

Data Analysis and Integration

Data profiling

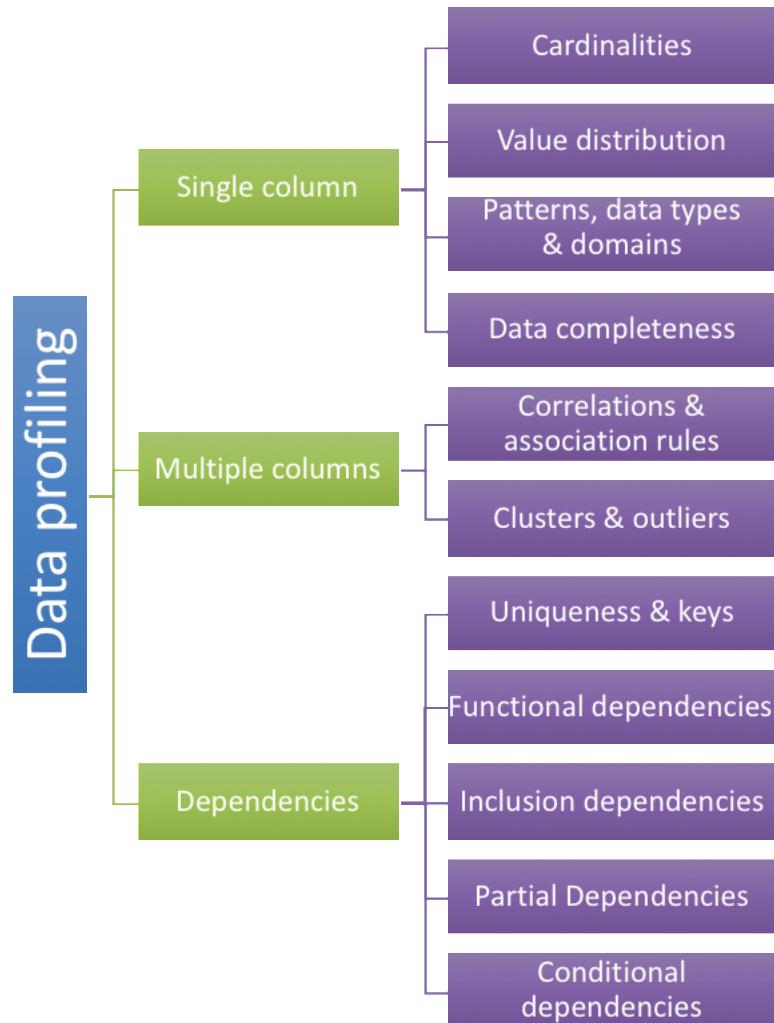
Data profiling

- 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.

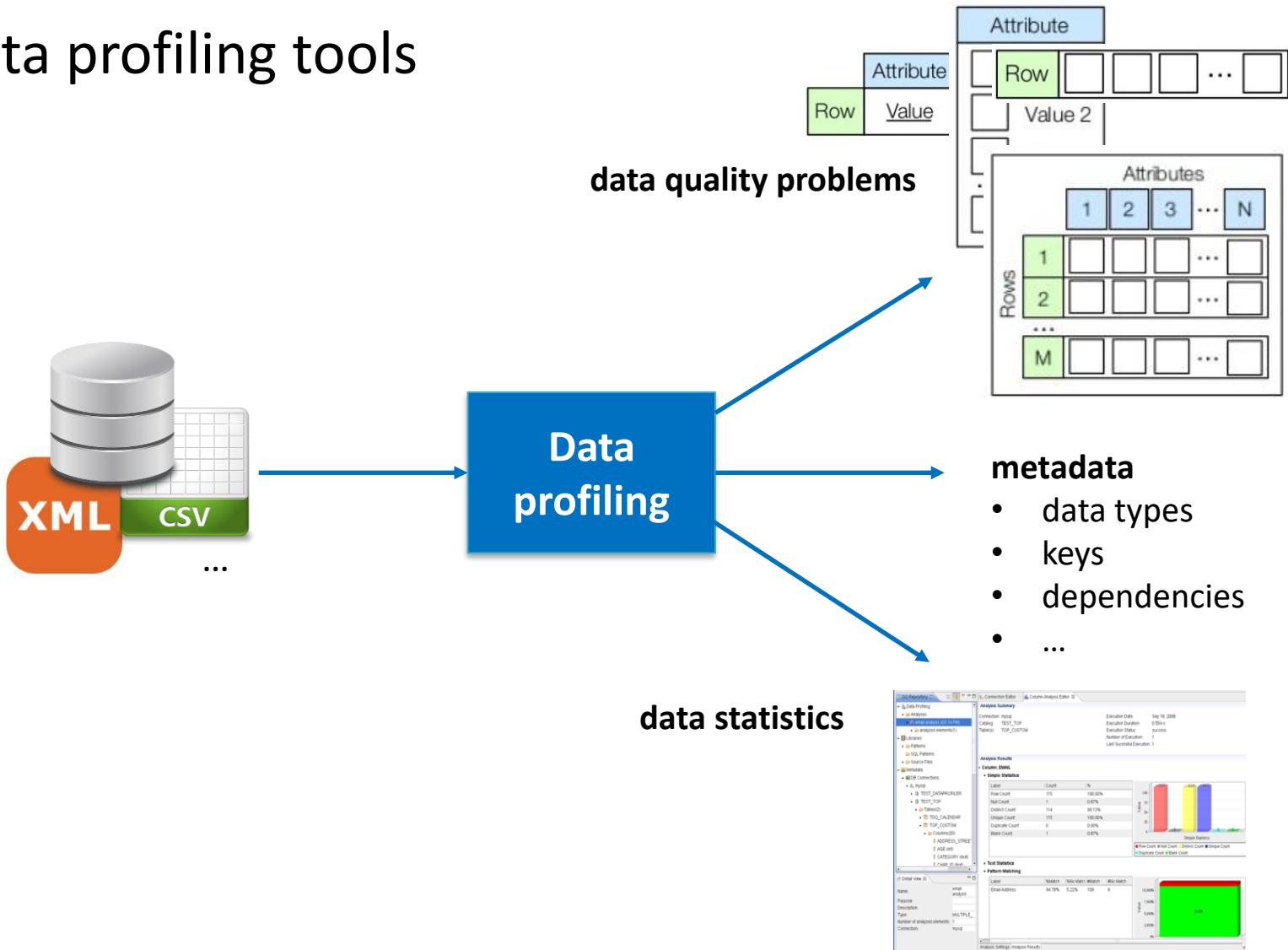
Data profiling



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

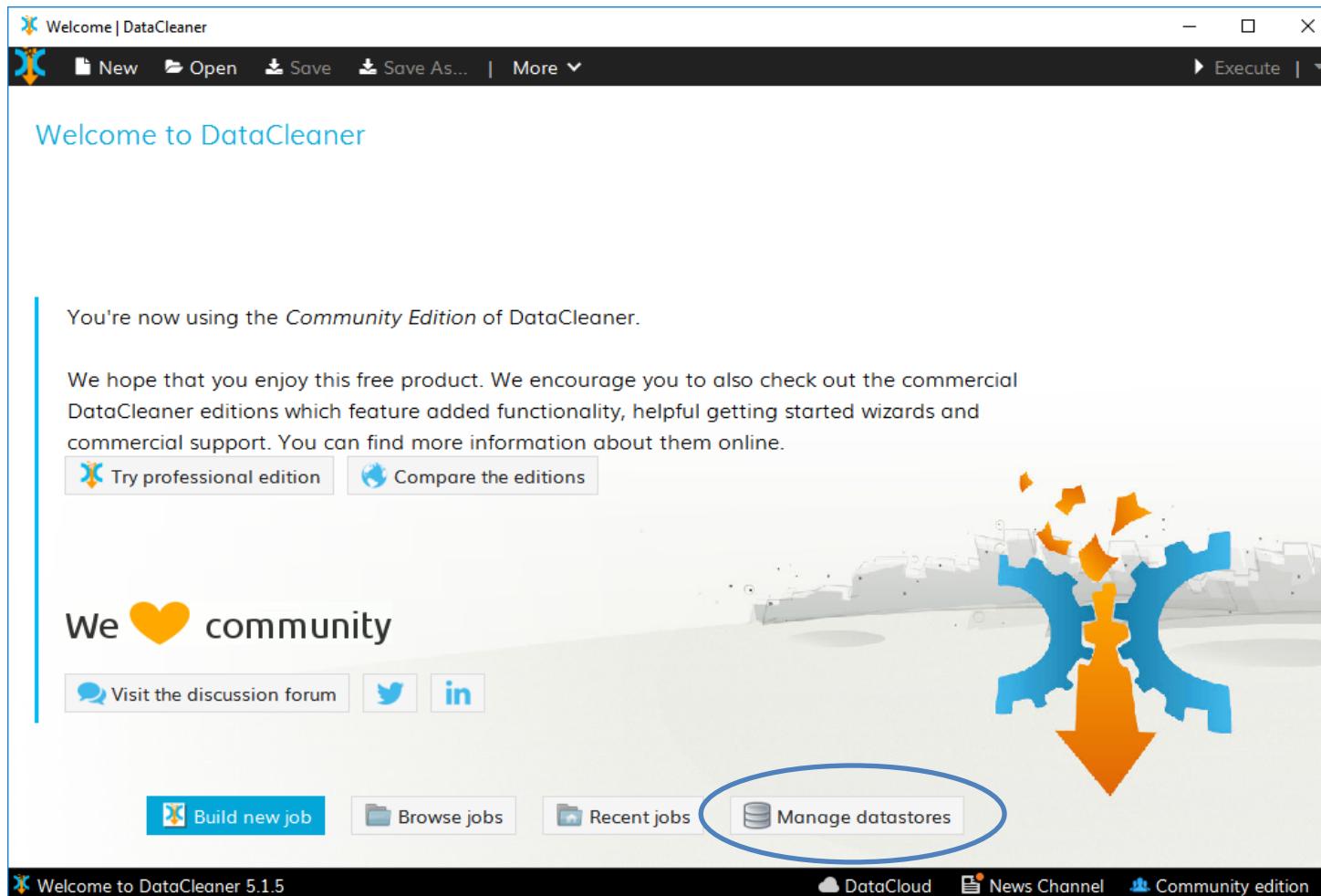
Data profiling

- Data profiling tools



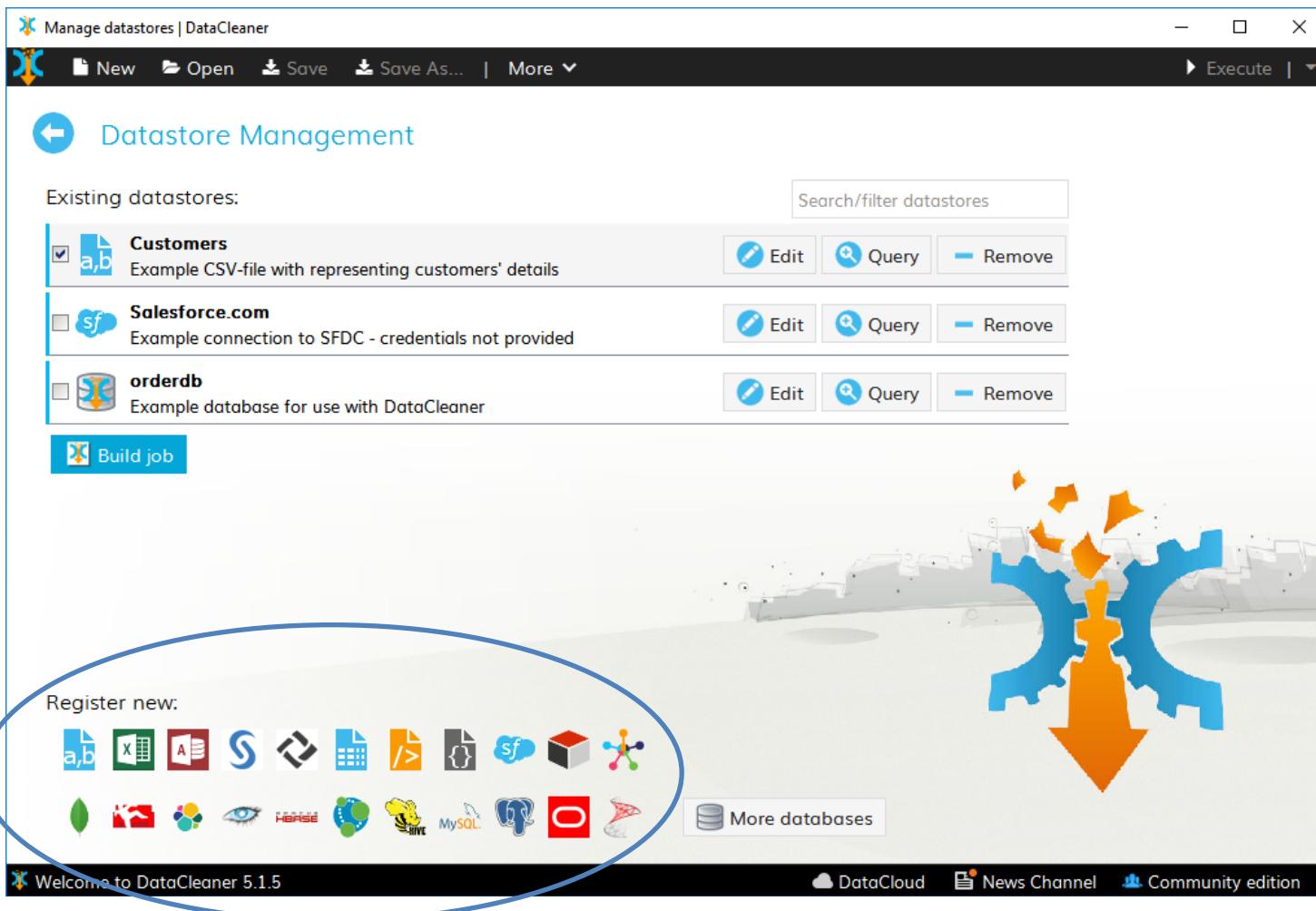
Data profiling

- Data profiling tools



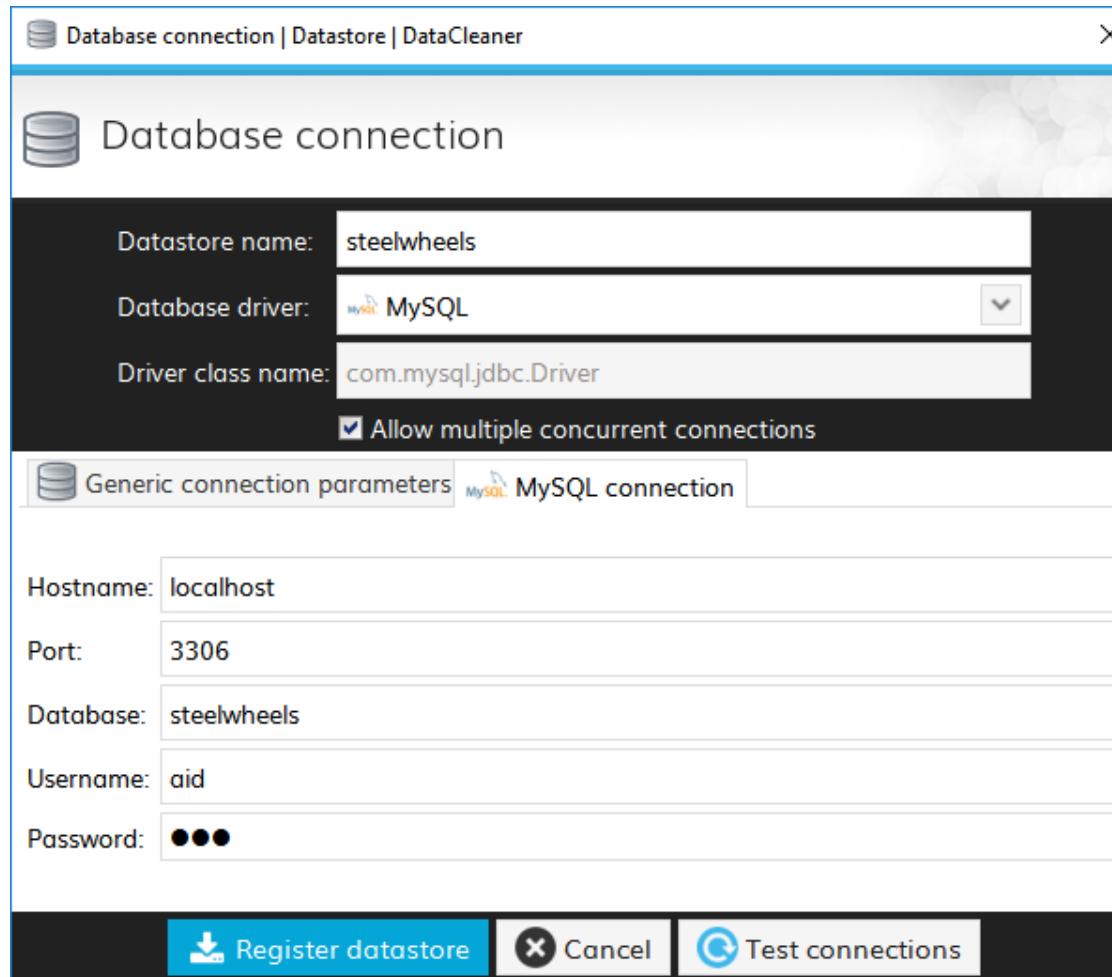
Data profiling

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



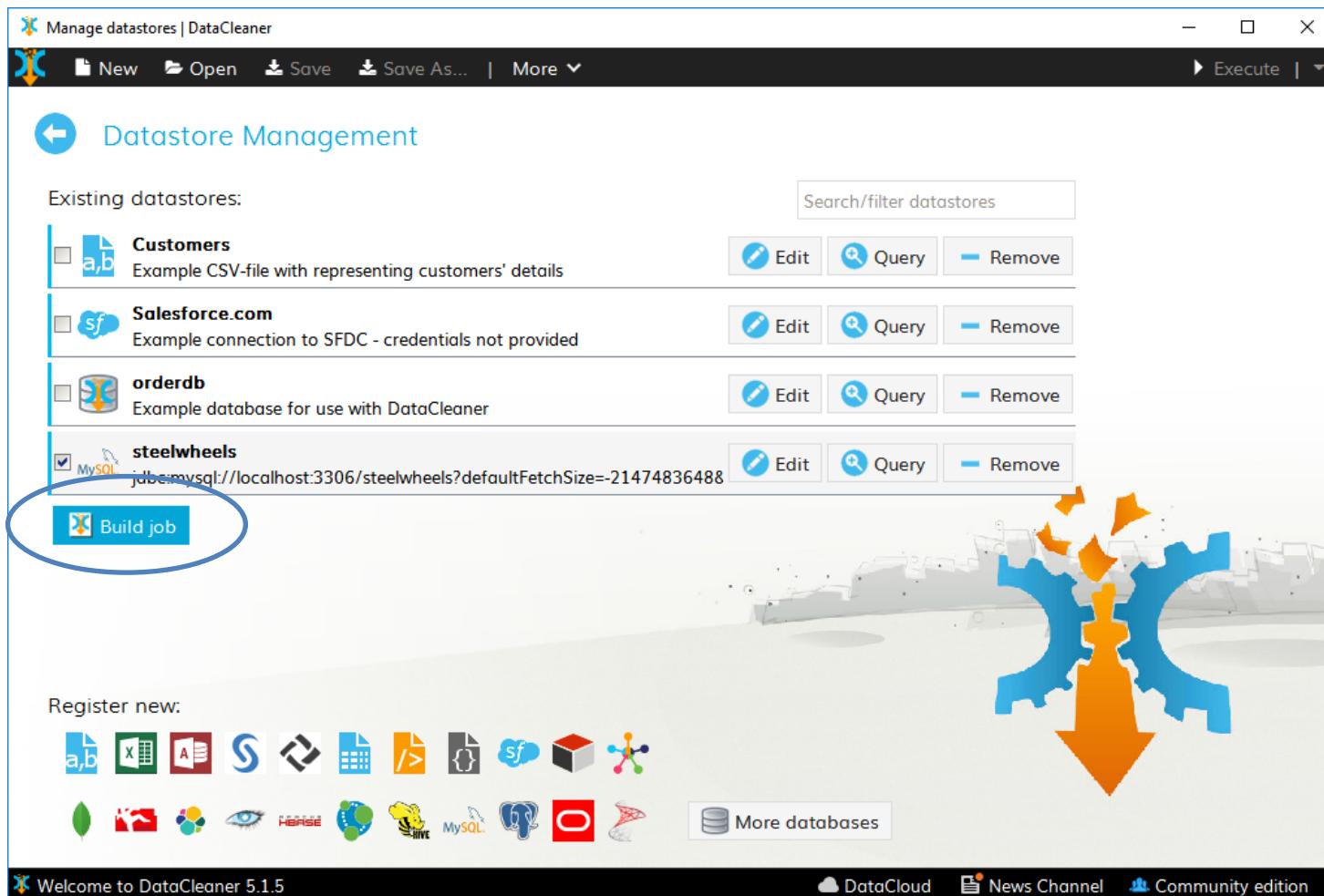
Data profiling

- Creating a new datastore



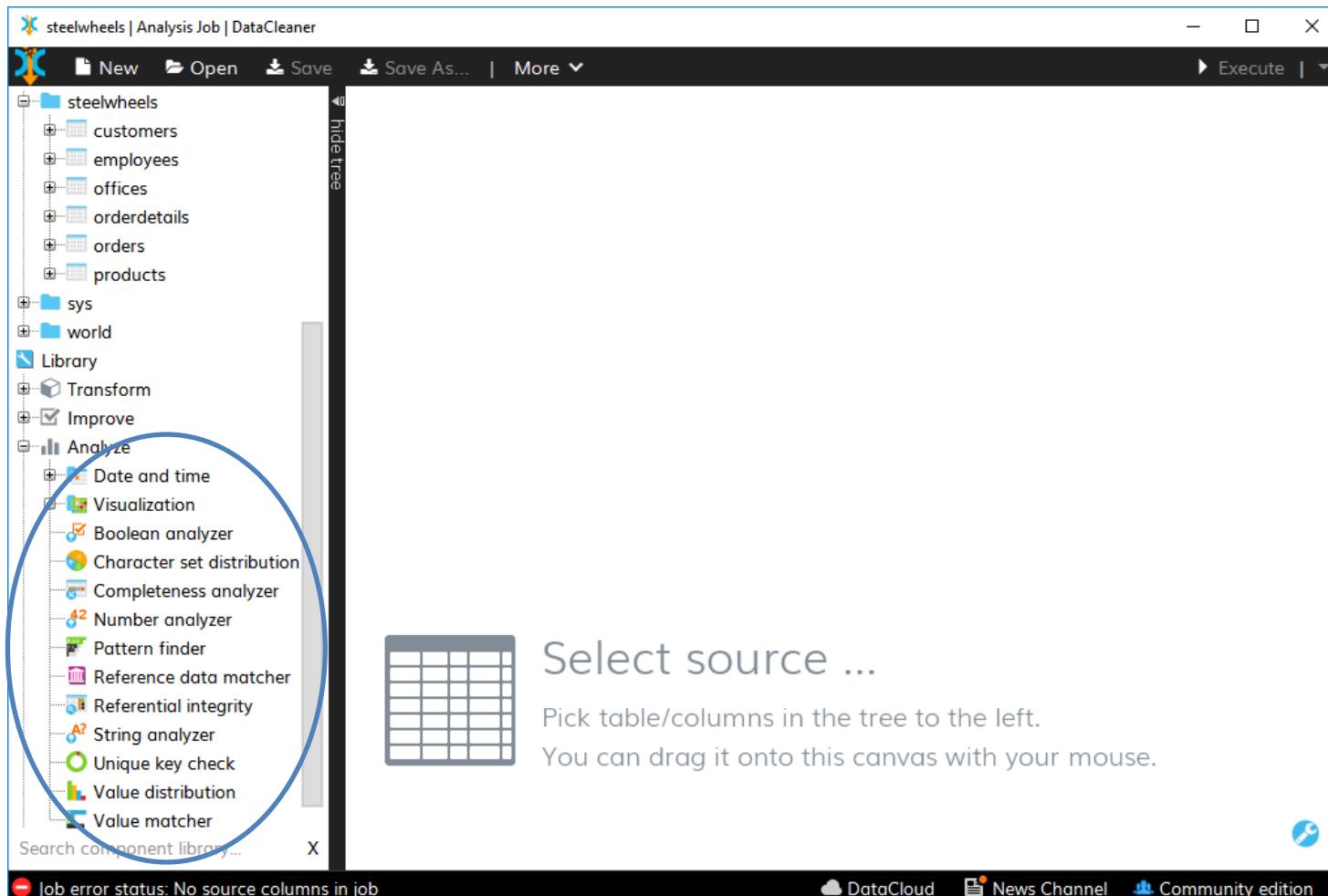
Data profiling

- Creating a new datastore



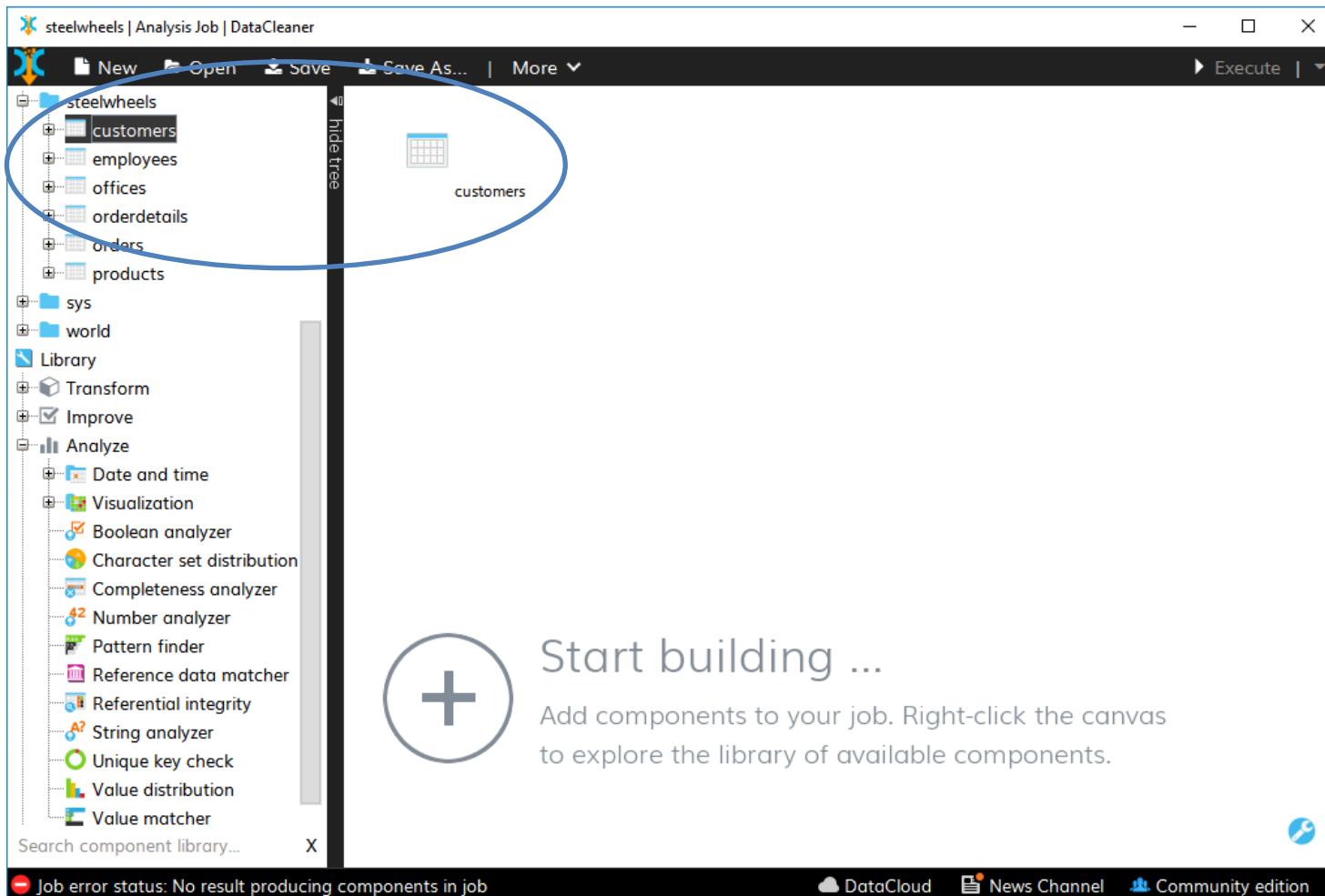
Data profiling

- Several options for data profiling tasks



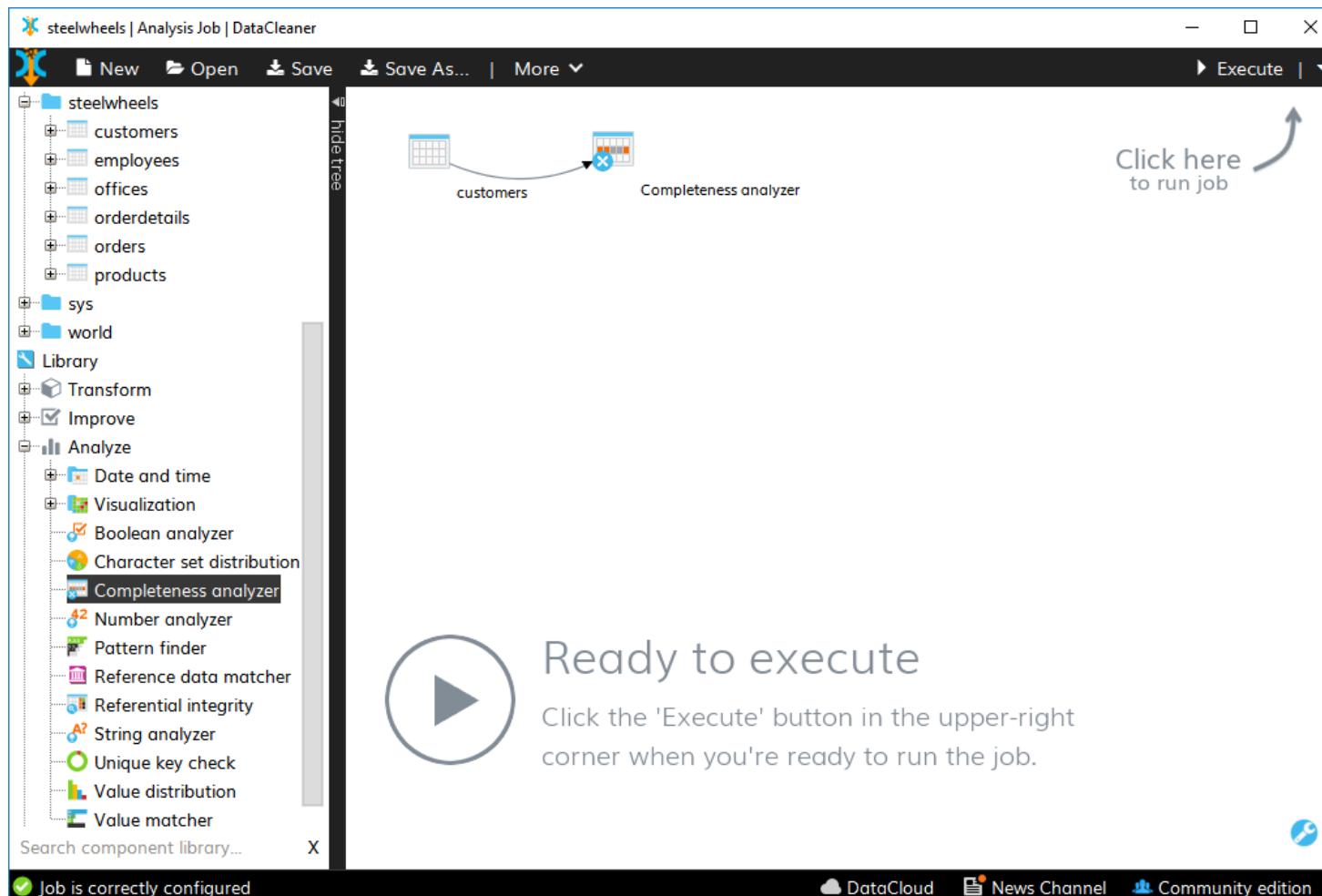
Data profiling

- Selecting the data source



Data profiling

- Completeness analysis



Ready to execute

Click the 'Execute' button in the upper-right corner when you're ready to run the job.

Data profiling

- Completeness analysis

The screenshot shows the 'Completeness analyzer' window in DataCleaner. The main panel is titled 'Input columns' and lists various fields for completeness checking. The 'STATE' field is selected, with a dropdown menu indicating the condition: 'Not <blank> or <null>'. Below this, the 'Required properties' section specifies 'Evaluation mode: When any field is incomplete, the record is incomplete'. An optional property is listed as '(1)'. At the bottom is a 'Close' button.

Completeness analyzer | DataCleaner

Completeness analyzer

Documentation Rename

Input columns

Values:
Values to check for completeness

CUSTOMERNUMBER
 CUSTOMERNAME
 CONTACTLASTNAME
 CONTACTFIRSTNAME
 PHONE
 ADDRESSLINE1
 ADDRESSLINE2
 CITY
 STATE
 POSTALCODE
 COUNTRY
 SALESREPLOYEEENUMBER
 CREDITLIMIT

Not <blank> or <null>

Required properties

Evaluation mode: When any field is incomplete, the record is incomplete

Optional properties (1)

Close

Data profiling

- Completeness analysis

The screenshot shows the steelwheels DataCleaner interface with the 'Completeness analyzer' selected. The main window displays a table of 'Incomplete records (74)' with columns: CUSTOMERNUMBER, ADDRESSLINE1, CITY, STATE, and COUNTRY. The 'STATE' column contains many null values, indicating incomplete data. A progress bar in the sidebar shows the analysis is at 100% completion.

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/ Moralzarjal, 86	Madrid	<null>	Spain
144	Berguvsvägen 8	Luleå	<null>	Sweden
145	Vinbæltet 34	København	<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

The screenshot shows the DataCleaner application window. The title bar reads "steelwheels | Analysis Job | DataCleaner". The menu bar includes "New", "Open", "Save", "Save As...", "More", "Execute", and "File". On the left, a tree view shows the project structure: "steelwheels" (customers, employees, offices, orderdetails, orders, products) and "sys" (world). Below this is the "Library" section with "Transform", "Improve", "Analyze" (Date and time, Visualization, Boolean analyzer, Character set distribution, Completeness analyzer, Number analyzer, Pattern finder, Reference data matcher, Referential integrity, String analyzer, Unique key check, Value distribution, Value matcher), and a "Search component library..." field. A status bar at the bottom says "Job is correctly configured". The main workspace displays a flow diagram: "customers" is connected to a "Completeness analyzer" icon, which then connects to a "Value distribution" icon. A large play button icon is centered in the workspace. Text on the right says "Ready to execute" and "Click the 'Execute' button in the upper-right corner when you're ready to run the job." A callout arrow points to the "Execute" button in the top right.

Data profiling

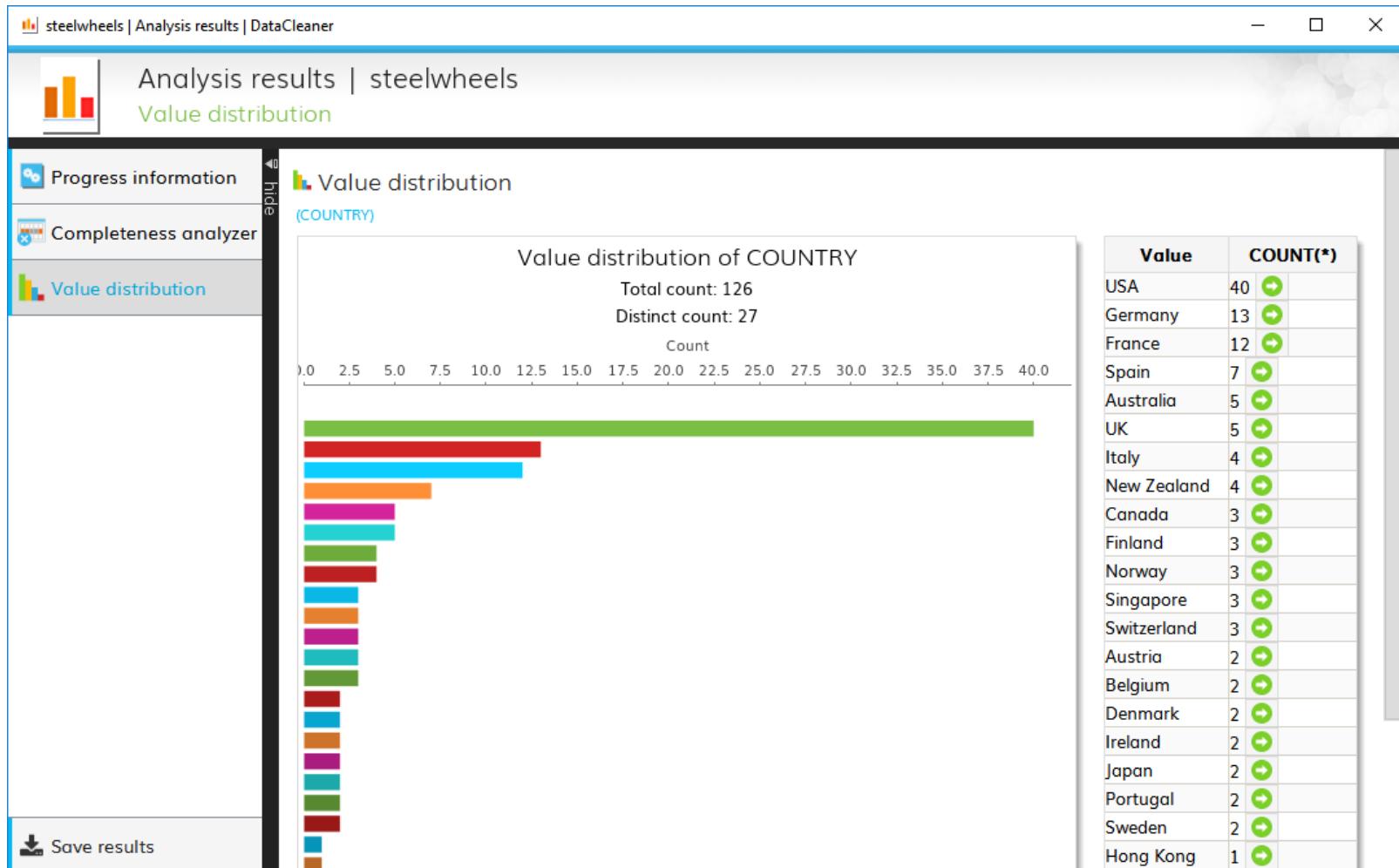
- Value distribution

The screenshot shows the 'Value distribution' dialog in DataCleaner. The title bar says 'Value distribution | DataCleaner'. The main area is titled 'Value distribution' with a bar chart icon. There are three buttons at the top right: 'Default scope', 'Documentation', and 'Rename'. Below is the 'Input columns' tab, which lists various columns with checkboxes. The 'COUNTRY' checkbox is checked and highlighted with a blue border. Other columns listed include CUSTOMERNUMBER, CUSTOMERNAME, CONTACTLASTNAME, CONTACTFIRSTNAME, PHONE, ADDRESSLINE1, ADDRESSLINE2, CITY, STATE, POSTALCODE, SALESREPRELOYEENUMBER, and CREDITLIMIT. At the bottom left of the dialog is a section titled 'Optional properties (5)' with a blue checkmark icon. At the bottom right is a 'Close' button with a red X icon.

Column	Status
CUSTOMERNUMBER	unchecked
CUSTOMERNAME	unchecked
CONTACTLASTNAME	unchecked
CONTACTFIRSTNAME	unchecked
PHONE	unchecked
ADDRESSLINE1	unchecked
ADDRESSLINE2	unchecked
CITY	unchecked
STATE	unchecked
POSTALCODE	unchecked
COUNTRY	checked
SALESREPRELOYEENUMBER	unchecked
CREDITLIMIT	unchecked

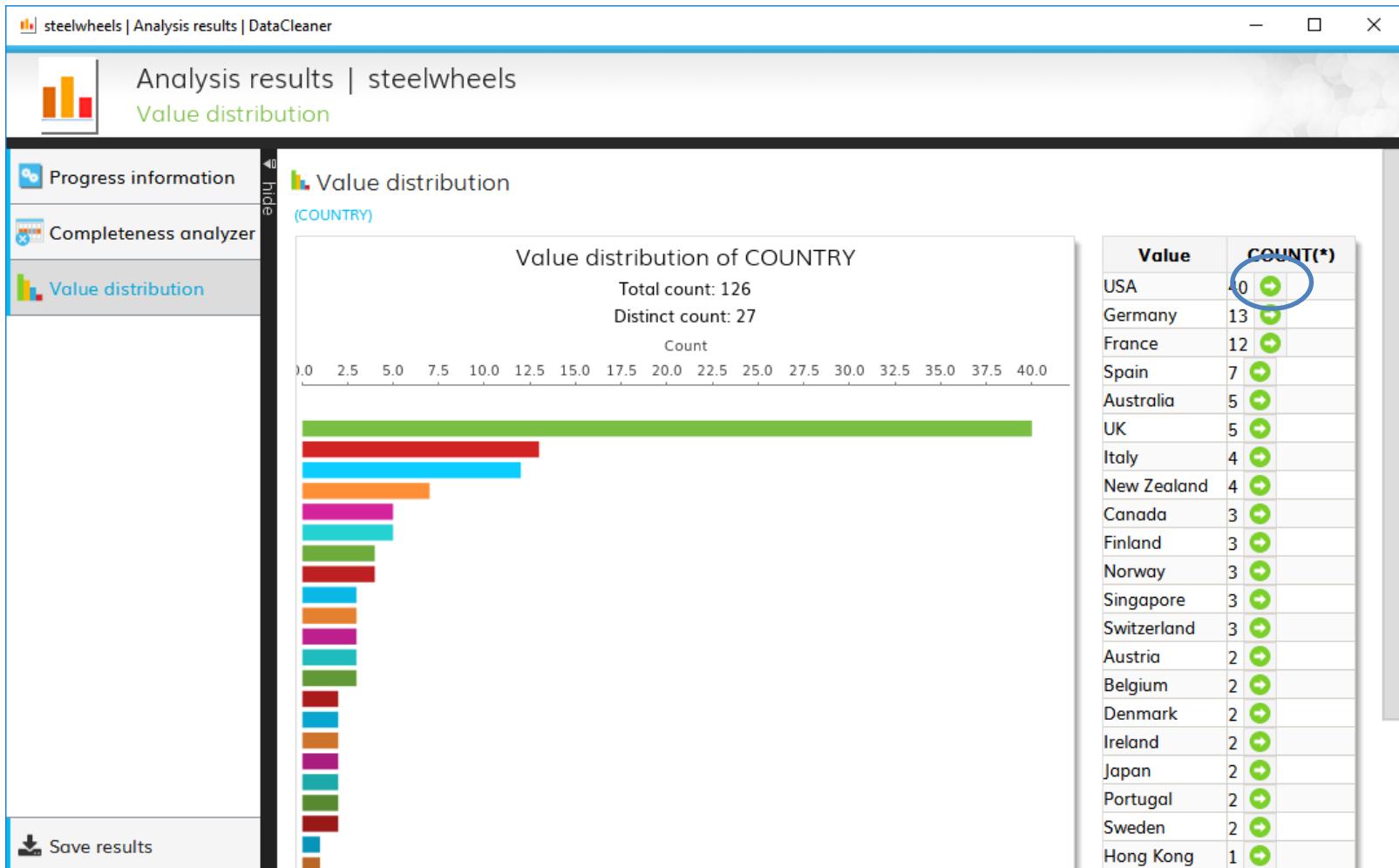
Data profiling

- Value distribution



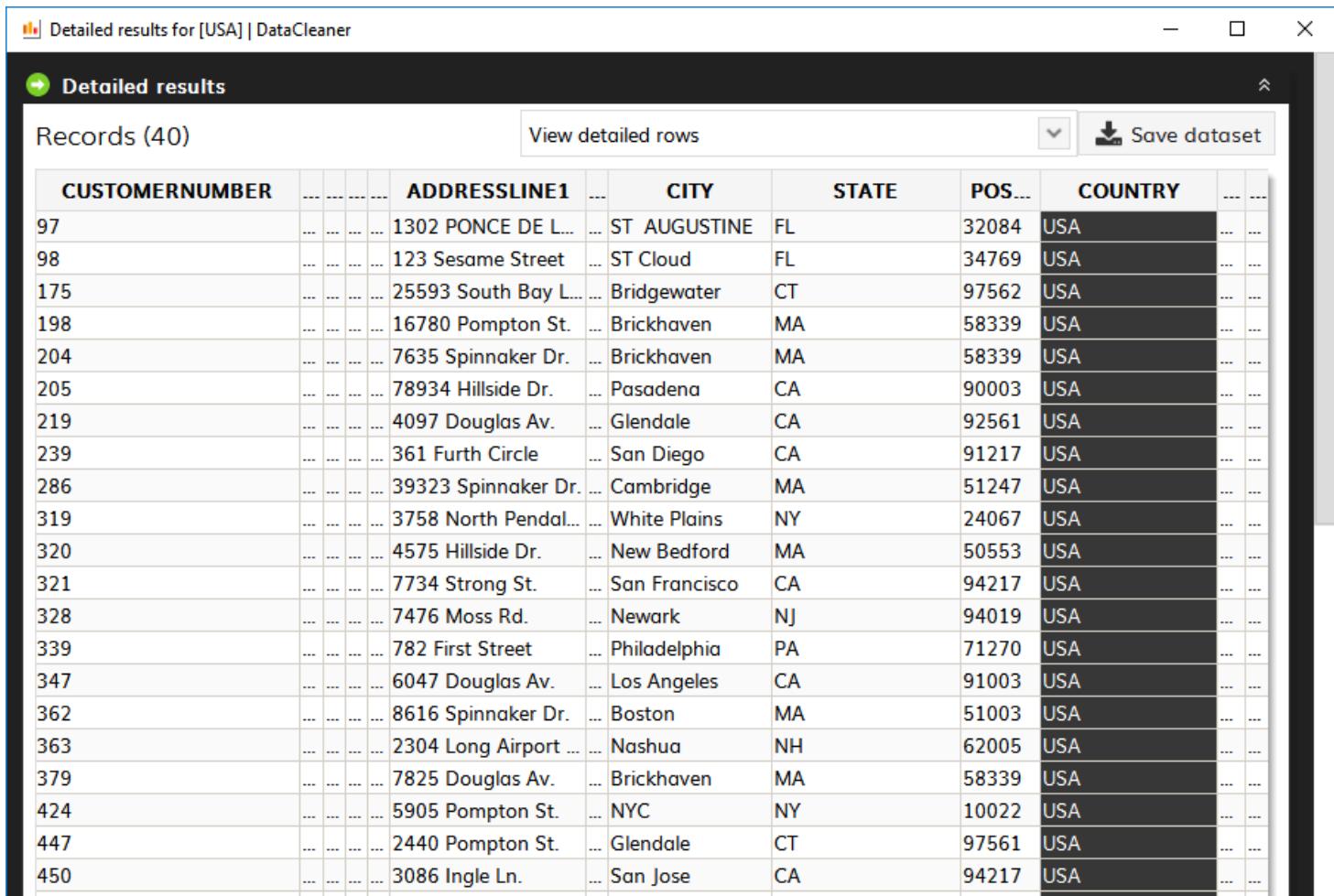
Data profiling

- Value distribution



Data profiling

- Value distribution – detailed results

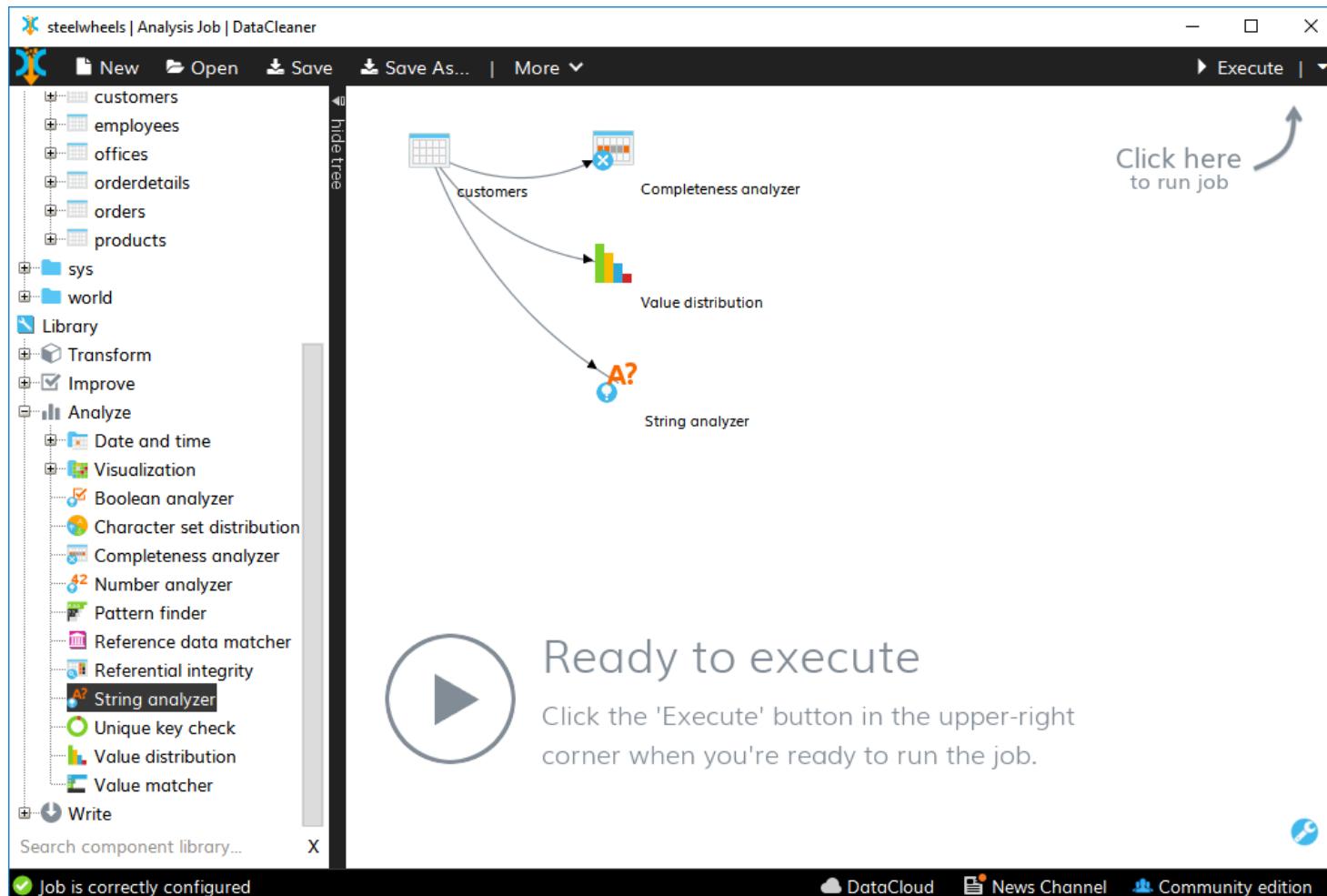


The screenshot shows a software window titled "Detailed results for [USA] | DataCleaner". The main area displays a table with 40 records, each containing customer information: CUSTOMERNUMBER, ADDRESSLINE1, CITY, STATE, POS..., and COUNTRY. The table is sorted by CUSTOMERNUMBER. The COUNTRY column consistently shows "USA".

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

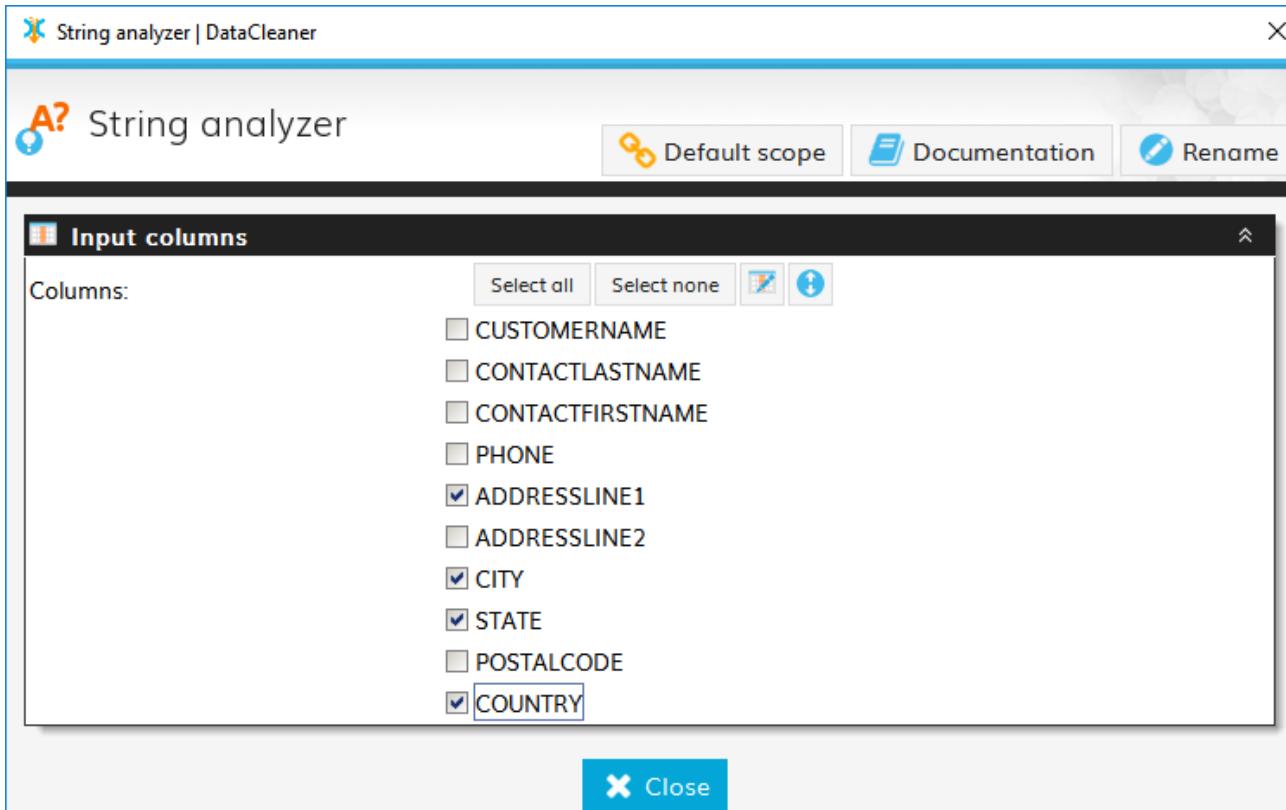
Data profiling

- String analysis



Data profiling

- String analysis



Data profiling

- String analysis

Analysis results | steelwheels
String analyzer

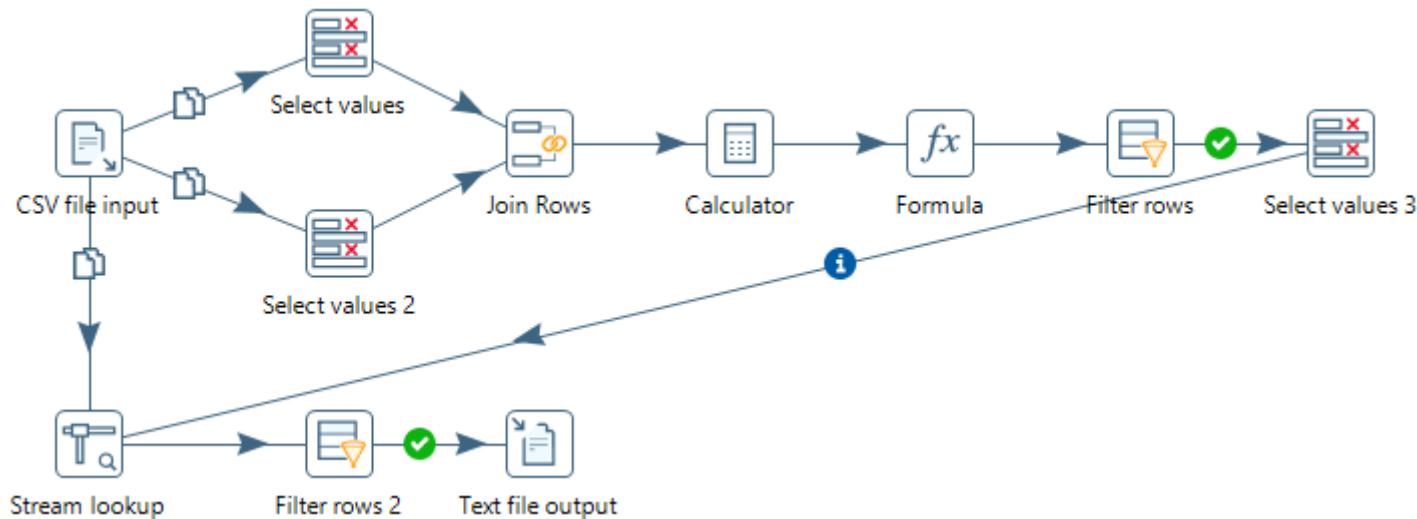
Progress information String analyzer (ADDRESSLINE1,CITY,STATE,COUNTRY)

	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

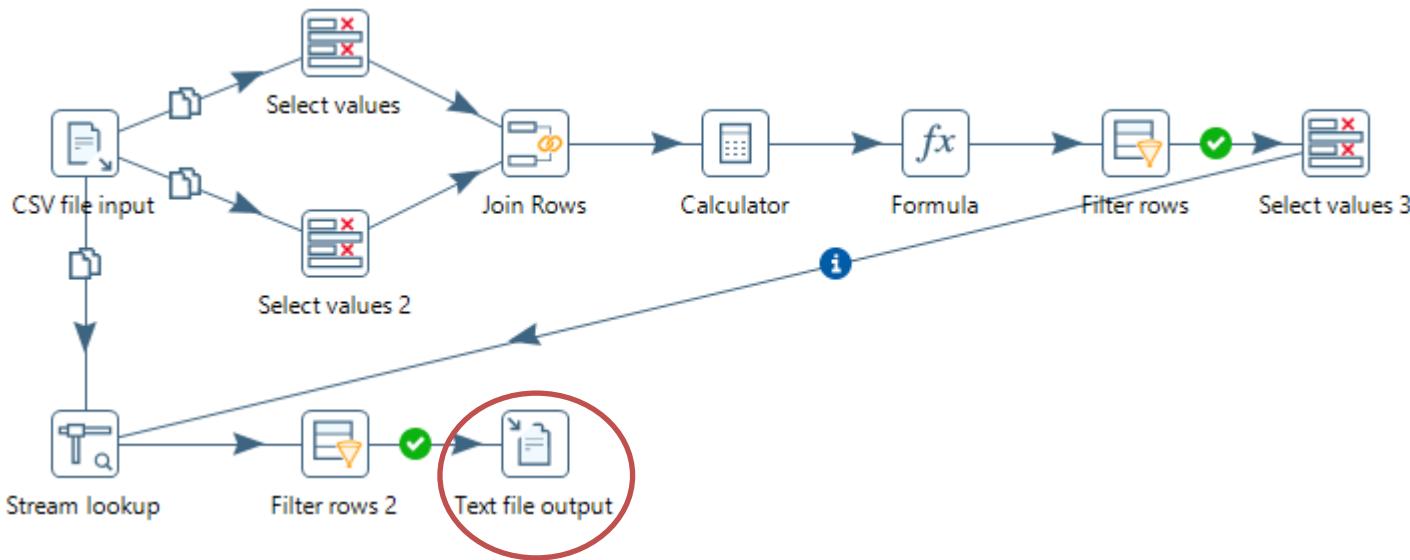
Save results

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

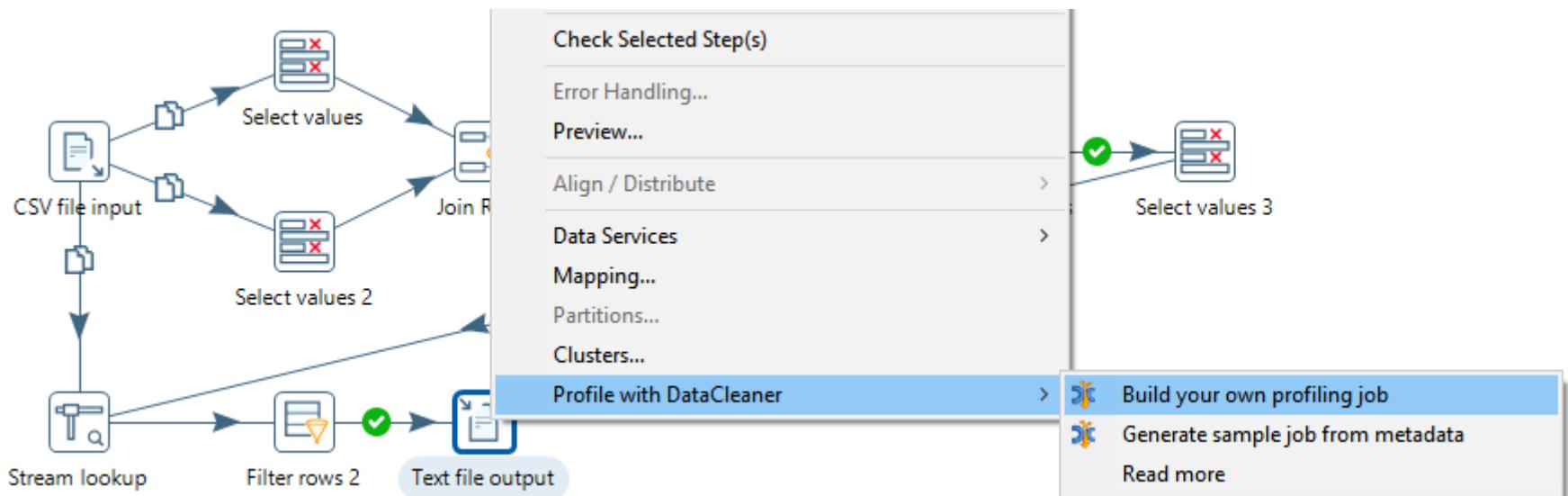


Rows of step: Text file output (64 rows)

#	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	Piestrzewicz	ul. Filtrowa 68	Warszawa	Poland	<null>
5	128	Roland	Keitel	Lyonerstr. 34	Frankfurt	Germany	<null>
6	141	Diego	Freyre	c/ Moralzarjal, 86	Madrid	Spain	<null>
7	144	Christina	Berglund	Berguvsvägen 8	Luleå	Sweden	<null>
8	145	Jytte	Petersen	Vinbæltet 34	København	Denmark	<null>
9	146	Mary	Saveley	2, rue du Commerce	Lyon	France	<null>

Data profiling

- Example
 - perform data profiling after duplicate elimination



Data profiling

The screenshot shows the DataCleaner application interface. The title bar indicates the file is named "datacleaner-job_8c154d46-a3af-11e7-b76a-6b232ec00831.xml | lab05 | Analysis Job | DataCleaner". The menu bar includes "New", "Open", "Save", "Save As...", "More", "Execute", and a close button. The left sidebar contains a tree view of the job structure under "lab05": "information_schema" and "lab05" (which contains "Text file output" with fields "customer_id", "first_name", "last_name", "address", "city", "country", and "duplicate"). Below this is a "Library" section with categories: Transform, Improve (selected), Analyze, Date and time, Visualization, Boolean analyzer, Character set distribution, Completeness analyzer, Number analyzer, Pattern finder, Reference data matcher, Referential integrity, String analyzer, Unique key check, Value distribution, and Value matcher. A search bar at the bottom left says "Search component library...". The main canvas area has a large plus sign icon and the text "Start building ...". Below the canvas, a message says "Add components to your job. Right-click the canvas to explore the library of available components." At the bottom, a status bar shows "Job error status: No result producing components in job", and icons for DataCloud, News Channel, and Community edition.

datacleaner-job_8c154d46-a3af-11e7-b76a-6b232ec00831.xml | lab05 | Analysis Job | DataCleaner

New Open Save Save As... More Execute X

lab05

information_schema

lab05

Text file output

customer_id
first_name
last_name
address
city
country
duplicate

Library

Transform

Improve

Analyze

Date and time

Visualization

Boolean analyzer

Character set distribution

Completeness analyzer

Number analyzer

Pattern finder

Reference data matcher

Referential integrity

String analyzer

Unique key check

Value distribution

Value matcher

Search component library... X

Start building ...

Add components to your job. Right-click the canvas to explore the library of available components.

Job error status: No result producing components in job

DataCloud News Channel Community edition

Data profiling

The screenshot shows the DataCleaner application interface. The title bar indicates the file is named "datacleaner-job_8c154d46-a3af-11e7-b76a-6b232ec00831.xml | lab05 | Analysis Job | DataCleaner". The menu bar includes "New", "Open", "Save", "Save As...", "More", "Execute", and a close button.

The left sidebar displays a tree view of the job structure:

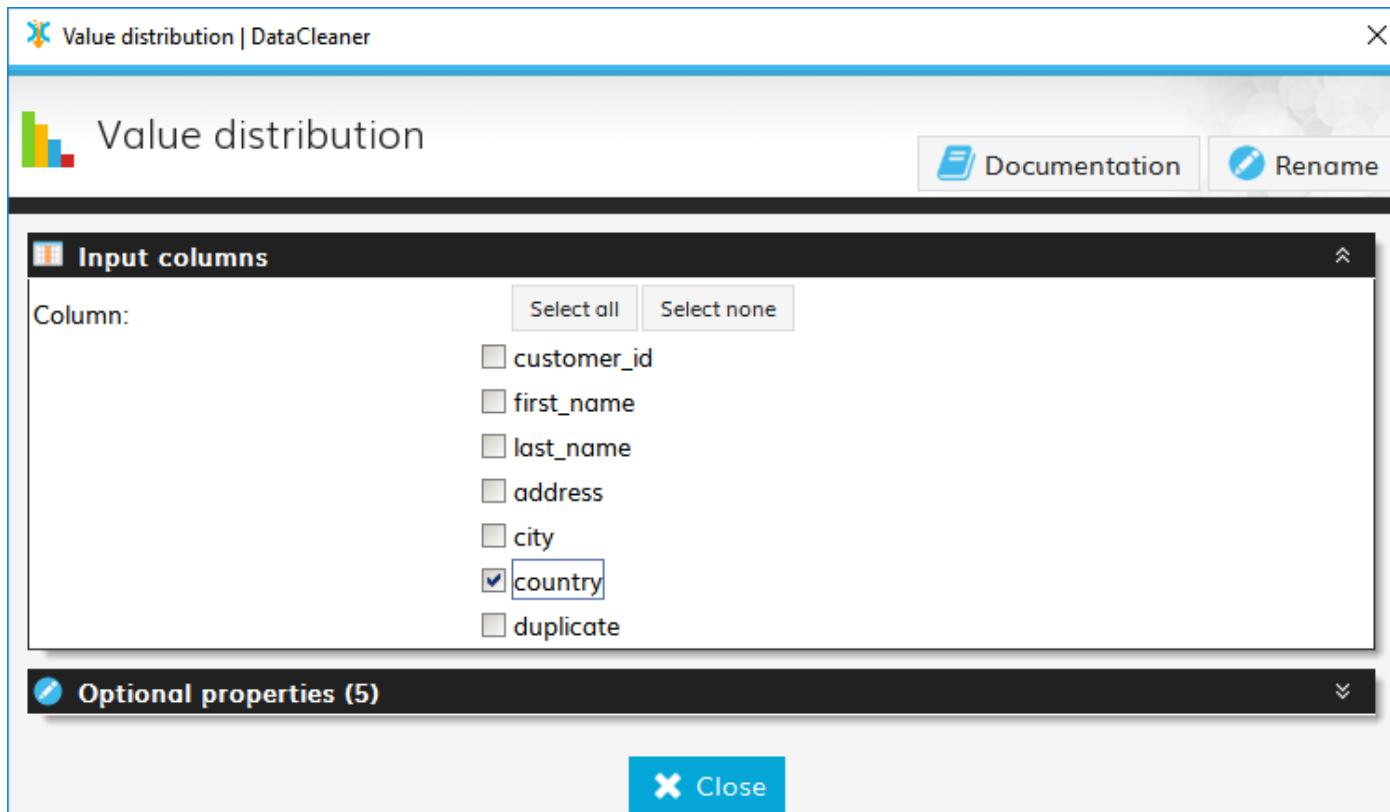
- Lab05
 - information_schema
 - lab05
 - Text file output
 - customer_id
 - first_name
 - last_name
 - address
 - city
 - country
 - duplicate
- Library
 - Transform
 - Improve
 - Analyze
 - Date and time
 - Visualization
 - Boolean analyzer
 - Character set distribution
 - Completeness analyzer
 - Number analyzer
 - Pattern finder
 - Reference data matcher
 - Referential integrity
 - String analyzer
 - Unique key check
 - Value distribution
 - Value matcher

- Search component library...

A large play button icon is positioned in the center-left area. To its right, the text "Ready to execute" is displayed above the instruction: "Click the 'Execute' button in the upper-right corner when you're ready to run the job." A callout bubble points to the "Execute" button in the top right of the window. A message at the bottom left states "Job is correctly configured" with a green checkmark icon. The bottom navigation bar includes links for "DataCloud", "News Channel", and "Community edition".

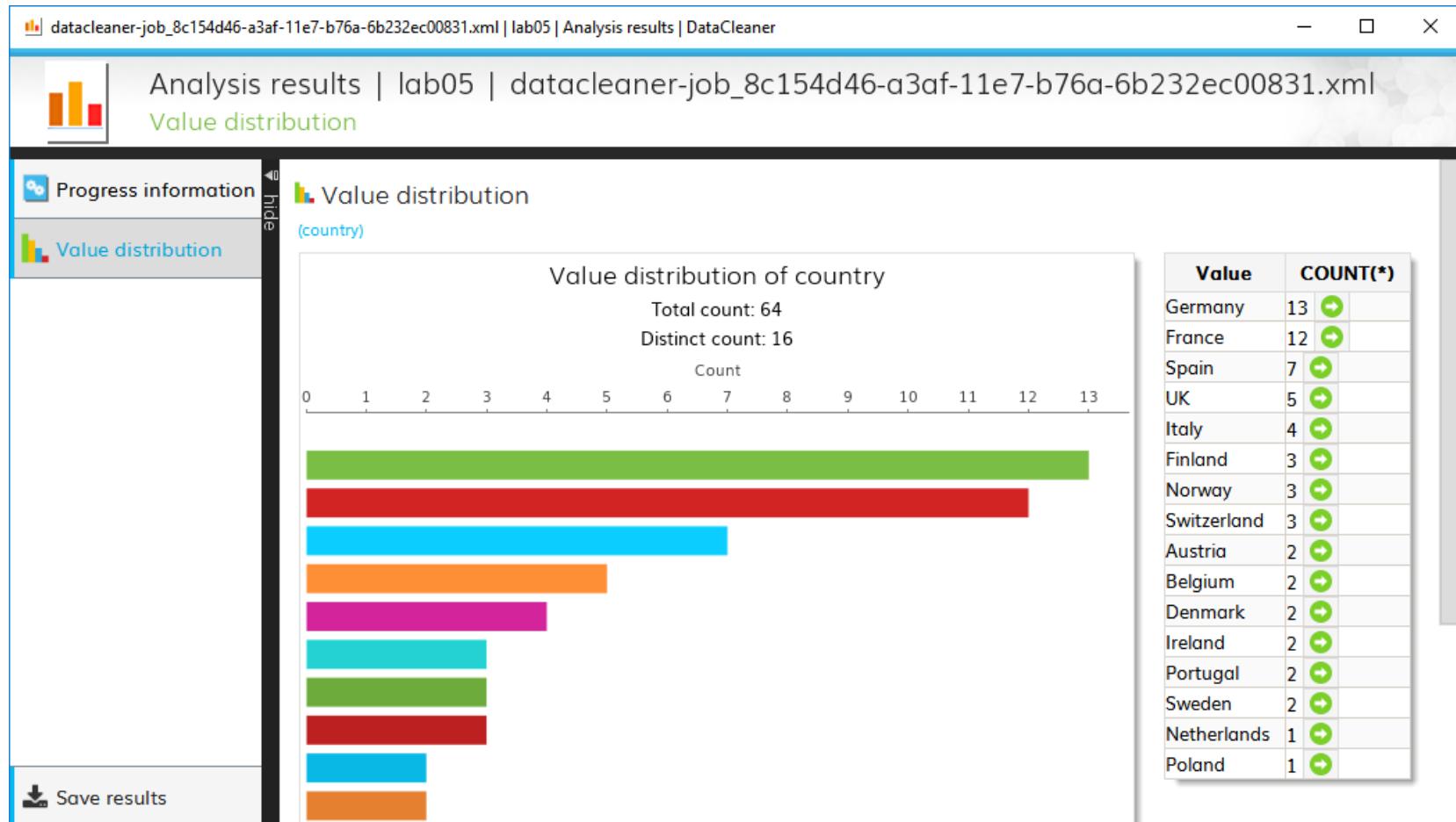
Data profiling

- Value distribution



Data profiling

- Value distribution



Data profiling

The screenshot shows the DataCleaner application window. The title bar reads "datacleaner-job_67a84087-a42e-11e7-b651-4379f2e87310.xml | lab05 | Analysis Job | DataCleaner". The menu bar includes "New", "Open", "Save", "Save As...", "More", "Execute", and a close button. The left sidebar displays a tree view of the job structure under "lab05": "Text file output" (with sub-items "customer_id", "first_name", "last_name", "address", "city", "country", "duplicate") and "Library" (with sub-items "Transform", "Improve", "Analyze" which contains "Date and time", "Visualization", "Boolean analyzer", "Character set distribution", "Completeness analyzer", "Number analyzer", "Pattern finder", "Reference data matcher", "Referential integrity", "String analyzer" (selected), "Unique key check", "Value distribution", and "Value matcher"). A search bar at the bottom says "Search component library...". A status bar at the bottom left says "Job is correctly configured". A large central area shows a flow diagram: "Text file output" feeds into "Value distribution" (represented by a bar chart icon) and "String analyzer" (represented by a question mark icon). A callout bubble points to the "String analyzer" with the text "A?". To the right of the flow diagram is a button labeled "Click here to run job" with an arrow pointing to it. Below the flow diagram is a large play button icon and the text "Ready to execute". Further down, it says "Click the 'Execute' button in the upper-right corner when you're ready to run the job." The bottom right corner has a gear icon.

datacleaner-job_67a84087-a42e-11e7-b651-4379f2e87310.xml | lab05 | Analysis Job | DataCleaner

New Open Save Save As... More Execute X

lab05

- information_schema
- lab05
 - Text file output
 - customer_id
 - first_name
 - last_name
 - address
 - city
 - country
 - duplicate

Library

- Transform
- Improve
- Analyze
 - Date and time
 - Visualization
 - Boolean analyzer
 - Character set distribution
 - Completeness analyzer
 - Number analyzer
 - Pattern finder
 - Reference data matcher
 - Referential integrity
 - String analyzer**
 - Unique key check
 - Value distribution
 - Value matcher

Search component library...

Job is correctly configured

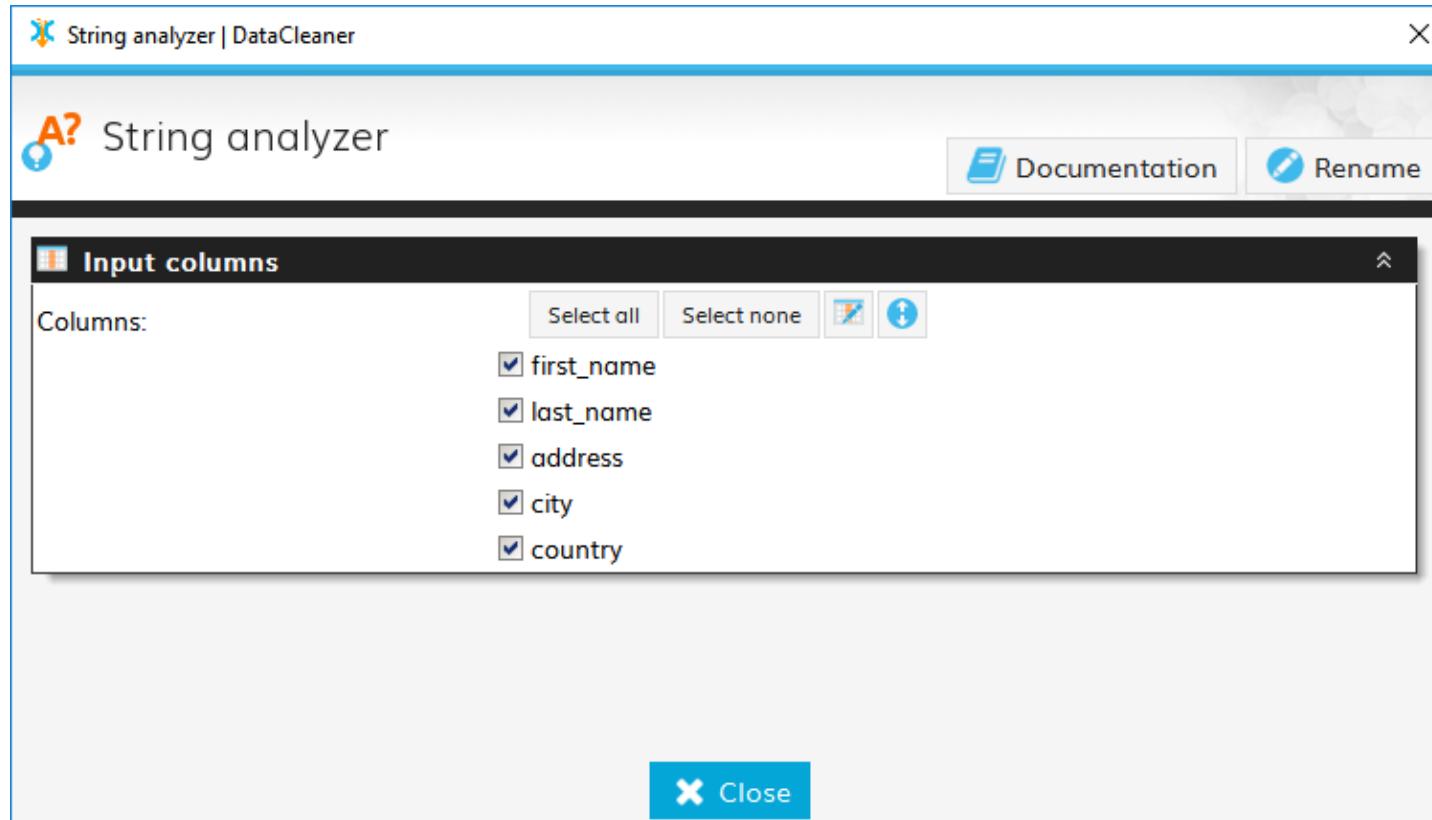
Ready to execute

Click the 'Execute' button in the upper-right corner when you're ready to run the job.

Cloud DataCloud News Channel Community edition

Data profiling

- String analysis



Data profiling

- String analysis

