



Talend Open Studio for Data Quality

Getting Started Guide

6.2.2

Adapted for v6.2.2. Supersedes previous releases.

Publication date: January 23, 2017

Copyright

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, Paracel 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

Chapter 1. Introduction to Talend Open Studio for Data Quality	1
1.1. Functional architecture of Talend Data Quality solutions	2
Chapter 2. Prerequisites to using Talend products	3
2.1. Memory requirements	4
2.2. Software requirements	4
2.3. Installing Java	4
2.4. Setting up the Java environment variable on Windows	5
2.5. Setting up the Java environment variable on Linux	5
2.6. Installing 7-Zip (Windows)	5
Chapter 3. Downloading and installing Talend Open Studio for Data Quality	7
3.1. Downloading Talend Open Studio for Data Quality	8
3.2. Installing Talend Open Studio for Data Quality	8
3.2.1. Extracting via 7-Zip (Windows recommended)	8
3.2.2. Extracting via Windows default unzipping tool	8
3.2.3. Extracting via the Linux GUI unzipper	9
Chapter 4. Configuring and setting up your Talend product	11
4.1. Launching the Studio for the first time	12
4.2. Installing additional packages	12
Chapter 5. Profiling data	13
5.1. Setting up input data	14
5.2. Identifying anomalies in data	14
5.2.1. Defining a column analysis	14
5.2.2. Creating the database connection	15
5.2.3. Setting system indicators	18
5.2.4. Setting patterns	20
5.2.5. Showing analysis results	21
5.3. Browsing non-match data	23
5.4. What's next	24



Chapter 1. Introduction to Talend Open Studio for Data Quality

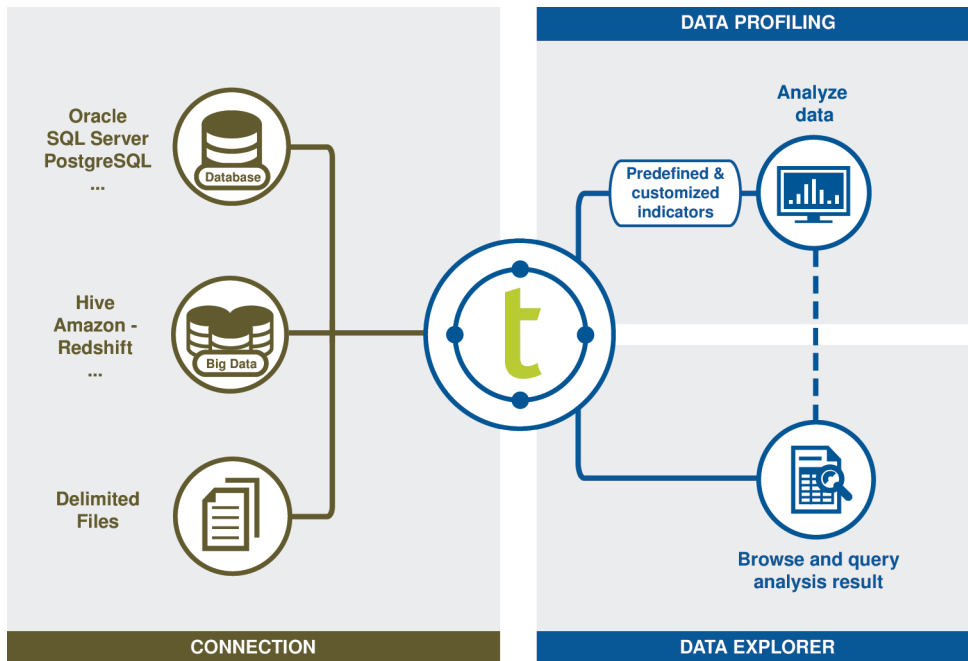
Talend Open Studio for Data Quality provides unified development and management tools to integrate and process all of your data with an easy to use, visual designer.

From the Studio, users can access and examine the data available in different data sources and collect statistics and information about this data.

1.1. Functional architecture of Talend Data Quality solutions

The Talend Open Studio for Data Quality functional architecture is an architectural model that identifies Talend Open Studio for Data Quality functions, interactions and corresponding IT needs. The overall architecture has been described by isolating specific functionalities in functional blocks.

The following chart illustrates the main architectural functional blocks.



The Studio comprises several main elements:

- A **Profiling** perspective where you can use predefined or customized patterns and indicators to analyze data stored in different data sources.
- A **Data Explorer** perspective where you can browse and query the results of the profiling analyses done on data.



Chapter 2. Prerequisites to using Talend products

This chapter provides basic software and hardware information required and recommended to get started with your *Talend* product:

- [Memory requirements.](#)
- [Software requirements.](#)

It also guides you to install and configure required and recommended third-party tools:

- [Installing Java.](#)
- [Setting up the Java environment variable on Windows](#) or [Setting up the Java environment variable on Linux.](#)
- [Installing 7-Zip \(Windows\).](#)

To successfully install the software, you need administrative access to your computer. To get administrative access, contact your Administrator.

2.1. Memory requirements

To make the most out of your *Talend* product, please consider the following memory and disk space usage:

Memory usage	3GB minimum, 4 GB recommended
Disk space	3GB

2.2. Software requirements

To make the most out of your *Talend* product, please consider the following system and software requirements:

Required software

- Operating System for Talend Studio:

Support type	Operating System	Version	Processor
Recommended	Microsoft Windows Professional	7	64-bit
Recommended	Linux Ubuntu	14.04	64-bit
Supported	Apple OS X	El Capitan/10.11	64-bit
		Yosemite/10.10	64-bit
		Mavericks/10.9	64-bit

- Java 8 JRE Oracle. See [Installing Java](#).

Optional software

- 7-Zip. See [Installing 7-Zip \(Windows\)](#).

2.3. Installing Java

To use your *Talend* product, you need Oracle Java Runtime Environment installed on your computer.

1. From the [Java SE Downloads](#) page, under **Java Platform, Standard Edition**, click the **JRE Download**.
2. From the **Java SE Runtime Environment 8 Downloads** page, click the radio button to **Accept License Agreement**.
3. Select the appropriate download for your Operating System.
4. Follow the Oracle installation steps to install Java.

When Java is installed on your computer, you need to set up the `JAVA_HOME` environment variable. For more information, see:

- [Setting up the Java environment variable on Windows](#).
- [Setting up the Java environment variable on Linux](#).

2.4. Setting up the Java environment variable on Windows

Prior to installing your *Talend* product, you have to set the `JAVA_HOME` and `Path` environment variables:

1. Go to the **Start Menu** of your computer, right-click on **Computer** and select **Properties**.
2. In the [Control Panel Home] window, click **Advanced system settings**.
3. In the [System Properties] window, click **Environment Variables...**
4. Under **System Variables**, click **New...** to create a variable. Name the variable `JAVA_HOME`, enter the path to the Java 8 JRE, and click **OK**.

Example of default JRE path: `C:\Program Files\Java\jre1.8.0_77`.

5. Under **System Variables**, select the **Path** variable and click **Edit...** to add the previously defined `JAVA_HOME` variable at the end of the `Path` environment variable, separated with semi colon.

Example: `<PathVariable>;%JAVA_HOME%\bin`.

2.5. Setting up the Java environment variable on Linux

Prior to installing your *Talend* product, you have to set the `JAVA_HOME` and `Path` environment variables:

1. Find the JRE installation home directory.

Example: `/usr/lib/jvm/jre1.8.0_65`

2. Export it in the `JAVA_HOME` environment variable.

Example:

```
export JAVA_HOME=/usr/lib/jvm/jre1.8.0_65
export PATH=$JAVA_HOME/bin:$PATH
```

3. Add these lines at the end of the user profiles in the `~/.profile` file or, as a superuser, at the end of the global profiles in the `/etc/profile` file.
4. Log on again.

2.6. Installing 7-Zip (Windows)

Talend recommends to install 7-Zip and to use it to extract the installation files: <http://www.7-zip.org/download.html>.

1. Download the 7-Zip installer corresponding to your Operating System.
2. Navigate to your local folder, locate and double-click the 7z exe file to install it.

The download will start automatically.



Chapter 3. Downloading and installing Talend Open Studio for Data Quality

Talend Open Studio for Data Quality is easy to install. After downloading it from *Talend's* Website, a simple unzipping will install it on your computer.

This chapter provides basic information useful to download and install it.

3.1. Downloading Talend Open Studio for Data Quality

Talend Open Studio for Data Quality is a free open source product that you can download directly from *Talend's* Website:

1. Go to *Talend Open Studio for Data Quality* [Download](#) page.
2. Click **DOWNLOAD FREE TOOL**.

The download will start automatically.

3.2. Installing Talend Open Studio for Data Quality

Installation is done by unzipping the TOS_DQ zip file previously downloaded.

This can be done either by using:

- 7Zip (Windows recommended): [Extracting via 7-Zip \(Windows recommended\)](#).
- Windows default unzipper: [Extracting via Windows default unzipping tool](#).
- Linux default unzipper (for a Linux based Operating System): [Extracting via the Linux GUI unzipper](#).

3.2.1. Extracting via 7-Zip (Windows recommended)

For Windows, *Talend* recommends you to install 7-Zip and use it to extract files. For more information, see [Installing 7-Zip \(Windows\)](#).

To install the studio, follow the steps below:

1. Navigate to your local folder, locate the **TOS** zip file and move it to another location with a path as short as possible and without any space character.

Example: *C:/Talend/*

2. Unzip it by right-clicking on the compressed file and selecting **7-Zip > Extract Here**.

3.2.2. Extracting via Windows default unzipping tool

If you do not want to use 7-Zip, you can use Windows default unzipping tool:

1. Unzip it by right-click the compressed file and select, **Extract All**.
2. Click on **Browse** and navigate to the *C: drive*.
3. Select **Make new folder** and name the folder *Talend*. Click **OK**.

4. Click on **Extract** to begin the installation.

3.2.3. Extracting via the Linux GUI unzipper

To install the studio, follow the steps below:

1. Navigate to your local folder, locate the **TOS** zip file and move it to another location with a path as short as possible and without any space character.

Example: *home/user/talend/*

2. Unzip it by right-clicking on the compressed file and selecting **Extract Here**.



Chapter 4. Configuring and setting up your Talend product

This chapter provides basic information required to configure and set up your *Talend* product, including:

- *Launching the Studio for the first time*
- *Installing additional packages*

4.1. Launching the Studio for the first time

The Studio installation directory contains binaries for several platforms including Mac OS X and Linux/Unix.

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

1. Double-click the executable file corresponding to your operating system, for example:
 - TOS_*-win-x86_64.exe, for Windows.
 - TOS_*-linux-gtk-x86_64, for Linux.
 - TOS_*-macosx-cocoa.app, for Mac.
2. In the **[User License Agreement]** dialog box that opens, read and accept the terms of the end user license agreement to proceed.
3. The *Talend Studio* opens briefly, then the **[Connect to TalendForge]** wizard opens. You can connect to it to benefit from the *Talend* community or **Skip this step**.

4.2. Installing additional packages

Talend recommends that you install additional packages, including third-party libraries and database drivers, as soon as you log in to your *Talend Studio* to allow you to fully benefit from the functionalities of the Studio.

1. When the **[Additional Talend Packages]** wizard opens, install additional packages by selecting the **Required** and **Optional third-party libraries** check boxes and clicking **Finish**.

This wizard opens 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.

For more information, see the section about installing additional packages in the *Talend Installation and Upgrade Guide*.

2. In the **[Download external modules]** window, click the **Accept all** button at the bottom of the wizard to accept all the licenses of the external modules used in the studio.

Depending on which libraries you selected, you may need to accept their license more than once.

Wait until all the libraries are installed before starting to use the studio.

3. If required, restart your *Talend Studio* for certain additional packages to take effect.



Chapter 5. Profiling data

This chapter takes the example of a company that provides movie rental and streaming video services, and shows how such a company could make use of Talend Open Studio for Data Quality.

You will work with data about your customers as you learn how to validate email addresses for customers and standardize phone numbers before sending them to the Customer Support System.

5.1. Setting up input data

The example in this document assumes that the customer data you want to profile is stored in a MySQL database. If you want to replicate the example and use the exact input data, you can download the *gettingstarted.sql* file of the customer data and then import it in a MySQL database.

Prerequisites:

- You have an access to a MySQL database.
- You have downloaded the *gettingstarted.sql* file by clicking [here](#) and stored it locally.

To import the customer data into a MySQL database:

1. Open the MySQL Workbench to launch an instance of the database.
2. From the menu bar, select **Server > Data Import** to open the import wizard wizard.
3. Select the **Import from Self-Contained File** option and browse to where you have stored the *gettingstarted.sql* file.
4. Select the schema to which you want to import the data, or click **New...** to define a new schema.
5. Click **Start Import** in the lower right corner.

The *gettingstarted* database is imported in the MySQL database.

5.2. Identifying anomalies in data

The use case explains how to use the **Profiling** perspective of the studio to analyze customer email addresses and phone numbers. It uses out-of-box indicators and patterns on the columns and shows the matching and non-matching address data.

You can then use the **Data Explorer** perspective to browse the non-matching data.

The sequence of profiling customer data involves the following steps:

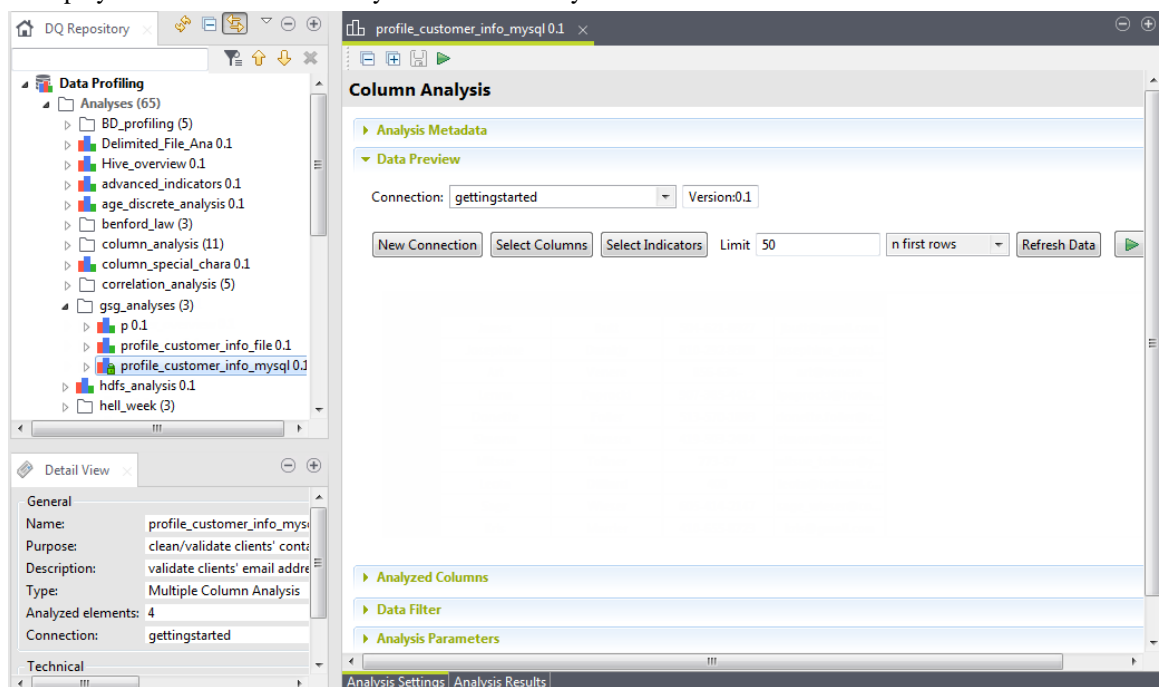
1. Create a column analysis on customer email addresses and phone numbers. For further information, see [Defining a column analysis](#).
2. Connect to the database which holds the customer data from the analysis editor. For further information, see [Creating the database connection](#).
3. Add indicators to provide simple statistics on data such as row , blank and duplicate counts. For further information, see [Setting system indicators](#).
4. Add standard patterns against which to match email addresses and phone numbers. For further information, see [Setting patterns](#).
5. Execute the analysis to show results in tables and charts. For further information, see [Showing analysis results](#).
6. Access a view of the analyzed data to see invalid records. For further information, see [Browsing non-match data](#).

5.2.1. Defining a column analysis

You want to create a column analysis from the **Profiling** perspective of the Studio to examine the *Email* and *Phone* columns in a MySQL databases and collect statistics on them. The analysis runs on several columns but each column is analyzed separately and independently.

1. In the **DQ Repository** tree view, right-click **Analyses** 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**.
3. In the **Name** field, enter a name for the analysis.
The **Name** field is mandatory. Do not use spaces or special characters in the analysis name.
4. Set a purpose and a description for the analysis, and click **Finish** to open the analysis editor.

The **Purpose** and **Description** fields are not mandatory, but you are advised to fill in this information which is displayed in **Detail View** when you select the analysis.



The new analysis is listed under the **Analysis** folder in the **DQ Repository** tree view.

5.2.2. Creating the database connection

Before you proceed to analyze customer data, stored in the MySQL database in this example, you must first set up the connection to the database.

Prerequisites:

- You have created a column analysis and opened the analysis editor. For further information, see [Defining a column analysis](#).
- You have imported the *gettingstarted* .sql file which holds the customer data into a MySQL database to which you have access. For further information, see [Setting up input data](#).

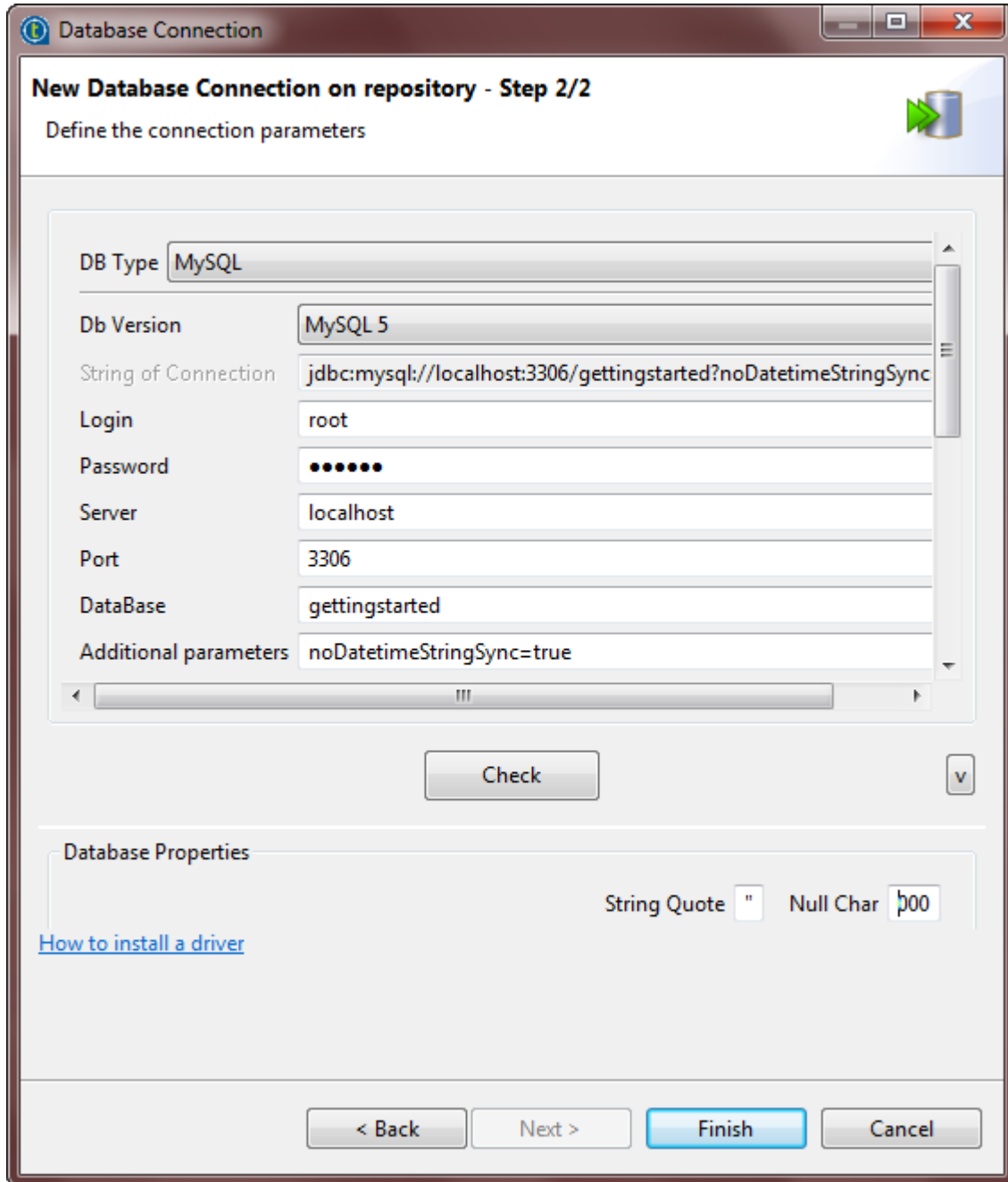
1. In the analysis editor, click the **New Connection** tab to open the **[Create New Connection]** wizard.
2. From the **Connection Type** list, select **DB connections** and click **Next**.
3. Enter a name for the database connection, *gettingstarted* in this example.

The **Name** field is mandatory. Do not use spaces or special characters in the connection name.

- Set a purpose and a description for the connection and click **Next**.

The **Purpose** and **Description** fields are not mandatory, but you are advised to fill in this information which is displayed in **Detail View** when you select the connection.

- From the **DB Type** and **DB Version** lists, select **MySQL** for this example and the database version respectively.

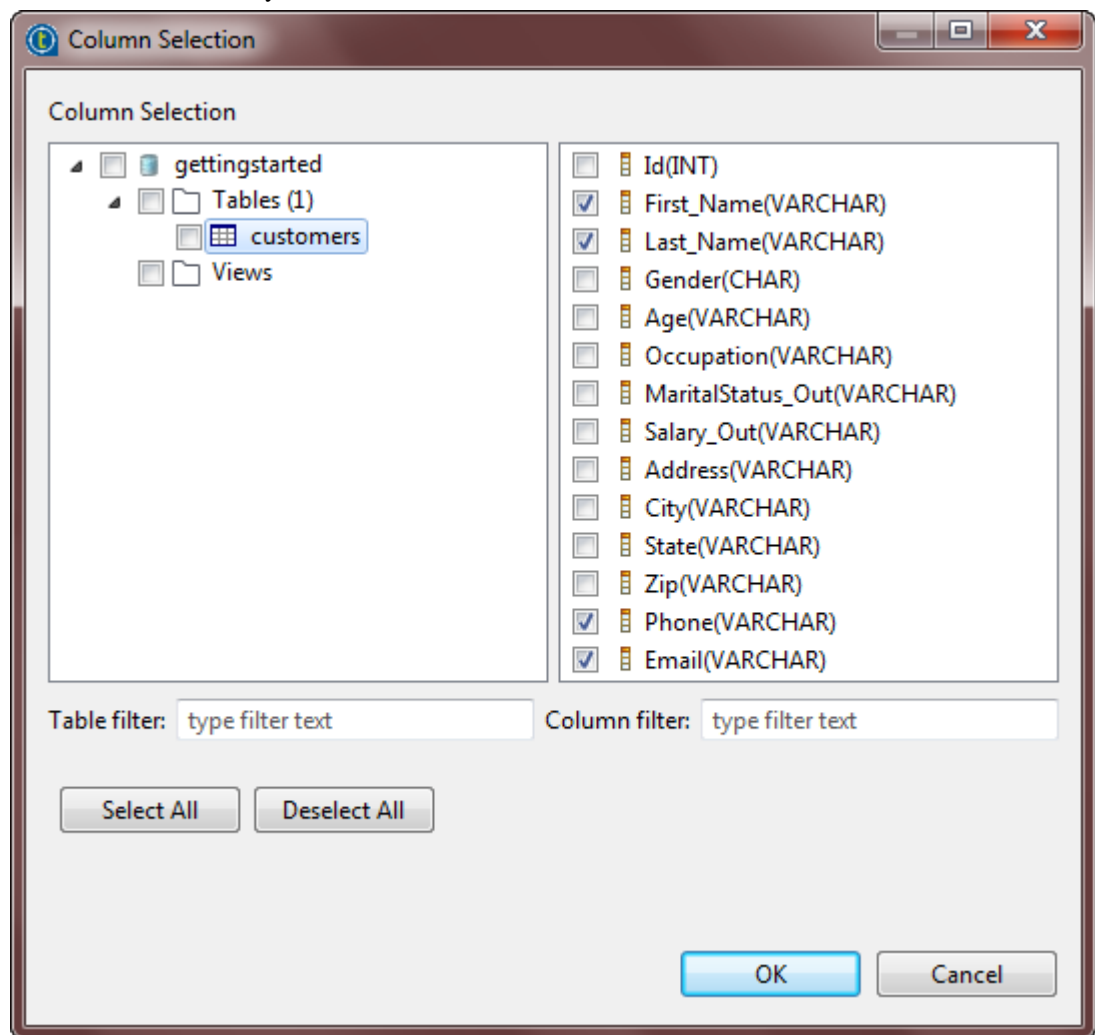


- Enter your login information and database detail in the corresponding fields and click the **Check** button to verify if your connection is successful.

If you have not already installed the database driver (.jar file) necessary to use the database, you will have a wizard prompting you to install the relative third-party module, click **Download and Install** and then close the wizard.

- Click **Finish** to create the database connection, list it under the **Metadata** node and open a new step in the wizard.

- Expand the database connection, click on the table name and select the check boxes of the columns on which you want to create the analysis.



- Click **OK** to close the wizard and list the columns in the analysis editor.

You can click **Refresh Data** to display the actual data in the analysis editor.

The screenshot shows the 'Column Analysis' tool interface. At the top, there are icons for home, add, save, and play. Below the title 'Column Analysis', there are sections for 'Analysis Metadata' and 'Data Preview'. The 'Data Preview' section shows a connection named 'gettingstarted' with version '0.1'. Below this are buttons for 'New Connection', 'Select Columns', 'Select Indicators', a 'Limit' field set to '50', a dropdown for 'n first rows', and a 'Refresh Data' button with a play icon. The main area displays a table with 10 rows and 4 columns: First_Name, Last_Name, Phone, and Email. Below the table are sections for 'Analyzed Columns', 'Data Filter', and 'Analysis Parameters'. At the bottom, there are tabs for 'Analysis Settings' and 'Analysis Results'.

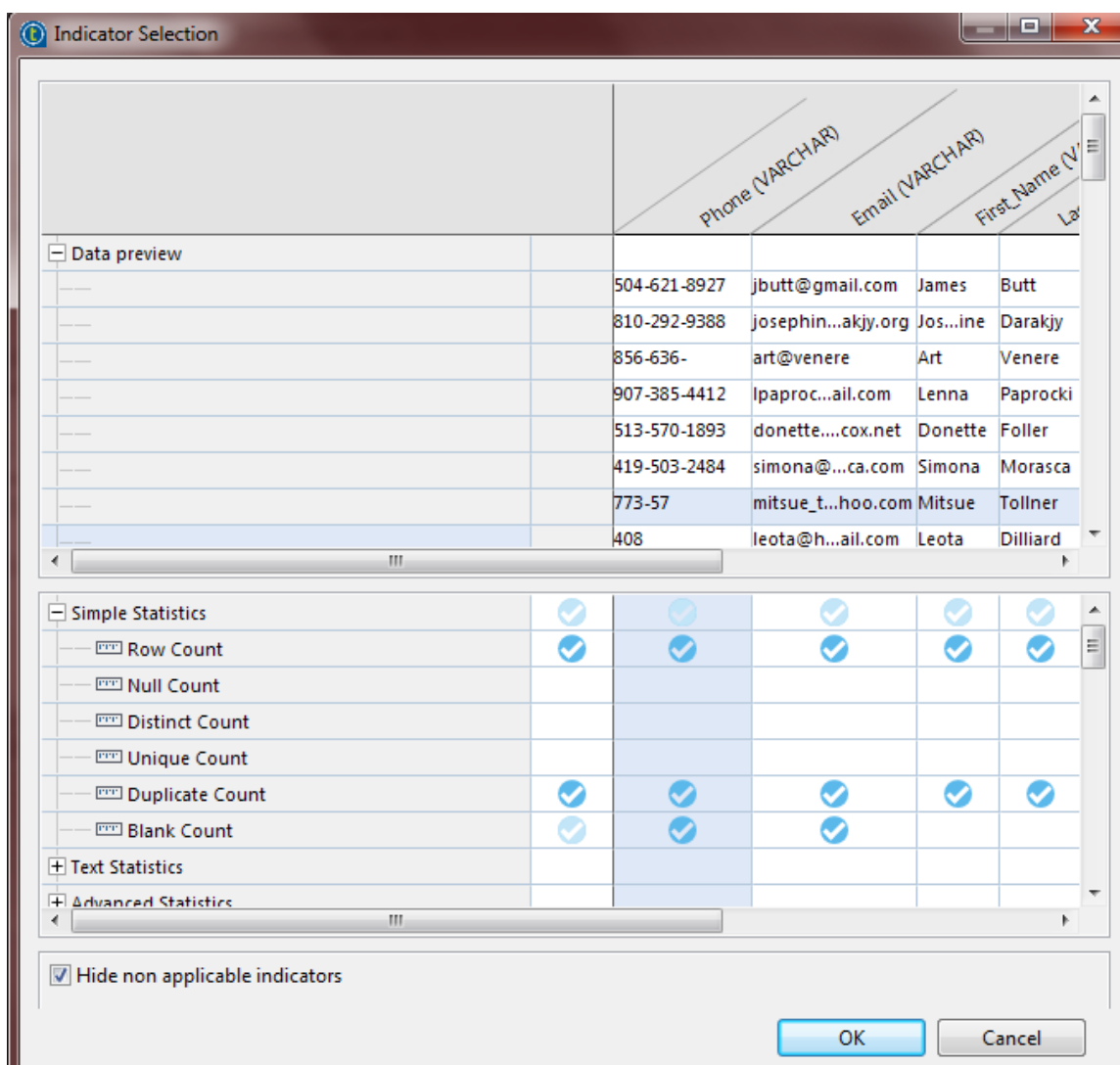
	First_Name	Last_Name	Phone	Email
1	James	Butt	504-621-8927	jbutt@gmail.com
2	Josephine	Darakjy	810-292-9388	josephine_darakj...
3	Art	Venere	856-636-	art@venere
4	Lenna	Paprocki	907-385-4412	lpaprocki@hotm...
5	Donette	Foller	513-570-1893	donette.foller@c...
6	Simona	Morasca	419-503-2484	simona@morasc...
7	Mitsue	Tollner	773-57	mitsue_tollner@y...
8	Leota	Dilliard	408	leota@hotmail.c...
9	Sage	Wieser	605-414-2147	sage_wieser@co...
10	Kris	Marrier	410-655-8723	kris@gmail.com

5.2.3. Setting system indicators

This column analysis uses out-of-box indicators to provide simple statistics such as row, blank and duplicate counts on the *Email* and *Phone* columns.

Prerequisites:

- You have created a column analysis and defined the connection to the database. For further information, see [Defining a column analysis](#) and [Creating the database connection](#) respectively.
1. In the **Data Preview** section in the analysis editor, click **Select indicators** to open the **[Indicator Selection]** dialog box.



- Expand **Simple Statistics** and select the row, duplicate and blank counts. Click **OK** to close the wizard.

You want to see the row, blank and duplicate counts in the *Email* and *Phone* columns to see how consistent the data is.

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

The screenshot shows the 'Column Analysis' window in Talend Studio. It features a tree view under 'Analyzed Columns' with sub-items for 'Phone (VARCHAR)' and 'Email (VARCHAR)'. Each sub-item has several indicators like 'Row Count', 'Duplicate Count', and 'Blank Count'. To the right, an 'Indicator settings' dialog box is open, showing the 'Indicator Thresholds' tab. The 'Upper threshold' field is set to 0.

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

Defining thresholds on the *Email* and *Phone* columns is very helpful as it will write in red the count of the duplicate and blank values in the analysis results.


For further information on indicator types and their usage when analyzing data, see *Talend Studio User Guide*.

5.2.4. Setting patterns

This column analysis uses predefined patterns to match the content of the *Email* and *Phone* columns against standard email and US phone patterns respectively. This defines the content, structure and quality of emails and phone numbers and give a percentage of the data that match the standard formats and the data that does not match.


Prerequisites:

- You have created a column analysis and defined the connection to the database. For further information, see [Defining a column analysis](#) and [Creating the database connection](#) respectively.

1. In the **Data Preview** section in the analysis editor, click the  icon next to the *Email* column to open the **[Pattern Selector]** dialog box.


Expand **Regex** > **internet**, select the **Email Address** check box and click **OK** to close the dialog box.

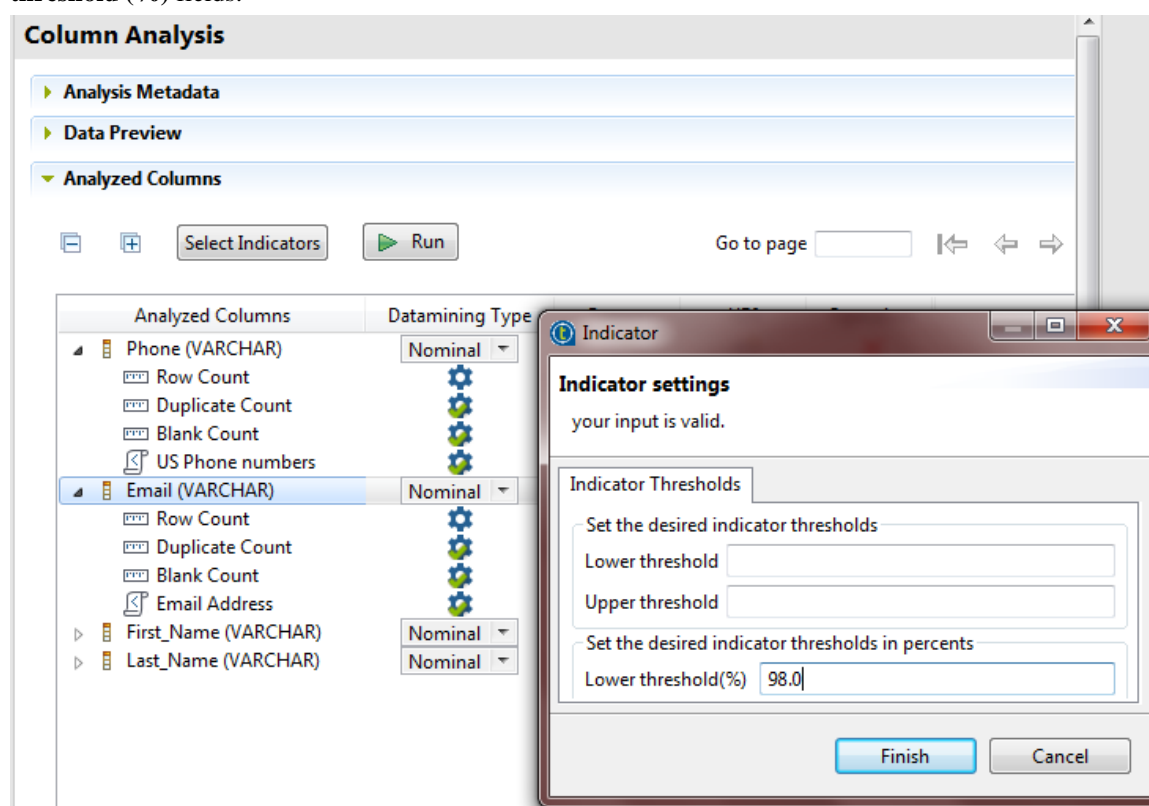
The pattern is added to the column in the **Analyzed Columns** section.

2. Click the  icon next to the *Phone* column to open the **[Pattern Selector]** dialog box.

Expand **Regex** > **phone**, select the **US phone numbers** check box and click **OK** to close the dialog box.

The pattern is added to the column in the **Analyzed Columns** section.

- Click the  icon next to the **Email Address** and **US phone numbers** patterns and set **98.0** in the **Lower threshold (%)** fields.



If the number of the records that match the patterns is fewer than 98%, it will be written in red in the analysis results.

For further information on pattern types and their usage when analyzing data, see *Talend Studio User Guide*.

5.2.5. Showing analysis results

Once you finalize creating the column analysis and setting the indicators and patterns, you can execute it and display analysis results in tables and charts.

Prerequisites:

- You have created a column analysis. For further information, see [Identifying anomalies in data](#).
- In the **Analysis Parameters**, select **java** from the **Execution engine** list to run the analysis with the Java engine.

For further information on the analysis execution engines, see *Talend Studio User Guide*.

- In the analysis editor, press **F6** to execute the analysis or click the **Run** button.

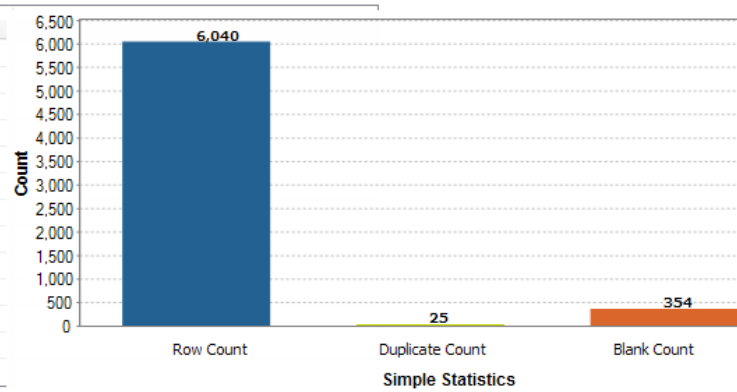
The editor switches to the **Analysis Results** view. The analysis results show the generated charts for the analyzed columns accompanied with tables that detail the statistic and pattern matching results.

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

▼ Column: customers.Email

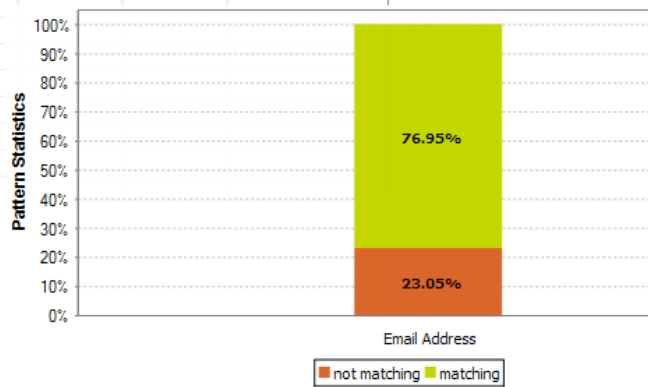
▼ Simple Statistics

Label	Count	%
Row Count	6040.00	100.00%
Duplicate Count	25.00	0.41%
Blank Count	354.00	5.86%



▼ Pattern Matching

Label	%Match	%No Match	#Match	#No Match
Email Address	76.95%	23.05%	4648.0	1392.0



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

Id	First_Name	Last_Name	Gender	Age	Occupation	MaritalSt...	Salary_Out	Address	City	State	Zip	Phone	Email
31	Ammie	Corrio	M	56+	Executive/Ma...	Divorced	> 200,000	74874 A...	Colum...	OH	43215	614-80...	
76	Moon	Parlato	M	35-44	Executive/Ma...	Divorced	100,000-149...	74989 B...	Wellsville	NY	14895	585-86...	
105	Rosio	Cork	M	45-49	Programmer	Single	> 200,000	4 10th S...	High P...	NC	27263	336-24...	
116	Glory	Kulzer	M	25-34	Technical/En...	Single	> 200,000	55892 J...	Owings...	MD	21117	410-22...	
152	Carissa	Batman	M	18-24	College/Grad...	Divorced	> 200,000	12270 C...	Eugene	OR	97401	541-32...	
157	Yolando	Luczki	M	35-44	Self-Employed	Single	100,000-149...	422 E 2...	Syracuse	NY	13214	315-30...	
170	Clay	Hoa	M	25-34	Lawyer	Married	100,000-149...	73 Saint...	Reno	NV	89502	775-50...	
192	Lemuel	Latzke	M	18-24	Academic/Ed...	Divorced	100,000-149...	70 Eucli...	Bohemia	NY	11716	631-74...	
193	Melodie	Knipp	F	45-49	Scientist	Married	100,000-149...	326 E M...	Thousa...	CA	91362	805-69...	
230	Ressie	Auffrey	M	45-49	Academic/Ed...	Single	100,000-149...	23 Palo ...	Miami	FL	33134	305-60...	
292	Filiberto	Tawil	M	35-44	Executive/Ma...	Married	100,000-149...	3882 W ...	Los An...	CA	90016	323-76...	

- In the **Analysis Results** view, right-click the result in the **Pattern Matching** of the *Email* column and select **View invalid values** for example.

A view opens listing all the invalid email addresses.

Email
alaine_bergesencox.net
allene_iturbide@cox
andra@gmail
arlene_klusman@gmail
art@venere
asergi@gmail
beatriz
beckie.silvestrini

5.4. What's next

You have learned how *Talend Studio* helps you profile your data and collect statistics and information about it in order to assess the quality level of the data according to defined set goals.

You have seen:

- How to use the **Profiling** perspective of the studio to analyze customer email addresses and phone numbers by using out-of-box indicators and patterns.
- How the analysis results show the matching and non-matching address records and how it is possible to browse such data.

Once you succeed with the simple procedures outlined in *Identifying anomalies in data*, you can start digging deeper to see in detail all the profiling capabilities of *Talend Studio*.

For further information about *Talend Studio*, see *Talend Studio User Guide* and *Talend Open Studio Components Reference Guide*.

To learn more about *Talend* products and solutions, visit www.talend.com.