase chars	1407	798	49	486
Digit chars	365	0	0	0
Diacritic chars	9	8	1	0
Non-letter chars	774	24	4	6
Word count	435	148	55	132
Max words	7	3	3	2
Min words	2	1	1	1

Data profiling

- Integration between tools
  - Pentaho Data Integration (PDI) with DataCleaner plugin
  - the output of any transformation step can be a data source for data profiling



Duplicate detection

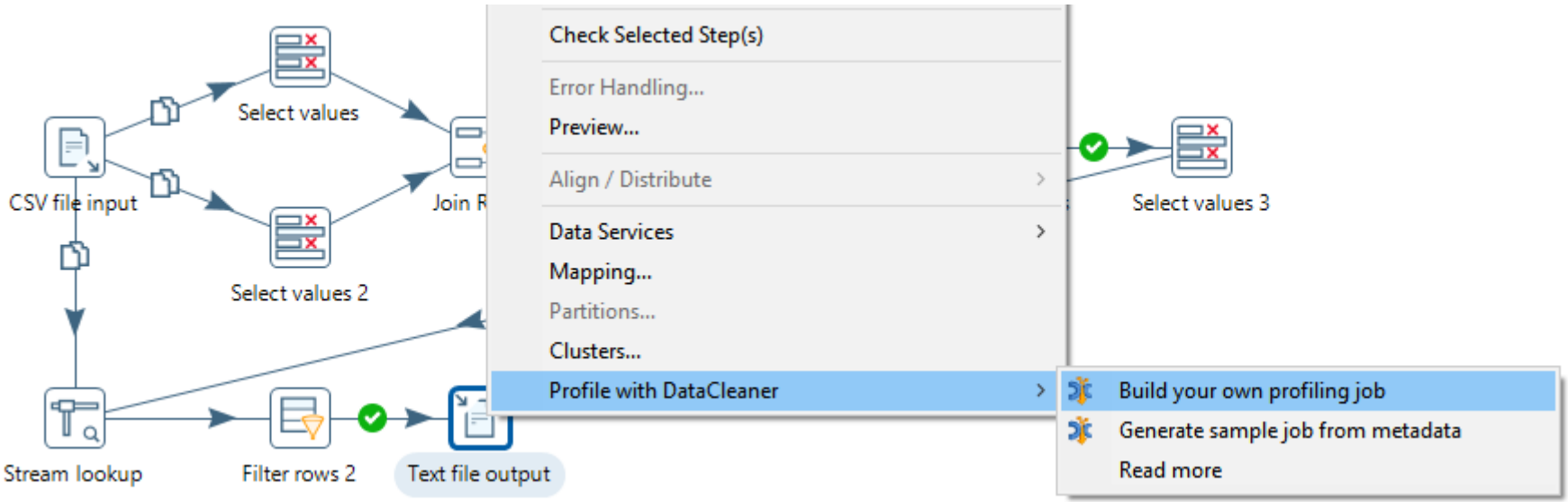


Rows of step: Text file output (64 rows)

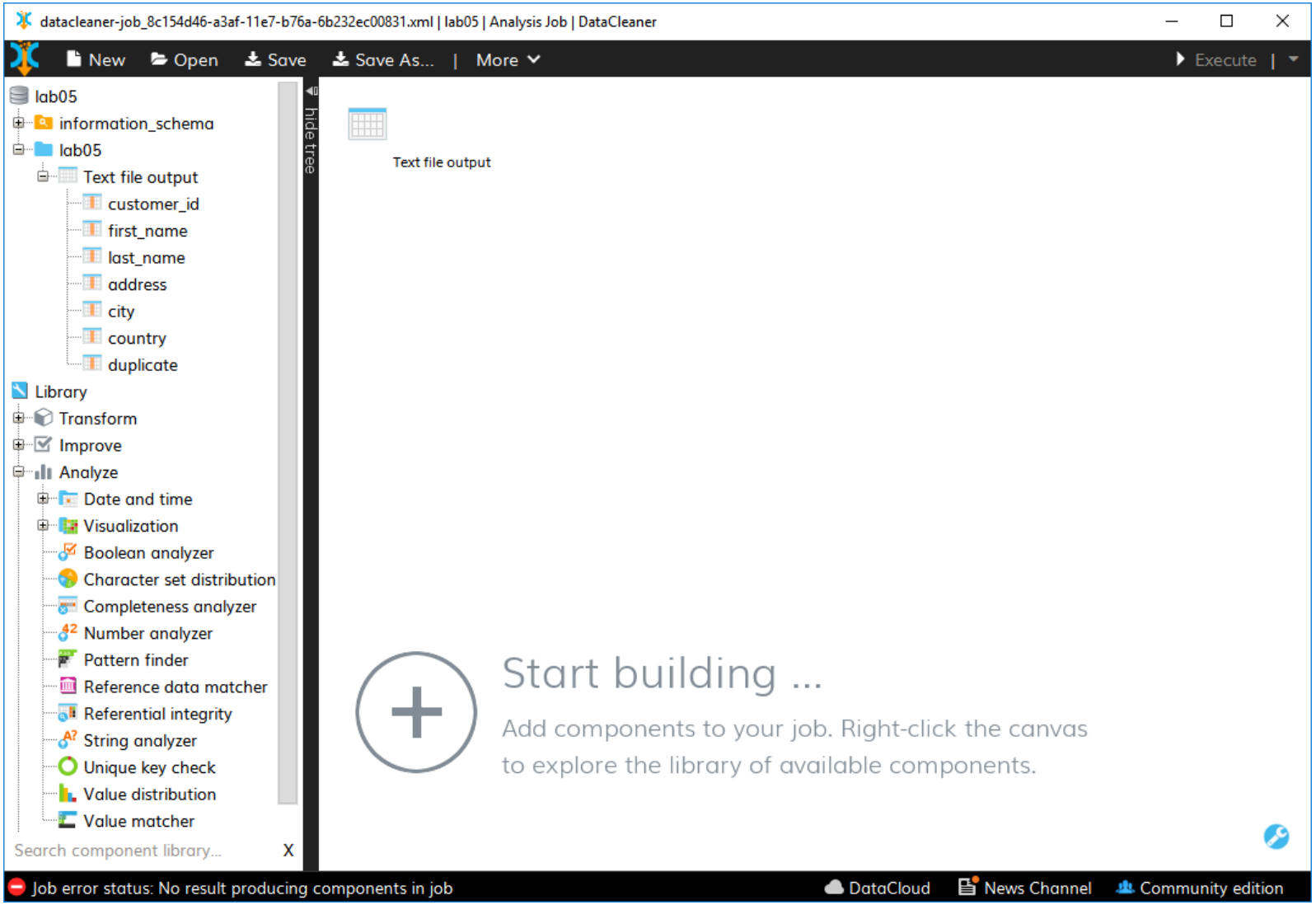
#	customer_id	first_name	last_name	address	city	country	duplicate
1	103	Carine	Schmitt	54, rue Royale	Nantes	France	<null>
2	119	Janine	Labrune	67, rue des Cinquante Otages	Nantes	France	<null>
3	121	Jonas	Bergulfsen	Erling Skakkes gate 78	Stavern	Norway	<null>
4	125	Zbygniew	Piestrzeniewicz	ul. Filtrowa 68	Warszawa	Poland	<null>
5	128	Roland	Keitel	Lyonerstr. 34	Frankfurt	Germany	<null>
6	141	Diego	Freyre	c/ Moralzarzal, 86	Madrid	Spain	<null>
7	144	Christina	Berglund	Berguvsvägen 8	Luleå	Sweden	<null>
8	145	Jytte	Petersen	Vinbæltet 34	Kobenhavn	Denmark	<null>
9	146	Mary	Saveley	2, rue du Commerce	Lyon	France	<null>

Data profiling

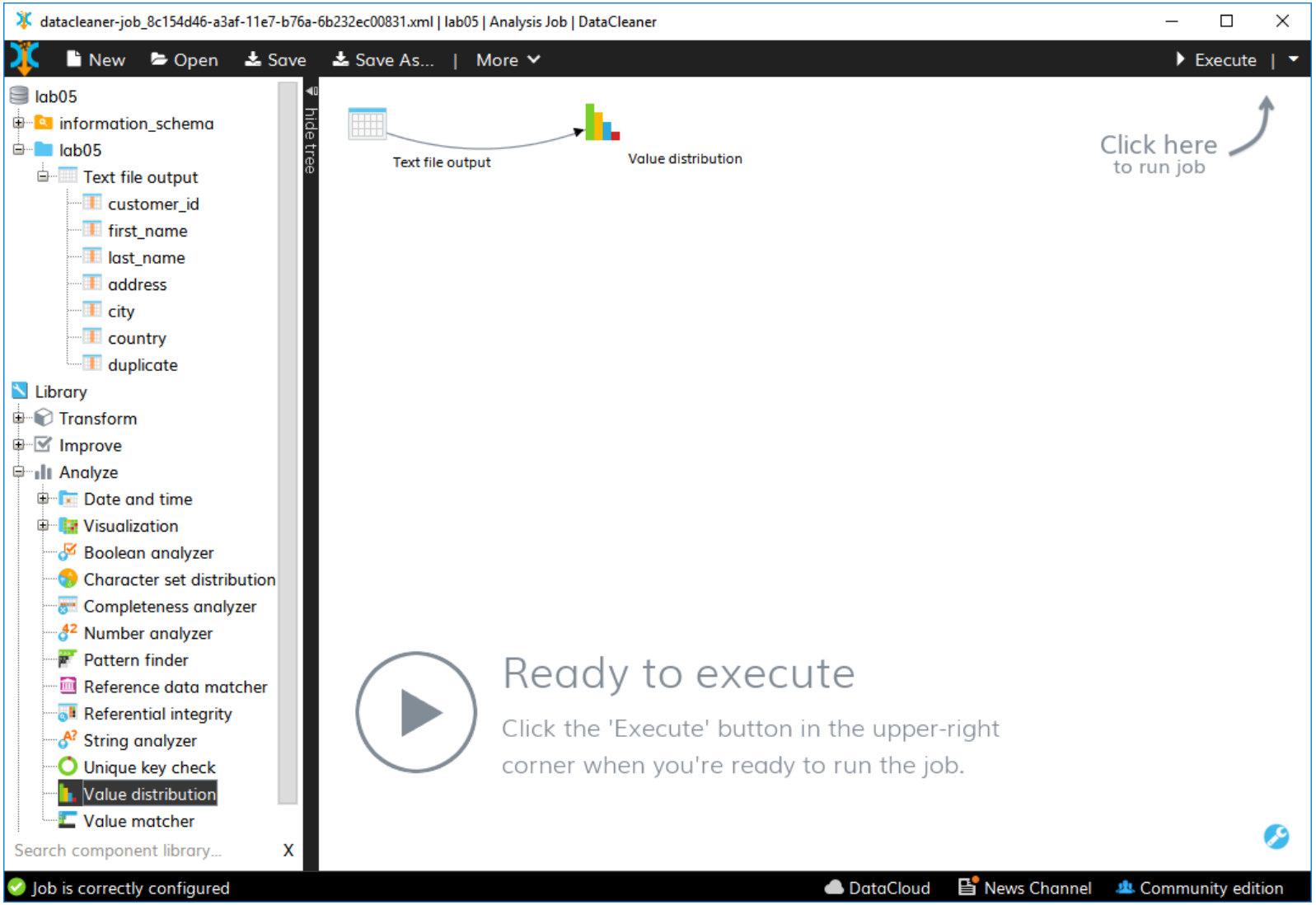
- Example
  - perform data profiling after duplicate elimination



Data profiling

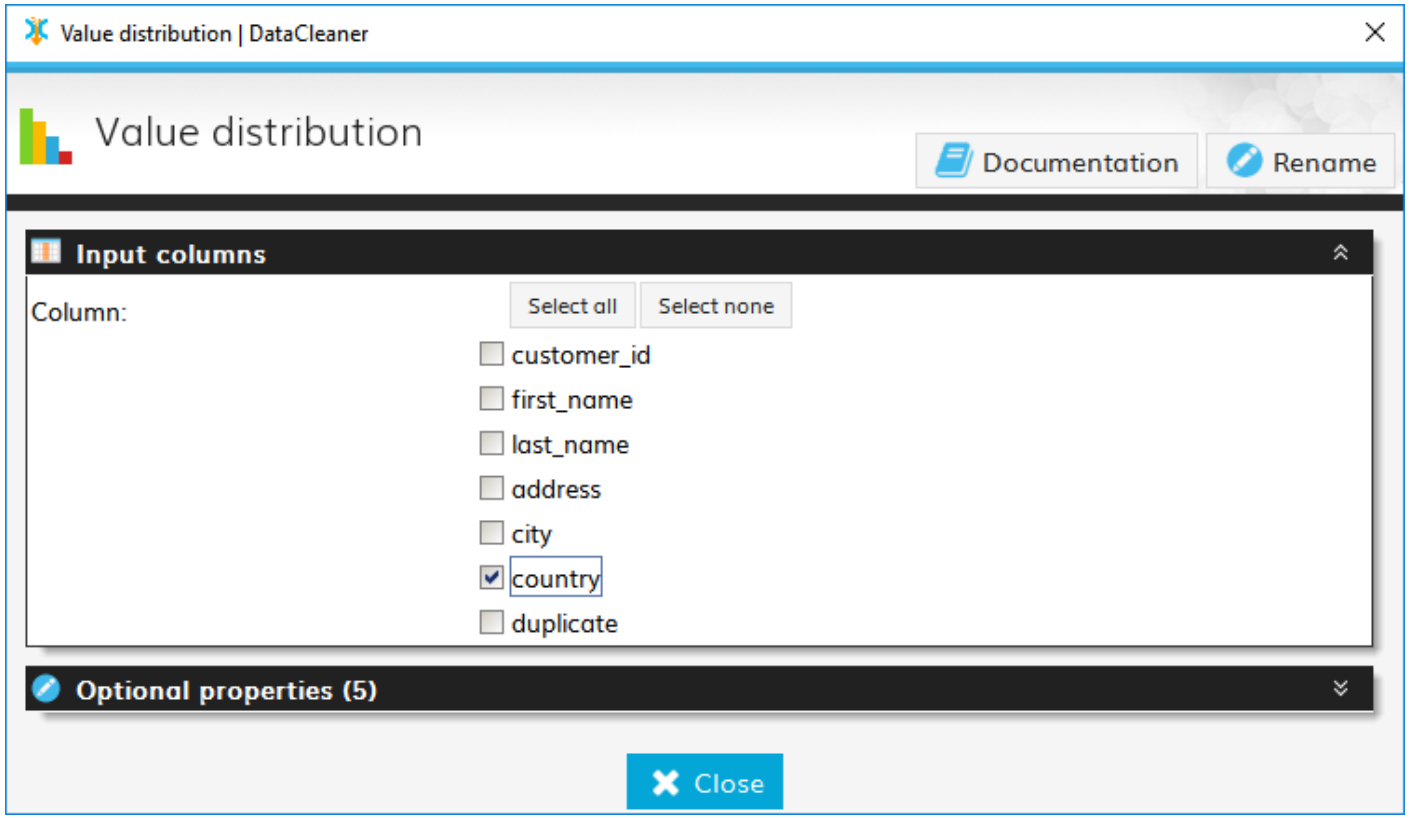


Data profiling



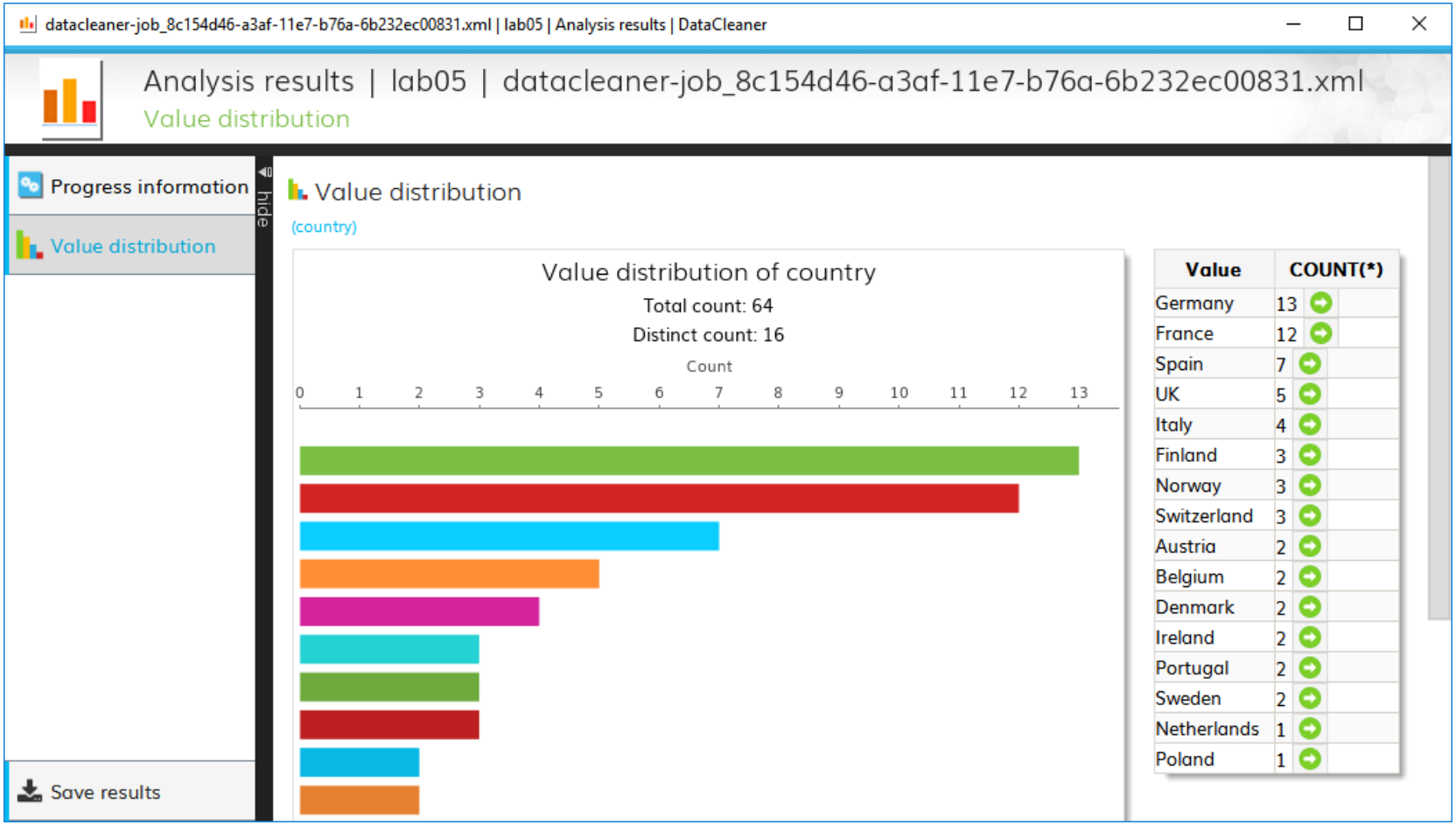
Data profiling

- Value distribution

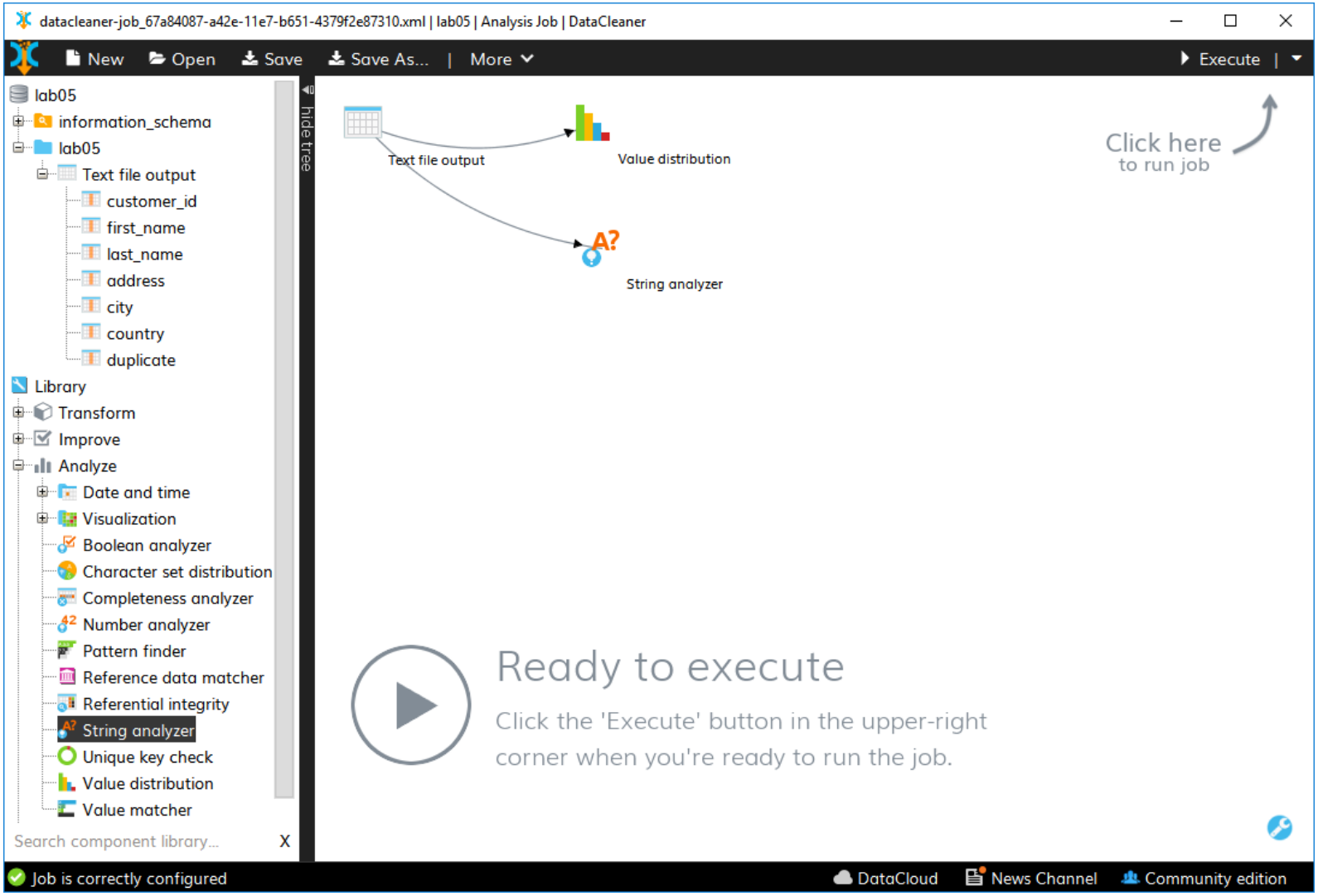


Data profiling

- Value distribution

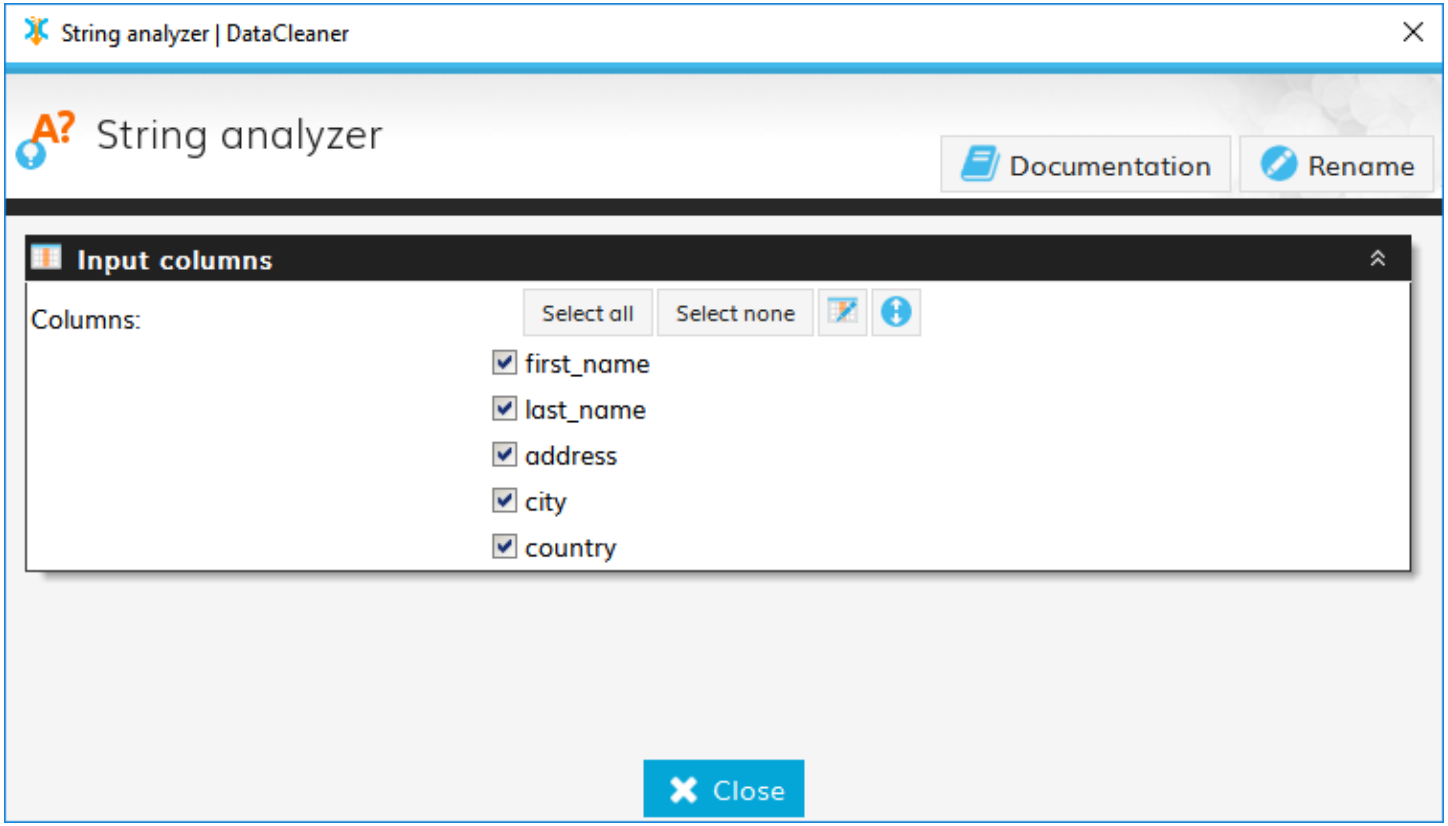


Data profiling



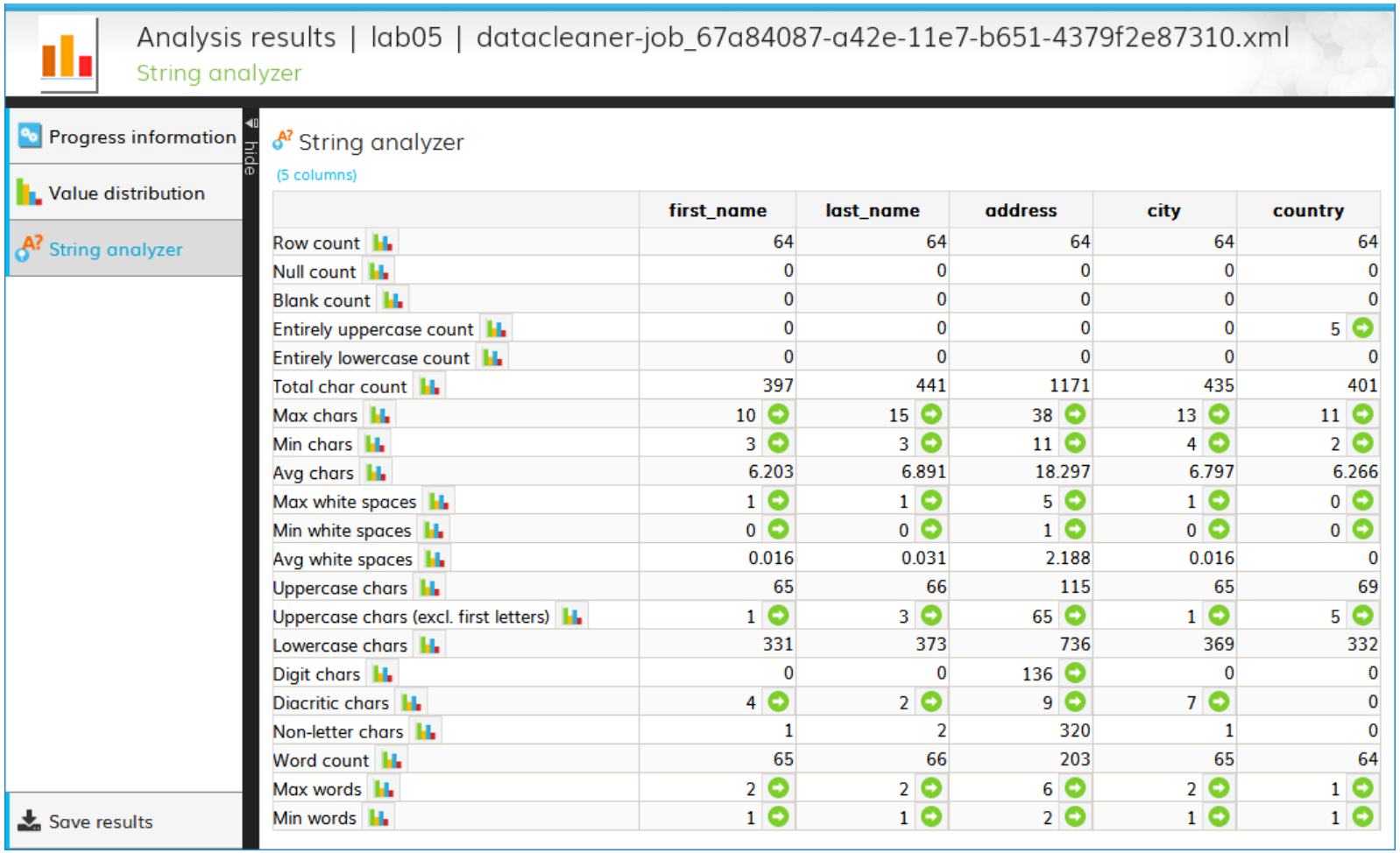
Data profiling

- String analysis



Data profiling

- String analysis



	first_name	last_name	address	city	country
Row count	64	64	64	64	64
Null count	0	0	0	0	0
Blank count	0	0	0	0	0
Entirely uppercase count	0	0	0	0	5
Entirely lowercase count	0	0	0	0	0
Total char count	397	441	1171	435	401
Max chars	10	15	38	13	11
Min chars	3	3	11	4	2
Avg chars	6.203	6.891	18.297	6.797	6.266
Max white spaces	1	1	5	1	0
Min white spaces	0	0	1	0	0
Avg white spaces	0.016	0.031	2.188	0.016	0
Uppercase chars	65	66	115	65	69
Uppercase chars (excl. first letters)	1	3	65	1	5
Lowercase chars	331	373	736	369	332
Digit chars	0	0	136	0	0
Diacritic chars	4	2	9	7	0
Non-letter chars	1	2	320	1	0
Word count	65	66	203	65	64
Max words	2	2	6	2	1
Min words	1	1	2	1	1