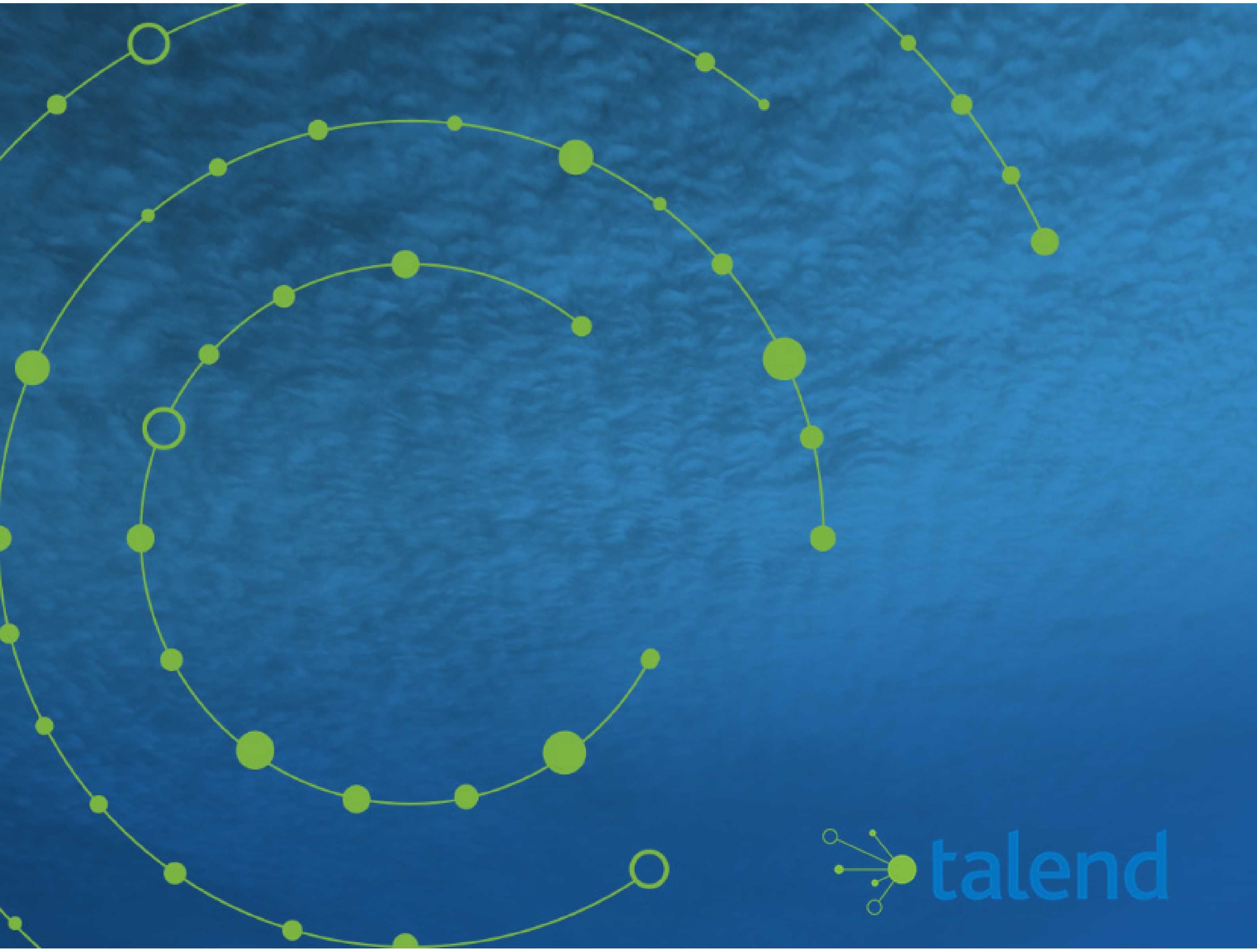


Data Quality Job and Analysis Examples

7.0.1

Contents

Copyleft

Adapted for 7.0.1. Supersedes previous releases.

Publication date: April 13, 2018

Notices

License Agreement

This product includes software developed at ASM, AntlR, Apache ActiveMQ, Apache Ant, Apache Axiom, Apache Axis, Apache Axis 2, Apache Chemistry, Apache Common Http Client, Apache Common Http Core, Apache Commons, Apache Commons Bcel, Apache Commons Lang, Apache Datafu, Apache Derby Database Engine and Embedded JDBC Driver, Apache Geronimo, Apache HCatalog, Apache Hadoop, Apache Hbase, Apache Hive, Apache HttpClient, Apache HttpComponents Client, Apache JAMES, Apache Log4j, Apache Neethi, Apache POI, Apache Pig, Apache Thrift, Apache Tomcat, Apache Xml-RPC, Apache Zookeeper, CSV Tools, DataNucleus, Doug Lea, Ezmorph, Google's phone number handling library, Guava: Google Core Libraries for Java, H2 Embedded Database and JDBC Driver, HighScale Lib, HsqlDB, JSON, JUnit, Jackson Java JSON-processor, Java API for RESTful Services, Java Universal Network Graph, Jaxb, Jaxen, Jetty, Joda-Time, Json Simple, MapDB, MetaStuff, Paraccel JDBC Driver, PostgreSQL JDBC Driver, Protocol Buffers - Google's data interchange format, Resty: A simple HTTP REST client for Java, SL4J: Simple Logging Facade for Java, SQLite JDBC Driver, The Castor Project, The Legion of the Bouncy Castle, Woden, Xalan-J, Xerces2, XmlBeans, XmlSchema Core, atinject. Licensed under their respective license.

Profiling customer data

Incorporating appropriate data quality tools in your business processes is vital at the beginning of any project and through the project plan in order to see what type of data quality you have and decide how and what data to resolve.

Suppose, for example, that you want to start a campaign for your sales and marketing groups, or you need to contact customers for billing and payment and your main source to contact appropriate people is email and postal addresses. Having consistent and correct address data is vital in such campaign to be able to reach all people.

This section provides an example of profiling US customer email and postal addresses.

Identifying data anomalies

The first step in this example is to profile the customer contact information in a MySQL database. The profiling results provides you with statistics about the values within each column.

How to profile address columns

You will use Talend Studio to analyze few customer columns including email and postal.

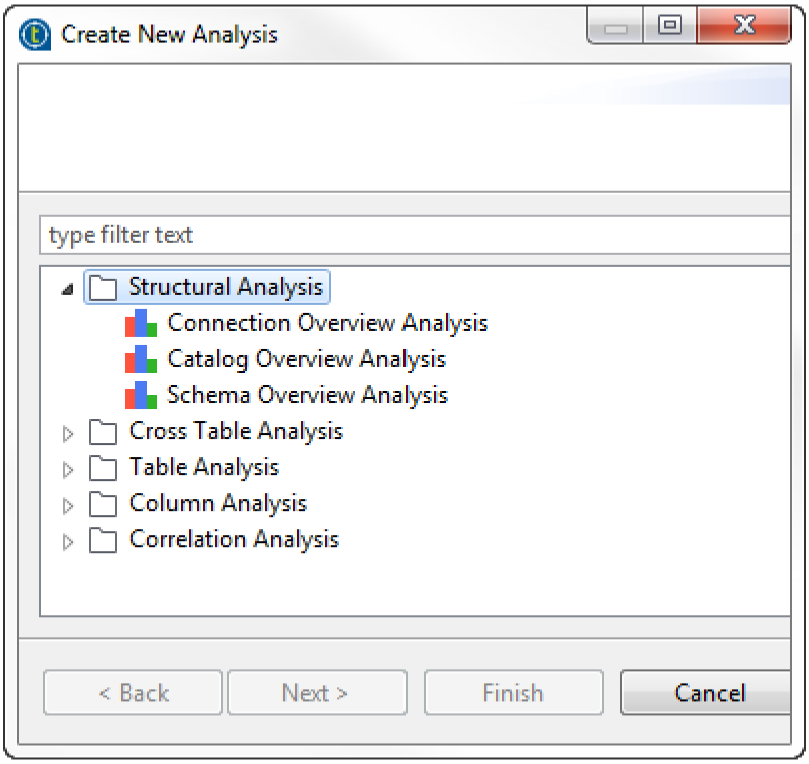
Using out-of-box indicators and patterns on these columns, you can show in the analysis results the matching and non-matching address data, the number of most frequent records for each distinct pattern and the row, duplicate and blank counts in each column.

Defining the column analysis

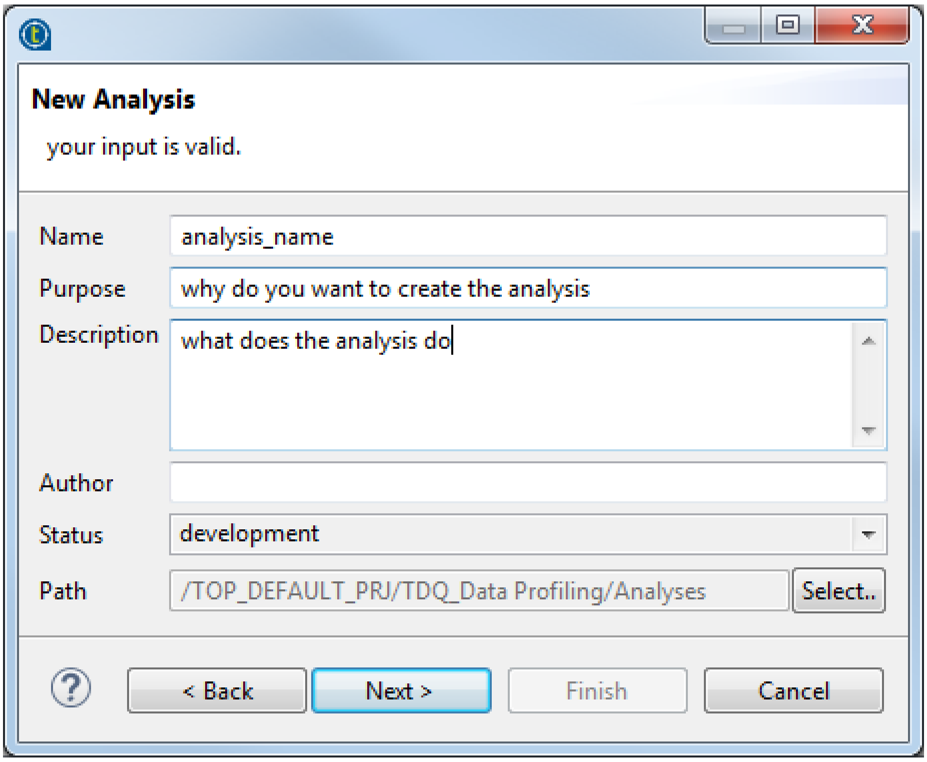
Procedure

1. In the **DQ Repository** tree view, right-click the **Analyses** folder and select **New Analysis**.

   The **Create New Analysis** wizard opens.



2. Start typing `Basic column analysis` in the search field, select **Basic Column Analysis** from the list and click **Next**.



New Analysis

your input is valid.

Name: analysis_name

Purpose: why do you want to create the analysis

Description: what does the analysis do

Author:

Status: development

Path: /TOP_DEFAULT_PRJ/TDQ_Data Profiling/Analyses

3. In the **Name** field, enter a name for the current column analysis.

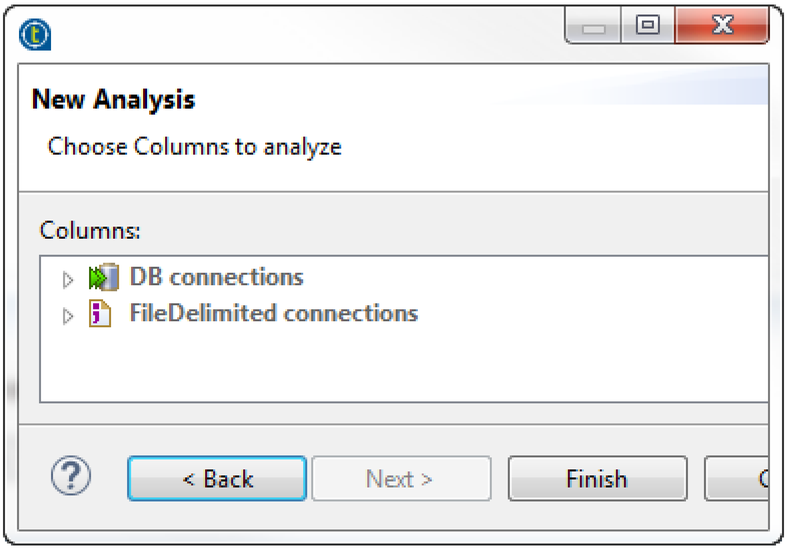
Note:

Avoid using special characters in the item names including:

"~", "!", "`", "#", "^", "&", "*", "\\", "/", "?", ":", ";", "\"", ".", "(", ")", "'", "¥", "'", """, "«", "»", "<", ">".

These characters are all replaced with "_" in the file system and you may end up creating duplicate items.

4. Set column analysis metadata (purpose, description and author name) in the corresponding fields and click **Next**.



New Analysis

Choose Columns to analyze

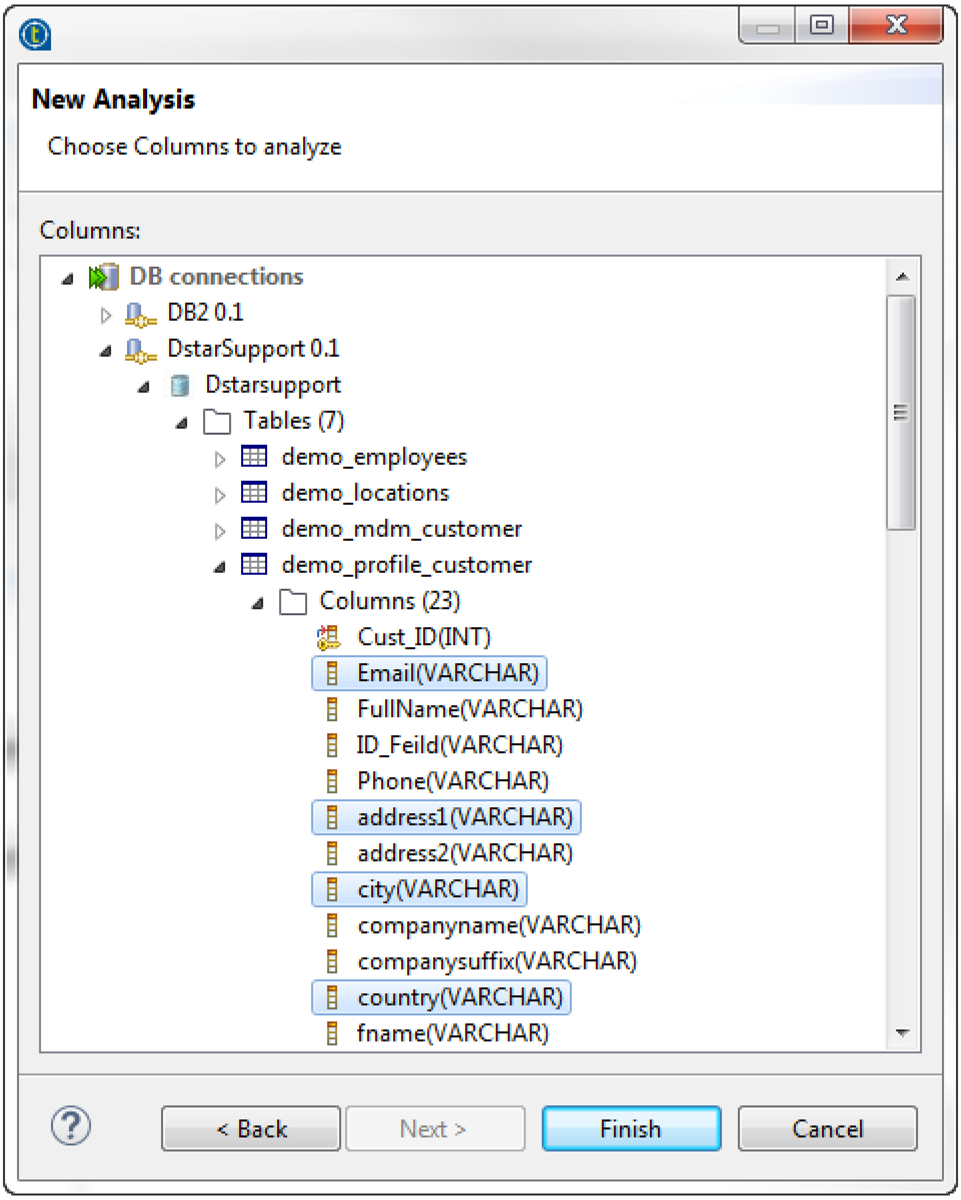
Columns:

- DB connections
- FileDelimited connections

Selecting the address columns and setting sample data

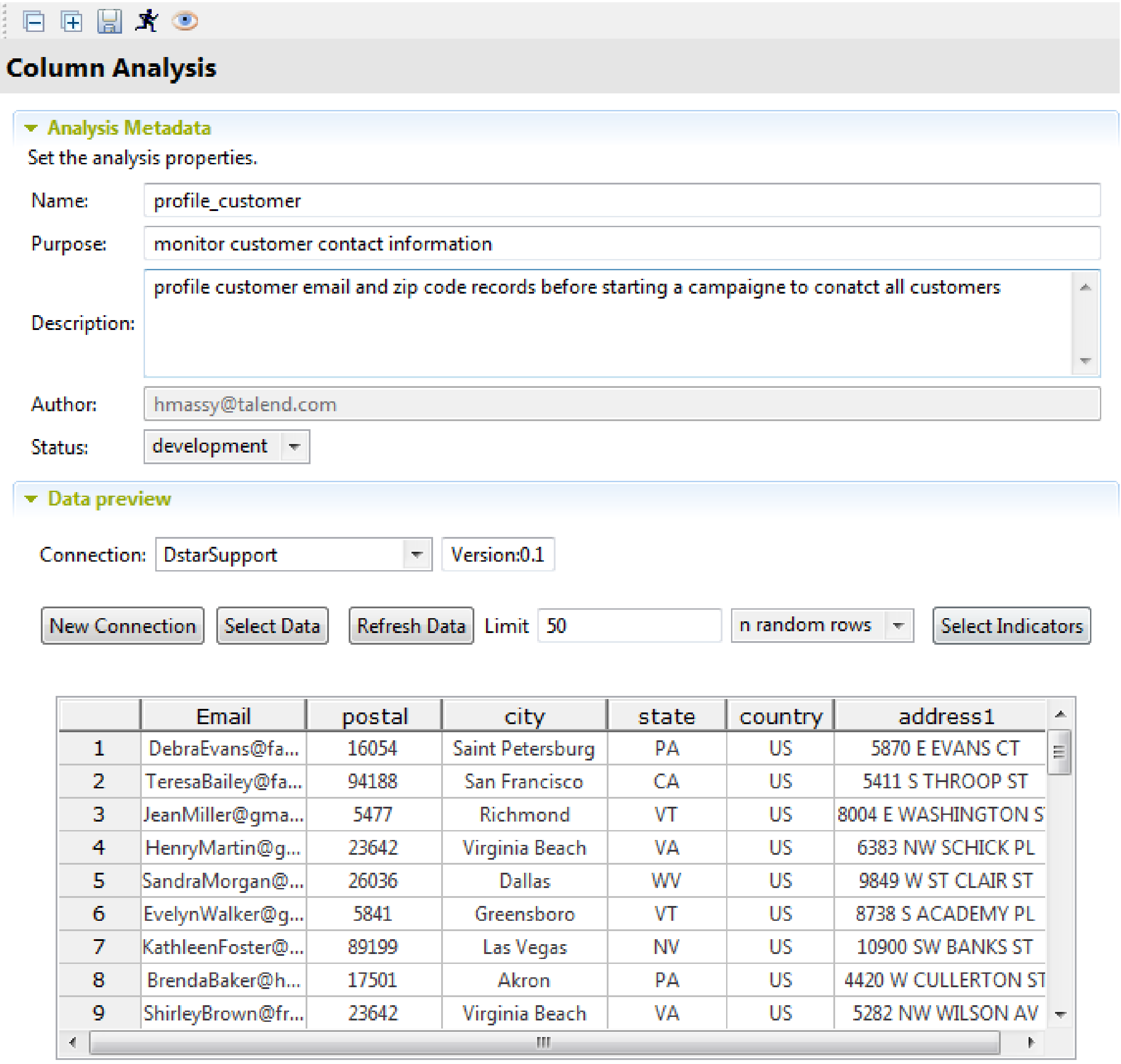
Procedure

1. Expand **DB connections** and browse to the address columns you want to analyze.



2. Select the columns and click **Finish** to close the wizard.

   A file for the newly created column analysis is listed under the **Analysis** node in the **DQ Repository** tree view, and the analysis editor opens with the analysis metadata.



Column Analysis

Analysis Metadata

Set the analysis properties.

Name: profile_customer

Purpose: monitor customer contact information

Description: profile customer email and zip code records before starting a campaigne to conatct all customers

Author: hmassy@talend.com

Status: development

Data preview

Connection: DstarSupport

Version:0.1

Limit 50

n random rows

	Email	postal	city	state	country	address1
1	DebraEvans@fa...	16054	Saint Petersburg	PA	US	5870 E EVANS CT
2	TeresaBailey@fa...	94188	San Francisco	CA	US	5411 S THROOP ST
3	JeanMiller@gma...	5477	Richmond	VT	US	8004 E WASHINGTON S
4	HenryMartin@g...	23642	Virginia Beach	VA	US	6383 NW SCHICK PL
5	SandraMorgan@...	26036	Dallas	WV	US	9849 W ST CLAIR ST
6	EvelynWalker@g...	5841	Greensboro	VT	US	8738 S ACADEMY PL
7	KathleenFoster@...	89199	Las Vegas	NV	US	10900 SW BANKS ST
8	BrendaBaker@h...	17501	Akron	PA	US	4420 W CULLERTON ST
9	ShirleyBrown@fr...	23642	Virginia Beach	VA	US	5282 NW WILSON AV

3. In the **Data preview** view, click **Refresh Data**.

   The data in the selected columns is displayed in the table.

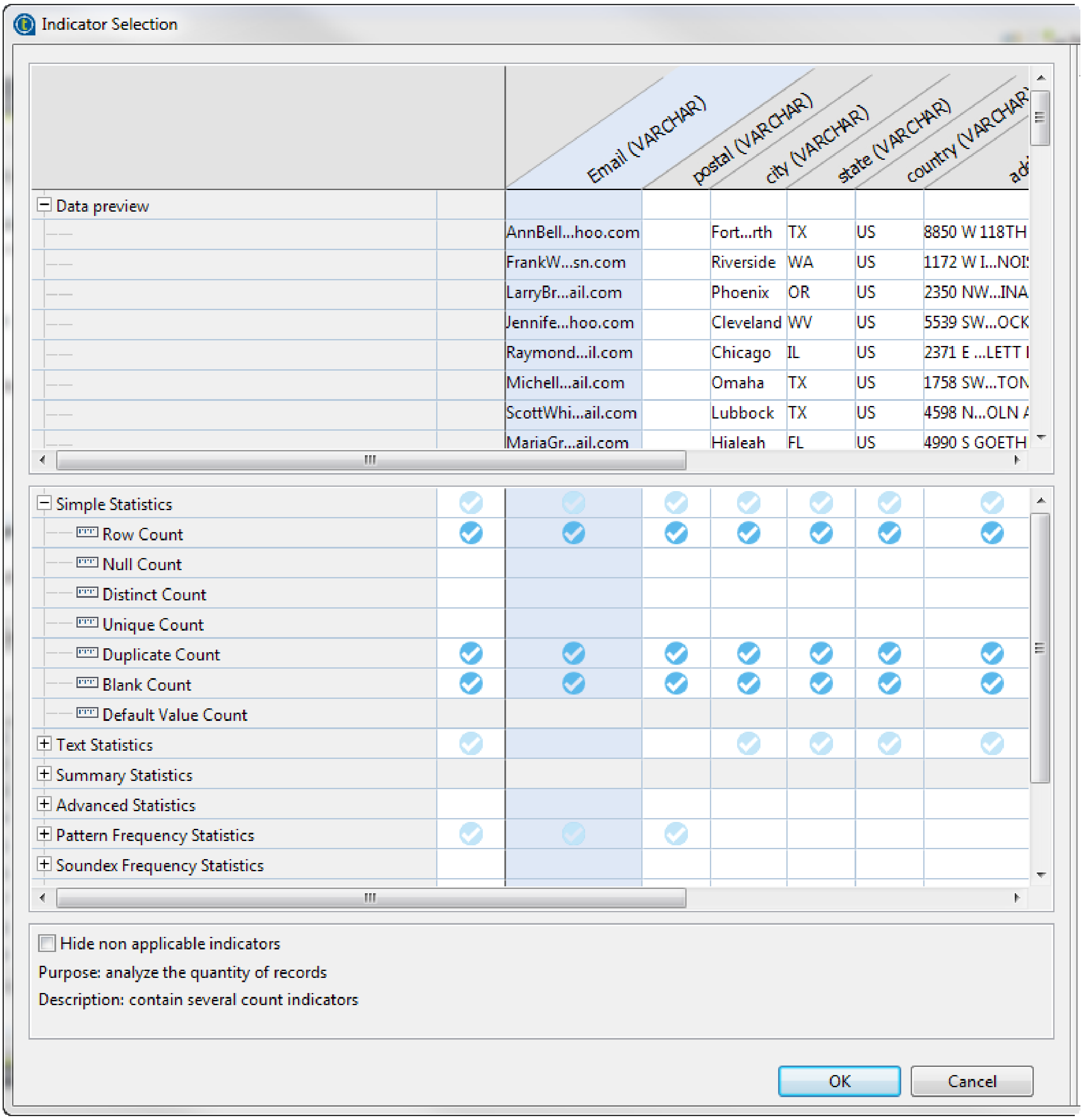
   You can change your data source and your selected columns by using the **New Connection** and **Select Data** buttons respectively.

4. In the **Limit** field, set to `50` the number for the data records you want to display in the table and use as sample data.
5. Select **n random rows** to list `50` random records from the selected columns.

Setting system indicators

Procedure

1. From the **Data preview** view in the analysis editor, click **Select indicators** to open the **Indicator Selection** dialog box.



Indicator Selection

	Email (VARCHAR)	postal (VARCHAR)	city (VARCHAR)	state (VARCHAR)	country (VARCHAR)	ad
Data preview						
	AnnBell...hoo.com		Fort...rth	TX	US	8850 W 118TH
	FrankW...sn.com		Riverside	WA	US	1172 W I...NOI
	LarryBr...ail.com		Phoenix	OR	US	2350 NW...INA
	Jennife...hoo.com		Cleveland	WV	US	5539 SW...OCK
	Raymond...il.com		Chicago	IL	US	2371 E ...LETT I
	Michell...ail.com		Omaha	TX	US	1758 SW...TON
	ScottWhi...ail.com		Lubbock	TX	US	4598 N...OLN
	MariaGr...ail.com		Hialeah	FL	US	4990 S GOETH

Simple Statistics	✓	✓	✓	✓	✓	✓	✓
Row Count	✓	✓	✓	✓	✓	✓	✓
Null Count							
Distinct Count							
Unique Count							
Duplicate Count	✓	✓	✓	✓	✓	✓	✓
Blank Count	✓	✓	✓	✓	✓	✓	✓
Default Value Count							
Text Statistics	✓			✓	✓	✓	✓
Summary Statistics							
Advanced Statistics							
Pattern Frequency Statistics	✓	✓	✓				
Soundex Frequency Statistics							

□ Hide non applicable indicators

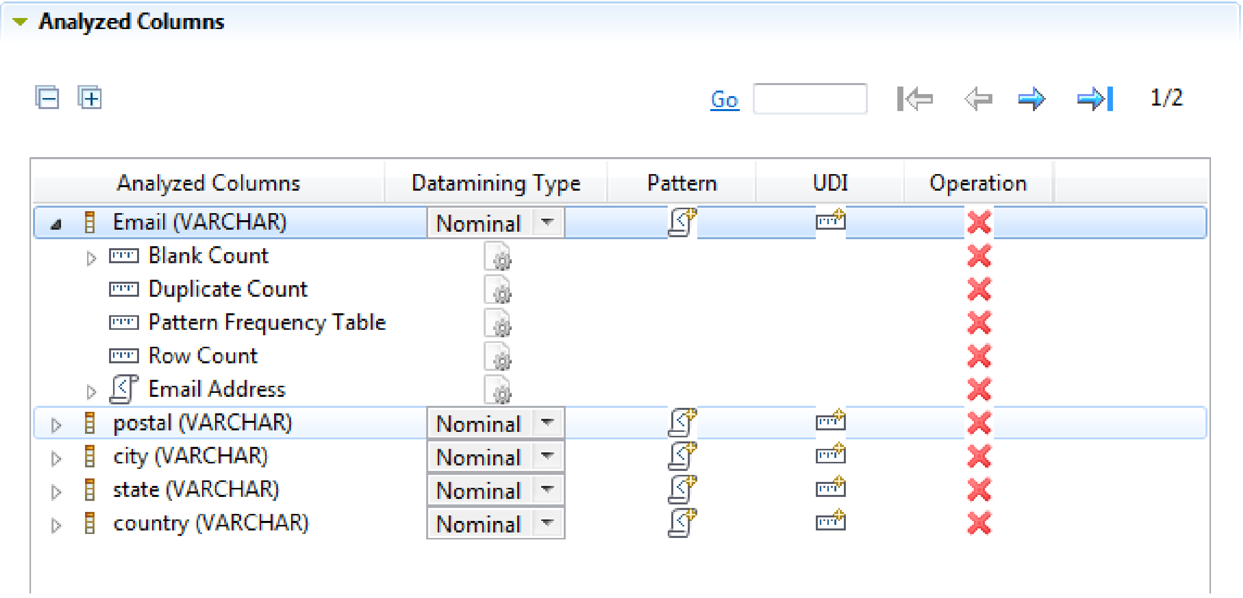
Purpose: analyze the quantity of records

Description: contain several count indicators

2. Click in the cells next to indicators names to set indicator parameters for the analyzed columns and click **OK**.

   You want to see the row, blank and duplicate counts in all columns to see how consistent the data is. Also you want to use the **Pattern Frequency Table** indicator on the email and postal columns in order to compute the number of most frequent records for each distinct pattern or value.

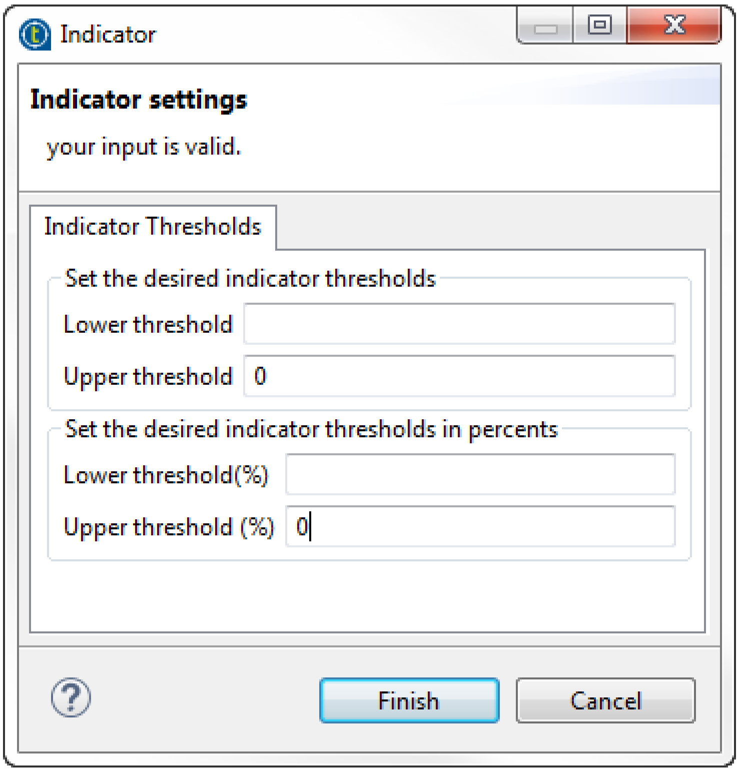
   Indicators are added accordingly to the columns in the **Analyzed Columns** view.



Analyzed Columns	Datamining Type	Pattern	UDI	Operation
Email (VARCHAR)	Nominal			✕
Blank Count				✕
Duplicate Count				✕
Pattern Frequency Table				✕
Row Count				✕
Email Address				✕
postal (VARCHAR)	Nominal			✕
city (VARCHAR)	Nominal			✕
state (VARCHAR)	Nominal			✕
country (VARCHAR)	Nominal			✕

3. Click the option icon ⚙ next to the **Blank Count** indicator and set `0` in the **Upper threshold** field.

   Defining thresholds on indicators is very helpful as it will write in red the count of the null values in the analysis results.



Indicator settings

your input is valid.

Set the desired indicator thresholds

Lower threshold:

Upper threshold: 0

Set the desired indicator thresholds in percents

Lower threshold(%):

Upper threshold (%): 0

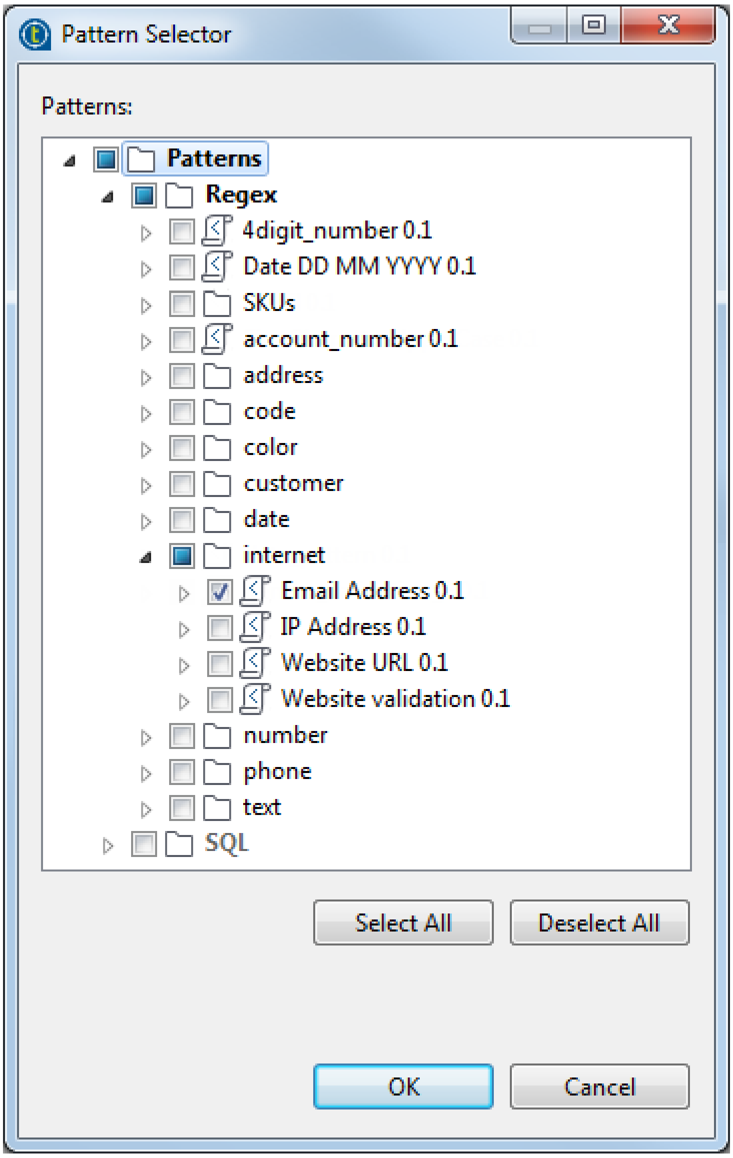
Setting patterns

You would want now to match the content of the email column against a standard email format and the postal column against a standard US zip code format.

This will define the content, structure and quality of emails and zip codes and give a percentage of the data that match the standard formats and the data that does not match.

Procedure

1. In the **Analyzed Columns** view, click the icon next to email.



2. In the **Pattern Selector** dialog box, expand **Regex** and browse to **Email Address** in the **internet** folder, and then click **OK**.
3. Click the option icon ⚙ next to the **Email Address** indicator and set `98.0` in the **Lower threshold (%)** field.

   If the number of the records that match the pattern is fewer than 98%, it will be written in red in the analysis results.
4. Do the same to add to the postal column the **US Zipcode Validation** pattern from the **address** folder.

   For further information on pattern types and their usage when analyzing data, see Talend Studio User Guide at https://help.talend.com.

Executing the analysis and displaying the profiling results

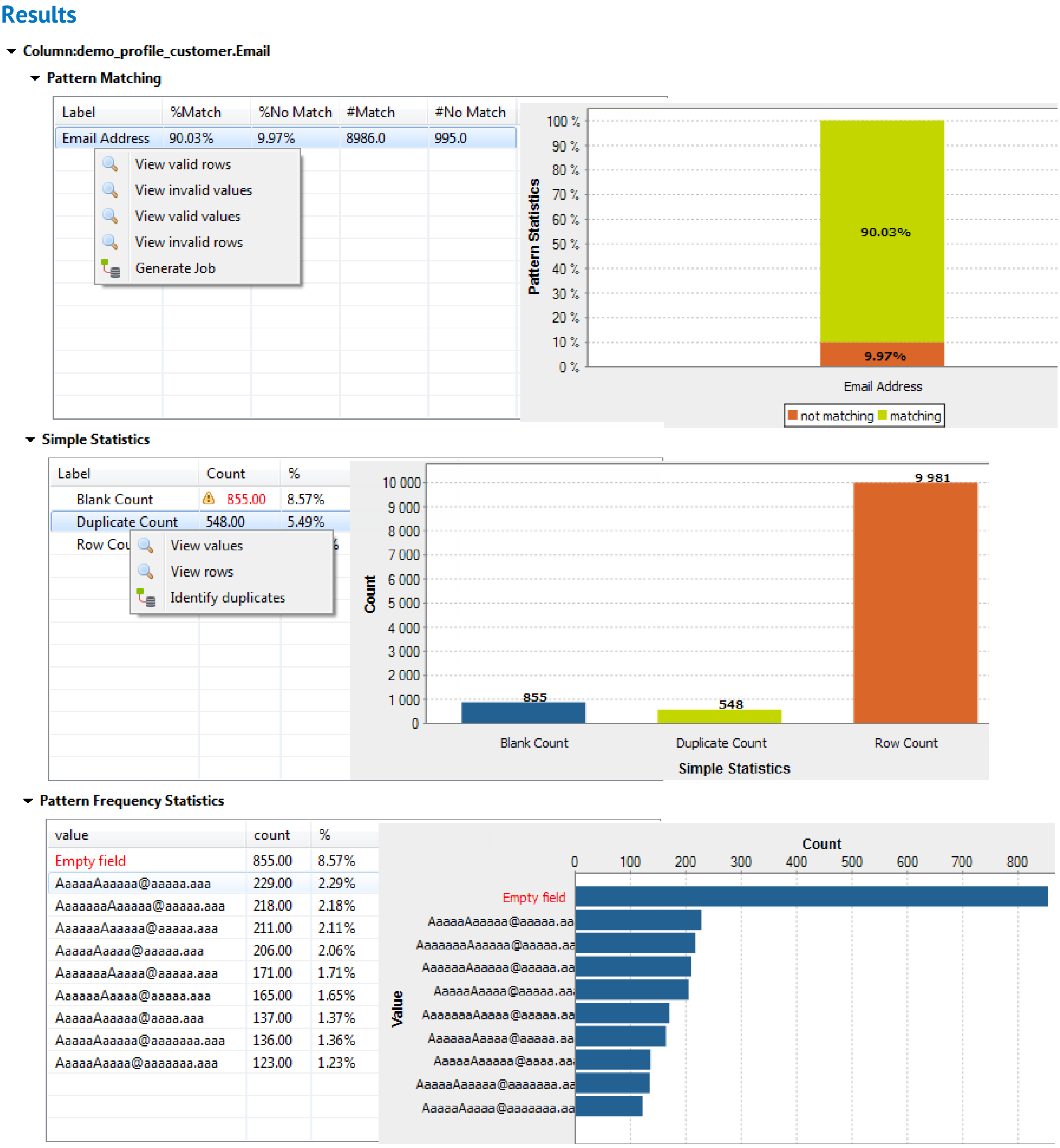
Procedure

1. Save the column analysis in the analysis editor and then press **F6** to execute it.

A group of graphics is displayed in the **Graphics** panel to the right of the analysis editor showing the results of the column analysis including those for pattern matching.

2. Click the **Analysis Results** tab at the bottom of the analysis editor to access a more detail result view.

   These results show the generated graphics for the analyzed columns accompanied with tables that detail the statistic and pattern matching results.



Results

Column:demo_profile_customer.Email

Pattern Matching

Label	%Match	%No Match	#Match	#No Match
Email Address	90.03%	9.97%	8986.0	995.0

Simple Statistics

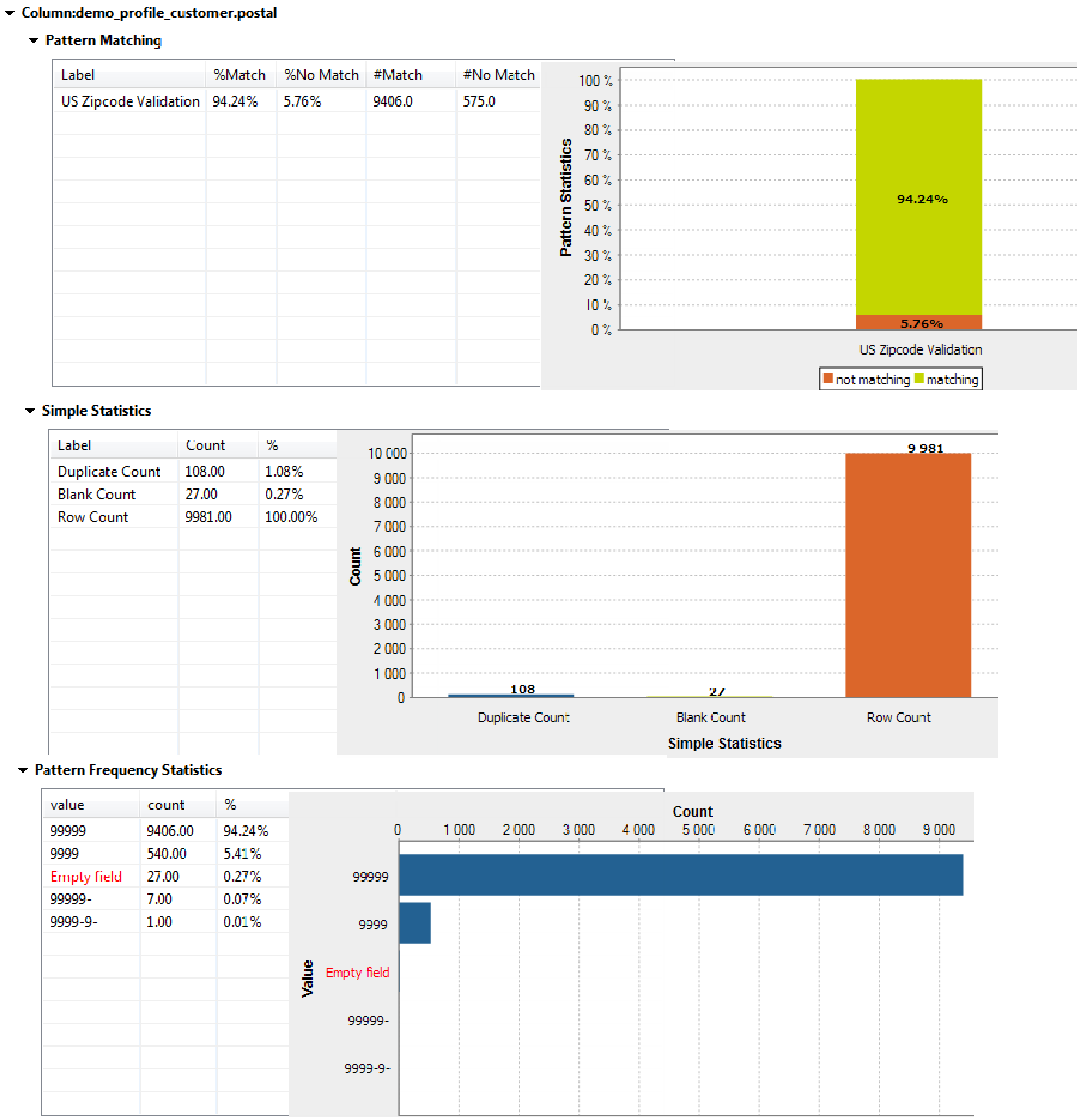
Label	Count	%
Blank Count	⚠ 855.00	8.57%
Duplicate Count	548.00	5.49%
Row Cou		

Pattern Frequency Statistics

value	count	%
Empty field	855.00	8.57%
AaaaaAaaaaa@aaaaa.aaa	229.00	2.29%
AaaaaaaAaaaaa@aaaaa.aaa	218.00	2.18%
AaaaaAaaaaa@aaaaa.aaa	211.00	2.11%
AaaaaAaaaa@aaaaa.aaa	206.00	2.06%
AaaaaaaAaaaa@aaaaa.aaa	171.00	1.71%
AaaaaAaaaa@aaaaa.aaa	165.00	1.65%
AaaaaAaaaaa@aaaa.aaa	137.00	1.37%
AaaaaAaaaaa@aaaaaaa.aaa	136.00	1.36%
AaaaaAaaaa@aaaaaaa.aaa	123.00	1.23%

The pattern matching results show that about 10% of the email records do not match the standard email pattern. The simple statistic results show that about 8% of the email records are blank and that about 5% are duplicates. And the pattern frequency results give the number of most frequent records for each distinct pattern. This shows that the data is not consistent and you need to correct and cleans the email data before starting your campaign.

The results for the postal column look as the following:



Column:demo_profile_customer.postal

Pattern Matching

Label	%Match	%No Match	#Match	#No Match
US Zipcode Validation	94.24%	5.76%	9406.0	575.0

Simple Statistics

Label	Count	%
Duplicate Count	108.00	1.08%
Blank Count	27.00	0.27%
Row Count	9981.00	100.00%

Pattern Frequency Statistics

value	count	%
99999	9406.00	94.24%
9999	540.00	5.41%
Empty field	27.00	0.27%
99999-	7.00	0.07%
9999-9-	1.00	0.01%

The result sets for the postal column give the count of the records that match and those that do not match a standard US zip code format. The results sets also give the blank and duplicate counts and the number of most frequent records for each distinct pattern. These results show that the data is not very consistent.

Then some percentage of the customers can not be contacted by either email or US mail service. These results show clearly that your data is not very consistent and that it needs to be corrected.

How to view analyzed data

After running the column analysis using the SQL engine and from the **Analysis Results** view of the analysis editor, you can right-click any of the rows/bars in the result tables/charts and access a view of the actual analyzed data.

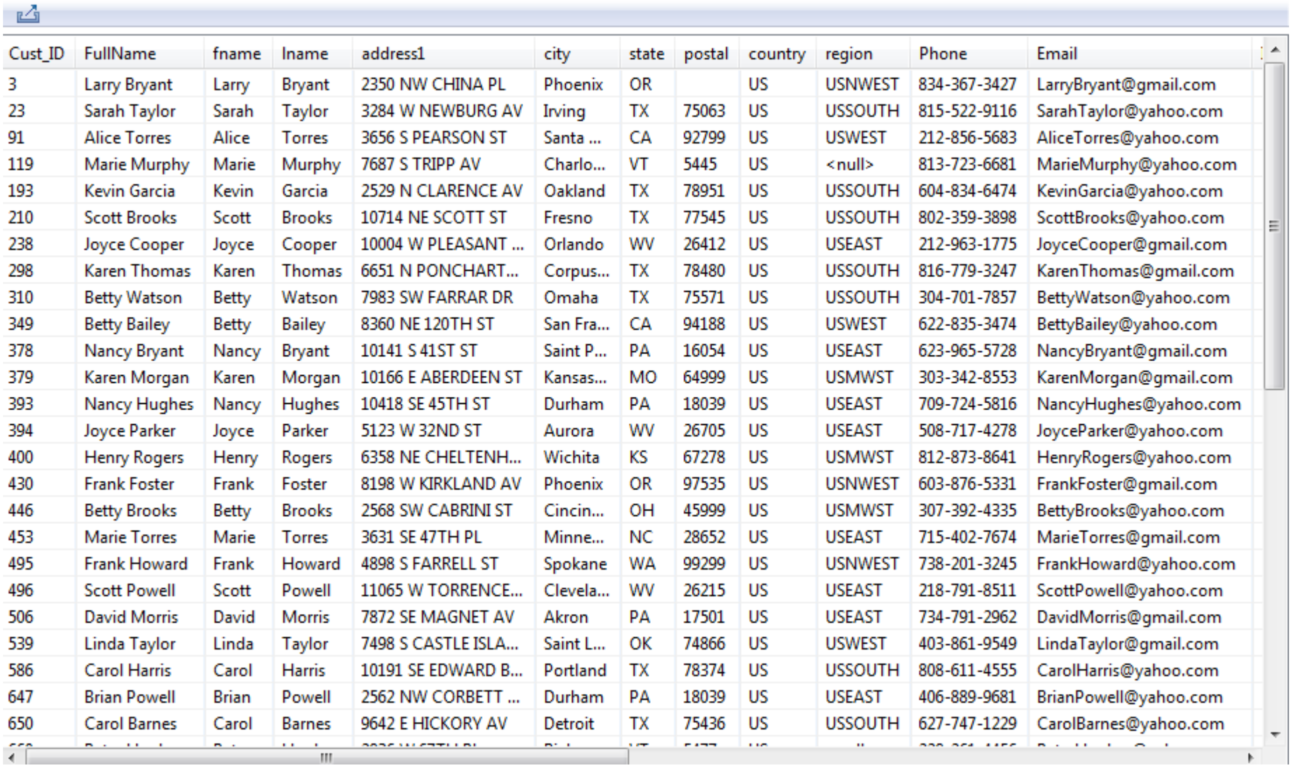
This could be very helpful to see invalid rows for example and start analyzing what needs to be done to clean such data.

Procedure

1. At the bottom of the analysis editor, click the **Analysis Results** tab to open a detailed view of the analysis results.
2. Right-click the data row in the statistic results of the email column and select **View rows** for example.

Results

The **Data Explorer** perspective opens listing the invalid rows in the email column.



Cust_ID	FullName	fname	lname	address1	city	state	postal	country	region	Phone	Email
3	Larry Bryant	Larry	Bryant	2350 NW CHINA PL	Phoenix	OR		US	USNWEST	834-367-3427	LarryBryant@gmail.com
23	Sarah Taylor	Sarah	Taylor	3284 W NEWBURG AV	Irving	TX	75063	US	USSOUTH	815-522-9116	SarahTaylor@yahoo.com
91	Alice Torres	Alice	Torres	3656 S PEARSON ST	Santa ...	CA	92799	US	USWEST	212-856-5683	AliceTorres@yahoo.com
119	Marie Murphy	Marie	Murphy	7687 S TRIPP AV	Charlo...	VT	5445	US	<null>	813-723-6681	MarieMurphy@yahoo.com
193	Kevin Garcia	Kevin	Garcia	2529 N CLARENCE AV	Oakland	TX	78951	US	USSOUTH	604-834-6474	KevinGarcia@yahoo.com
210	Scott Brooks	Scott	Brooks	10714 NE SCOTT ST	Fresno	TX	77545	US	USSOUTH	802-359-3898	ScottBrooks@yahoo.com
238	Joyce Cooper	Joyce	Cooper	10004 W PLEASANT ...	Orlando	WV	26412	US	USEAST	212-963-1775	JoyceCooper@gmail.com
298	Karen Thomas	Karen	Thomas	6651 N PONCHART...	Corpus...	TX	78480	US	USSOUTH	816-779-3247	KarenThomas@gmail.com
310	Betty Watson	Betty	Watson	7983 SW FARRAR DR	Omaha	TX	75571	US	USSOUTH	304-701-7857	BettyWatson@yahoo.com
349	Betty Bailey	Betty	Bailey	8360 NE 120TH ST	San Fra...	CA	94188	US	USWEST	622-835-3474	BettyBailey@yahoo.com
378	Nancy Bryant	Nancy	Bryant	10141 S 41ST ST	Saint P...	PA	16054	US	USEAST	623-965-5728	NancyBryant@gmail.com
379	Karen Morgan	Karen	Morgan	10166 E ABERDEEN ST	Kansas...	MO	64999	US	USMWST	303-342-8553	KarenMorgan@gmail.com
393	Nancy Hughes	Nancy	Hughes	10418 SE 45TH ST	Durham	PA	18039	US	USEAST	709-724-5816	NancyHughes@yahoo.com
394	Joyce Parker	Joyce	Parker	5123 W 32ND ST	Aurora	WV	26705	US	USEAST	508-717-4278	JoyceParker@yahoo.com
400	Henry Rogers	Henry	Rogers	6358 NE CHELTENH...	Wichita	KS	67278	US	USMWST	812-873-8641	HenryRogers@yahoo.com
430	Frank Foster	Frank	Foster	8198 W KIRKLAND AV	Phoenix	OR	97535	US	USNWEST	603-876-5331	FrankFoster@gmail.com
446	Betty Brooks	Betty	Brooks	2568 SW CABRINI ST	Cincin...	OH	45999	US	USMWST	307-392-4335	BettyBrooks@yahoo.com
453	Marie Torres	Marie	Torres	3631 SE 47TH PL	Minne...	NC	28652	US	USEAST	715-402-7674	MarieTorres@gmail.com
495	Frank Howard	Frank	Howard	4898 S FARRELL ST	Spokane	WA	99299	US	USNWEST	738-201-3245	FrankHoward@yahoo.com
496	Scott Powell	Scott	Powell	11065 W TORRENCE...	Clevela...	WV	26215	US	USEAST	218-791-8511	ScottPowell@yahoo.com
506	David Morris	David	Morris	7872 SE MAGNET AV	Akron	PA	17501	US	USEAST	734-791-2962	DavidMorris@gmail.com
539	Linda Taylor	Linda	Taylor	7498 S CASTLE ISLA...	Saint L...	OK	74866	US	USWEST	403-861-9549	LindaTaylor@gmail.com
586	Carol Harris	Carol	Harris	10191 SE EDWARD B...	Portland	TX	78374	US	USSOUTH	808-611-4555	CarolHarris@yahoo.com
647	Brian Powell	Brian	Powell	2562 NW CORBETT ...	Durham	PA	18039	US	USEAST	406-889-9681	BrianPowell@yahoo.com
650	Carol Barnes	Carol	Barnes	9642 E HICKORY AV	Detroit	TX	75436	US	USSOUTH	627-747-1229	CarolBarnes@yahoo.com