Talend Open Studio for Data Quality

Getting Started Guide

6.0.1

Adapted for v6.0.1. Supersedes previous releases.

Publication date: September 10, 2015

Copyleft

This documentation is provided under the terms of the Creative Commons Public License (CCPL).

For more information about what you can and cannot do with this documentation in accordance with the CCPL, please read: http://creativecommons.org/licenses/by-nc-sa/2.0/

Notices

Talend is a trademark of Talend, Inc.

All brands, product names, company names, trademarks and service marks are the properties of their respective owners.

License Agreement

The software described in this documentation is licensed under the Apache License, Version 2.0 (the "License"); you may not use this software except in compliance with the License. You may obtain a copy of the License at http://www.apache.org/licenses/LICENSE-2.0.html. Unless required by applicable law or agreed to in writing, software distributed under the License is distributed on an "AS IS" BASIS, WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied. See the License for the specific language governing permissions and limitations under the License.

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.

Table of Contents

Preface

1. General information

1.1. Purpose

This guide aims at helping users get started with the *Talend Open Studio for Data Quality* quickly. For detailed explanations on features and functions of the *Talend Open Studio for Data Quality*, see the other documentation delivered with the *Talend Open Studio for Data Quality*.

Information presented in this document applies to *Talend Open Studio for Data Quality* **6.0.1**.

1.2. Audience

This guide is for users and administrators of *Talend Open Studio for Data Quality*.

The layout of GUI screens provided in this document may vary slightly from your actual GUI.

1.3. Typographical conventions

This guide uses the following typographical conventions:

- text in **bold:** window and dialog box buttons and fields, keyboard keys, menus, and menu and options,

- text in **[bold]:** window, wizard, and dialog box titles,

- text in `courier`: system parameters typed in by the user,

- text in *italics*: file, schema, column, row, and variable names,

- text in *italics*: file, schema, column, row, and variable names,

- The icon indicates an item that provides additional information about an important point. It is also used to add comments related to a table or a figure,

- The icon indicates a message that gives information about the execution requirements or recommendation type. It is also used to refer to situations or information the end-user needs to be aware of or pay special attention to.

- `Any command is highlighted with a grey background or code typeface.`

2. Feedback and Support

Your feedback is valuable. Do not hesitate to give your input, make suggestions or requests regarding this documentation or product and find support from the **Talend** team, on **Talend**'s Forum website at:

http://talendforge.org/forum

Chapter 1. Getting Started with Talend Studio

This chapter provides basic information required to get started with *Talend Studio*, including launching *Talend Studio* and creating projects.

1.1. Launching Talend Studio

This section guides you through the basics for launching *Talend Studio* for the first time and opening your first project in the Studio, and provides information on setting up a project.

## 1.1.1. How to launch the Studio for the first time

To open *Talend Studio* for the first time, complete the following:

1. Uncompress the *Talend Studio* zip file and, in the folder, double-click the executable file corresponding to your operating system.

    > The Studio zip archive contains binaries for several platforms including Mac OS X and Linux/Unix.

2. In the **[User License Agreement]** dialog box that opens, read and accept the terms of the end user license agreement to proceed.

3. Click **Start now!** to open *Talend Studio* main window, which displays a welcome page that provides useful tips for beginners on how to get started with the Studio. Clicking an underlined link brings you to the corresponding tab view or opens the corresponding dialog box.
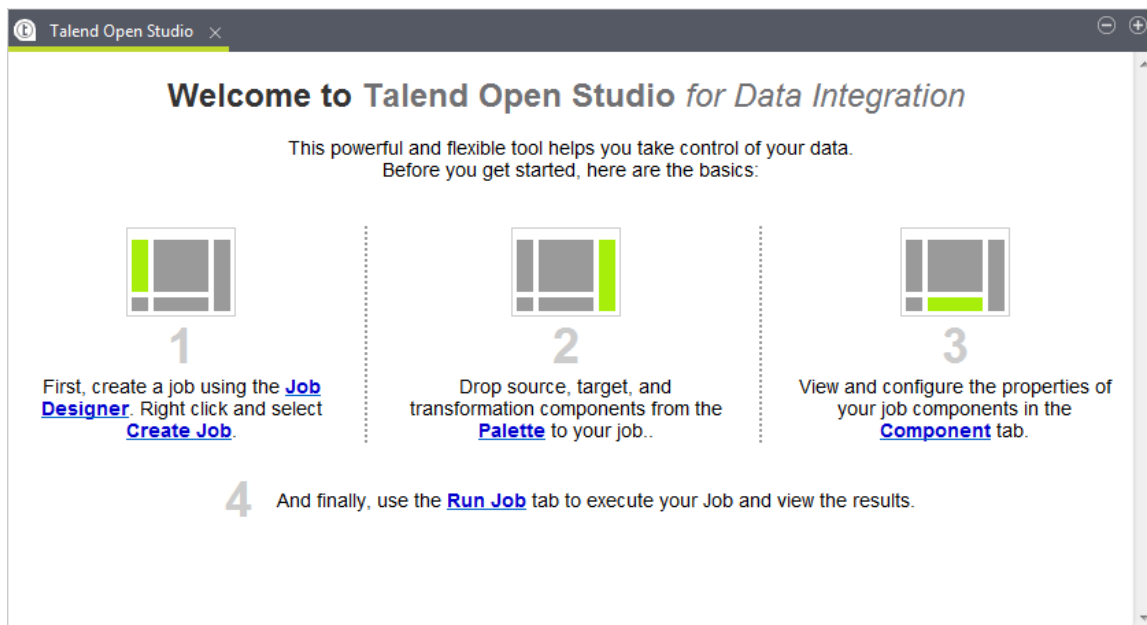
> Depending on you are using, the product information displayed in your Studio may differ slightly from what is shown above.

4. When the **[Additional Talend Packages]** wizard opens, install additional packages such as language packs if needed. For more information, see the section about installing additional packages in the *Talend Installation and Upgrade Guide*.

    You can skip this installation step and close the wizard by clicking **Cancel**.

    This wizard appears each time you launch the studio if any additional package is available for installation unless you select the **Do not show this again** check box. You can also display this wizard by selecting **Help** > **Install Additional Packages** from the menu bar.

1.1.2. How to connect to TalendForge

Every fourth time you launch *Talend Studio*, until you are connected to the **Talend** Community, the **[Connect to TalendForge[** dialog box opens, inviting you to connect to the **Talend** Community so that you can check, download, install external components and upload your own components to the **Talend** Community to share with other **Talend** users directly in the **Exchange** view of your Job designer in the Studio.

To learn more about the **Talend** Community, click the **TalendForge Terms of Use** link. For more information on using and sharing community components, see the section on how to download/upload **Talend** community components of your Studio User Guide.

If you want to connect to the **Talend** Community later, click **Skip this Step** to continue launching the Studio without setting up a connection to the **Talend** Community.

1. By default, the Studio will automatically collect product usage data and send the data periodically to servers hosted by **Talend** for product usage analysis and sharing purposes only. If you do not want the Studio to do so, clear the **I want to help to improve Talend by sharing anonymous usage statistics** check box.

    You can also turn on or off usage data collection from the **[Preferences]** dialog box (**Talend** > **Usage Data Collector**). For more information, see the section on setting *Talend Studio* preferences of your Studio User Guide.

2. Fill in the required information, select the **I Agree to the TalendForge Terms of Use** check box, and click **CREATE ACCOUNT** to create your account and connect to the **Talend** Community automatically and continue launching the Studio.

    Be assured that any personal information you may provide to **Talend** will never be transmitted to third parties nor used for any purpose other than joining and logging in to the **Talend** Community and being informed of **Talend** latest updates.
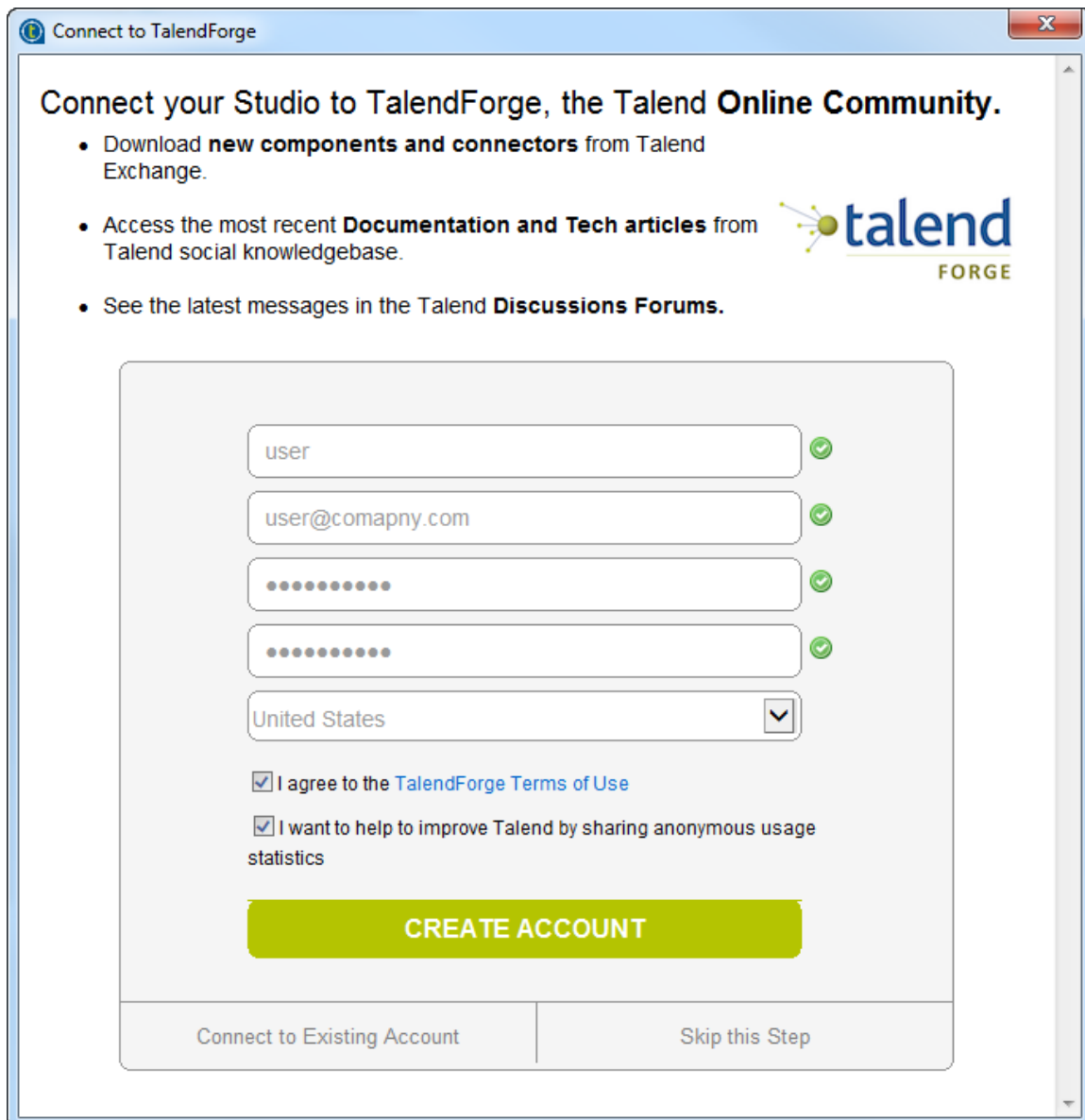
If you already have created an account at http://www.talendforge.org, click **Connect to Existing Account**, fill in your user name and password, and click **CONNECT TO MY ACCOUNT** to sign in the **Talend** Community and continue launching the Studio.

This page will not appear again when the Studio starts up once you successfully connect to the **Talend** Community. To show this page again, select **Talend** > **Exchange** from the **[Preferences]** dialog box, and click Sign In. For more information, see the section on setting *Talend Studio* preferences of your Studio User Guide.

1.2. Multi-perspective approach

Talend Studio offers a comprehensive set of tools and functions for all its key capabilities including data and application integration, data profiling and master data management. These tools are all accessible from different perspectives within the studio.

## 1.2.1. Switching between different perspectives

There are different ways to switch between different perspectives in the studio. They are as follows:

To switch between perspectives using quick access icons, do the following:

• Click the quick access icon in the top left corner of the studio to switch between the perspectives.

Alternatively, you may switch between perspectives using the menu bar:

1. On the menu bar, click **Window > Perspective**.
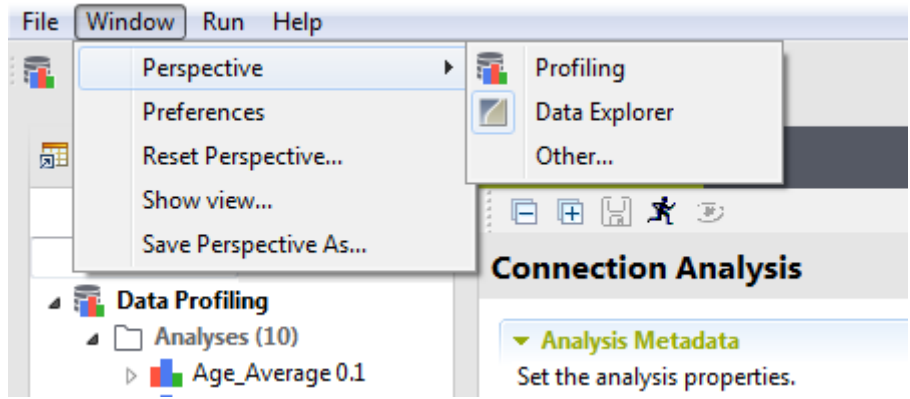
2. Select from the list:

| Item	to...
Profiling	open the data profiler perspective where you can examine data available in different data sources.
Data Explorer	open the data explorer perspective where you can browse and query analyzed data.
Other...	open a dialog box from which you can select to open different perspectives that extend the studio functionalities.

It is also possible, using the **Window - Show view...** combination, to show views from other perspectives in the open perspective.
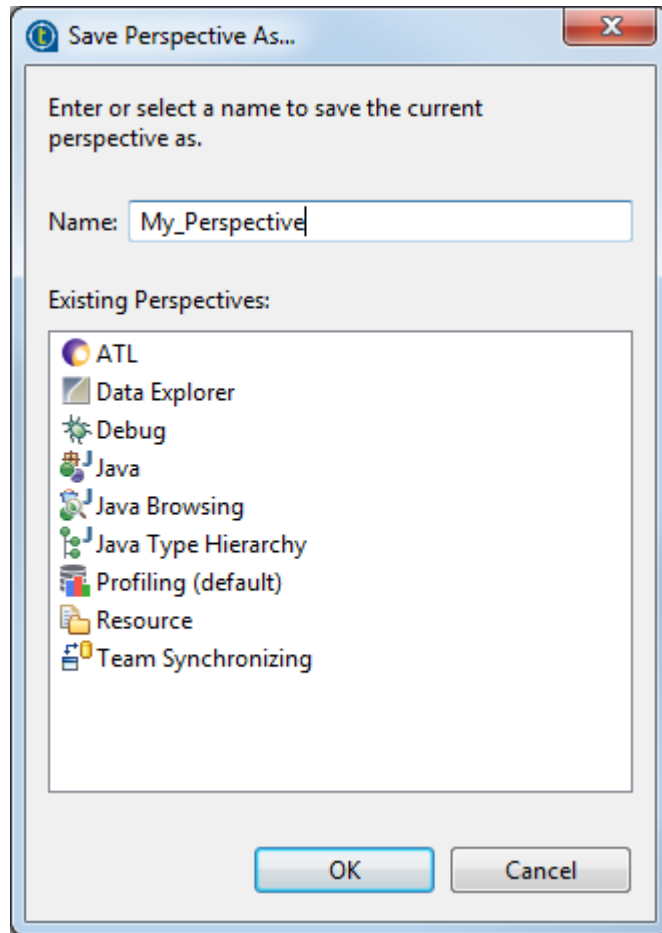
1.2.2. Saving the configuration of a perspective

You can save the configuration of your current perspective in order to list it as a new perspective in the perspective dialog box.

To save the configuration of the current perspective, do the following:

1. On the menu bar, click **Window > Save Perspective As....**
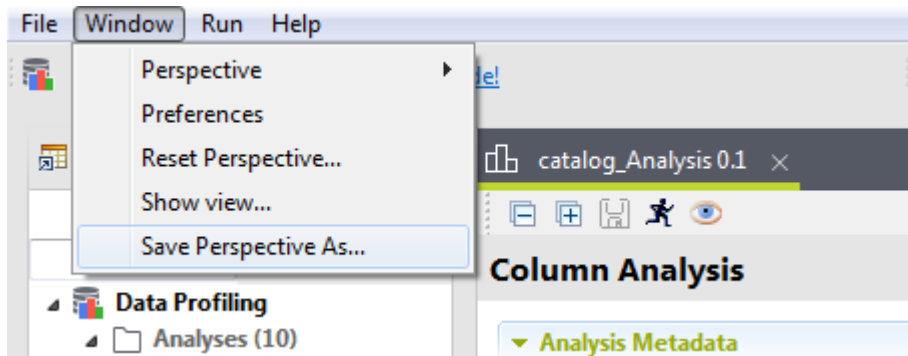
2. In the **Name** field, enter a name.

3. Click **OK** .

The current perspective is saved as a new perspective under the new name.

You can open this perspective any time by selecting it from the **[Open Perspective]** dialog box. For further information, see *Switching between different perspectives*.

Chapter 2. Profiling data

This chapter aims at users of *Talend Data Quality* who seek a real-life use case to help them take full control over data quality products.

It describes how to use the **Profiling** perspective in *Talend Studio* to profile data.

2.1. 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 sails 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.
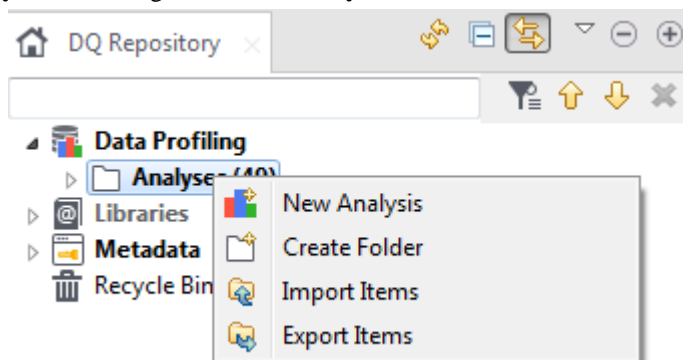
2.1.1. Identifying data anomalies

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.

2.1.1.1. How to profile address columns

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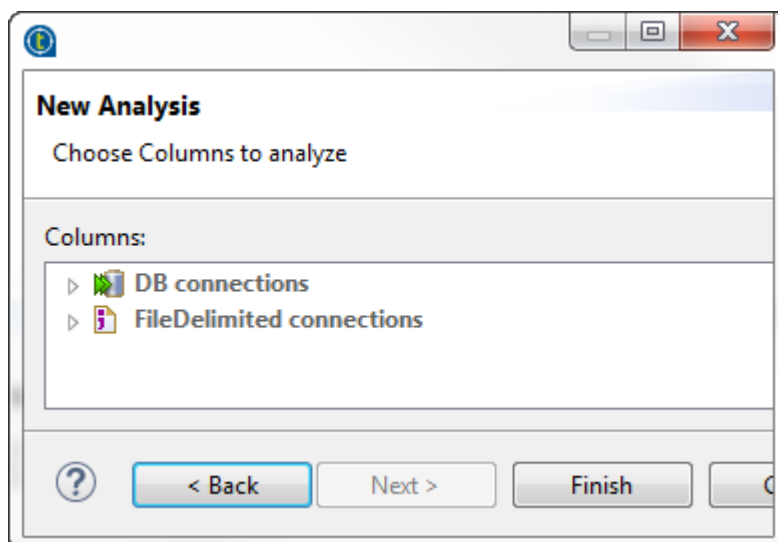
1.   In the **DQ Repository** tree view, right-click the **Analysis** folder and select **New Analysis**.

   The **[Create New Analysis]** wizard opens.

---

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

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

> 💡 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**.
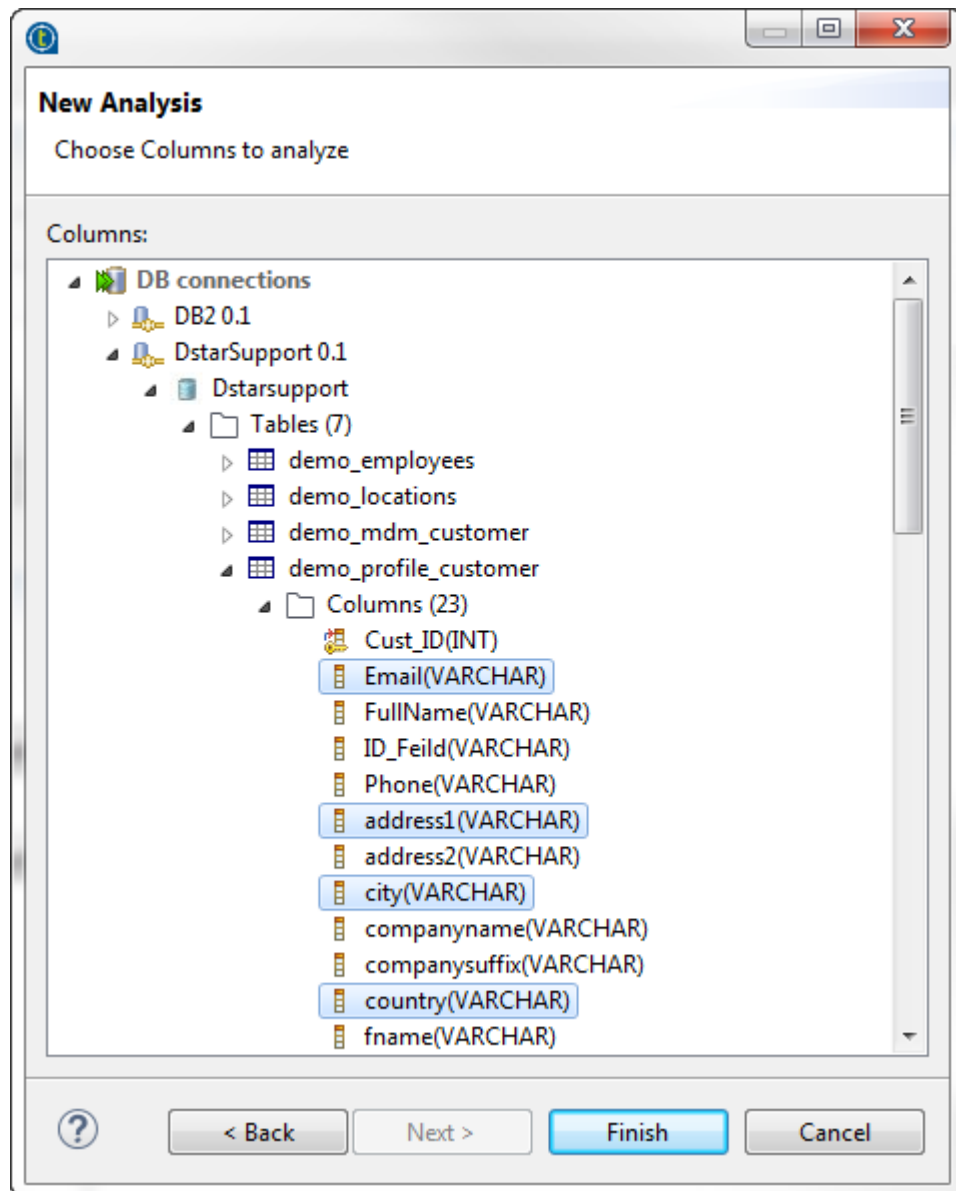

Selecting the address columns and setting sample data

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.

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

1. From the **Data preview** view in the analysis editor, click **Select indicators** to open the **[Indicator Selection]** dialog box.
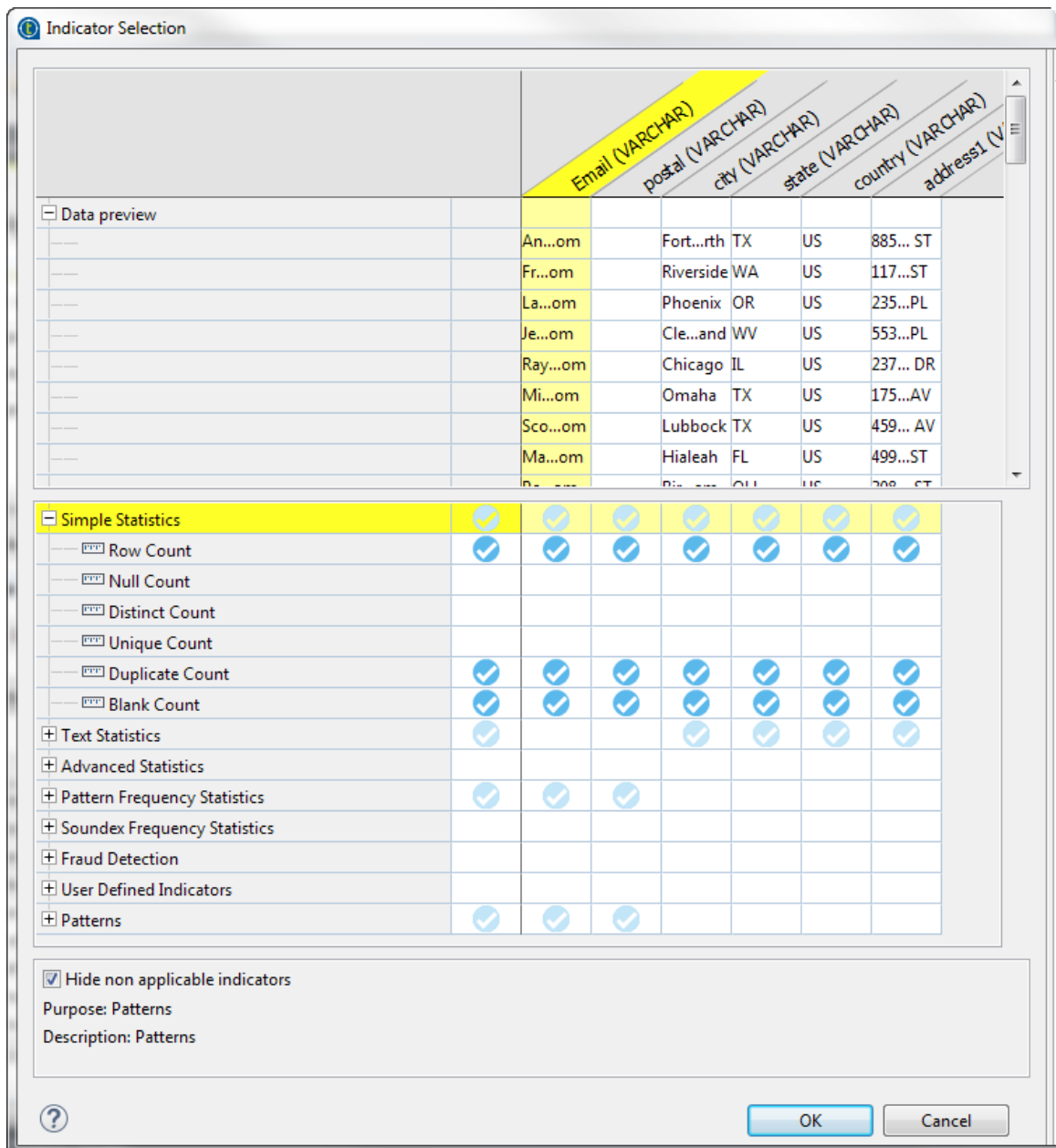
2. Click in the cells next to indicators names to set indicator parameters for the analyzed columns and click **OK**.

    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.

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.

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.

1.

In the **Analyzed Columns** view, click the icon next to *email*.
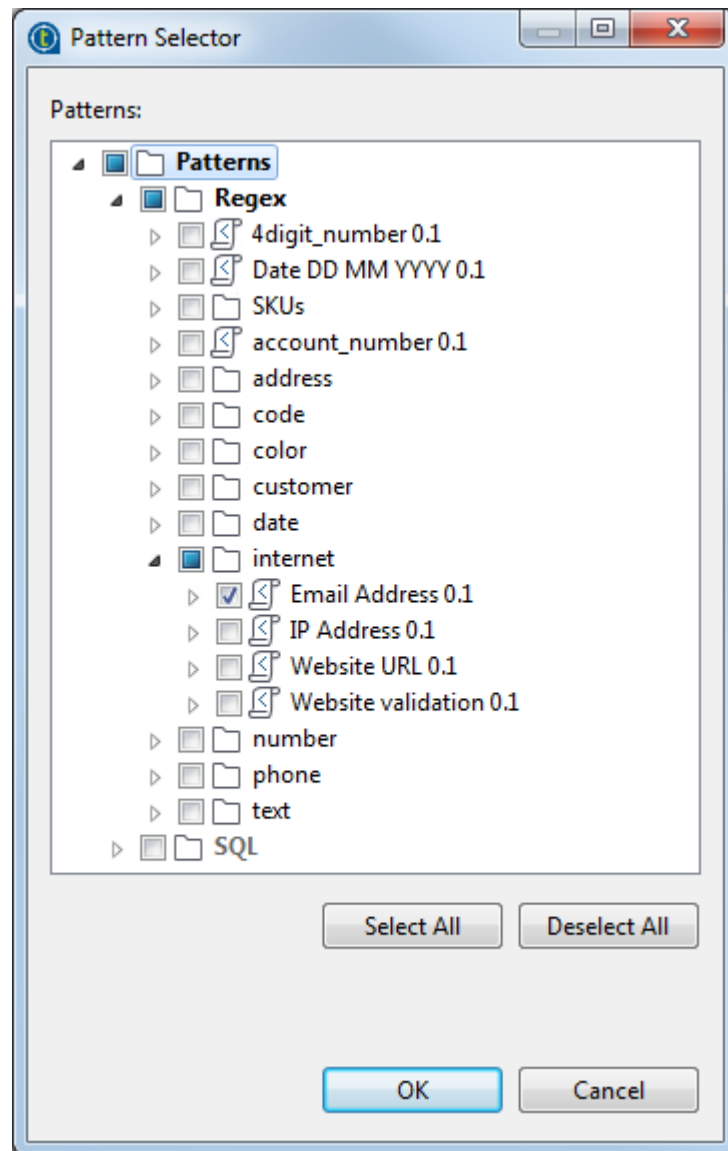
2. In the **[Pattern Selector]** dialog box, expand **Regex** and browse to **Email Address** in the **internet** folder, and then click **OK**.

3. Click the option icon [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.

Executing the analysis and displaying the profiling results

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.

The results for the *email* column look as the following:

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	94.34%	5.66%	9426.0	565.0

▼ **Simple Statistics**

Label	Count	%
Duplicate Count	108.00	1.08%
Blank Count	17.00	0.17%
Row Count	9991.00	100.00%
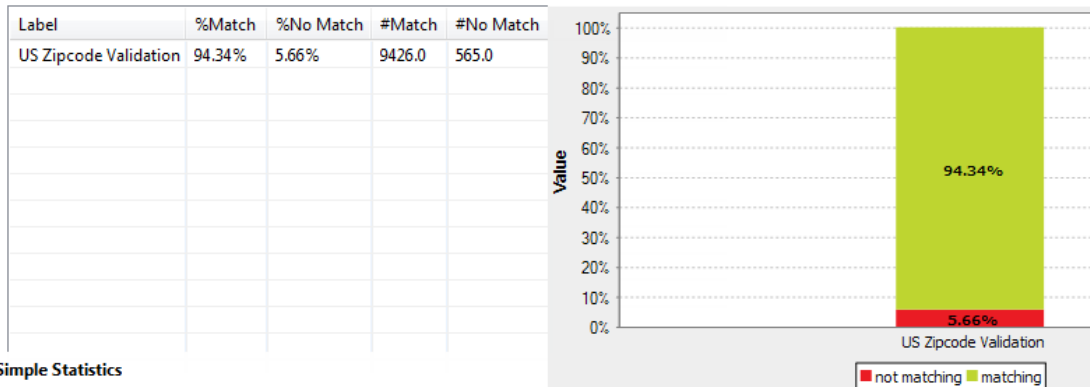
▼ **Pattern Frequency Statistics**

value	count	%
99999	9426.00	94.34%
9999	540.00	5.40%
Empty field	17.00	0.17%
99999-	7.00	0.07%
9999-9-	1.00	0.01%
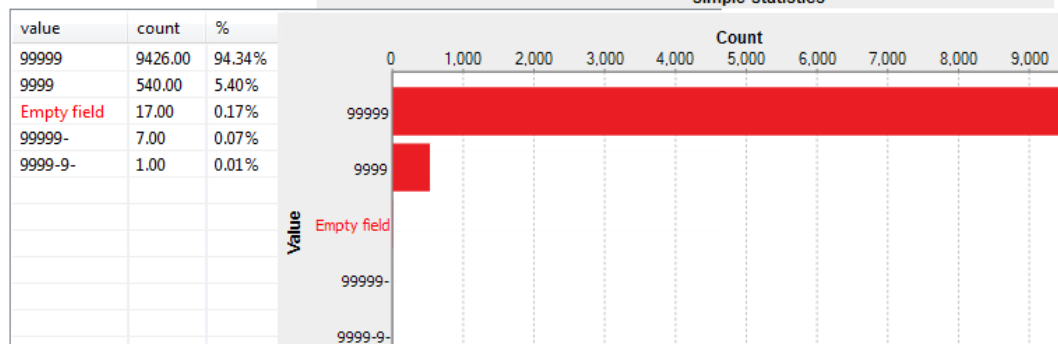
The result sets for the *postal* column give the count of the records that match and those that do not match a standard US zip code format. The results sets also give the blank and duplicate counts and the number of most frequent records for each distinct pattern. These results show that the data is not very consistent.

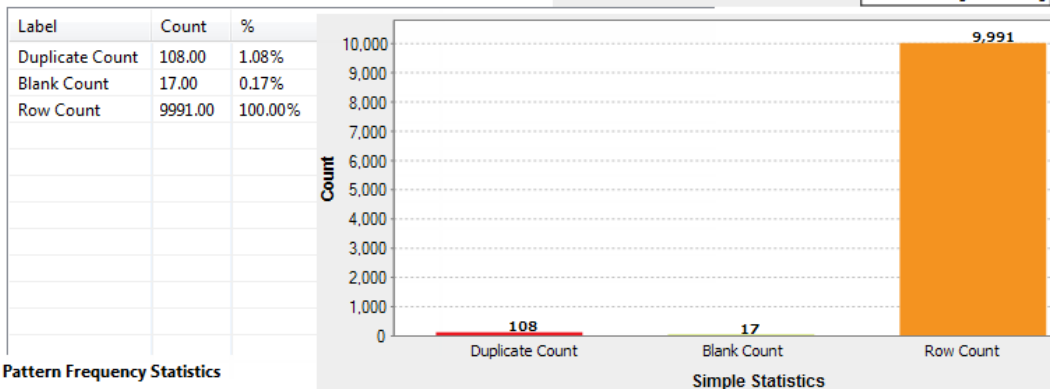Then some percentage of the customers can not be contacted by either email or US mail service. These results show clearly that your data is not very consistent and that it needs to be corrected.

2.1.1.2. 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.

To view and export the analyzed data, do the following:

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.

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