



Talend Open Profiler 3.X

User Guide

Version 3.2_a

Adapted for **Talend Open Profiler** v3.2.x. Supersedes previous User Guide releases.

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/>

Talend Open Profiler

User Guide i

Preface	vii
Purpose	vii
Audience	vii
Typographical conventions	vii
History of changes	vii
Feedback and Support	viii

CHAPTER 1

Overview of data profiling management 1

1.1 Main concepts	2
1.1.1 The problems data profiling addresses	2
1.1.2 How does data profiling promote better data quality	2
1.2 About Talend Open Profiler	2
1.2.1 What is Talend Open Profiler	2
1.2.2 A simple-to-use data profiler	3
1.2.3 Core features of Talend Open Profiler	3
Metadata repository	3
Patterns	3
Indicators	4
Simple statistics	4
Text statistics	4
Summary statistics	5
Advanced statistics	5
Pattern frequency statistics	5
Soundex frequency statistics	5

CHAPTER 2

Data profiling management procedures 7

2.1 Setting up the display parameters of all editors	8
2.2 Managing database connections	9
2.2.1 How to create a new database connection	9
2.2.2 How to filter tables/views in a database connection	12
2.2.3 How to open a database connection	13
2.2.4 How to delete a database connection	14
2.3 Comparing tree-view metadata structures with database structures	15
2.3.1 How to compare catalog and schema lists	15
2.3.2 How to compare table lists	17
2.3.3 How to compare column lists	18
2.4 Synchronizing tree-view metadata structures with database structures	19
2.4.1 How to synchronize catalog and schema lists	19
2.4.2 How to synchronize table lists	20
2.4.3 How to synchronize column lists	21
2.5 Managing database content analyses	21
2.5.1 How to create a database content analysis	21
2.5.2 How to create a database content analysis directly from the DB connection	25
2.5.3 How to create a catalog analysis	26
2.5.4 How to create a schema analysis	30
2.6 Managing the analysis of a set of columns	33

2.6.1 How to analyze a set of columns	33
2.6.2 How to analyze a set of columns in shortcut procedures	41
2.6.3 Data mining types	41
Nominal	42
Interval	42
Unstructured text	42
Other	42
2.7 Creating column comparison analysis	42
2.8 Managing column correlation analysis	46
2.8.1 How to create numerical correlation analysis	46
2.8.2 How to create time correlation analysis	51
2.8.3 How to create nominal correlation analysis	55
2.9 Managing table analyses	60
2.9.1 How to create a table analysis with DQ rules	60
2.9.2 How to create a column functional dependency analysis	65
2.10 Adding a task to an item	69
2.10.1 How to add a task to a column in a database connection	70
2.10.2 How to add a task to an item in a specific analysis context	71
2.10.3 How to add a task to an indicator in a column analysis	72
2.10.4 How to delete a completed task	73
2.11 Generic procedures for all types of analyses	75
2.11.1 How to open an analysis	75
2.11.2 How to delete an analysis	76
2.11.3 How to execute an analysis	76
2.11.4 How to duplicate an analysis	77
2.11.5 How to add a task to an analysis	77

CHAPTER 3

Advanced analysis procedures	79
3.1 Managing data quality rules	80
3.1.1 How to create a DQ rule	80
3.1.2 How to open a DQ rule	82
3.2 Managing patterns	83
3.2.1 How to declare a regular expression in a specific database	83
3.2.2 How to edit or delete the User-Defined Function	84
3.2.3 How to create a new pattern	85
3.2.4 How to add patterns to analyzed columns	88
3.2.5 How to analyze a set of columns with pattern indicators	89
3.2.6 How to edit a pattern in the analyzed column	90
3.2.7 How to edit a pattern	92
3.2.8 How to delete a pattern	93
How to delete a pattern from the analyzed column:	93
How to delete a pattern from the DQ Repository	93
3.2.9 How to duplicate a pattern	94
3.2.10 How to import patterns from a csv file	94
3.2.11 How to import patterns from Talend Exchange	96
3.2.12 How to export patterns	97
3.2.13 How to export patterns to Talend Exchange	99
3.3 Managing indicators	102
3.3.1 How to create a user-defined indicator	102
3.3.2 How to edit the definition of an indicator	104
3.3.3 How to duplicate an indicator	106
3.3.4 How to export user-defined indicators to a csv file	106
3.3.5 How to export user-defined indicators to Talend Exchange	108
3.3.6 How to export system indicator to a definition file	108
3.3.7 How to import user-defined indicators from a csv file	109

3.3.8 How to import system indicators from a definition file	110
3.3.9 How to import indicators from Talend Exchange	111
3.3.10 How to set indicators for the columns to analyze	112
3.3.11 How to set options for indicators	113
3.3.12 Indicators parameters	114

APPENDIX A

Talend Open Profiler management GUI	117
A.1 Main window of Talend Open Profiler	118
A.2 Menu bar of Talend Open Profiler	119
A.3 Toolbar of Talend Open Profiler	119
A.4 Tree view of Talend Open Profiler	120
A.5 Detailed View of Talend Open Profiler	120
A.6 Design workspace of Talend Open Profiler	121
A.7 Tab panel of the column analysis editor	121
A.8 How to select a task from Talend Open Profiler management GUI	123
A.9 Cheat Sheets of Talend Open Profiler	124

APPENDIX B

Data Explorer management GUI	125
B.1 Main window of the Data Explorer	126
B.2 Menu bar of the Data Explorer	126
B.3 Toolbar of the Data Explorer	127
B.4 Connections view	127
B.5 SQL History view	127
B.6 SQL Editor view	128
B.7 Database Structure view	129
B.8 Database Detail view	130



Preface

Purpose

This User Guide explains how to manage **Talend Open Profiler** functions in a normal operational context.

Information presented in this document applies to **Talend Open Profiler** releases beginning with **3.2.x**.

Audience



This guide is for business users, database administrators and data analysts in charge of checking the quality of data and collecting statistics and information about that data.



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

Typographical conventions

This guide uses the following typographical conventions:

- text in **bold**: window and wizard buttons and fields, keyboard keys, menus and menu options,
- text in **[bold]**: window, wizard and dialog box titles,
- text in `courier`: system parameters selected by the user,
- 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 need to be aware of or pay special attention to.

History of changes

The below table lists the changes made in this release of the **Talend Open Profiler** User Guide.

Version	Date	History of Change
v3.1_a	27/04/2009	Updates in Talend Open Profiler User Guide include: -importing patterns from Talend Exchange -exporting patterns to Talend Exchange -new indicators -new analysis types -DQRules
v3.1_b	00/00/2009	No updates. Documentation not published.

Version	Date	History of Change
v3.1_c	27/04/2009	Updates concerns only Talend Data Quality
v3.1_d	27/07/2009	<p>Updates in Talend Open Profiler User Guide include:</p> <ul style="list-style-type: none"> -Updating all the sections in Appendix A that talk about TDQ graphical interface to add the correct figures and text after the integration with Talend Integration Suite Studio. -Replacing many captures with updated ones to synchronize the figures in the User Guide with the GUI. -Adding few new sections in different chapters: Declaring regular patterns function for some databases, Adding a task to an Indicator, Editing an Indicator, How to delete a completed task, Setting up the display of connection and analyses editors and Data mining types. -Deleting all sections talking about "How to add a task..." in different chapters and replaced them with one general section "Adding a task to an item" in chapter 2.
v3.2_a	20/10/2009	<p>Updates in Talend Open Profiler User Guide include:</p> <ul style="list-style-type: none"> -Added "default value count" to Simple Statistics. -Added a new "data type mining" section. -Added few new sections in different chapters, for example How to import and export indicators from and to Talend Exchange, How to duplicate indicators, How to create user-defined indicators and How to generate a Job to alert for threshold violation etc. -Changed many screen captures in different chapters to show the new or modified options and correcting the related text accordingly.

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

Overview of data profiling management

This chapter introduces data profiling and provides the basics for managing data profiling from [Talend Open Profiler](#).

1.1 Main concepts

Beginning a data-driven initiative in your enterprise without first understanding the enterprise data will lead to great losses. Data improvement efforts must start with an understanding of the integrity of the data of the enterprise.

1.1.1 The problems data profiling addresses

Data profiling is the process of examining the data available in existing data sources (for example, databases or applications) and collecting statistics and information about this data. Data profiling helps to assess the quality level of the data contained in the information system according to defined set goals.

If data is of a poor quality, or managed in structures that cannot be integrated to meet the needs of the enterprise, business processes and decision-making suffer.

Traditional approaches to data analysis cannot answer all the questions that need to be asked. It is not always easy to quantify the exact cost of having projects and business processes undermined by data quality issues, but it is accepted that they impact profitability across an enterprise.

A clear, up-front picture of all the potential issues is essential to plan projects effectively. Data analyzing, data cleansing and data transformation requirements must be understood before timescales and costs are finalized, and not after.

The ability to reduce and eliminate these overheads justifies the adoption of a Data Profiling technology.

1.1.2 How does data profiling promote better data quality

The first step in improving the quality of data is to “profile” or evaluate that data. Data profiling helps you understand the data you manage and the rules that govern that data. Without this knowledge, no effective data management plan can be formulated.

Compared to manual analysis techniques, data profiling technology improves the enterprise’s ability to meet the challenge of managing data quality and to address the data quality challenges faced during data migrations and data integrations.

1.2 About Talend Open Profiler

This section introduces **Talend Open Profiler** and lists its key features.

1.2.1 What is Talend Open Profiler

Talend Open Profiler is a data profiling tool that defines the content, structure and quality of highly complex data structures. It analyzes data on an ongoing basis, and analyzes changes to source data over time to improve its quality.

Talend Open Profiler helps you discover and understand the quality of your data. With **Talend Open Profiler**, you can carry out accurate data profiling processes and thus reduce the time and

resources you need to find problematic data. This data profiling tool allows you to identify potential problems before beginning data-intensive projects such as data integration.

Its comprehensive data profiling features will help you enhance and accelerate your data analysis projects.

1.2.2 A simple-to-use data profiler

Talend Open Profiler is a sophisticated yet simple-to-use and easy to implement data profiler. It centralizes a:

- data profiler, for more information about the data profiler, see *Talend Open Profiler management GUI on page 117*.
- data automated explorer, for more information about the data explorer, see *Data Explorer management GUI on page 125*.
- pattern manager, for more information about the pattern manager, see *Patterns on page 3* and *Indicators on page 4*.
- metadata manager, for more information about the metadata manager, see *Metadata repository on page 3*.

For more information about the pattern and metadata managers, see *Core features of Talend Open Profiler on page 3*.

1.2.3 Core features of Talend Open Profiler

This section describes the basic features of **Talend Open Profiler**.

Metadata repository

Talend Open Profiler connects to databases to analyze their structure (catalogs, schemas and tables), and stores the description of their metadata in its metadata repository. You can then use this metadata to set up matrices and indicators.

For more information, see *Managing database connections on page 9* and *Managing database content analyses on page 21*.

Patterns

Patterns are sets of strings against which you can define the content, structure and quality of high complex data.

Talend Open Profiler lists two types of patterns under the **Patterns** folder in the **DQ Repository** tree view area. The first is a list of predefined regular patterns (regular expressions), and the second is SQL patterns, which are the patterns you add using a `LIKE` clause.

With **Talend Open Profiler**, you can carry out column analyses using the above mentioned patterns. These pattern-based analyses illustrate the frequencies of various data patterns found in the values of the analyzed columns.

For more information, see *Managing the analysis of a set of columns on page 33*.

Talend Open Profiler makes it possible for you as well to create your own regular expressions to use it later in column analyses.

For more information about patterns, see *Managing patterns on page 83*.

In **Talend Open Profiler**, you can generate graphs for any of the regular expressions. You can also display in the **Analysis Results** view tables that write in words the generated graphs. From those graphs and analysis results you can easily determine the percentage of invalid values based on the listed patterns.

For more information, see *Tab panel of the column analysis editor on page 121*.

Indicators

Indicators are the results achieved through the implementation of different patterns. They can represent the results of data matching and different other operations.

With **Talend Open Profiler**, you can define a set of indicators on columns of database tables that need to be analyzed or monitored. These indicators can range from simple or advanced statistics to text strings analysis, including summary data and statistical distributions of records.

The below sections describe the indicators you can set with **Talend Open Profiler**.

For more information about how to set indicators for columns in tables, see *Managing indicators on page 102*.

Simple statistics

They provide simple statistics on the number of records falling in certain categories, including the number of rows, the number of null values, the number of distinct and unique values, the number of duplicates, or the number of blank fields.

- **Distinct count:** counts the number of distinct values of your column.
- **Unique count:** counts the number of distinct values with only one occurrence. It is necessarily less or equal to Distinct counts.
- **Duplicate count:** counts the number of values appearing more than once. You have the relation: Duplicate count + Unique count = Distinct count. For example, a,a,a,b,b,c,d,e => 9 values, 5 distinct values, 3 unique values, 2 duplicate values.
- **Row count:** counts the number of rows.
- **Null count:** counts the number of null rows.
- **Blank count:** counts the number of blank rows. A “blank” is a non null textual data that contains only white space. Note that Oracle does not distinguish between the empty string and the null value.
- **Default value count:** counts the number of default values.

Text statistics

They analyze the characteristics of textual fields in the columns, including minimum, maximum and average length.

- **Min length:** computes the minimal length of a text field.
- **Max length:** computes the maximal length of a text field.

- Average length: computes the average length of a field.

You can set parameters for any of the above three indicators to avoid counting blank or null data.

Summary statistics

They perform statistical analyses on numeric data, including the computation of location measures such as the mean and the average, the computation of statistical dispersions such as the inter quartile range and the range.

- Mean: computes the average of the records.
- Median: computes the value separating the higher half of a sample, a population, or a probability distribution from the lower half.
- Inter quartile range: computes the difference between the third and first quartiles.
- Range: computes the difference between the highest and lowest records.

Advanced statistics

They determine the most probable and the most frequent values and builds frequency tables.

- Mode: computes the most probable value. For numerical data or continuous data, you can set bins in the parameters of this indicator. It is different from the “average” and the “median”. It is good for addressing categorical attributes.
- Frequency table: computes the number of most frequent records for each distinct record.
- Low frequency table: computes the number of less frequent records for each distinct record.

Pattern frequency statistics

Indicators in this group determine the most and less frequent patterns.

- Pattern frequency table: computes the number of most frequent records for each distinct pattern.
- Pattern low frequency table: computes the number of less frequent records for each distinct pattern.

Soundex frequency statistics

Indicators in this group use the Soundex algorithm built in the DBMS.

They index records by sounds. This way, records with the same pronunciation (only English pronunciation) are encoded to the same representation so that they can be matched despite minor differences in spelling.

- Soundex frequency table: computes the number of most frequent distinct records relative to the total number of records having the same pronunciation.
- Soundex low frequency table: computes the number of less frequent distinct records relative to the total number of records having the same pronunciation.



CHAPTER 2

Data profiling management procedures

This chapter provides the information you need to perform data profiling management GUI procedures.

Before starting data profiling management procedures, you need to be familiar with **Talend Open Profiler** Graphical User Interface (GUI). For more information, see *Talend Open Profiler management GUI on page 117*.

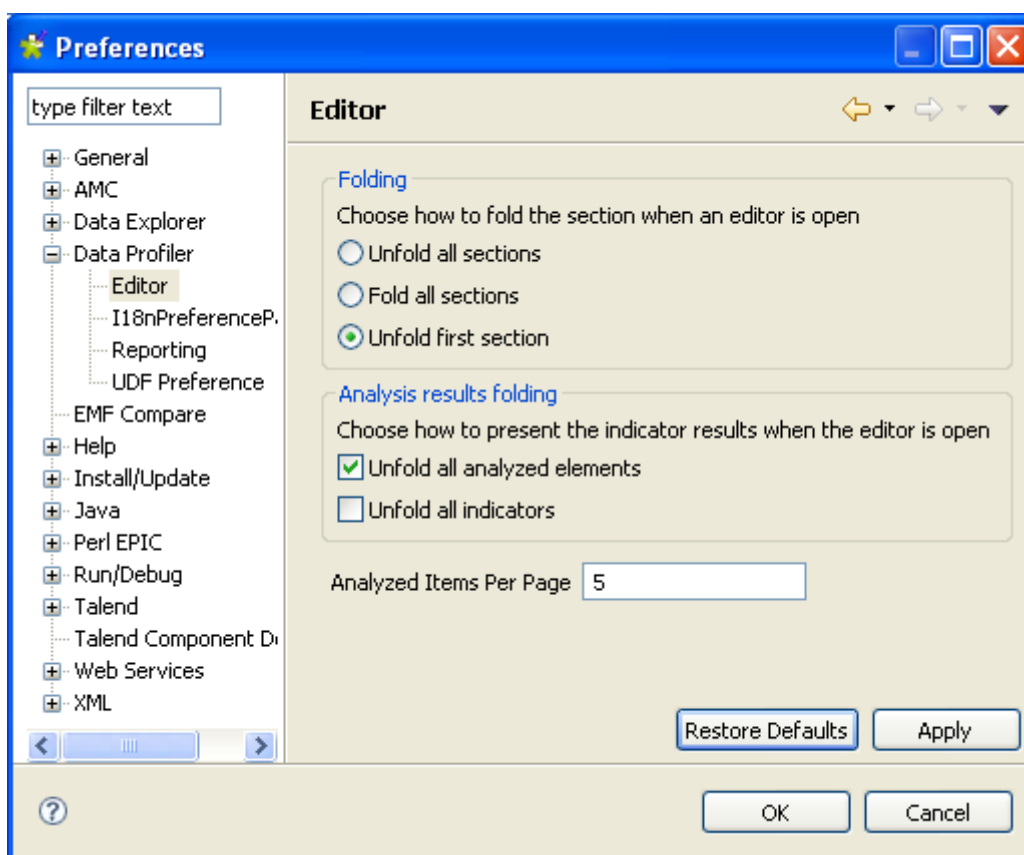
2.1 Setting up the display parameters of all editors

Talend Open Profiler enables you to decide what sections to fold by default when you open any of the connection or analysis editors. It offers the possibility as well to set up the display of all analysis results.

Prerequisite(s): **Talend Open Profiler** main window is open.

To set up editors display:

- On the menu bar, select **Window - Preferences** to display the **[Preferences]** dialog box.
- Expand **Data Profiler** and select **Editor**.



- In the **Folding** area, select the check boxe(s) corresponding to the display mode you want to set for the different sections in all the editors.
- In the **Analysis results folding** area, select the check boxes corresponding to the display mode you want to set for the analysis statistics results in the **Analysis Results** view of the Column Analysis editor.
- In the **Analyzed Items Per Page** field, set the number for the analyzed items you want to group in each page.



You can always click the **Restore Defaults** tab on the **[Preferences]** dialog box to bring back the default values.

- Click **Apply** and then **Ok** to validate the changes and close the **[Preferences]** dialog box.

While carrying on different analyses, all corresponding editors will open with the display mode you set in the [preferences] dialog box.

2.2 Managing database connections

You can use **Talend Open Profiler** to create a connection on your DataBase Management System (DBMS) and display the content of all available databases in the **DQ Repository** tree view.



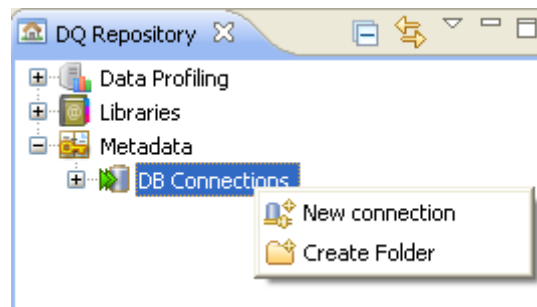
The logical and physical structure of data differs from one database to another in relational databases. The highest level structure “Catalog” followed by “Schema” and finally by “Table” is not applicable to all database types. Thus connection to different databases are reflected by different tree levels and different icons in the **DQ Repository** tree view.

2.2.1 How to create a new database connection

Prerequisite(s): **Talend Open Profiler** main window is open.

To create a new database connection:

- In the **DQ Repository** tree view, expand the **Metadata** folder.
- Right-click **DB Connections** and select **New connection**.



The [Database Connection] wizard opens.

Database Connection

New Database Connection on repository - Step 1/2

Define the properties

Name: SQL_Connection

Purpose: Connecting to MySQL database

Description:

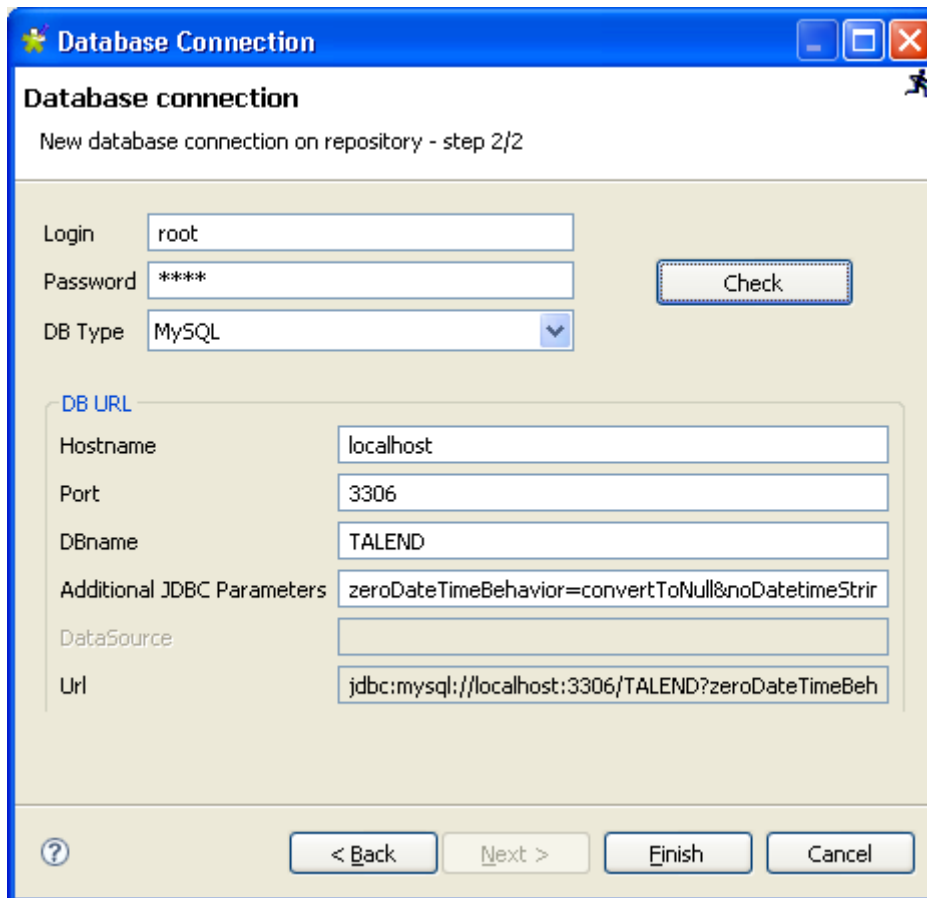
Author: user@company.com

Status: Draft

Path: /FIRSTPROJECT/TDQ_Metadata/DB Connection Select..

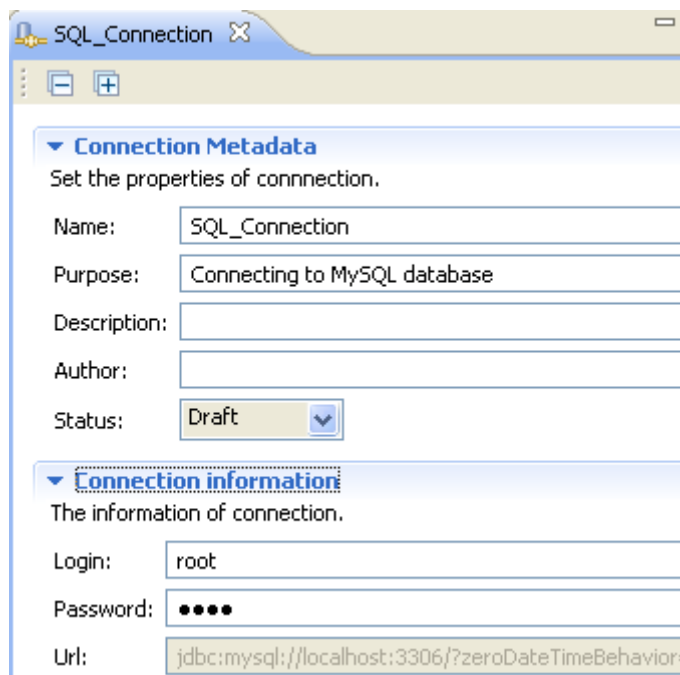
< Back Next > Finish Cancel

- In the **Name** field, enter a name for this new database connection.
- If needed, set other connection metadata (purpose, description and author name) in the corresponding fields and click **Next** to open a new view in the wizard.



- Enter your login and password in their corresponding fields.
- On the **DB Type** list, select the database to connect to.
- In the **DB URL** panel, set your connection parameters.
- Click the **Check** button to verify if your connection is successful.
- Click **Finish** to close the **[Database Connection]** wizard.

A folder for the created MySQL database connection shows under **DB Connection** in the **DQ Repository** tree view, and the Connection Analysis editor opens with the defined metadata.



From the Connection Analysis editor, you can:

- click **Connection information** to display the connection parameters for the relevant database.
- click the **Check** button to check the status of your current connection.

2.2.2 How to filter tables/views in a database connection

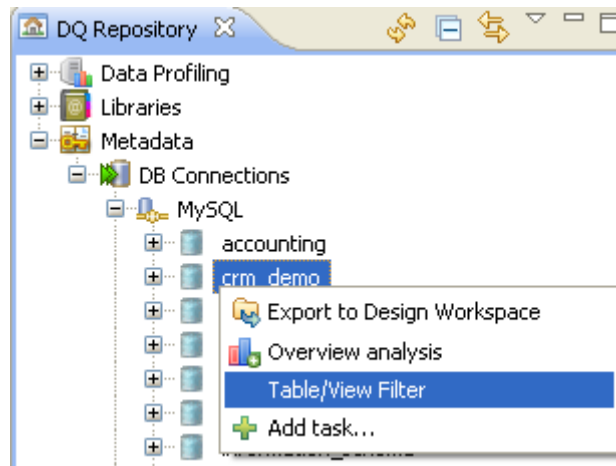
Talend Open Profiler enables you to filter the tables/views to list under any database connection.

This option is very helpful when the number of tables in the database **Talend Open Profiler** is connecting to is very big. If so, a message displays prompting you to set a table filter on the database connection in order to list only defined tables in the **DQ Repository** tree view.

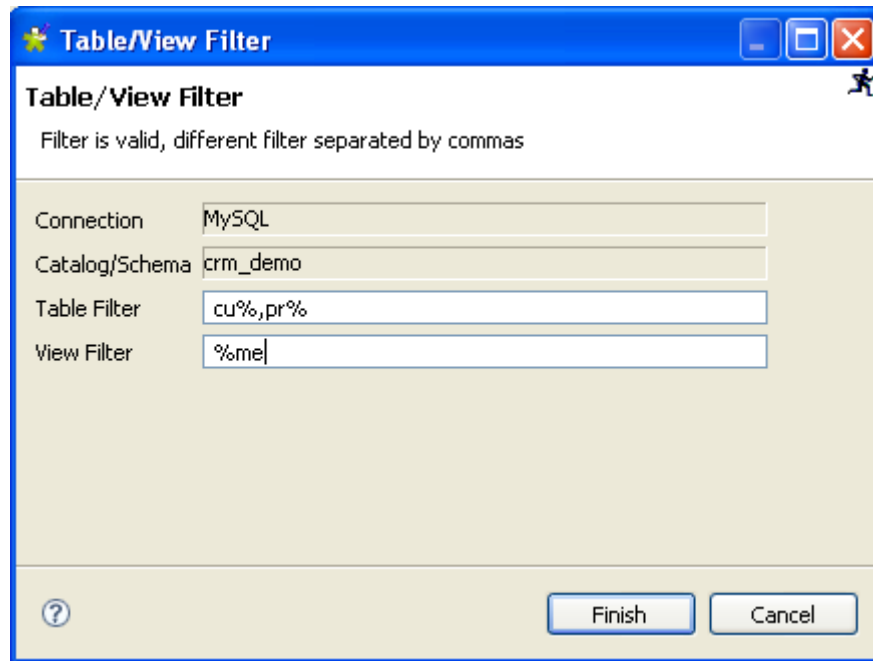
Prerequisite(s): **Talend Open Profiler** main window is open. You have already created a DB connection.

To filter tables/views in a database connection:

- In the **DQ Repository** tree view, expand **Metadata** and **DB Connection** in succession.
- Expand the database connection you want to filter its tables/views and right-click the desired catalog.



- Select **Table/View Filter** from the list to display the corresponding dialog box.



- Set a table and a view filter in the corresponding fields and click **Finish** to close the dialog box.

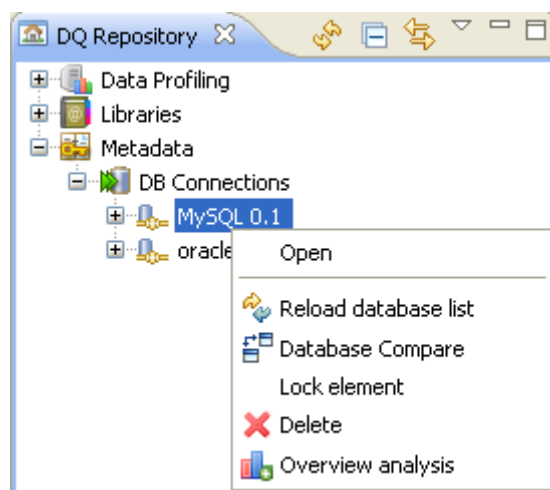
Only tables/views that match the filter you set are listed in the **DQ Repository** tree view.

2.2.3 How to open a database connection

Prerequisite(s): **Talend Open Profiler** main window is open. You have already created a DB connection.

To open a database connection:

- In the **DQ Repository** tree view, expand the **Metadata** and the **DB Connection** folders in succession.
- Either, double-click the database connection you want to open, or
- Right-click the database connection and select **Open** in the drop-down list.



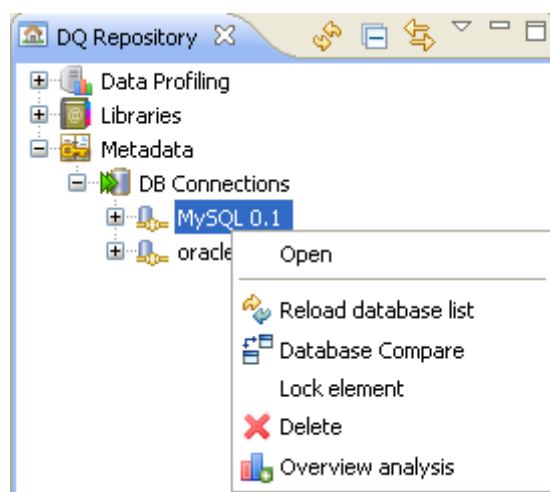
The Connection Analysis editor for the selected database connection displays.

2.2.4 How to delete a database connection

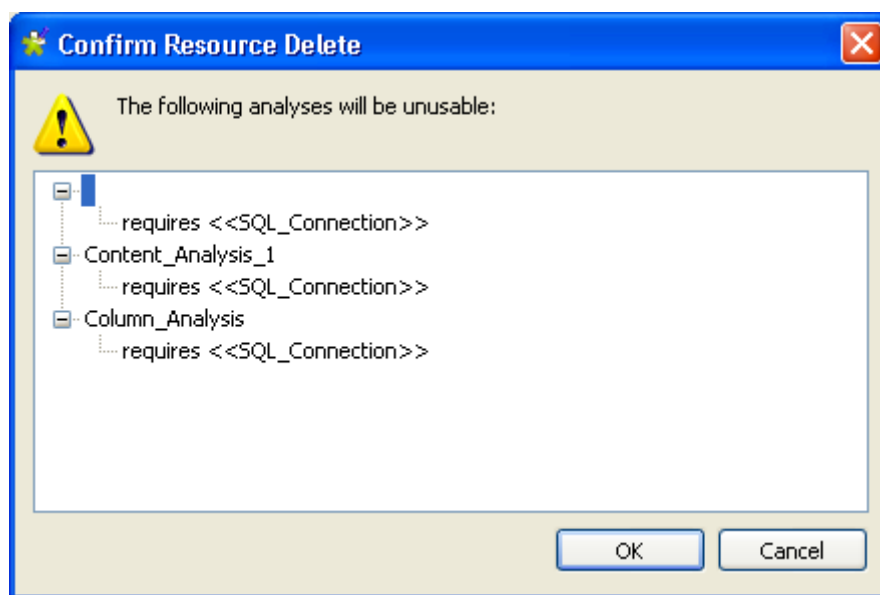
Prerequisite(s): **Talend Open Profiler** main window is open.

To delete a database connection:

- In the **DQ Repository** tree view, expand the **Metadata** and the **DB Connection** folders in succession.
- Right-click the database connection you want to delete and select **Delete** in the drop-down list.



The [**Confirm Resource Delete**] dialog box displays listing all analyses done on the selected database connection. It alerts you that if you delete the connection, all the analyses will become unusable although they will still show in the **DQ Repository** tree view.



You can then either confirm the deletion operation or cancel it.

2.3 Comparing tree-view metadata structures with database structures

Talend Open Profiler can quickly and accurately compare metadata lists displayed in the **DQ Repository** tree view with the database structures you created the connection on to indicate any incoherences.

Talend Open Profiler takes a metadata list in the **DQ Repository** tree view and compares it to the database trying to locate all structure differences and display these differences in the **Compare** view.

You can later, if necessary, synchronize the metadata structure in the tree view with the database structure. For more information, see *Synchronizing tree-view metadata structures with database structures on page 19*.

You can perform the structure comparison at the following three different levels:

- **DB connection** level to compare the catalog and schema lists,
- the **Tables** folder level to compare the list of tables,
- the **Column** folder level to compare the list of columns.

2.3.1 How to compare catalog and schema lists

Prerequisite(s): **Talend Open Profiler** main window is open. You have already created a DB connection.

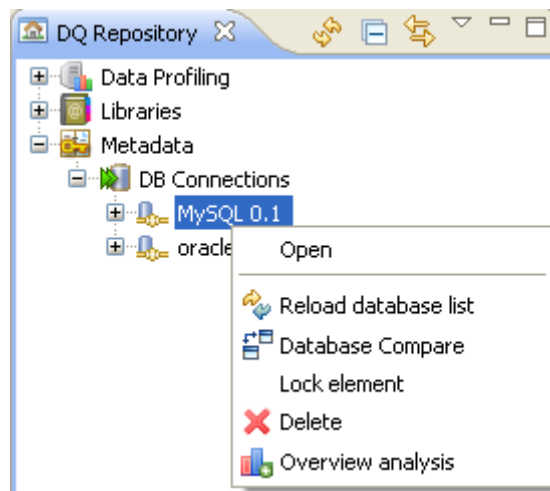
To compare the catalog and schema lists:

- In the **DQ Repository** tree view, expand the **Metadata** and the **DB connection** folders in succession.

Data profiling management procedures

Comparing tree-view metadata structures with database structures

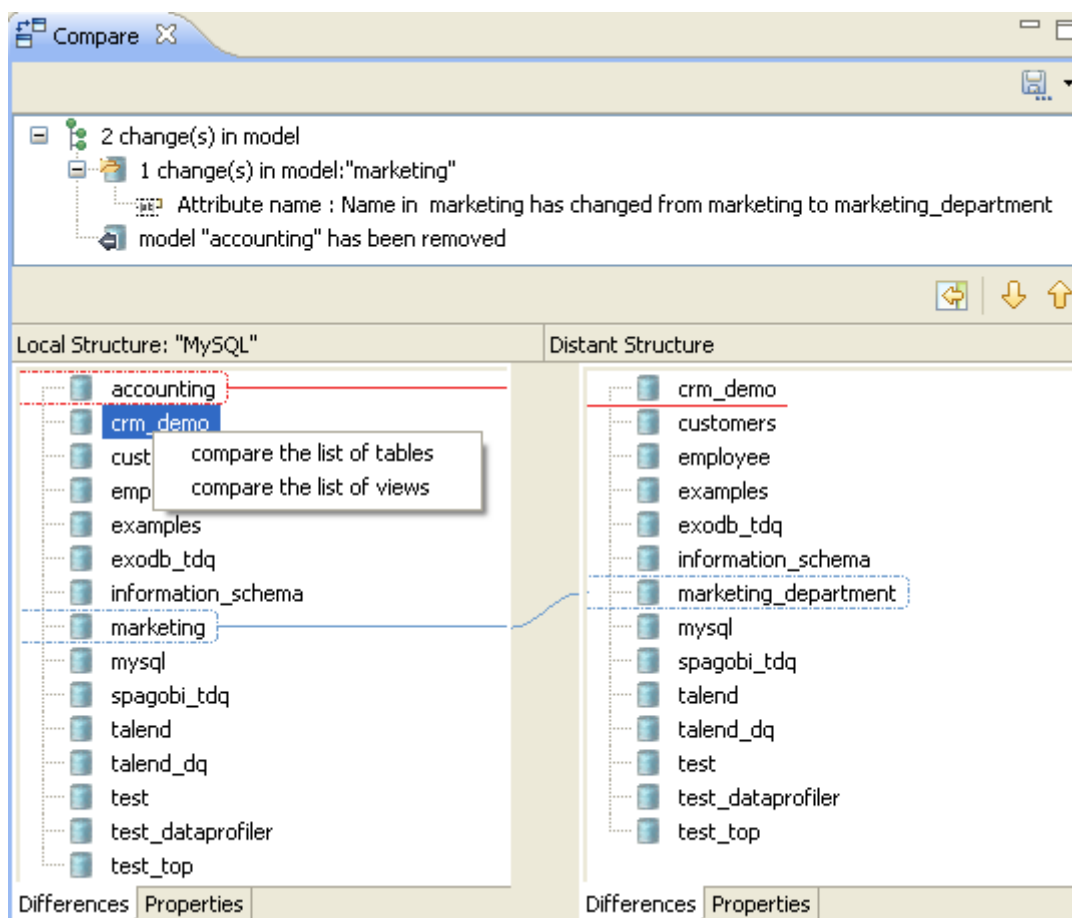
- Right-click the DB connection for which you want to compare the metadata structure with the database structure and select **Database Compare**.



A progress information pop-up opens to confirm that the operation is in progress.

If needed, click the **Cancel** button on the pop-up to stop the operation.

The **Compare** view opens displaying any differences between the metadata structure and the database structure.



- If needed, right-click a specific catalog in the **Compare** view to display a drop-down list where you can select **Compare the list of tables** or **Compare the list of views** to display respectively the table list or the view list of the selected catalog.



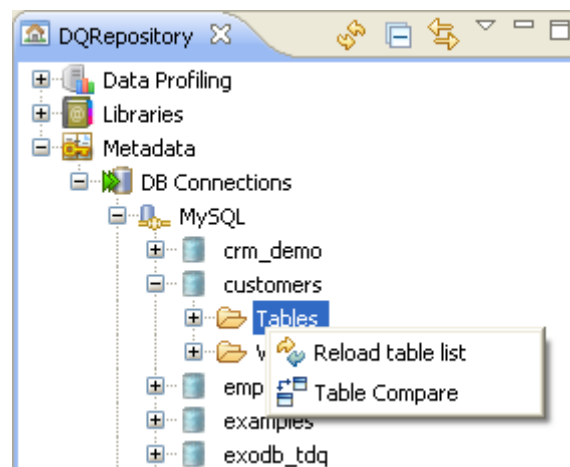
If you select a specific catalog in the **Compare** list and press the **T** or **V** keys on your keyboard, you can display respectively the table or view lists of the selected catalog.

2.3.2 How to compare table lists

Prerequisite(s): **Talend Open Profiler** main window is open. You have already created a DB connection.

To compare a table list:

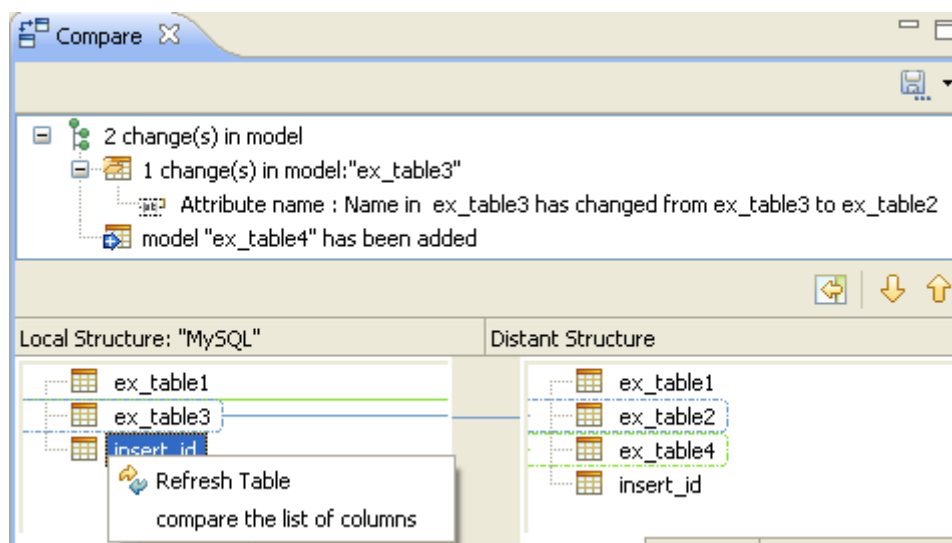
- In the **DQ Repository** tree view, expand the **Metadata** and the **DB connection** folders in succession and browse through the entities in your database connection to reach the **Table** folder you want to compare with that of the database.
- Right-click the **Tables** folder and select **Table Compare**.



A progress information pop-up opens to confirm that the operation is in progress.

If needed, click the **Cancel** button on the pop-up to stop the operation.

The **Compare** view opens displaying any differences between the table lists in the tree view and the actual database.



- If needed, right-click a specific table in the **Compare** view to display a drop-down list where you can select **Compare the list of columns** to display the columns list of the selected table.



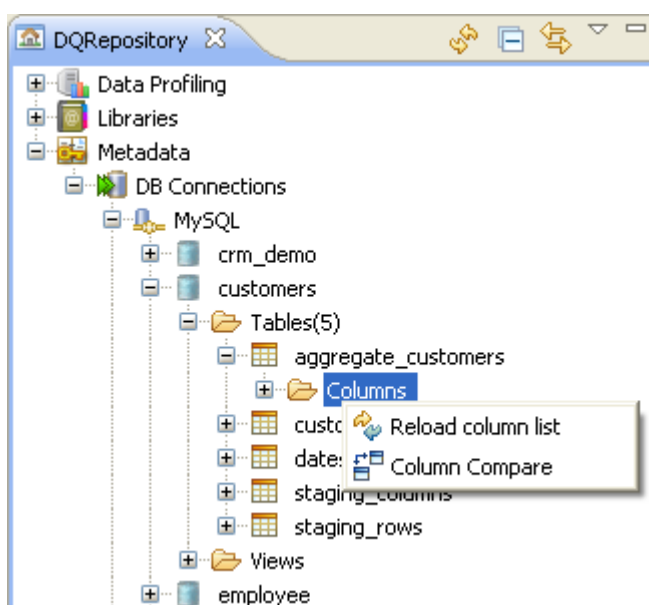
If you select a specific table in the **Compare** list and press the **C** key on your keyboard, you can display the column list of the selected table.

2.3.3 How to compare column lists

Prerequisite(s): **Talend Open Profiler** main window is open. You have already created a DB connection.

To compare a column list:

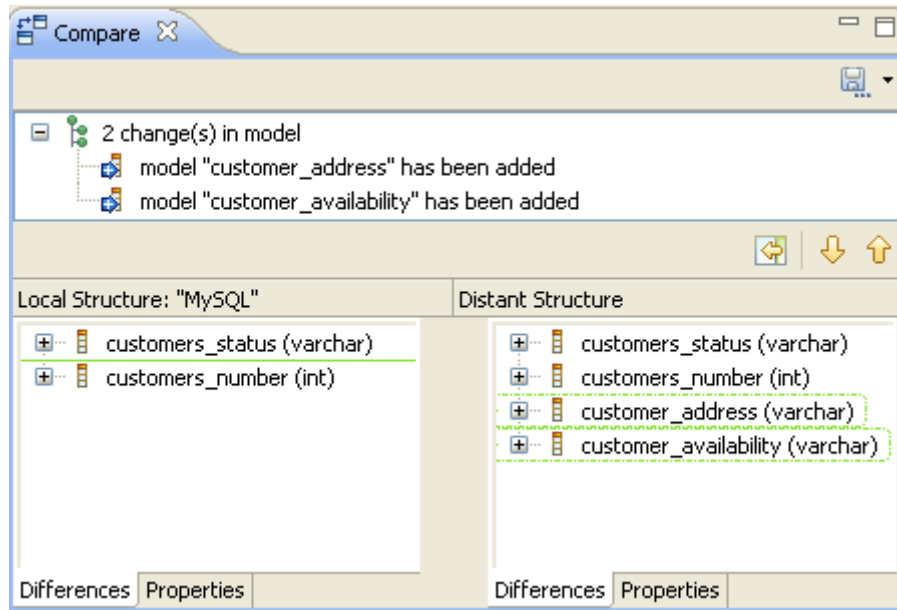
- In the **DQ Repository** tree view, expand the **Metadata** and the **DB connection** folders in succession and browse through the entities in your database connection to reach the **Columns** folder you want to compare with that of the database.
- Right-click the **Columns** folder and select **Column Compare**.



A progress information pop-up opens to confirm that the operation is in progress.

If needed, click the **Cancel** button on the pop-up to stop the operation.

The **Compare** view opens displaying any differences between the column list in the tree view and the database.



2.4 Synchronizing tree-view metadata structures with database structures

In **Talend Open Profiler** you can synchronize metadata lists displayed in the **DQ Repository** tree view with the database structures to eliminate any incoherences. You can perform synchronization at the following three different levels:

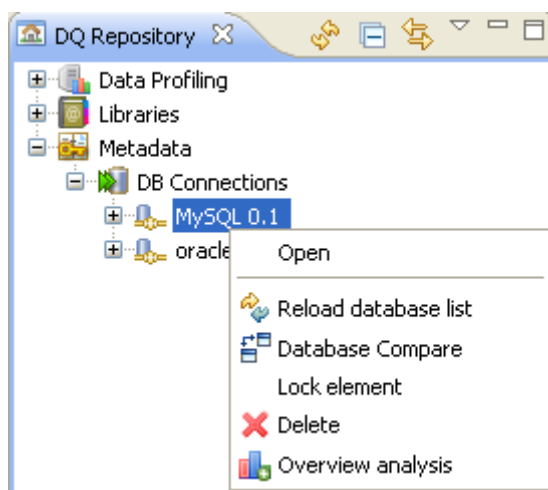
- **DB connection** level to refresh the catalog and schema lists,
- the **Tables** folder level to refresh the list of tables,
- the **Column** folder level to refresh the list of columns.

2.4.1 How to synchronize catalog and schema lists

Prerequisite(s): **Talend Open Profiler** main window is open. You have already created a DB connection.

To synchronize the catalog and schema lists:

- In the **DQ Repository** tree view, expand the **Metadata** and the **DB connection** folders in succession.
- Right-click the DB connection you want to synchronize with the database and select **Reload database list**.



A progress information pop-up opens to confirm that the operation is in progress.

If needed, click the **Cancel** button on the pop-up to stop the synchronization.

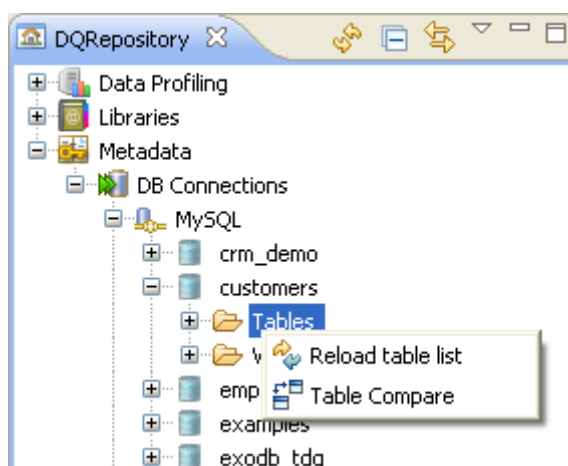
The selected DB connection is updated with the new catalogs and schemas, if any.

2.4.2 How to synchronize table lists

Prerequisite(s): **Talend Open Profiler** main window is open. You have already created a DB connection.

To synchronize a table list:

- In the **DQ Repository** tree view, expand the **Metadata** and the **DB connection** folders in succession and browse through the entities in your database connection to reach the **Table** folder you want to synchronize with the database.
- Right-click the **Tables** folder and select **Reload table list**.



A progress information pop-up opens to confirm that the operation is in progress.

If needed, click the **Cancel** button on the pop-up to stop the synchronization.

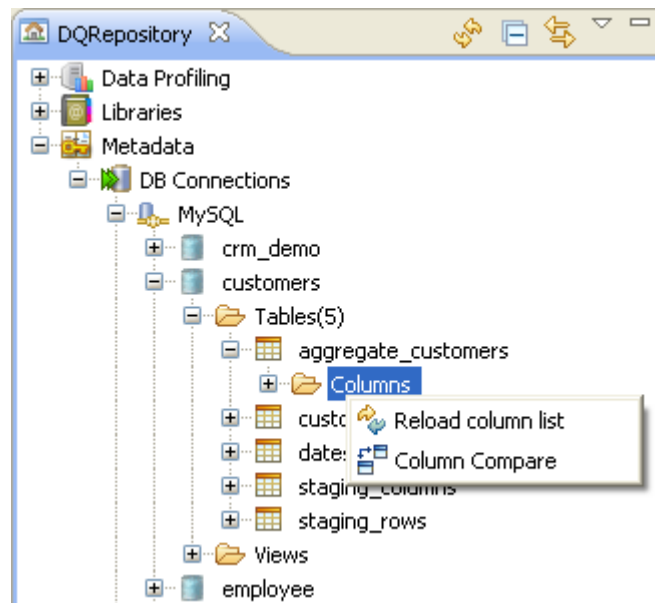
The selected table list is updated with the new tables in the database, if any.

2.4.3 How to synchronize column lists

Prerequisite(s): **Talend Open Profiler** main window is open. You have already created a DB connection.

To synchronize a column list:

- In the **DQ Repository** tree view, expand the **Metadata** and the **DB connection** folders in succession and browse through the entities in your database connection to reach the **Columns** folder you want to synchronize with the database.
- Right-click the **Columns** folder and select **Reload column list**.



A progress information pop-up opens to confirm that the operation is in progress.

If needed, click the **Cancel** button on the pop-up to stop the synchronization.

The selected column list is updated with the new column in the database, if any.

2.5 Managing database content analyses

You can analyze the content of a database to have an overview of the number of:

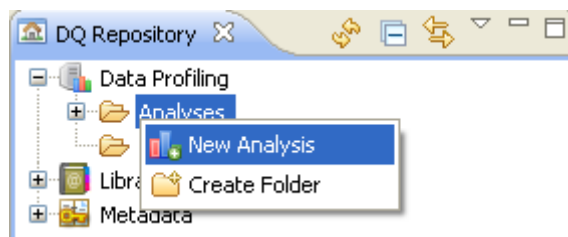
- tables in the database,
- rows per table,
- indexes and primary keys.

2.5.1 How to create a database content analysis

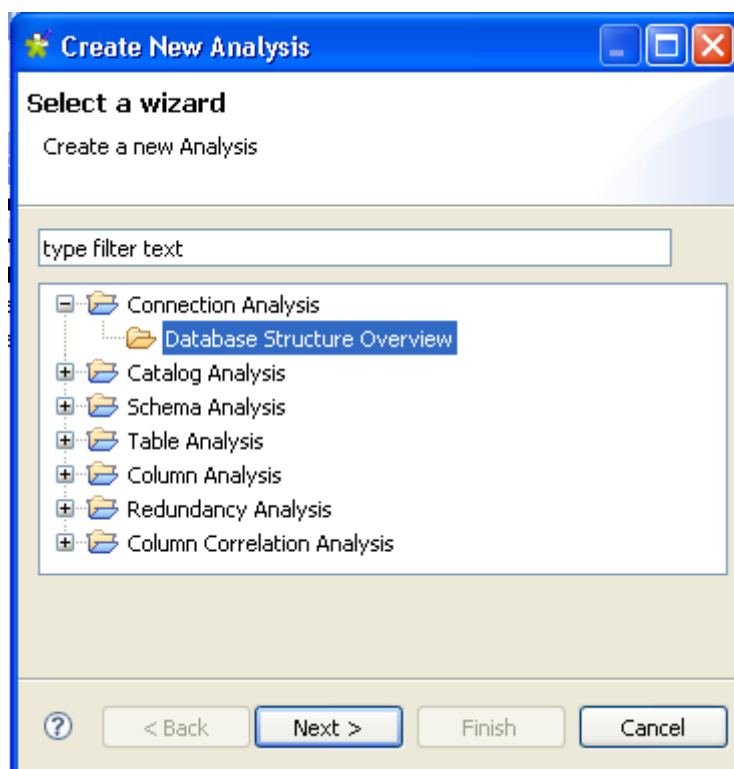
Prerequisite(s): **Talend Open Profiler** main window is open.

To create a database content analysis:

- In the **DQ Repository** tree view, expand the **Data Profiling** folder.
- Right-click the **Analysis** folder and select **New Analysis**.



The [Create New Analysis] wizard opens.



- Expand the **Connection Analysis** node and click **Database Structure Overview**.
- Click the **Next** button to open a new view on the wizard.

New Analysis
Add an analysis in the repository

Name: Analysis_Name|

Purpose: Why do you want to do the analysis?

Description: analysis description

Author: user@company.com

Status: Draft

Path: /TALENDEMOSJAVA/TDQ_Data Profiling/Analy Select..

Type: Multiple Column Analysis

< Back Next > Finish Cancel

- In the **Name** field, enter a name for the current analysis.
- If needed, set the analysis metadata (purpose, description and author name) in the corresponding fields and click **Next** to open a new view.

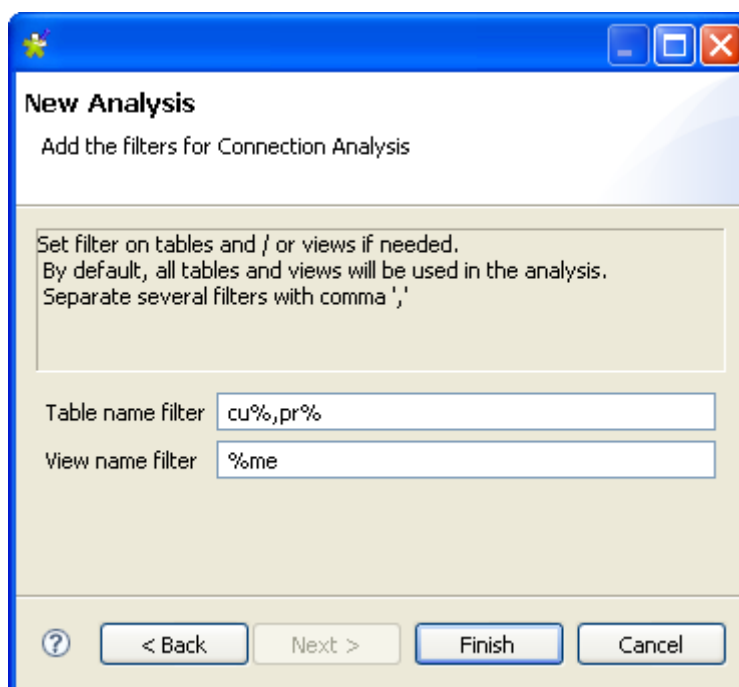
New Analysis
Choose a connection to analyze

Connections:

- TOP_DEFAULT_PRJ
 - TDQ_Metadata
 - DB Connections
 - Oracle
 - MySQL

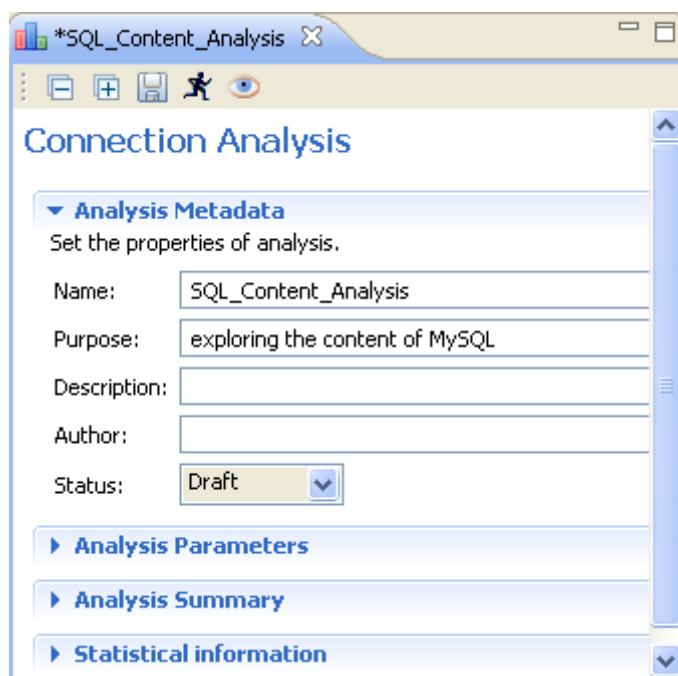
< Back Next > Finish Cancel

- Expand **DB Connections** in succession and select a connection to analyze, if more than one exist.
- Click **Next** to open a new view on the wizard.



- If needed, set filters on tables and/or views in their corresponding fields using the SQL language. By default, the analysis will include all tables and views in the database.
- Click **Finish** to close the **[Create New Analysis]** wizard.

A folder for the newly created analysis shows under **Analysis** in the **DQ Repository** tree view, and the Connection Analysis editor opens with the defined metadata.



- If needed, click **Analysis Parameters** to check/modify filters on table and/or views, if any.
- If needed, click **Analysis Summary** to display all the parameters of the current analysis along with the current analysis execution status.

- Press **F6** to execute the current analysis.
 A progress information pop-up opens to confirm that the operation is in progress.
 Analysis results are stored in the **Statistical informations** view.
- Click **Statistical informations** to display analytical information about the content of the relevant database.

Catalog	#rows	#tables	#rows/table	#views	#rows/view	#keys	#indexes
accounting	0	0	NaN	0	NaN	0	0
crm_demo	25977	18	1443.17	0	NaN	13	13
customers	244	6	40.67	0	NaN	2	2
employee	16	3	5.33	0	NaN	1	1
examples	3	3	1.00	0	NaN	3	3
exodb_tdq	185	21	8.81	0	NaN	21	21
information_schema	0	0	NaN	0	NaN	0	0
marketing	0	0	NaN	0	NaN	0	0
mysql	1981	23	86.13	0	NaN	42	48
spagobi_tdq	609	37	16.46	0	NaN	59	120
talend	7	1	7.00	0	NaN	1	1
talend_dq	7890	12	657.50	17	17.18	16	22
test	10	1	10.00	0	NaN	0	0
test_dataprofiler	203	13	15.62	0	NaN	0	0
test_top	201	1	201.00	0	NaN	1	1

Table	#rows	#keys	#indexes
aggregate_c...	8	0	0
customer	9	1	1
dates	0	0	0
new table	0	1	1
staging_colu...	33	0	0
staging_rows	194	0	0

View



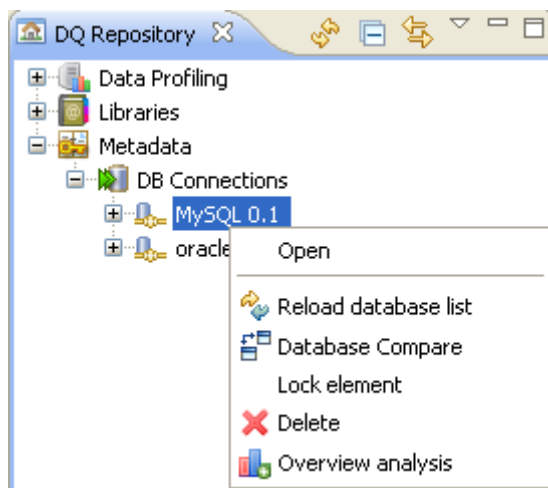
You can click a catalog or a schema in the displayed analytical table to open a result list that detail all tables included in the catalog or schema with a summary of their content.



You can sort alphabetically data listed in catalogs or schemas by simply clicking any column header in the analytical table. You can also sort alphabetically all columns in the result table doing the same.

2.5.2 How to create a database content analysis directly from the DB connection

In **Talend Open Profiler**, you can create an analysis of the content of a given database directly from the drop-down list of a DB connection folder by selecting **Overview analysis**.



This way, you do not have to choose in the **[Create New Analysis]** wizard either the type of analysis you want to carry out or the DB connection to analyze. Otherwise, all other procedural steps are exactly the same as in *How to create a database content analysis on page 21*.

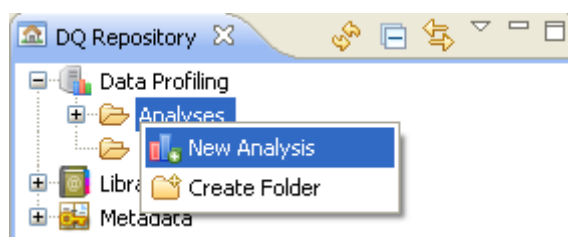
2.5.3 How to create a catalog analysis

You can use **Talend Open Profiler** to analyze one specific catalog in a database, if this entity is used in the physical structure of the database. The result of the analysis gives analytical information about the content of this catalog, for example number of rows, number of tables, number of rows per table and so on.

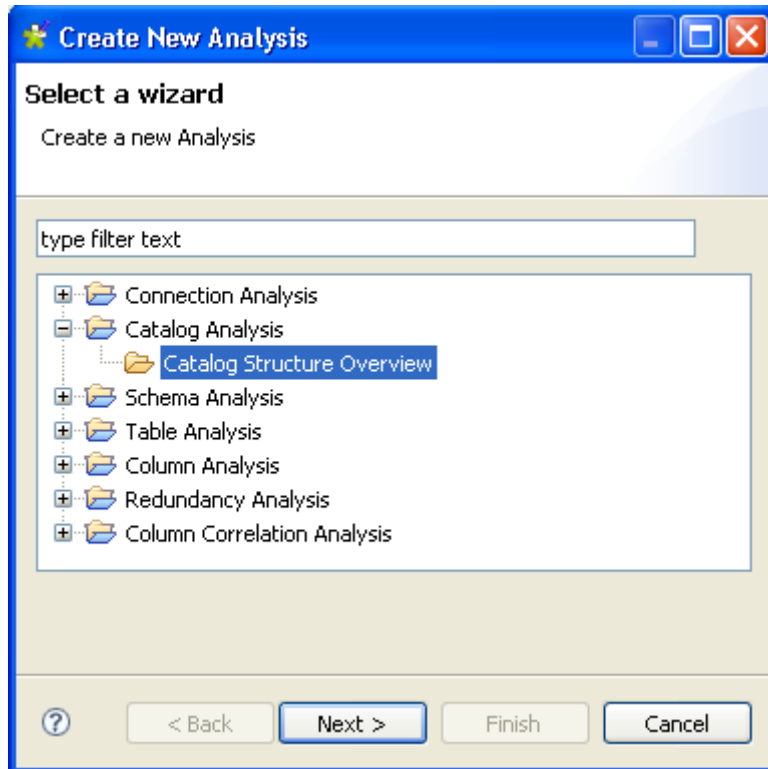
Prerequisite(s): **Talend Open Profiler** main window is open. At least one database connection has been created to connect to a database that uses the “catalog” entity to structure its data, for example the MySQL database.

To create a catalog analysis:

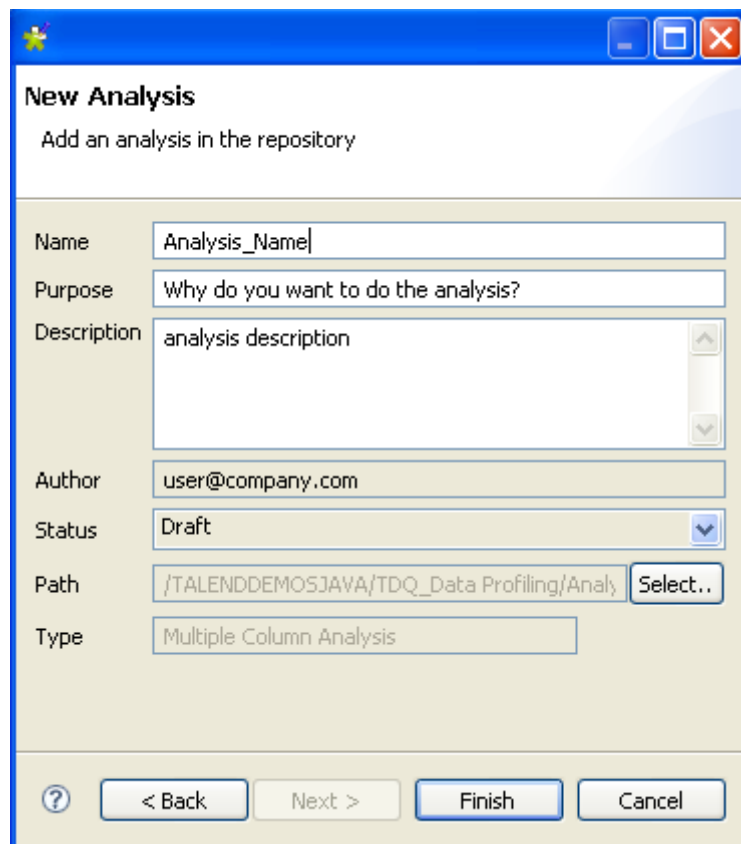
- In the **DQ Repository** tree view, expand the **Data Profiling** folder.
- Right-click the **Analysis** folder and select **New Analysis**.



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

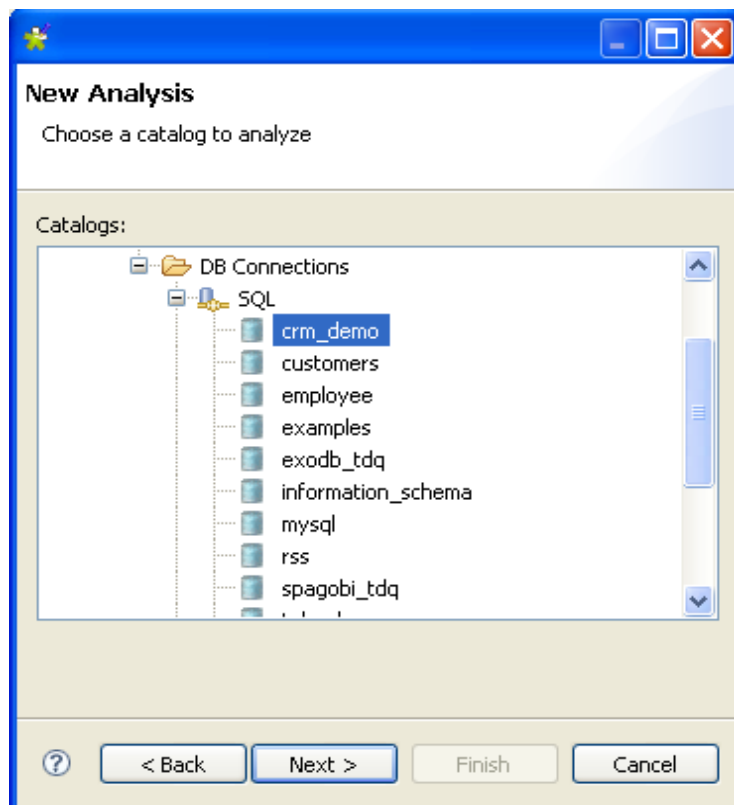


- Expand the **Catalog Analysis** node and click **Catalog Structure Overview**.
- Click the **Next** button to open a new view on the wizard.

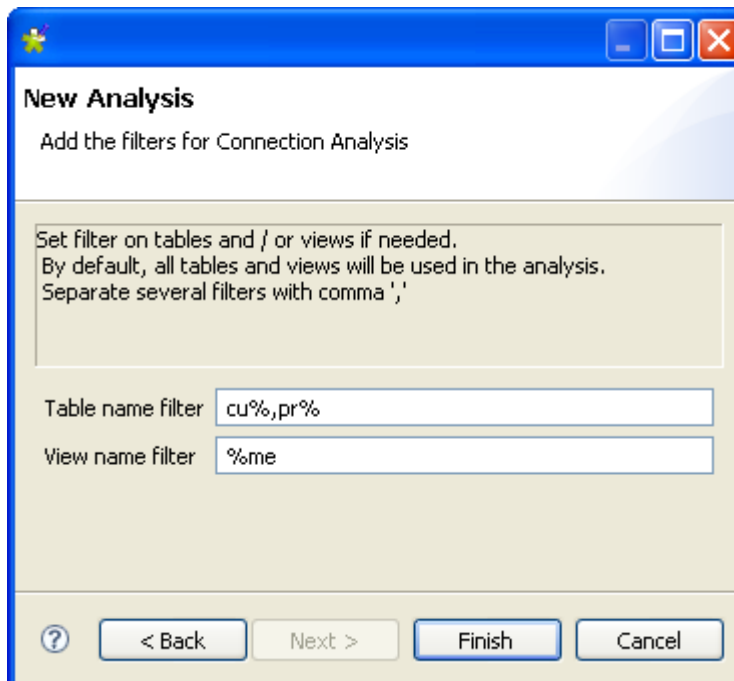


- In the **Name** field, enter a name for the current analysis.

- If needed, set the analysis metadata (purpose, description and author name) in the corresponding fields and click **Next** to open a new view on the wizard.



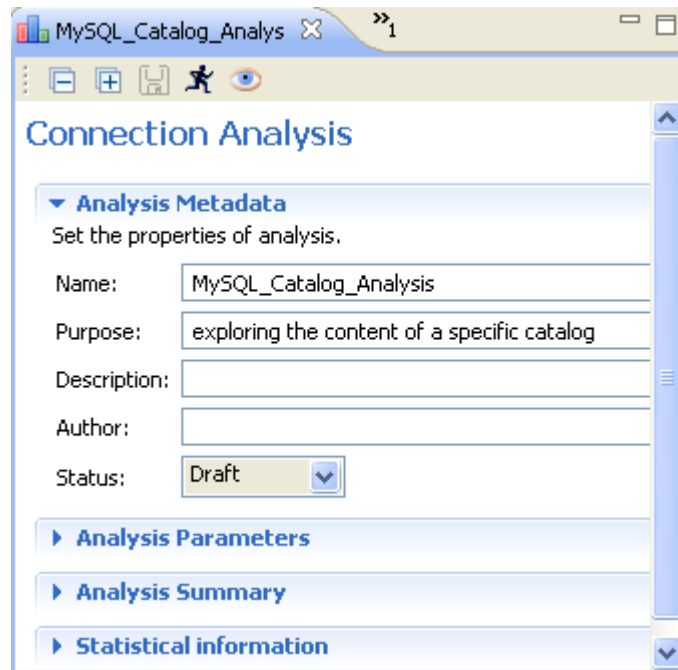
- Expand in succession **DB Connections** and the database that include catalog entities in its physical structure and select a catalog to analyze.
- Click **Next** to open a new view on the wizard.



- If needed, set filters on tables and/or views in their corresponding fields using the SQL language. By default, the analysis will include all tables and views in the catalog.

- Click **Finish** to close the [Create New Analysis] wizard.

A folder for the newly created analysis shows under **Analysis** in the **DQ Repository** tree view, and the Catalog Analysis editor opens with the defined metadata.



- If needed, click **Analysis Parameters** to check/modify filters on table and/or views, if any.
- If needed, click **Analysis Summary** to display all the parameters of the current analysis along with the current analysis execution status.
- Click the **Run** button or press **F6** to execute the current analysis. A progress information pop-up opens to confirm that the operation is in progress. Analysis results are stored in the **Statistical informations** panel.
- Click **Statistical informations** to display analytical information about the content of the relevant catalog.

Statistical informations							
Catalog	#rows	#tables	#rows/table	#views	#rows/view	#keys	#indexes
crm_demo	25977	18	1443.17	0	NaN	13	13

Table	#rows	#keys	#indexes
account	11	1	1
currency	72	2	2
customer	10281	1	1
department	12	1	1
employee	1155	1	1
inventory_fa...	5000	0	0
position	18	1	1
product	1560	1	1
product_class	110	0	0



You can click any catalog in the analytical table to have a detailed view of the tables included in the selected catalog with a summary of their content.



You can sort alphabetically data listed in the result table by simply clicking any column header in the result table.

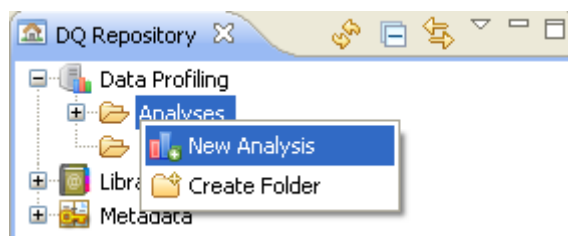
2.5.4 How to create a schema analysis

You can use **Talend Open Profiler** to analyze one specific schema in a database, if this entity is used in the physical structure of the database. The result of the analysis gives analytical information about the content of this schema, for example number of rows, number of tables, number of rows per table and so on.

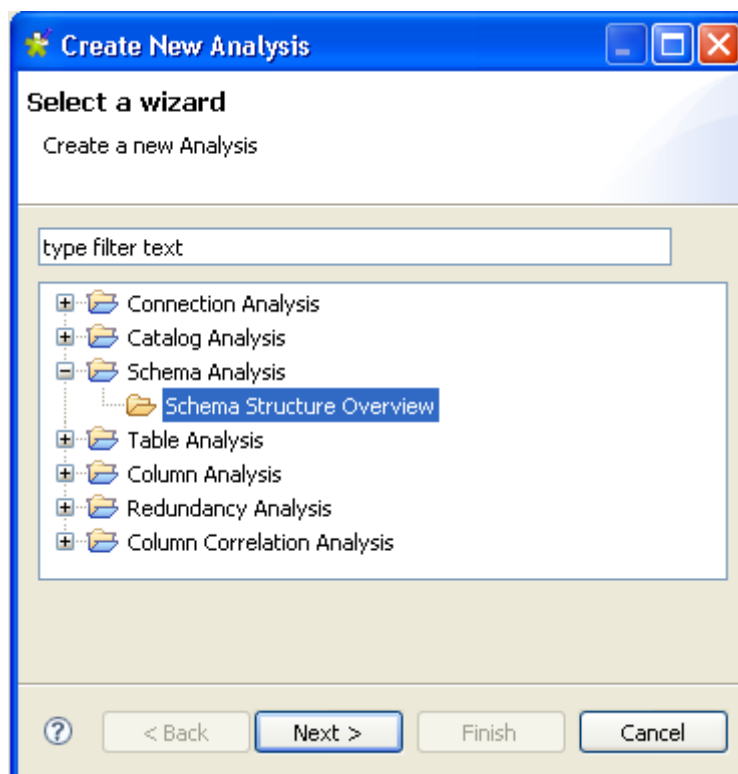
Prerequisite(s): **Talend Open Profiler** main window is open. At least one database connection has been created to connect to a database that uses the “schema” entity to structure its data, for example the Oracle database.

To create a schema analysis:

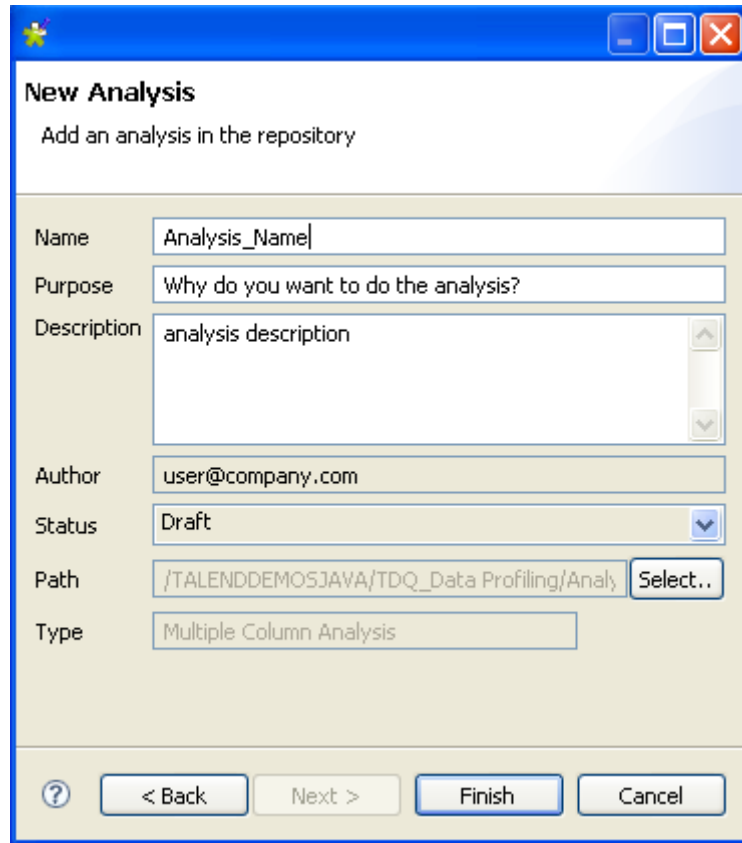
- In the **DQ Repository** tree view, expand the **Data Profiling** folder.
- Right-click the **Analysis** folder and select **New Analysis**.



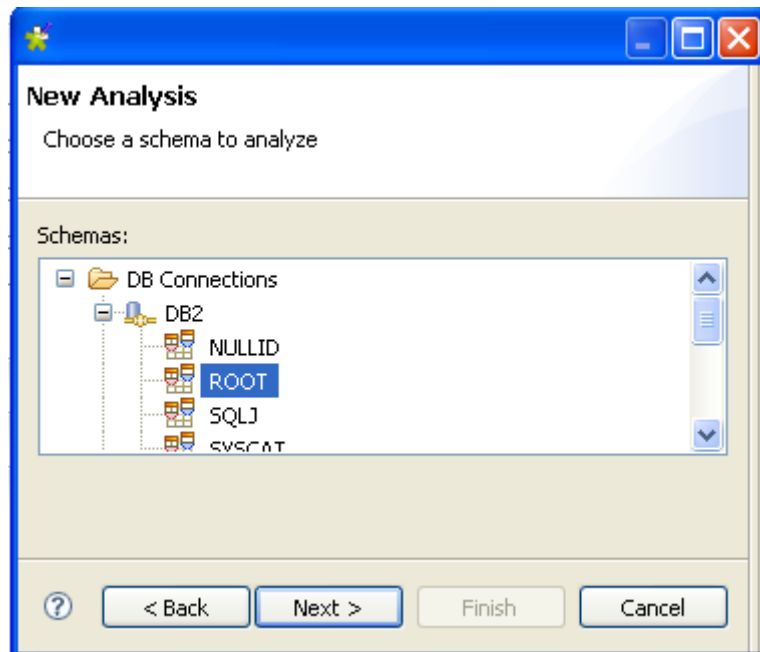
The [Create New Analysis] wizard opens.



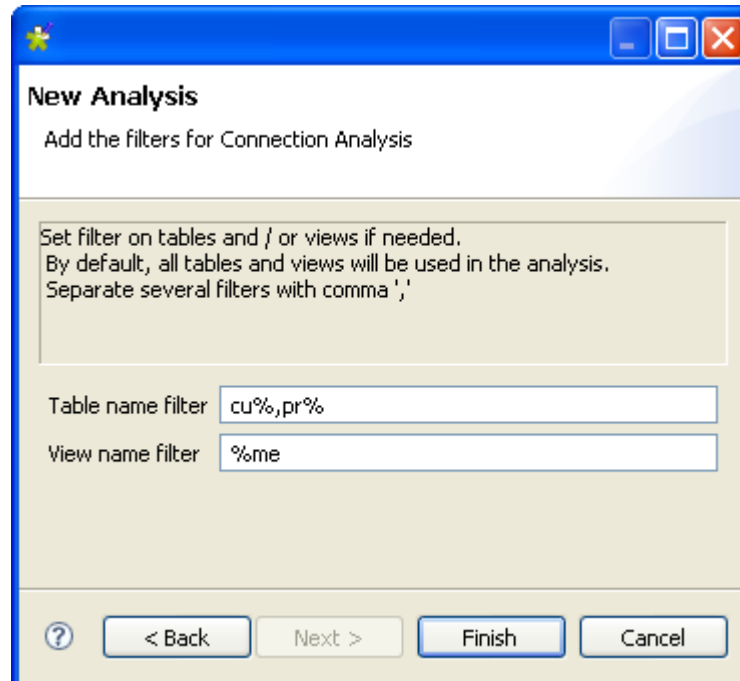
- Expand the **Schema Analysis** node and click **Overview Analysis**.
- Click the **Next** button to open a new wizard.



- In the **Name** field, enter a name for the current analysis.
- If needed, set the analysis metadata (purpose, description and author name) in the corresponding fields and click **Next** to open a new view on the wizard.

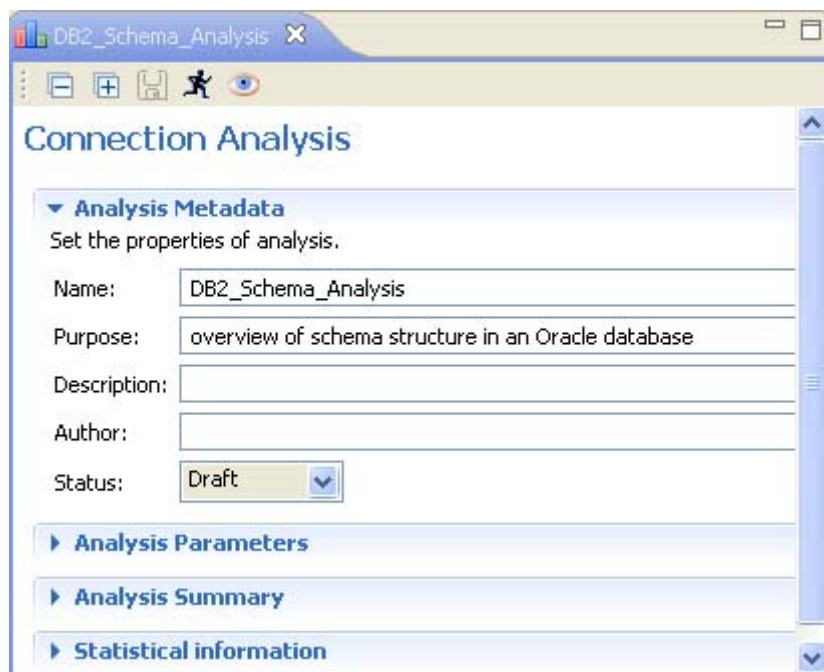


- Expand in succession **DB Connections** and the database that include schema entities in its physical structure and select a schema to analyze.
- Click **Next** to open a new view on the wizard.



- If needed, set filters on tables and/or views in their corresponding fields using the SQL language. By default, the analysis will include all tables and views in the catalog.
- Click **Finish** to close the **[Create New Analysis]** wizard.

A folder for the newly created analysis shows under **Analysis** in the **DQ Repository** tree view, and the Schema Analysis editor opens with the defined metadata.



- If needed, click **Analysis Parameters** to check/modify filters on table and/or views, if any.
- If needed, click **Analysis Summary** to display all the parameters of the current analysis along with the current analysis execution status.

- Click the **Run** button or press **F6** to execute the current analysis.
A progress information pop-up opens to confirm that the operation is in progress.
Analysis results are stored in the **Statistical informations** panel.
- Click **Statistical informations** to display analytical information about the content of the relevant catalog.

▼ **Statistical informations**

Schema	#rows	#tables	#rows/table	#views	#rows/view	#keys	#indexes
ROOT	8455	388	21.79	0	NaN	1	389

Table	#rows	#keys	#indexes
FEATURE3271	0	0	1
SCDDEST	6	0	1
SCDTEST	4	0	1
TABLE06U8HB	0	0	1
TABLE0G5MGA	0	0	1
TABLE0G7J6T	0	0	1
TABLE1NHCNH	0	0	1

View



You can click a schema in the list to have a detailed view of the tables included in the selected schema with a summary of their content.



You can sort alphabetically data listed in the result table by simply clicking any column header in the result table.

2.6 Managing the analysis of a set of columns

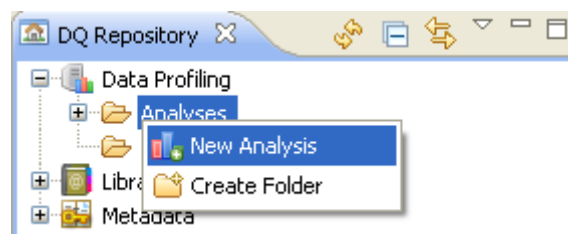
Talend Open Profiler enables you to analyze the content of a set of columns using the Java or the SQL execution engine.

2.6.1 How to analyze a set of columns

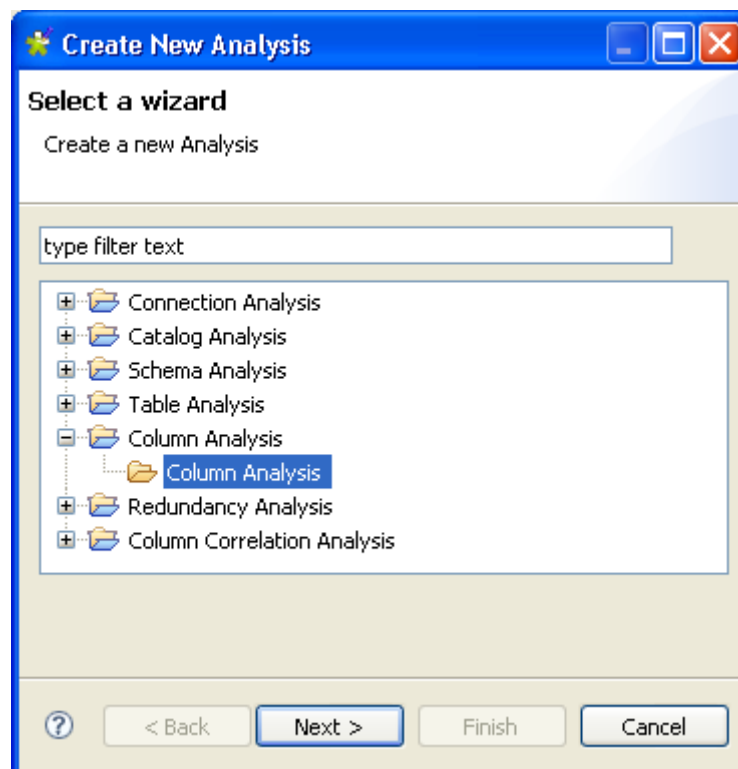
Prerequisite(s): **Talend Open Profiler** main window is open.

To analyze a set of columns:

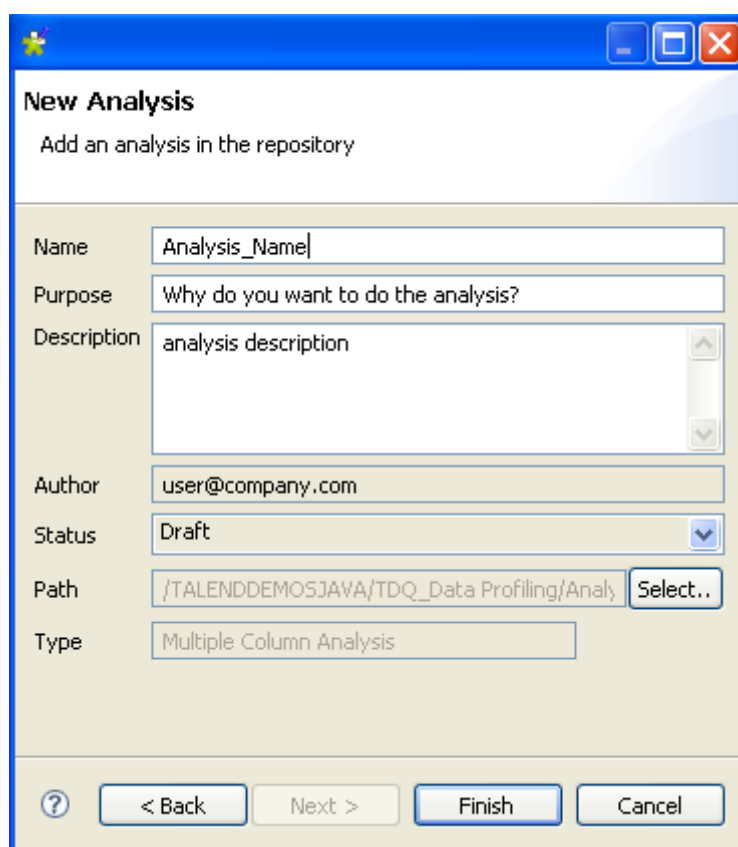
- In the **DQ Repository** tree view, expand the **Data Profiling** folder.
- Right-click the **Analysis** folder and select **New Analysis**.



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



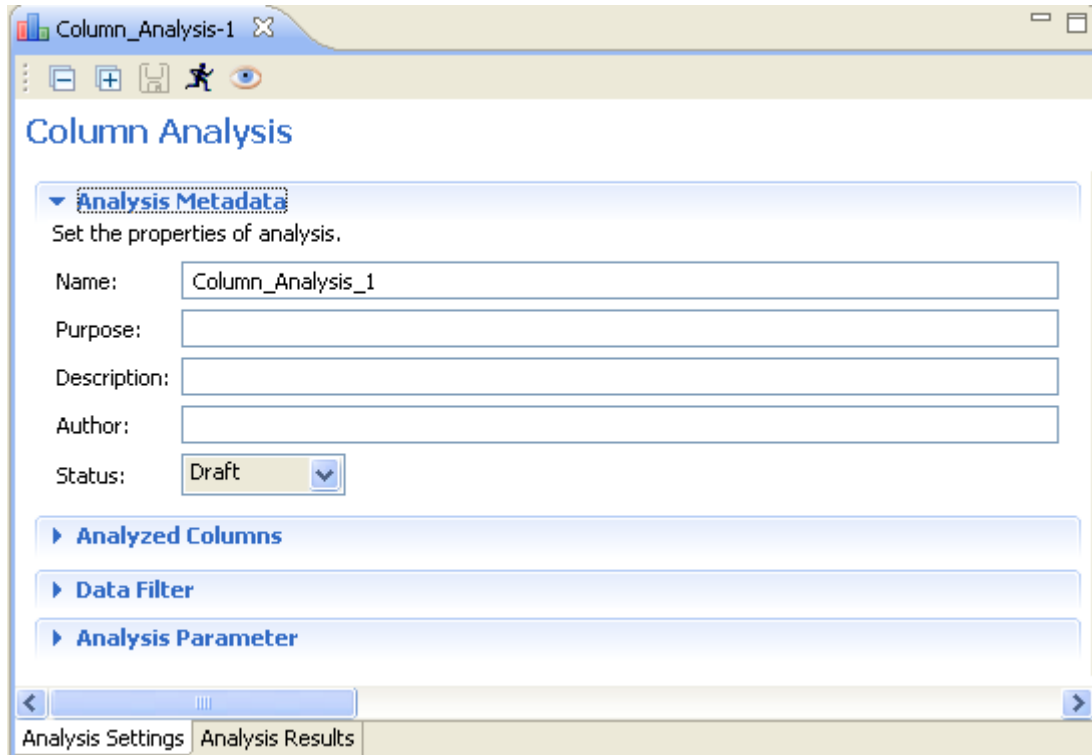
- Expand the **Column Analysis** folder and click **Column Analysis**.
- Click the **Next** button to open a new view on the wizard.



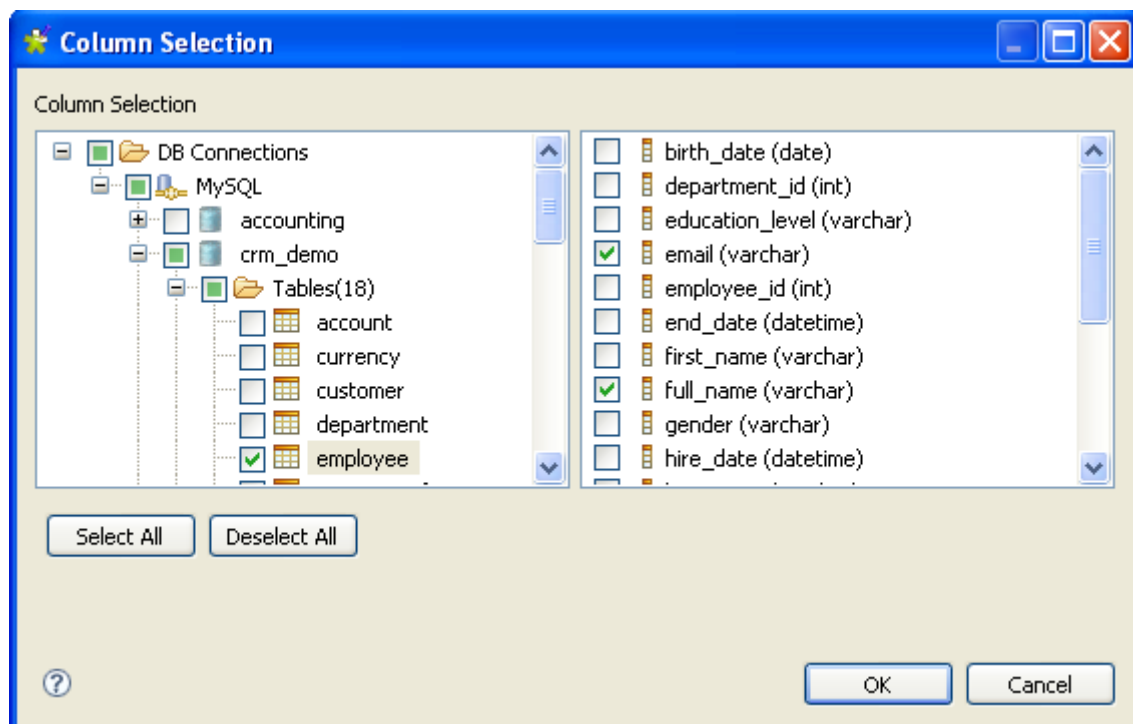
- In the **Name** field, enter a name for the current column analysis.

- If needed, set column analysis metadata (purpose, description and author name) in the corresponding fields and click **Finish** to close the [Create New Analysis] wizard.

A folder for the newly created column analysis shows under **Analysis** in the **DQ Repository** tree view, and the Column Analysis editor opens with the defined analysis metadata.



- Click **Analyzed Columns** to display the **Analyzed Columns** view.
- In the **Analyzed Columns** view, click **Select column to analyze** to open the [Column Selection] dialog box.



- Expand **DB Connections** and browse through the entities in your database connection to reach the table you want to analyze.
- Click the table name to display all its columns in the right-hand panel of the **[Column Selection]** dialog box.
- In the column list, select the check boxes of the column(s) you want to analyze and click **OK** to close the dialog box.

The selected columns display in the **Analyzed Column** view of the Column Analysis editor.

▼ Analyzed Columns

Connection: MySQL

[Select columns to analyze](#)

[Select indicators for each column](#)

Analyzed Columns	Datamining Type	Pattern	Operation
email (varchar)	Nominal		
full_name (varchar)	Nominal		



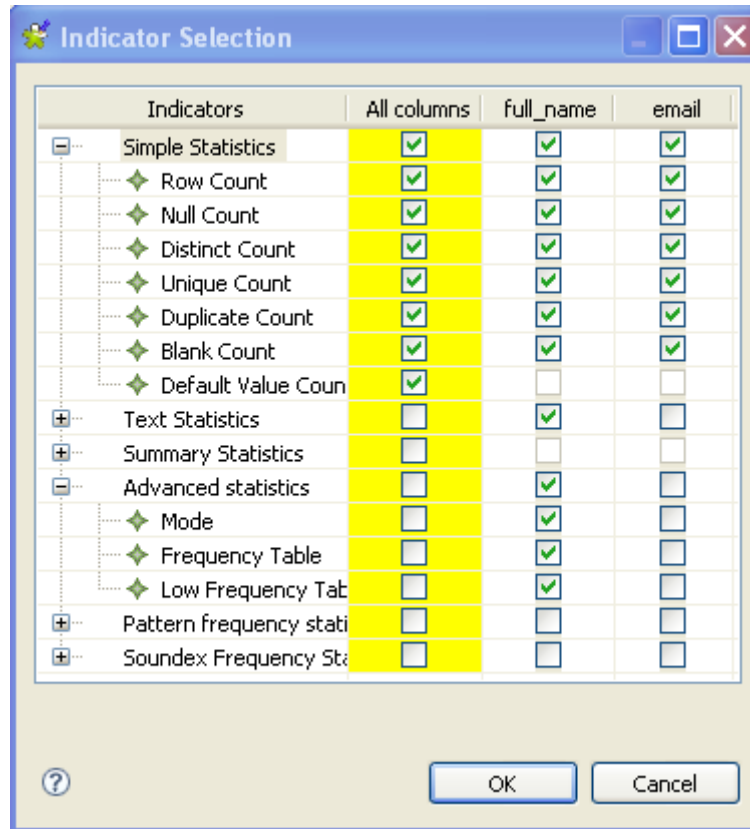
If one of the columns you want to analyze is a primary or a foreign key, its data mining type will automatically become Nominal when you list it in the **Analyzed Columns** view.

For more information on data mining types in **Talend Open Profiler**, see *Data mining types on page 41*.



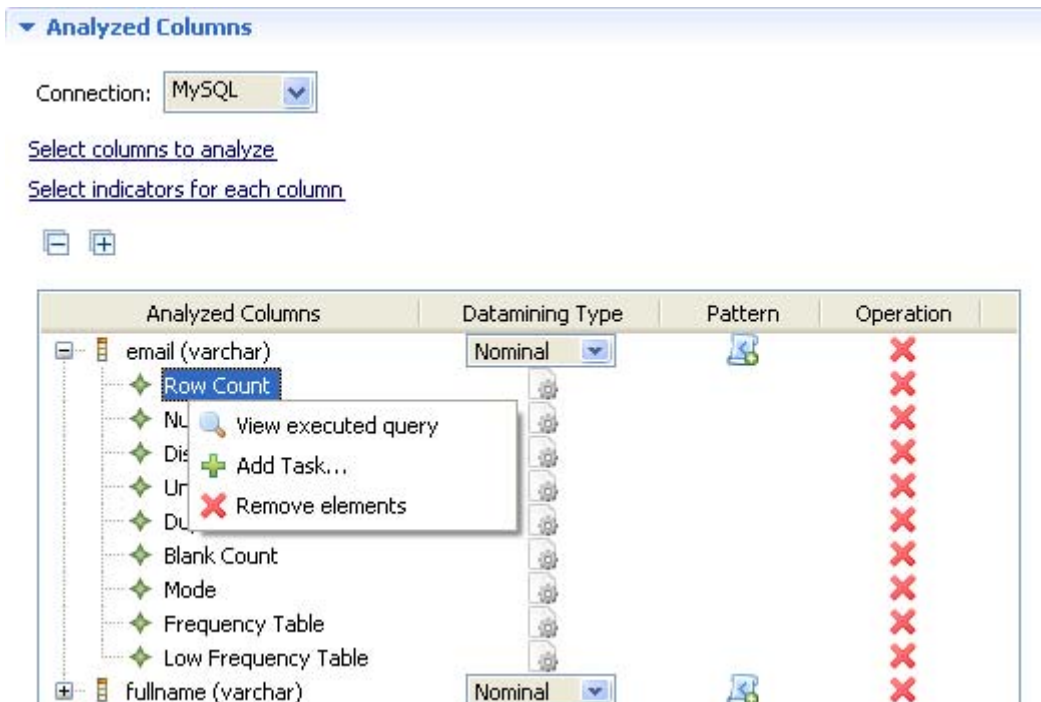
You can change your database connection by selecting another database from the **Connection** list. If the columns listed in the **Analyzed Columns** view do not exist in the new database connection you want to set, you will receive a warning message that enables you to continue or cancel the operation.

- Click **Select indicators for each column** to open the **[Indicator Selection]** dialog box.




- Set indicator parameters for the analyzed columns as needed and click **OK** to close the dialog box.

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





You can view the executed query for each of the attached indicators if you right-click the indicators and then select the corresponding option from the list. However, when you use the Java engine, SQL queries will not be accessible and thus clicking this option will display a warning message.

- If needed, click the option icon  to open a dialog box where you can set options for each indicator. For more information about indicators management, see *Managing indicators on page 102*.
- If needed, click the **add pattern** icon to add patterns to indicator. For more information, see *How to analyze a set of columns with pattern indicators on page 89*, and for more information about patterns management, see *Managing patterns on page 83*.
- Click **Data Filter** in the Column Analysis editor to display its view and filter data through SQL “WHERE” clauses, if needed.
- Click **Analysis Parameter** in the Column Analysis editor to display its view and then select from the **Execution engine** list the engine you want to use, Java or SQL, for the defined patterns.



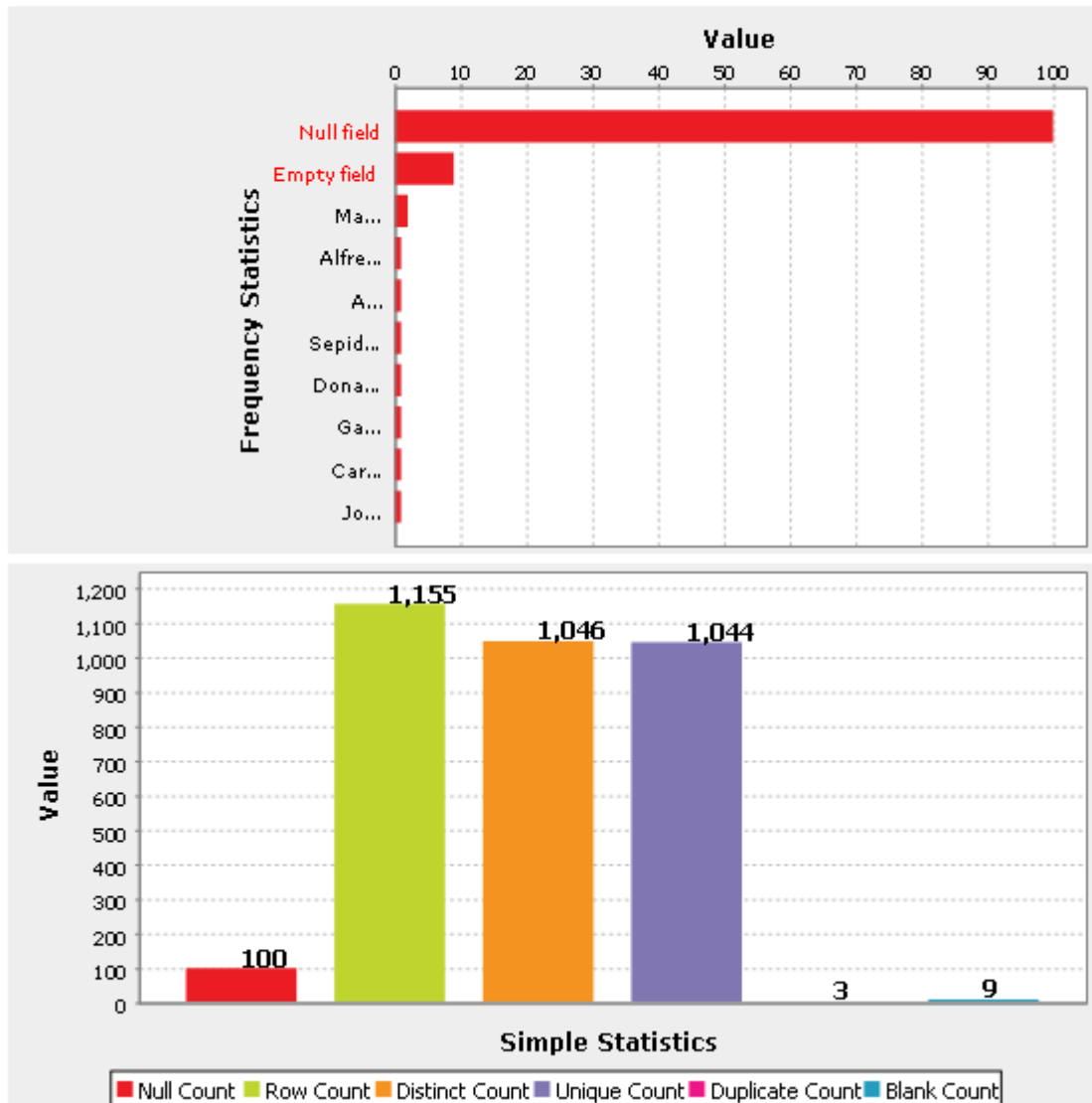
If you select the Java engine, the system will look for Java regular expressions first, if none is found, it looks for SQL regular expressions.

- Click the save icon on the toolbar of the Column Analysis editor and then press **F6** to execute the column analysis.

The **Graphics** panel to the right of the Column Analysis editor displays a group of graphic(s), each corresponding to the group of the indicators set for each analyzed column.

Below are graphics representing the Frequency Statistics and Simple Statistics for the first analyzed column, *email*, in the above procedure.

Column: email



To view the different graphics associated with all analyzed columns, you need to navigate through the different pages in the **Graphics** panel using the toolbar on the upper-right corner.

To access a more detailed view of the analysis results:

- Click the **Analysis Results** tab at the bottom of the Column Analysis editor to open the corresponding view.
- Click the **Analysis Result** tab and then the name of the analyzed column from which you want to display to display the detailed results.

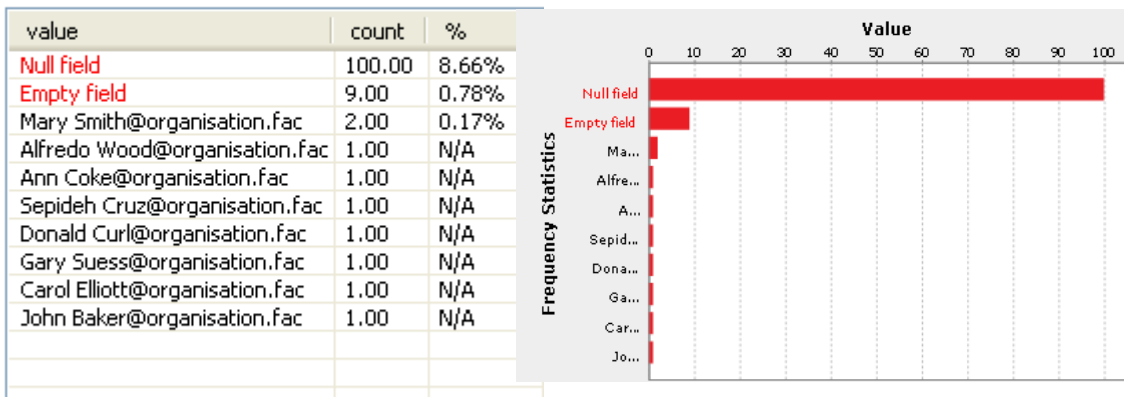
The detailed analysis results view shows the generated graphics for the analyzed columns accompanied with tables that detail the statistic results.

Below are the tables that accompany the Frequency and Simple Statistics graphics in the **Analysis Results** view for the analyzed *email* column.

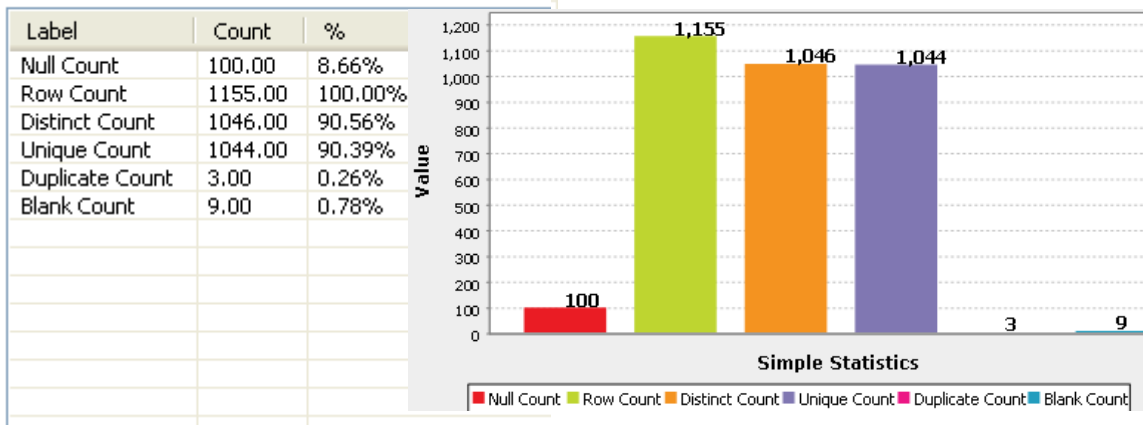
▼ Column:email

▶ Mode Indicator

▼ Frequency Statistics



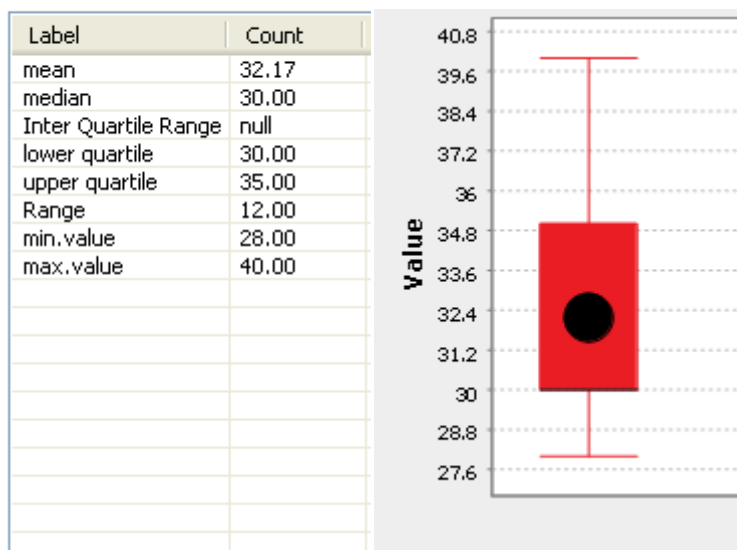
▼ Simple Statistics



Below is an example of a Summary Statistics table and graphic for an *age* analyzed column.

▼ Column:age

▼ Summary Statistics



2.6.2 How to analyze a set of columns in shortcut procedures

In **Talend Open Profiler**, you can analyze a set of columns more directly and quickly from:

- the table name file under the relevant **DB Connection** folder,
- the column name file under the relevant **DB Connection** folder.

When you analyze a set of columns directly from the table name file in the DB connection, you do not need to choose in the **[Create New Analysis]** wizard the type of analysis you want to carry out. However, you still need to set indicators for each analyzed column. Otherwise, all other procedural steps are exactly the same as in *How to analyze a set of columns on page 33*.

When you analyze a set of columns directly from the column name file in the DB connection, you do not need to:

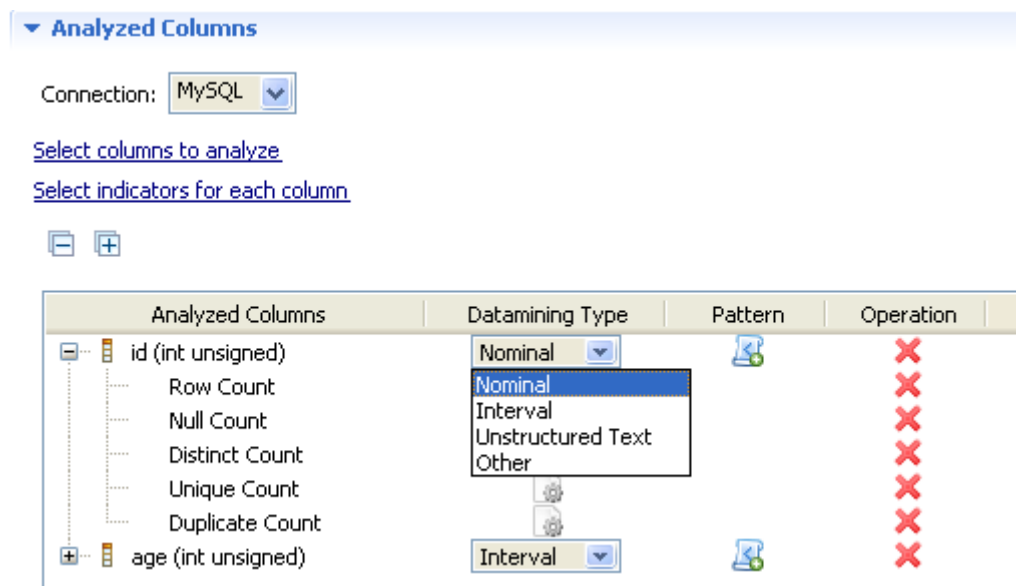
- choose in the **[Create New Analysis]** wizard the type of analysis you want to carry out,
- set indicators for each analyzed column.

Otherwise, all other procedural steps are exactly the same as in *How to analyze a set of columns on page 33*.

When you analyze a set of columns from the column name file under the relevant DB connection folder, you can directly create any of the following analyses for the selected column name(s): standard analysis, simple analysis, nominal value analysis and discrete analysis.

2.6.3 Data mining types

When you create a column analysis in **Talend Open Profiler**, you can see a **Datamining Type** box next to each of the columns that you set to be analyzed in the **Analyzed Columns** view of the open Column Analysis editor. The selected type in the box represents the data mining type of the associated column.



These data mining types help **Talend Open Profiler** choosing the appropriate metrics for the associated column since not all indicators (or metrics) can be computed on all data types.

Available data mining types in **Talend Open Profiler** are: Nominal, Interval, Unstructured Text and Other. Those data mining types are described in the below sections.

Nominal

Nominal data is categorical data which values/observations can be assigned a code in the form of a number where the numbers are simply labels. You can count but not order or measure nominal data.

In **Talend Open Profiler**, the mining type of textual data is set to nominal. For example, a column called *WEATHER* with the values: *sun*, *cloud* and *rain* is nominal.

And a column called *POSTAL_CODE* that have the values *52200* and *75014* is nominal as well in spite of the numerical values. Such data is of nominal type because it identifies a postal code in France. Computing mathematical quantities such as the average on these data is non sense. In such a case, you should set the data mining type of this column to “nominal”, because there is currently no way in **Talend Open Profiler** to automatically guess the correct type of data in cases like this.

The same is true for primary or foreign-key data. Keys are most of the time represented by numerical data, but their data mining type is Nominal.

Interval

This data mining type is used for numerical data and time data. Averages can be computed on this kind of data. In databases, sometimes numerical quantities are stored in textual fields. In **Talend Open Profiler**, it is possible to declare the data mining type of a textual column (e.g. a column of type *VARCHAR*) as Interval. In that case, the data should be treated as numerical data and summary statistics should be available.

Unstructured text

Unstructured Text is a new data mining type introduced by **Talend Open Profiler**. This data mining type is dedicated to handle unstructured textual data.

For example, the data mining type of a column called *COMMENT* that contains commentary text can not be Nominal, since the text in it is unstructured. Still, we could be interested in seeing the duplicate values of such a column and here comes the need for such a new data mining type.

Other

Other is another new data mining type in **Talend Open Profiler**. This type designs the data that **Talend Open Profiler** does not know how to handle yet.

2.7 Creating column comparison analysis

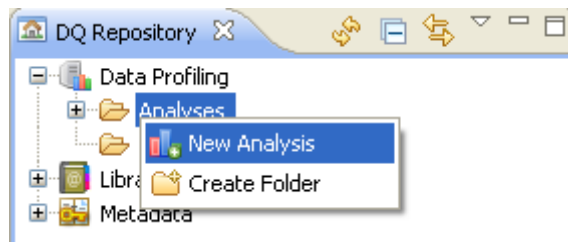
Talend Open Profiler allows you to better explore the relationship between tables as well as the quality of data through comparing identical columns in tables.

Prerequisite(s): **Talend Open Profiler** main window is open.

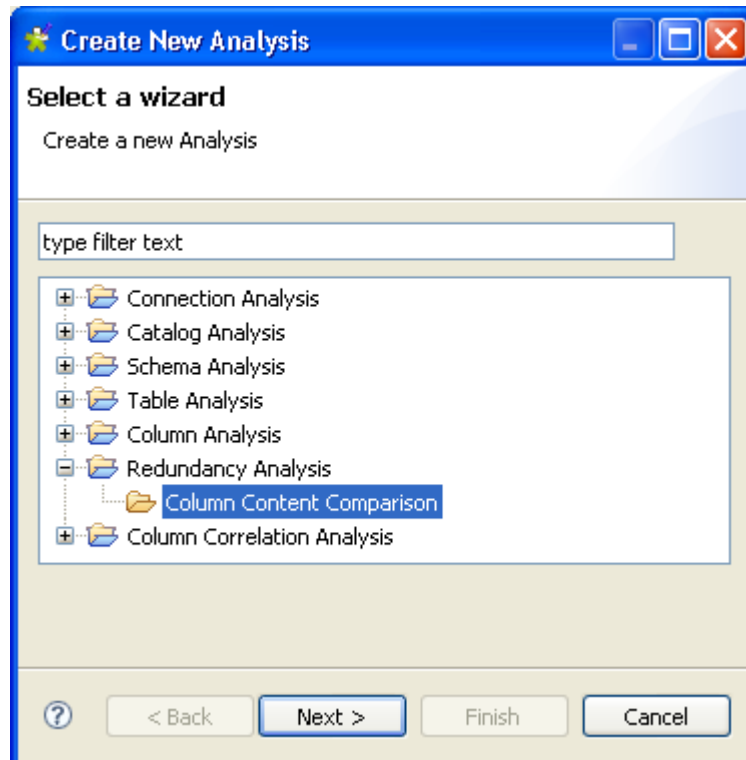
To create an analysis comparing two sets of columns:

- In the **DQ Repository** tree view, expand the **Data Profiling** folder.

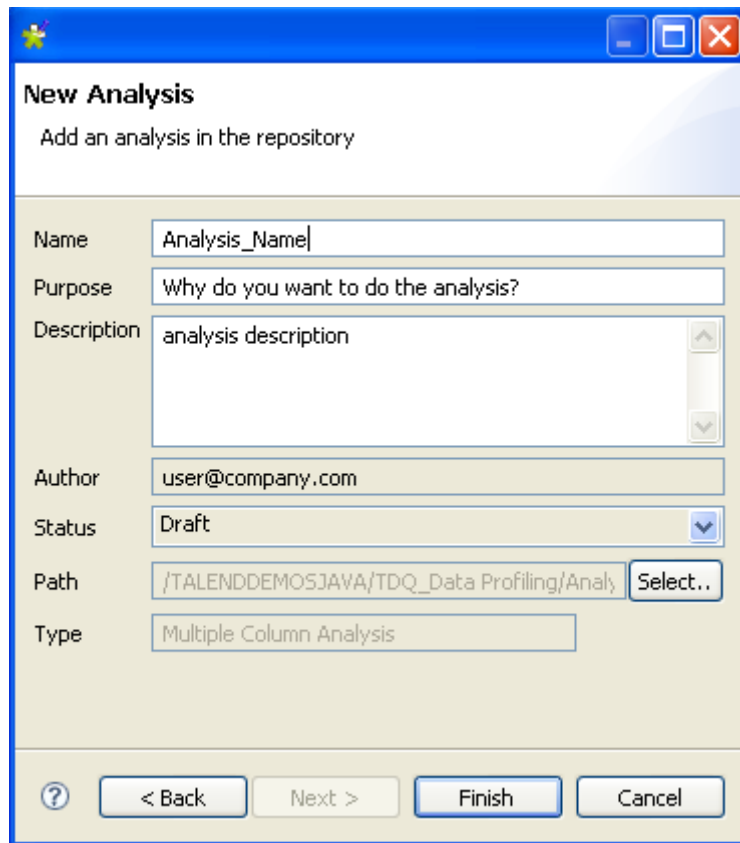
- Right-click the **Analysis** folder and select **New Analysis**.



The [Create New Analysis] wizard opens.

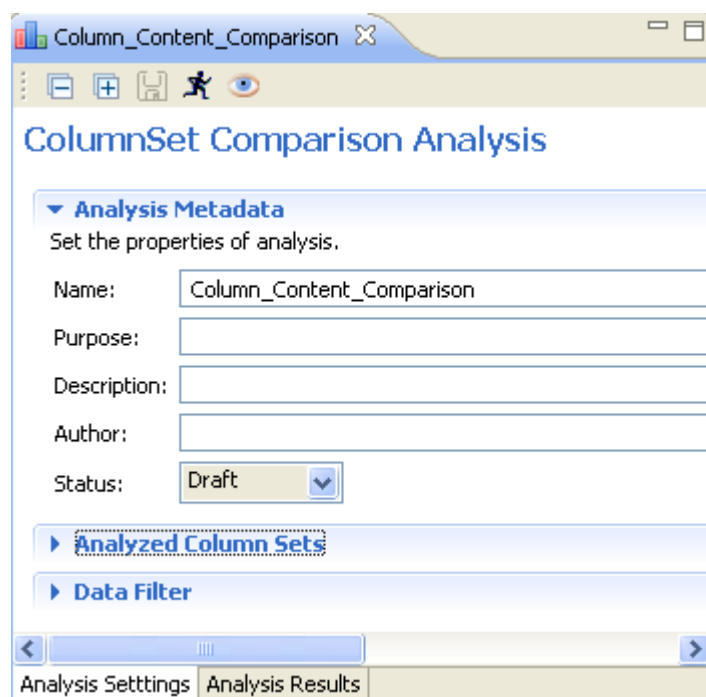


- Expand the **Redundancy Analysis** folder and select **Column Content Comparison**.
- Click the **Next** button to open a new view on the wizard.



- In the **Name** field, enter a name for the current analysis.
- If needed, set the analysis metadata (purpose, description and author name) in the corresponding fields and click **Finish** to close the [**Create New Analysis**] wizard.

A folder for the newly created analysis shows under **Analysis** in the **DQ Repository** tree view, and the Column Comparison Analysis editor opens with the defined analysis metadata.



- Click **Analyzed Column Sets** to display the view where to analyze two sets of identical columns.

▼ Analyzed Column Sets

Select tables or columns to compare.
 For tables comparison, select one table from A elements and one table from B elements.
 For columns comparison, select one or several columns from A and the same number of columns from B.

Compute only number of A rows not in B

▼ Left Columns

Select columns for A Set

☰ c1

☰ c2

▼ Right Columns

Select columns for B Set

☰ c1

☰ c2

- Click **Select columns for the A set** to open the [Column Selection] dialog box and select the first set of columns, or drag it directly from the **DQ Repository** tree view to the left column panel.
- Do the same to select the second set of columns or drag it to the right column panel.



To compute only the number of rows from the A set not present in the B set, select the **Compute only number of A rows not in B** check box.

- If needed, click **Data Filter** in the Column Comparison Analysis editor to display the view where you can set a filter on each of the column sets.

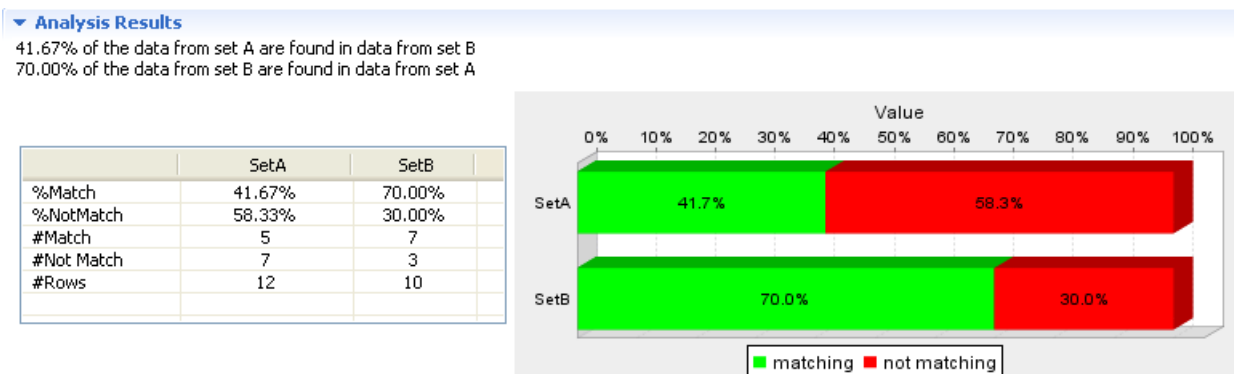
▼ Data Filter

Edit the data filter, left for ColumnSet A and right for ColumnSet B:

Where

Where

- Press **F6** to execute the column comparison analysis.
- Click the **Analysis Results** tab at the bottom of the Column Comparison Analysis editor to display the column comparison analysis results.



2.8 Managing column correlation analysis

Talend Open Profiler provides the possibility to explore relationships between two or more columns so that these relationships give a new interpretation in the case of correlation matrices.

Two types of column correlation analysis are possible. For more information, see *How to create numerical correlation analysis on page 46* and *How to create time correlation analysis on page 51*.

2.8.1 How to create numerical correlation analysis

This type of analysis analyzes correlation between nominal and interval columns and gives the result in a kind of a bubble chart.

A bubble chart is created for each selected numeric column. In a bubble chart, each bubble represents a distinct record of the nominal column. For example, a column "WHEATHER" with 3 distinct nominal instances "sunny" (11 records), "cloudy" (3 records), "rainy" (30records) will generate a bubble chart with 3 bubbles.

The vertical axis represents the average of the numeric column and the horizontal axis represents the number of records of each nominal instance. For example, a numerical column could be the "TEMPERATURE" in degrees Celsius. The average temperature would be 27 for the "sunny" instances, 23 for the "cloudy" instances and 17 for the "rainy" instances.

The more the bubble is near the left axis, the less confident we are in the average of the numeric column. For the "cloudy" example, there are only 3 records, hence the bubble is near the left axis. We cannot be confident in the average with only 3 records. When looking for data quality issues, these bubbles could indicate problematic values.

The bubbles near the top of the chart and those near the bottom of the chart may suggest data quality issues too. A too high or too low temperature in average could indicate a bad measure of the temperature.

In fact, in such a chart, outlier bubbles must be further investigated.

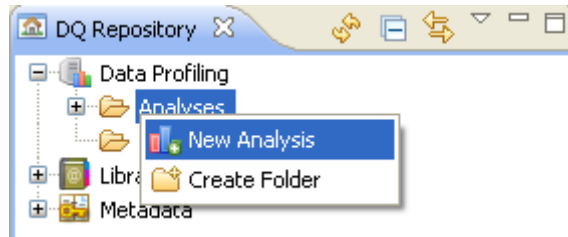
Another information in this chart is the size of the bubble. It represents the number of null numeric values. The more there are null values (e.g. in the "TEMPERATURE" column), the bigger will be the bubble size.

When several nominal columns are selected, the order of the columns plays a crucial role in this analysis. A series of bubbles (with one color) is displayed for the average temperature and the weather. Another series of bubbles is displayed for the average temperature and each record of the weather and the windy column.

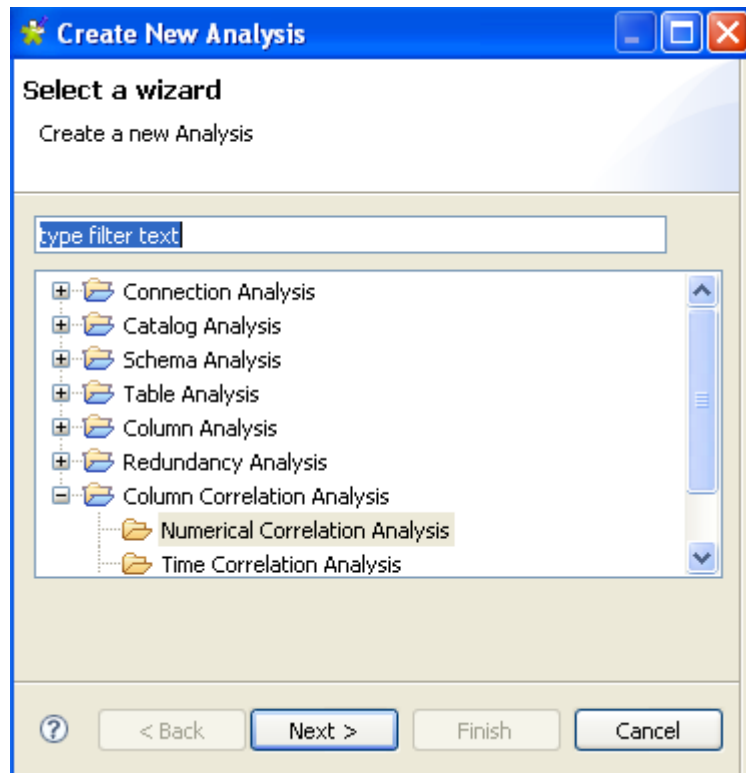
Prerequisite(s): **Talend Open Profiler** main window is open.

To create a numerical correlation analysis:

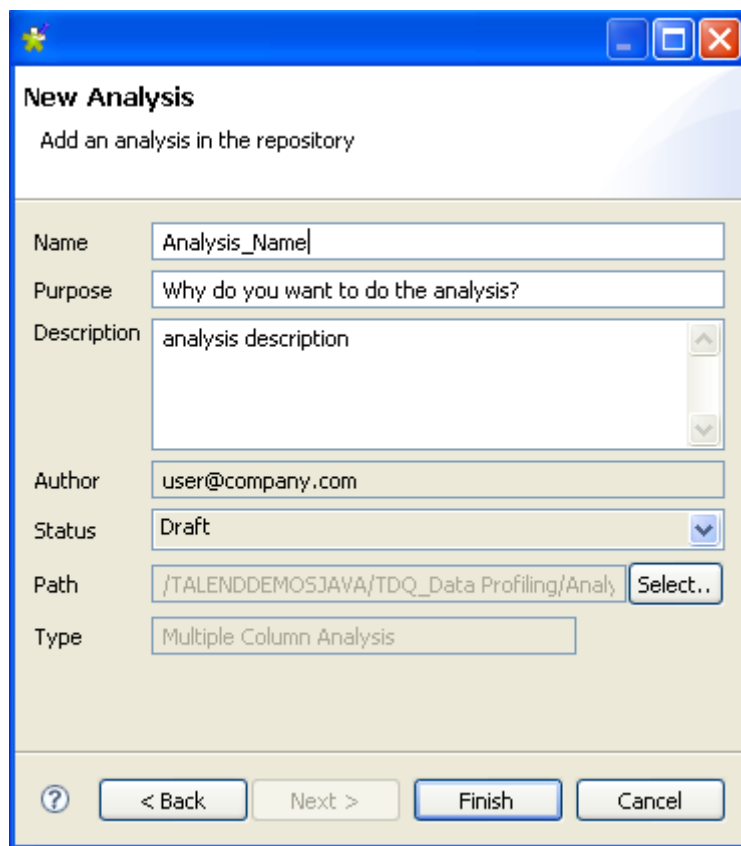
- In the **DQ Repository** tree view, expand the **Data Profiling** folder.
- Right-click the **Analysis** folder and select **New Analysis**.



The [Create New Analysis] wizard opens.



- Expand the **Column Correlation Analysis** folder and select **Numerical Correlation Analysis**.
- Click the **Next** button to open a new view on the wizard.



New Analysis
Add an analysis in the repository

Name: Analysis_Name|

Purpose: Why do you want to do the analysis?

Description: analysis description

Author: user@company.com

Status: Draft

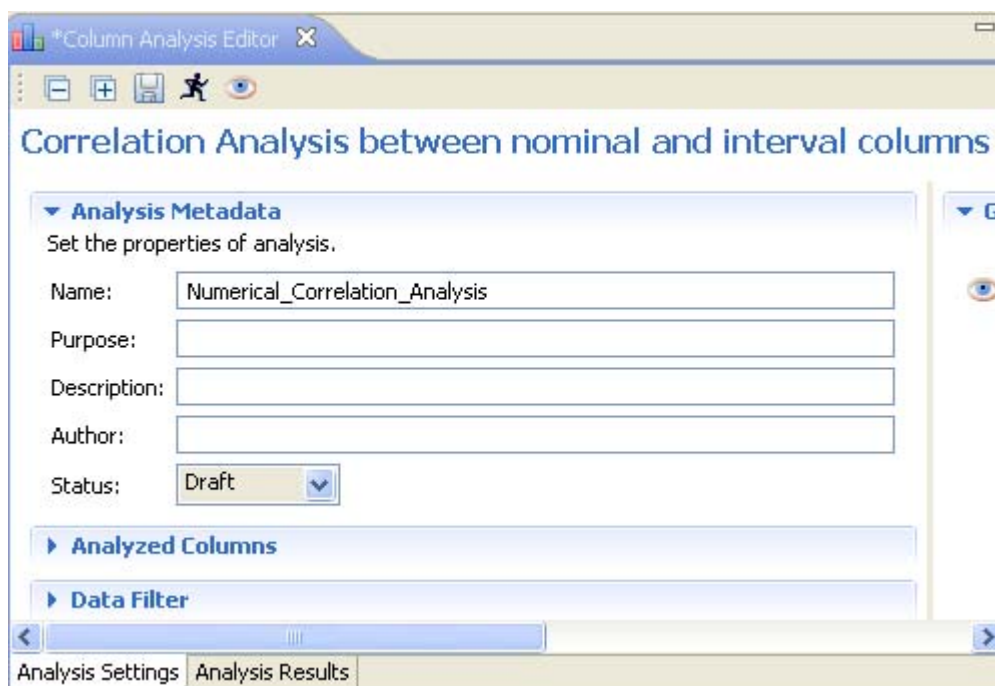
Path: /TALENDDDEMOSJAVA/TDQ_Data Profiling/Analy Select..

Type: Multiple Column Analysis

? < Back Next > Finish Cancel

- In the **Name** field, enter a name for the current analysis.
- If needed, set the analysis metadata (purpose, description and author name) in the corresponding fields and click **Finish** to close the **[Create New Analysis]** wizard.

A folder for the newly created analysis shows under **Analysis** in the **DQ Repository** tree view, and the Column Analysis editor opens with the defined analysis metadata.



*Column Analysis Editor

Correlation Analysis between nominal and interval columns

Analysis Metadata
Set the properties of analysis.

Name: Numerical_Correlation_Analysis

Purpose:

Description:

Author:

Status: Draft

Analyzed Columns

Data Filter

Analysis Settings Analysis Results

- Click **Analyzed Columns** to display the corresponding view.

▼ **Analyzed Columns**

Connection:

[Select columns to analyze](#)

Analyzed Columns	Datamining Type	Operation
STATE (varchar)	Nominal	✗
AGE (int)	Interval	✗
COMPANY (varchar)	Nominal	✗

- Click **Select columns to analyze** to open the [Column Selection] dialog box and select the columns, or drag them directly from the **DQ Repository** tree view into the **Analyzed Columns** area.



You can change your database connection by selecting another database from the **Connection** list. If the columns listed in the **Analyzed Columns** view do not exist in the new database connection you want to set, you will receive a warning message that enables you to continue or cancel the operation.

- If needed, click **Data Filter** in the Column Analysis editor to display the view where you can set a filter on the data of the analyzed columns.
- Press **F6** to execute the column comparison analysis and display the graphical result in the **Graphics** panel to the right of the editor.

▼ **Graphics**

[Refresh the graphics](#)

[-] **Column: AGE**

Count	Average (COMPANY STATE)	Average (COMPANY)
0-10	0-100	0-100
105	-	60

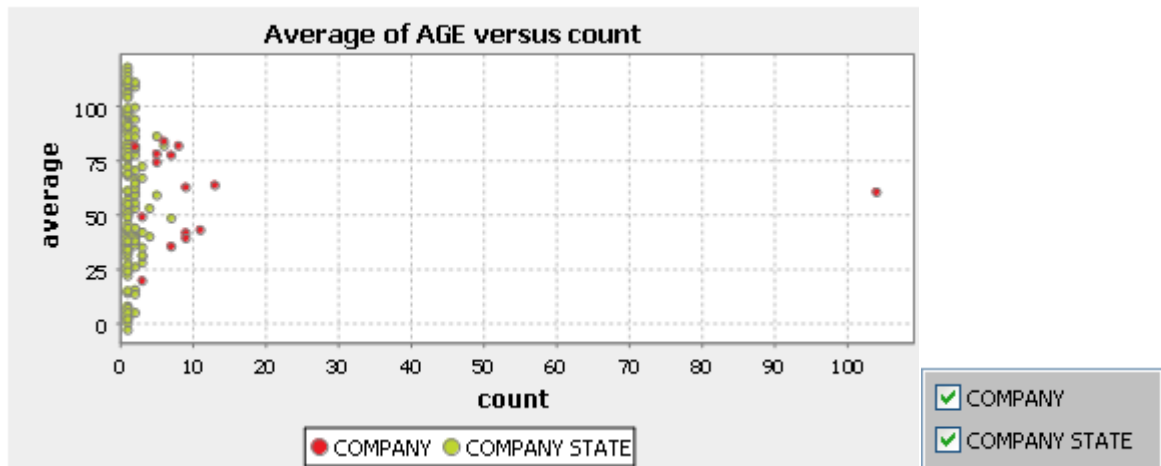
To access a more detailed view of the analysis results:

- Click the **Analysis Results** tab at the bottom of the Column Analysis editor to open the corresponding view.

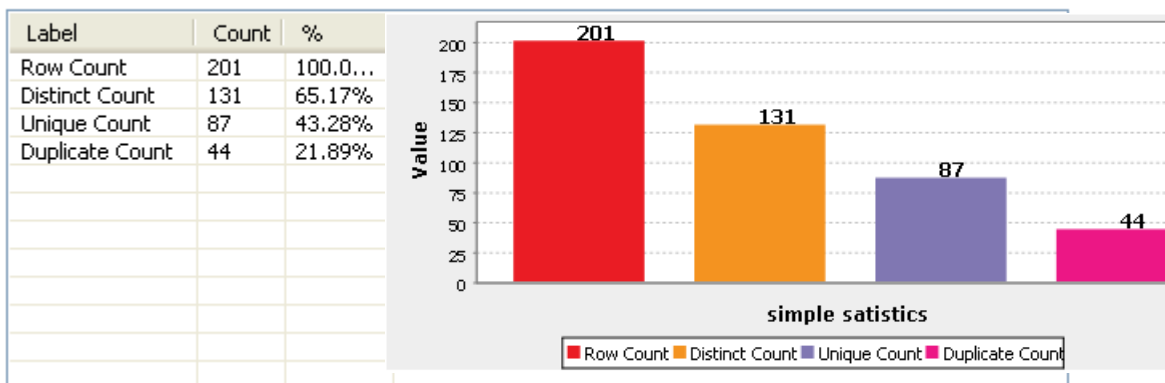
- Click the **Analysis Result** tab to display the analysis more detailed results including a tabular explanation of the graphic, simple statistics information and lists of analyzed data in the **Graphics**, **Simple Statistics** and **Data** sections.

▼ Graphics

☐ Column: AGE



▼ Simple Statistics



▼ Data

COMPANY	STATE	AVG(AGE)	COUNT(AGE)	SUM(CASE WHEN AGE IS NULL THEN 1 ELSE 0 END)	COUNT(*)
	Alabama	109.0000	2	0	2
	Alaska	81.8333	6	0	6
Altavista	Alaska	36.0000	1	0	1
Yahoo	Alaska	109.0000	1	0	1
	Arizona	99.0000	1	0	1
Google	Arizona	35.0000	1	0	1
Adobe	Arkansas	39.0000	2	0	2
Lycos	Arkansas	76.0000	1	0	1
Macromedia	Arkansas	83.0000	1	0	1
Yahoo	Arkansas	104.0000	1	0	1
	California	31.0000	3	0	3
Google	California	57.0000	1	0	1



In the **Graphics** section, you can clear the check box of the value(s) you want to hide in the graphic.



In the **Data** section, you can sort the data listed in the result table by simply clicking any column header in the result table.

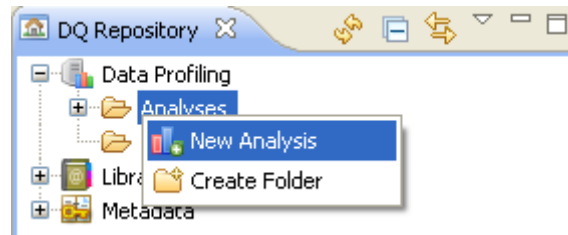
2.8.2 How to create time correlation analysis

This type of analysis analyzes correlation between nominal and date columns and gives the result in a gantt chart that illustrates the start and finish dates of each variable.

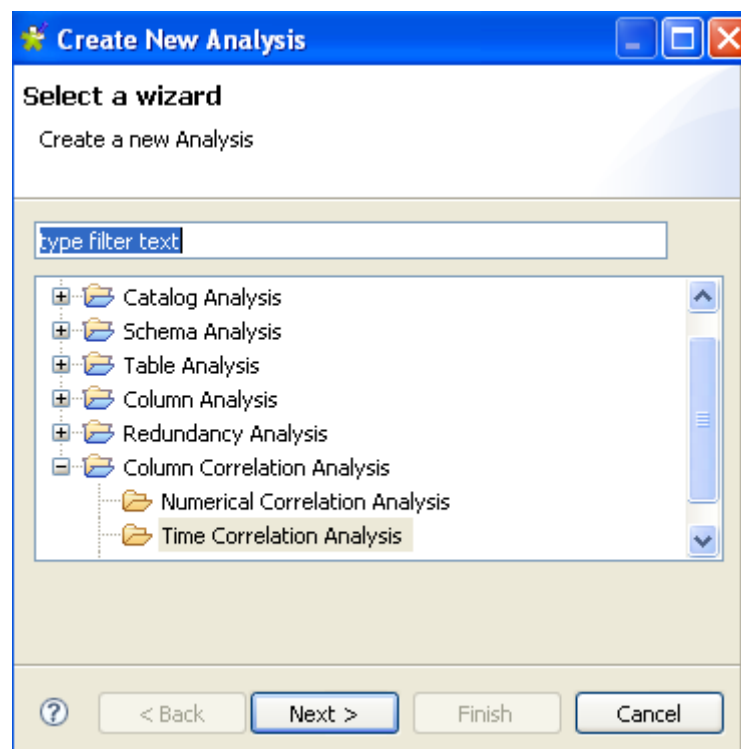
Prerequisite(s): **Talend Open Profiler** main window is open.

To create a time correlation analysis:

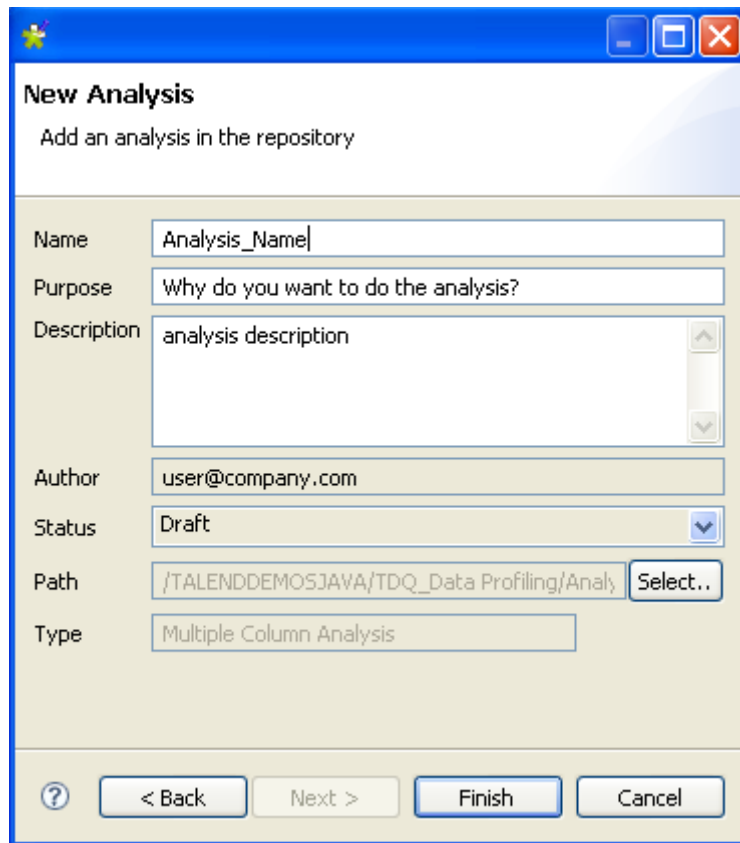
- In the **DQ Repository** tree view, expand the **Data Profiling** folder.
- Right-click the **Analysis** folder and select **New Analysis**.



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

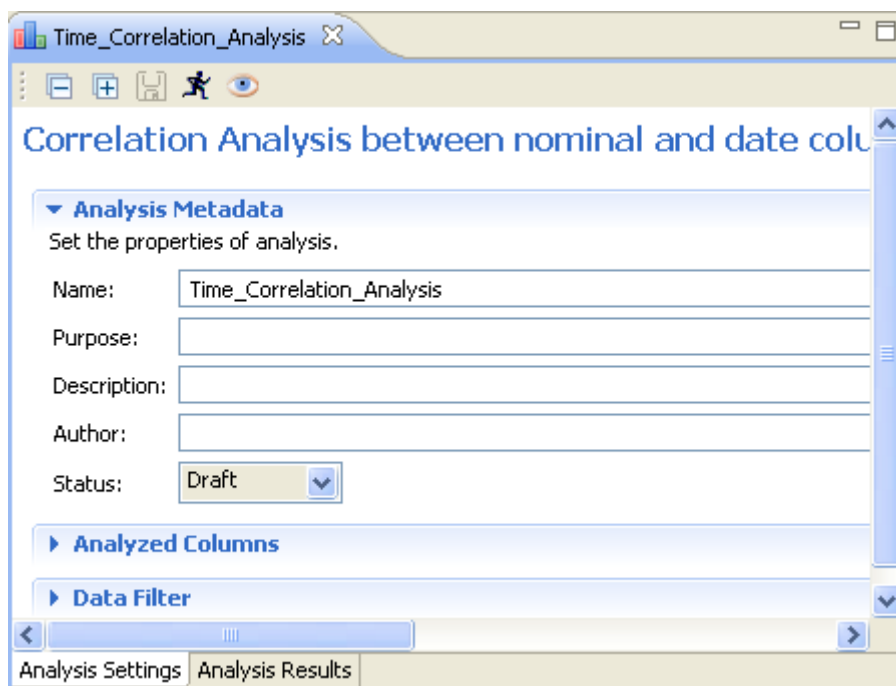


- Expand the **Column Correlation Analysis** folder and select **Time Correlation Analysis**.
- Click the **Next** button to open a new view on the wizard.



- In the **Name** field, enter a name for the current analysis.
- If needed, set the analysis metadata (purpose, description and author name) in the corresponding fields and click **Finish** to close the **[Create New Analysis]** wizard.

A folder for the newly created analysis shows under **Analysis** in the **DQ Repository** tree view, and the Time Correlation Analysis editor opens with the defined analysis metadata.



- Click **Analyzed Columns** to display the corresponding view.

▼ **Analyzed Columns**

Connection: MySQL

[Select columns to analyze](#)

Analyzed Columns	Datamining Type	Operation
SUBSCRIPTION_DATE (timesta	Interval	✗
COMPANY (varchar)	Nominal	✗

- Click **Select columns to analyze** to open the [Column Selection] dialog box and select the columns, or drag them directly from the **DQ Repository** tree view into the panel.



You can change your database connection by selecting another database from the **Connection** list. If the columns listed in the **Analyzed Columns** view do not exist in the new database connection you want to set, you will receive a warning message that enables you to continue or cancel the operation.

- If needed, click **Data Filter** in the Time Correlation Analysis editor to display the view where you can set a filter on the analyzed column set.
- Press **F6** to execute the column comparison analysis and display the graphical result in the **Graphics** panel to the right of the editor.

▼ **Graphics**

[Refresh the graphics](#)

☐ **Column: SUBSCRIPTION_DATE**

The gantt chart displays the minimal and maximal dates for each record of the selected nominal column. The x-axis represents time from Jul-2007 to Jul-2009. The y-axis lists companies: Adobe, Altavista, Apple Systems, Borland, Cakewalk, Chami, Finale, Google, Lavasoft, Lycos, Macromedia, Microsoft, Sibelius, and Yahoo. Red horizontal bars represent the date ranges for each company. A legend at the bottom indicates that the red bars represent the 'COMPANY' category.

This gantt chart displays the minimal and maximal dates for each record of the selected nominal column. It also highlights the bars which contain null values

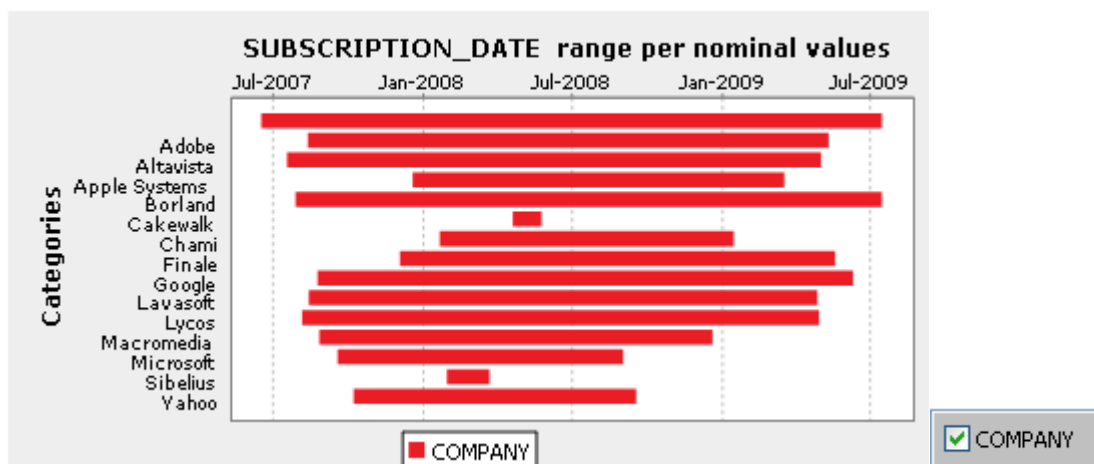
To access a more detailed view of the analysis results:

- Click the **Analysis Results** tab at the bottom of the Column Analysis editor to open the corresponding view.

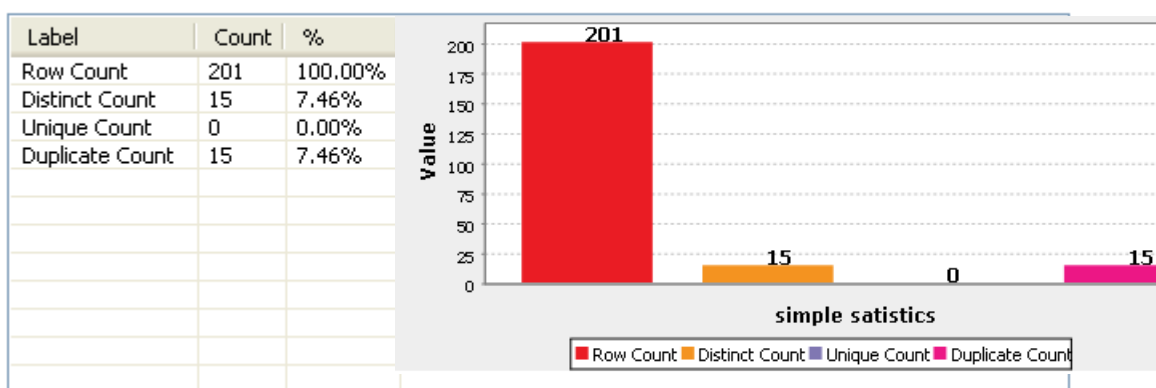
- Click the **Analysis Result** tab to display the analysis more detailed results including a tabular explication of the graphic, simple statistics information and lists of analyzed data in the **Graphics**, **Simple Statistics** and **Data** sections.

▼ Graphics

☐ Column: SUBSCRIPTION_DATE



▼ Simple Statistics



▼ Data

COMP...	MIN(SUBSCRIPTION_DATE)	MAX(SUBSCRIPTION_DATE)	COUNT(SUBSC...	SUM(CASE WHE...	COUNT(*)
	2007-06-17 15:39:55.0	2009-07-16 16:14:21.0	104	0	104
Adobe	2007-08-13 15:56:18.0	2009-05-12 10:03:17.0	13	0	13
Altavista	2007-07-19 04:08:00.0	2009-05-03 02:17:57.0	9	0	9
Apple Systems	2007-12-19 23:30:31.0	2009-03-18 22:15:52.0	5	0	5
Borland	2007-07-29 20:01:52.0	2009-07-16 16:14:21.0	11	0	11
Cakewalk	2008-04-20 12:08:08.0	2008-05-25 23:18:33.0	2	0	2
Chami	2008-01-22 01:52:15.0	2009-01-16 02:12:48.0	3	0	3
Finale	2007-12-04 10:45:59.0	2009-05-20 08:49:17.0	7	0	7
Google	2007-08-25 17:34:46.0	2009-06-11 08:16:29.0	9	0	9
Lavasoft	2007-08-14 19:27:58.0	2009-04-28 11:36:13.0	9	0	9
Lycos	2007-08-06 14:09:03.0	2009-04-30 11:46:22.0	8	0	8



In the **Graphics** section, you can clear the check box of the value(s) you want to hide in the graphic.



In the **Data** section, you can sort the data listed in the result table by simply clicking any column header in the result table.

2.8.3 How to create nominal correlation analysis

This type of analysis analyzes minimal correlations between nominal columns in the same table and gives the result in a chart.

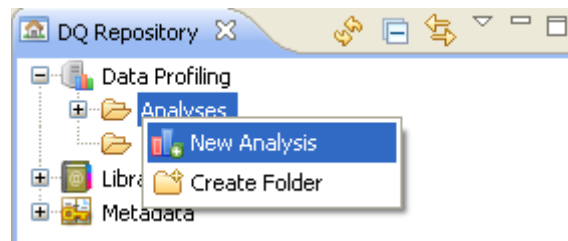
In the chart, each column will be represented by a node that has a given color. The correlations between the nominal values are represented by lines. The thicker the line is, the weaker the association is. Thicker lines can identify problems or correlations that need special attention. However, you can always inverse edge weight, that is give larger edge thickness to higher correlation, by selecting the **Inverse Edge Weight** check box below the nominal correlation chart.

The correlations in the chart are always pairwise correlations: show associations between pairs of columns.

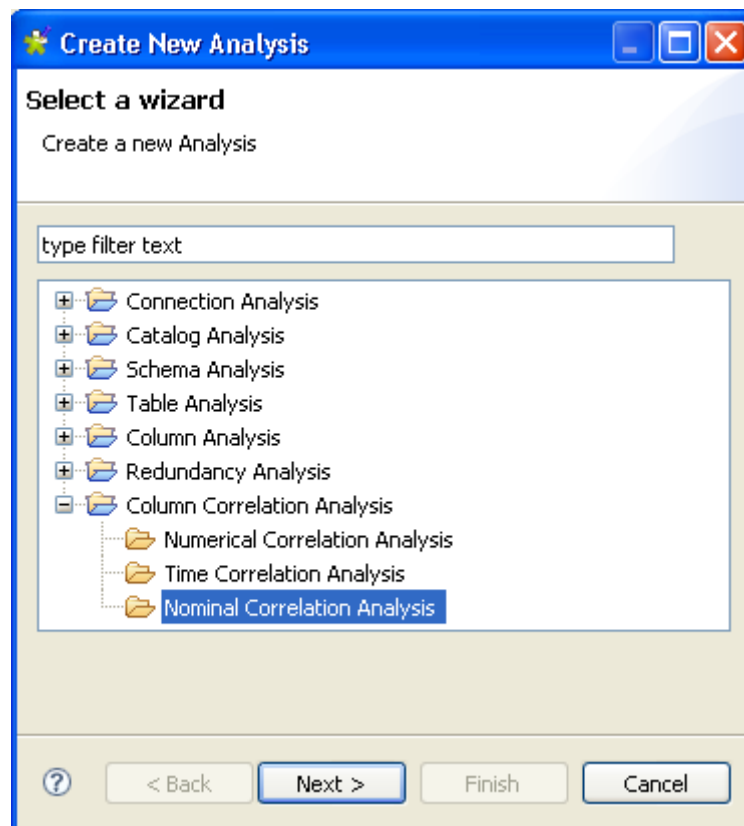
Prerequisite(s): Talend Open Profiler main window is open.

To create a nominal correlation analysis:

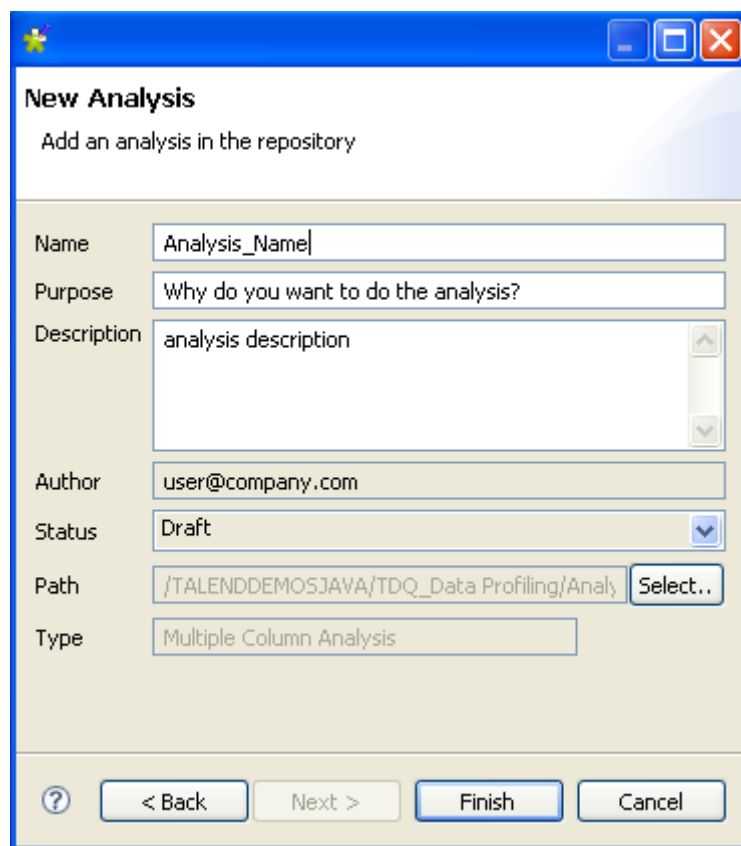
- In the **DQ Repository** tree view, expand the **Data Profiling** folder.
- Right-click the **Analysis** folder and select **New Analysis**.



The [Create New Analysis] wizard opens.



- Expand **Column Correlation Analysis** and select **Nominal Correlation Analysis**.
- Click the **Next** button to open a new view on the wizard.



New Analysis
Add an analysis in the repository

Name: Analysis_Name

Purpose: Why do you want to do the analysis?

Description: analysis description

Author: user@company.com

Status: Draft

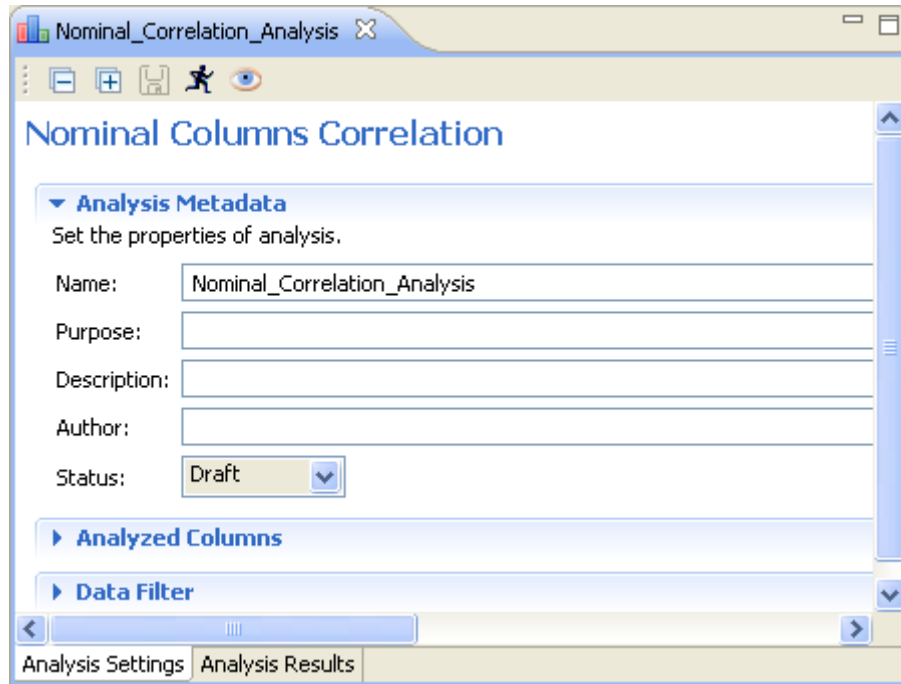
Path: /TALENDDemosJAVA/TDQ_Data Profiling/Anal Select..

Type: Multiple Column Analysis

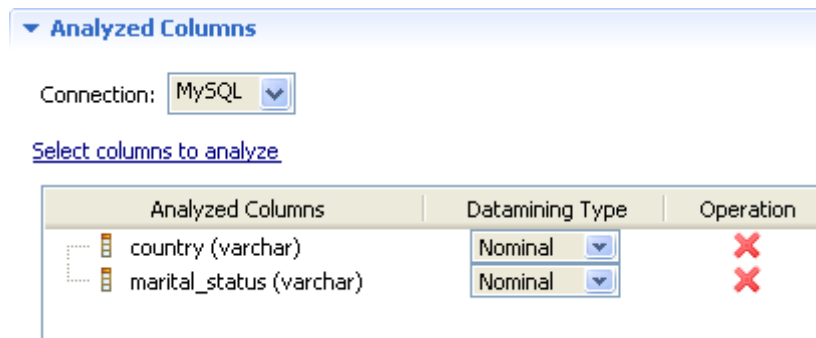
< Back Next > Finish Cancel

- In the **Name** field, enter a name for the current analysis.
- If needed, set the analysis metadata (purpose, description and author name) in the corresponding fields and click **Finish** to close the **[Create New Analysis]** wizard.

A folder for the newly created analysis shows under **Analysis** in the **DQ Repository** tree view, and the Nominal Correlation Analysis editor opens with the defined analysis metadata.



- Click **Analyzed Columns** to display the corresponding view.



- Click **Select columns to analyze** to open the [Column Selection] dialog box and select as many nominal columns as you want, or drag them directly from the **DQ Repository** tree view into the **Analyzed Columns** view.

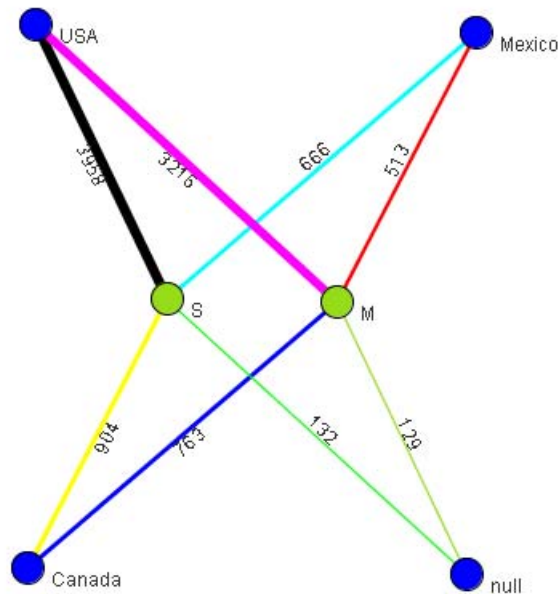


If you select too many columns, the analysis result chart will be very difficult to read.



You can change your database connection by selecting another database from the **Connection** list. If the columns listed in the **Analyzed Columns** view do not exist in the new database connection you want to set, you will receive a warning message that enables you to continue or cancel the operation.

- If needed, click **Data Filter** in the Nominal Correlation Analysis editor to display the view where you can set a filter on the data of the analyzed columns.
- Press **F6** to execute the nominal correlation analysis and display the graphical result in the **Graphics** panel to the right of the editor.



In the above chart, each value in the *country* and *marital-status* columns is represented by a node that has a given color. Nominal correlation analysis is carried out to see the relationship between the number of married or single people and the country they live in. Correlations are represented by lines, the thicker the line is, the higher the association is, since the **Inverse Edge Weight** check box is selected.



If you right-click anywhere in the chart and select **Show in full screen**, you open the chart in a separate window.

The buttons below the chart help you manage the chart display. The following table describes these buttons and their usage.

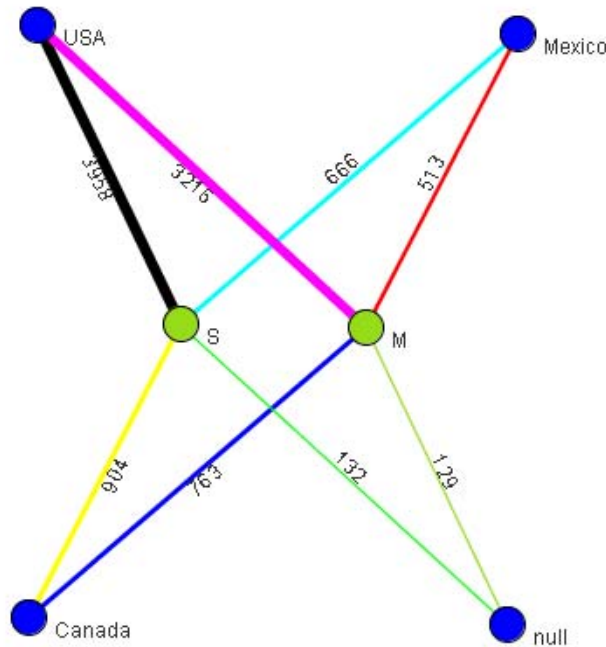
Button	Description
Filter Edge Weight	Move the slider to the right to (filter out edges with small weight) visualize the more important edges.
plus and minus	Click the plus/minus buttons to respectively zoom in and zoom out the chart size.
Reset	Click to put the chart back to its initial state.
Inverse Edge Weight	By default, the thicker the line is, the weaker the correlation is. Select this check box to inverse the current edge weight, that is give larger edge thickness to higher correlation.
Picking	Select this check box to be able to pick any node and drag it to anywhere in the chart.
Save Layout	Click this button to save the chart layout.
Restore Layout	Click this button to restore the chart to its previously saved layout.

To access a more detailed view of the analysis results:

- Click the **Analysis Results** tab at the bottom of the Column Analysis editor to open the corresponding view.

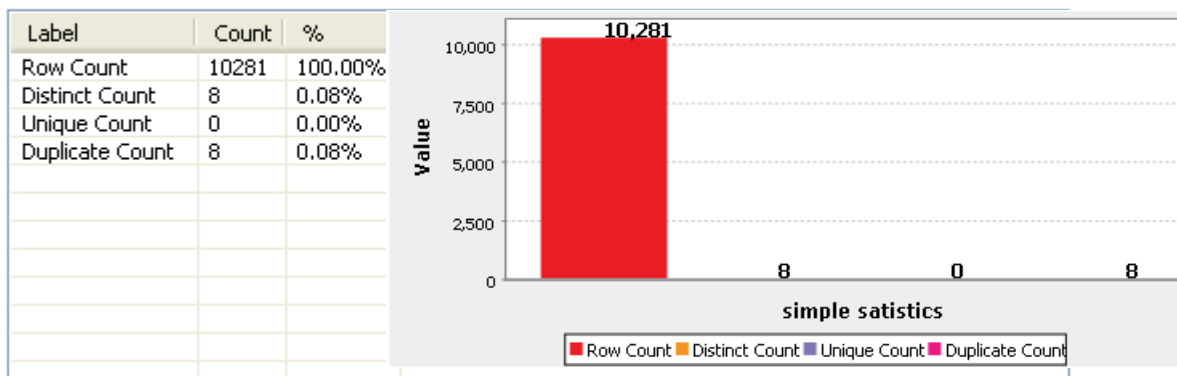
- Click the **Analysis Result** tab to display the analysis more detailed results including a tabular explication of the graphic, simple statistics information and lists of analyzed data in the **Graphics**, **Simple Statistics** and **Data** sections.

▼ Graphics



Filter edge weight inverse edge weight Picking

▼ Simple Statistics



▼ Data

marital_status	country	COUNT(*)
M	null	129
S	null	132
M	Canada	763
S	Canada	904
M	Mexico	513



In the **Data** section, you can sort the data listed in the result table by simply clicking any column header in the result table.

2.9 Managing table analyses

Talend Open Profiler allows you to better explore the quality of data in a database table through either:

- adding predefined data quality rules as indicators to table analysis, or
- detecting anomalies in column dependencies.

The following two sections describe in details each of the above procedures.

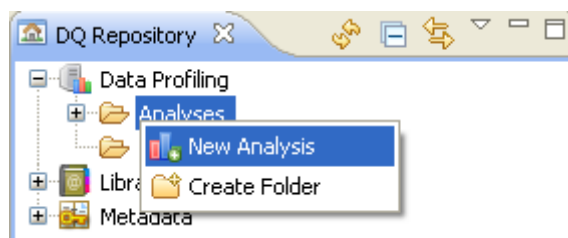
For more information about data quality rules, see *Managing data quality rules on page 80*.

2.9.1 How to create a table analysis with DQ rules

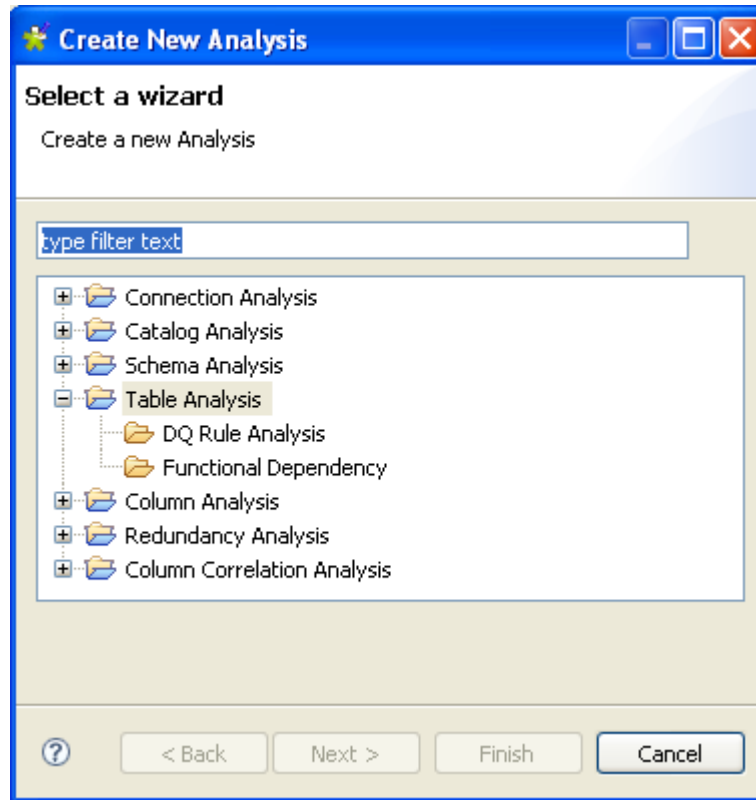
Prerequisite(s): **Talend Open Profiler** main window is open. At least one DQ rule is created. For more information about creating DQ rules, see *How to create a DQ rule on page 80*.

To create a table analysis using a predefined data quality rule:

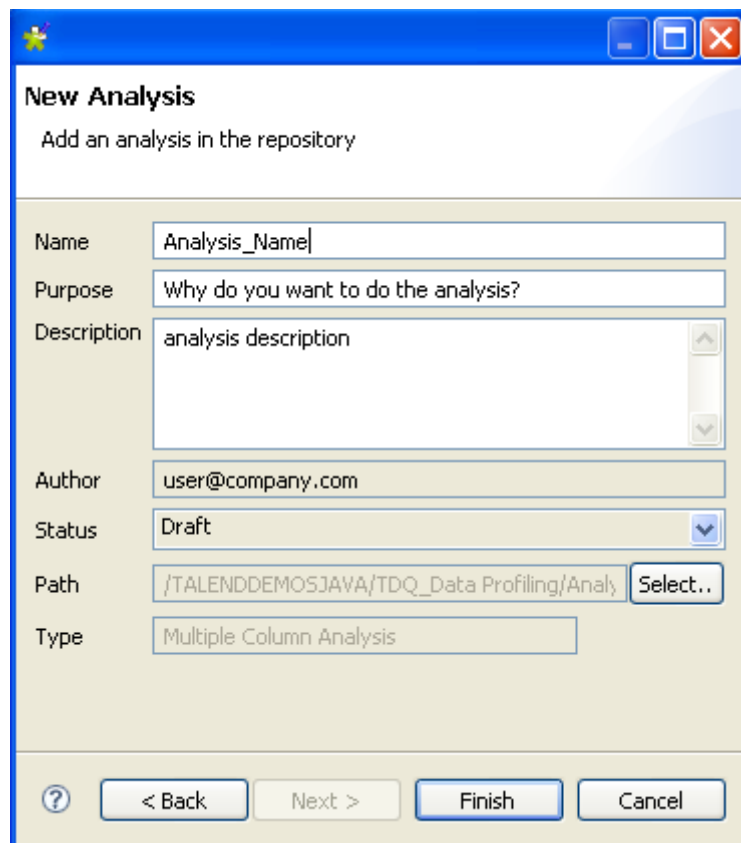
- In the **DQ Repository** tree view, expand the **Data Profiling** folder.
- Right-click the **Analysis** folder and select **New Analysis**.



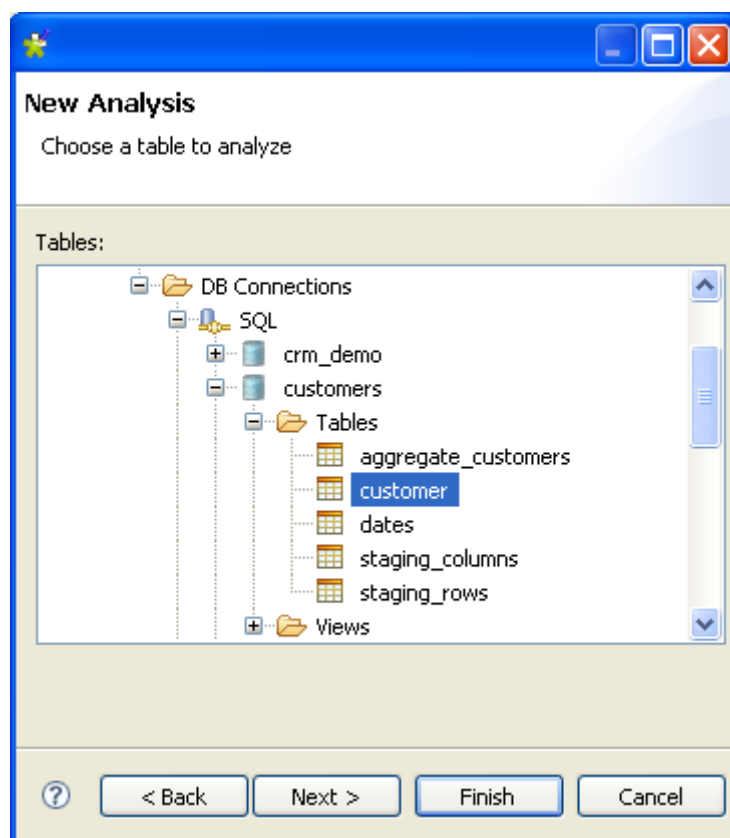
The **[Create New Analysis]** wizard opens.



- Expand the **Table Analysis** folder and select **DQ Rule Analysis**.
- Click the **Next** button to open a new view on the wizard.



- In the **Name** field, enter a name for the current analysis.
- If needed, set the analysis metadata (purpose, description and author name) in the corresponding fields and click **Next** to open a new view on the wizard.

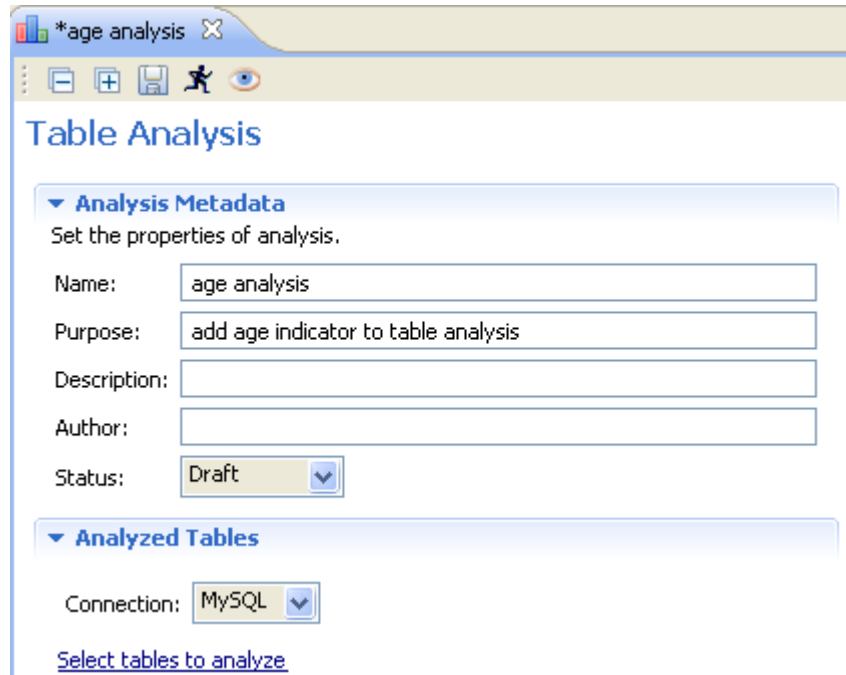


- Expand **DB Connections**, browse to the table(s) to be analyzed and select it/them.
- Click **Finish** to close the **[Create New Analysis]** wizard.

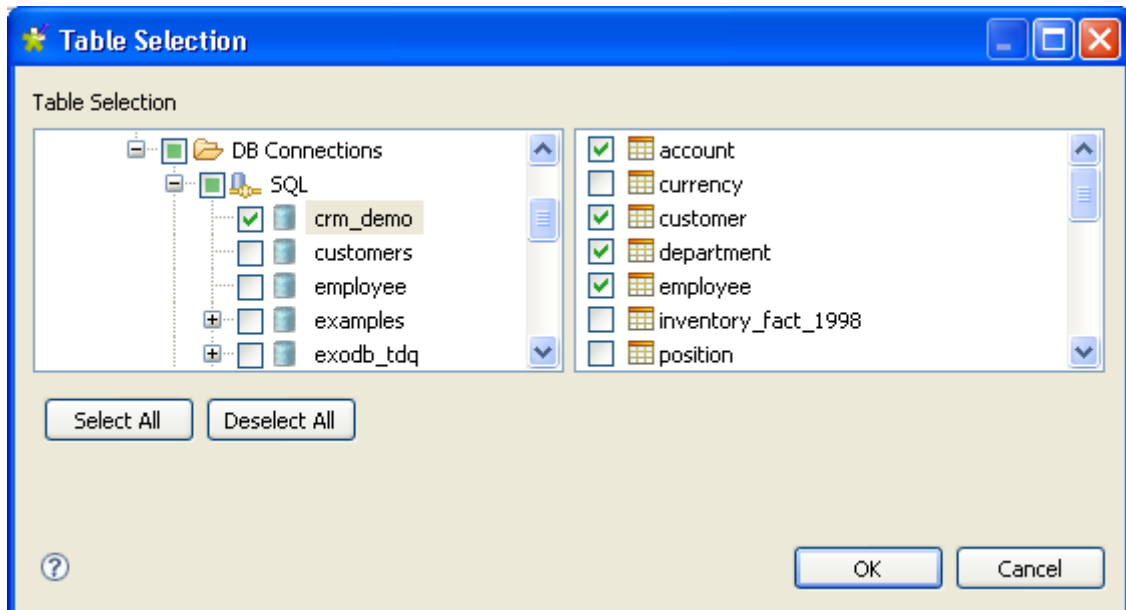


You can directly select the data quality rule you want to add to the current analysis by clicking the **Next** button in the **[New Analysis]** wizard or you can do that at later stage in the **Analyzed Tables** view as shown in the following steps.

A folder for the newly created table analysis shows under **Analysis** in the **DQ Repository** tree view, and the Table Analysis editor opens with the defined metadata.



- Click the **Analyzed Tables** tab to open the **Analyzed Tables** view.
- Click **Select tables to analyze** to open the [Table Selection] dialog box and modify the selection and/or select new tables.

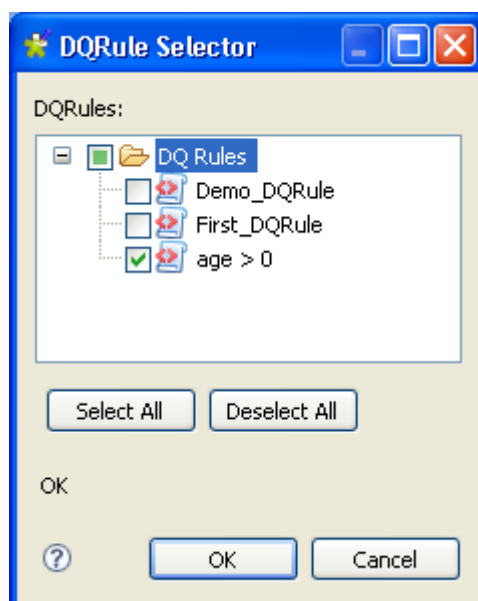


- Expand **DB Connections** and browse to the table(s) you want to analyze.
- Select the tables' check boxes and click **OK** to close the dialog box. The selected tables display in the **Analyzed Tables** view.



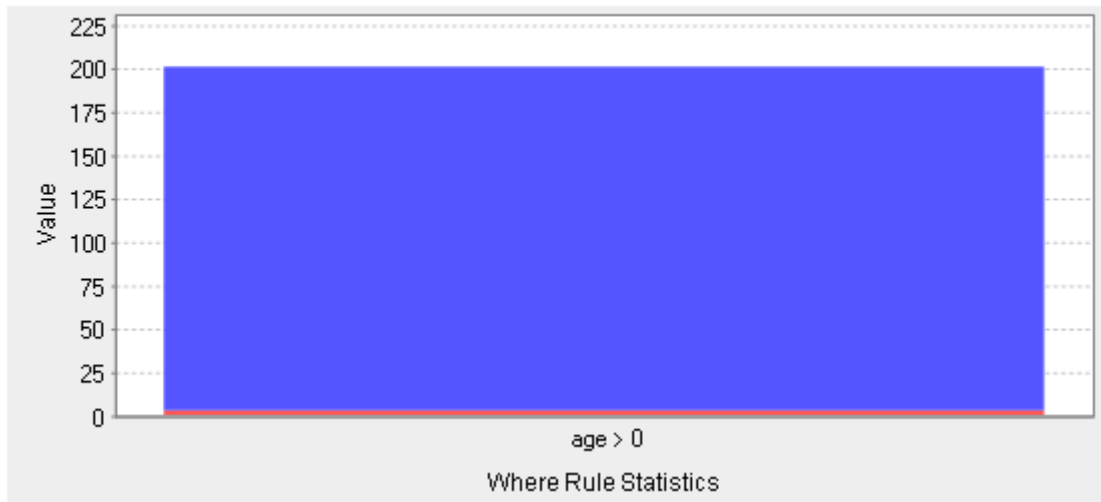
You can change your database connection by selecting another database from the **Connection** list. If the tables listed in the **Analyzed Tables** view do not exist in the new database connection you want to set, you will receive a warning message that enables you to continue or cancel the operation.

- Click the **DQ Rule** icon next to the table name where you want to add the DQ rule to display the **[DQRule Selector]** dialog box.



- Expand the **DQ Rules** folder and select the check box(es) of the predefined DQ rule(s) you want to use on the corresponding table and click **OK** to close the dialog box.
- If needed, click **Data Filter** in the Table Analysis editor to display the view where you can set a filter on the data of the analyzed tables.
- Press **F6** to execute the current analysis.
A progress information pop-up opens to confirm that the operation is in progress. Table analysis results display in the **Graphics** panel to the right.

Table:top_custom



The current analysis evaluated age records in the selected tables against the set DQ rule and returned the results indicating “dirty” age records by a red line.

2.9.2 How to create a column functional dependency analysis

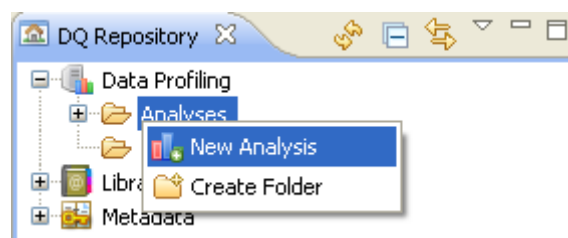
This type of analysis helps you to detect anomalies in your column dependencies through defining columns as either “determinant” or “dependent” and then analyzing values in dependant columns against those in determinant columns. This type of analysis detects to what extent a value in a determinant column functionally determines another value in a dependant column.

This can help you identify problems in your data, such as values that are not valid. For example, if you analyze the dependency between a column that contains United States Zip Codes and a column that contains states in the United States. The same Zip Code should always have the same state. Running the functional dependency analysis on these two columns will show if there are any violations of this dependency.

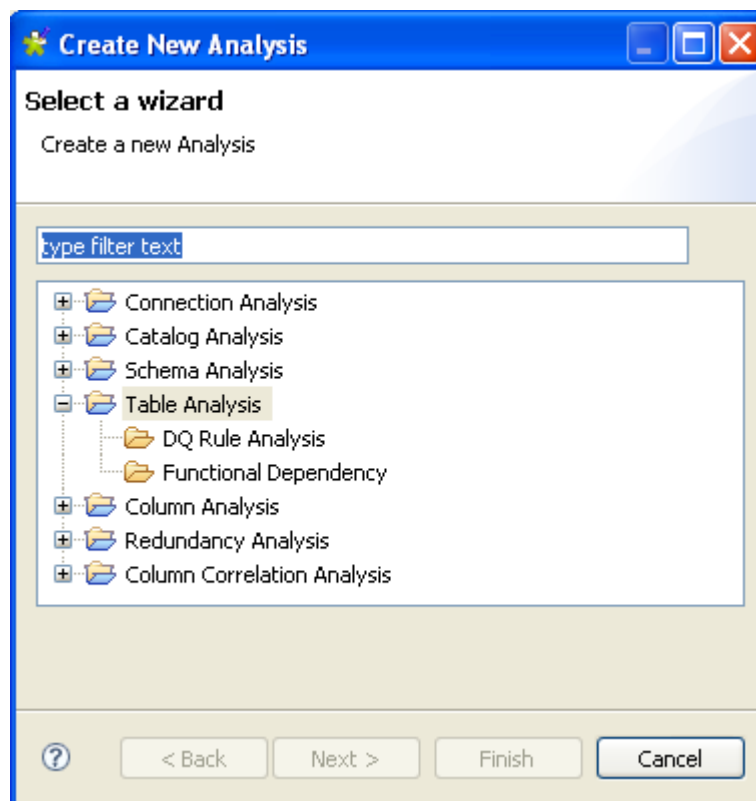
Prerequisite(s): Talend Open Profiler main window is open.

To create a column functional dependency analysis:

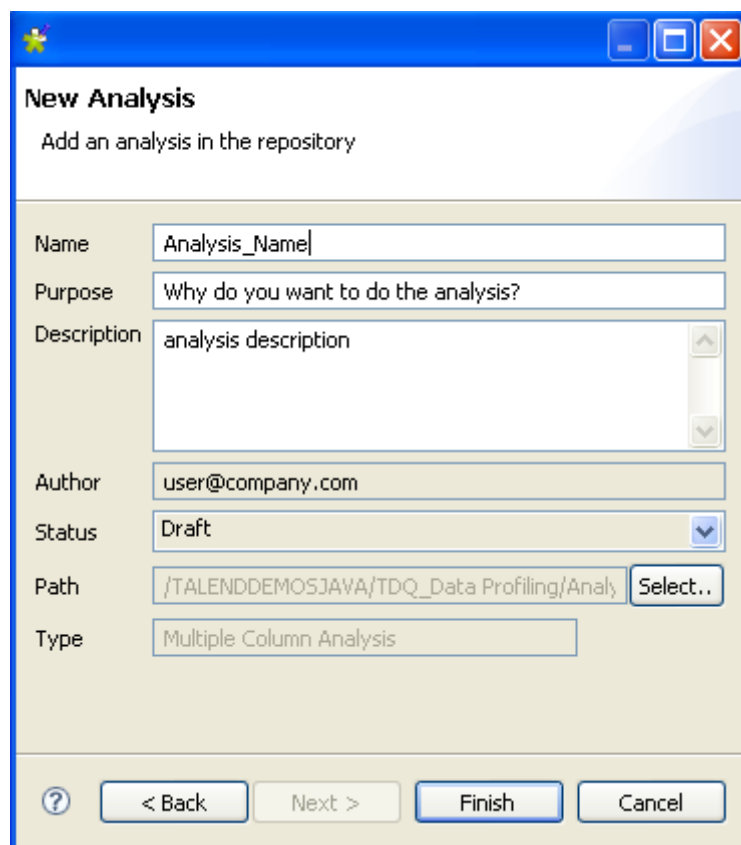
- In the **DQ Repository** tree view, expand the **Data Profiling** folder.
- Right-click the **Analysis** folder and select **New Analysis**.



The [Create New Analysis] wizard opens.

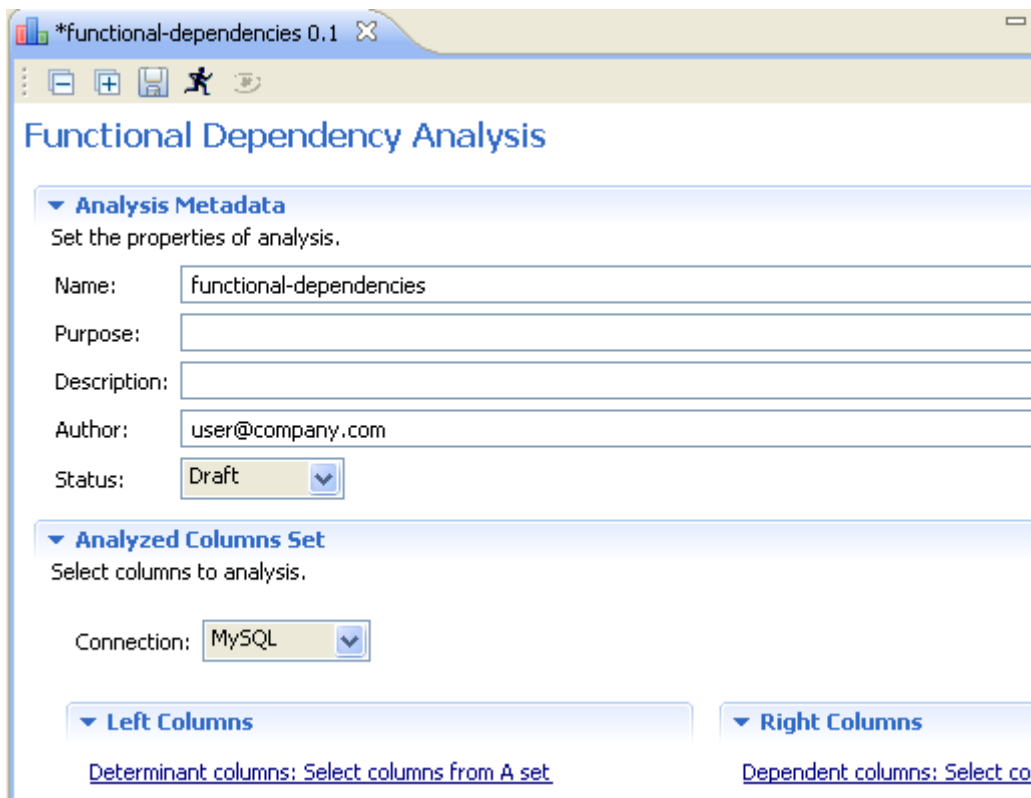


- Expand the **Table Analysis** folder and select **Functional Dependency**.
- Click the **Next** button to open a new view on the wizard.



- In the **Name** field, enter a name for the current analysis.
- If needed, set the analysis metadata (purpose, description and author name) in the corresponding fields and click **Next** to open a new view on the wizard.
- Click **Finish** to close the **[Create New Analysis]** wizard.

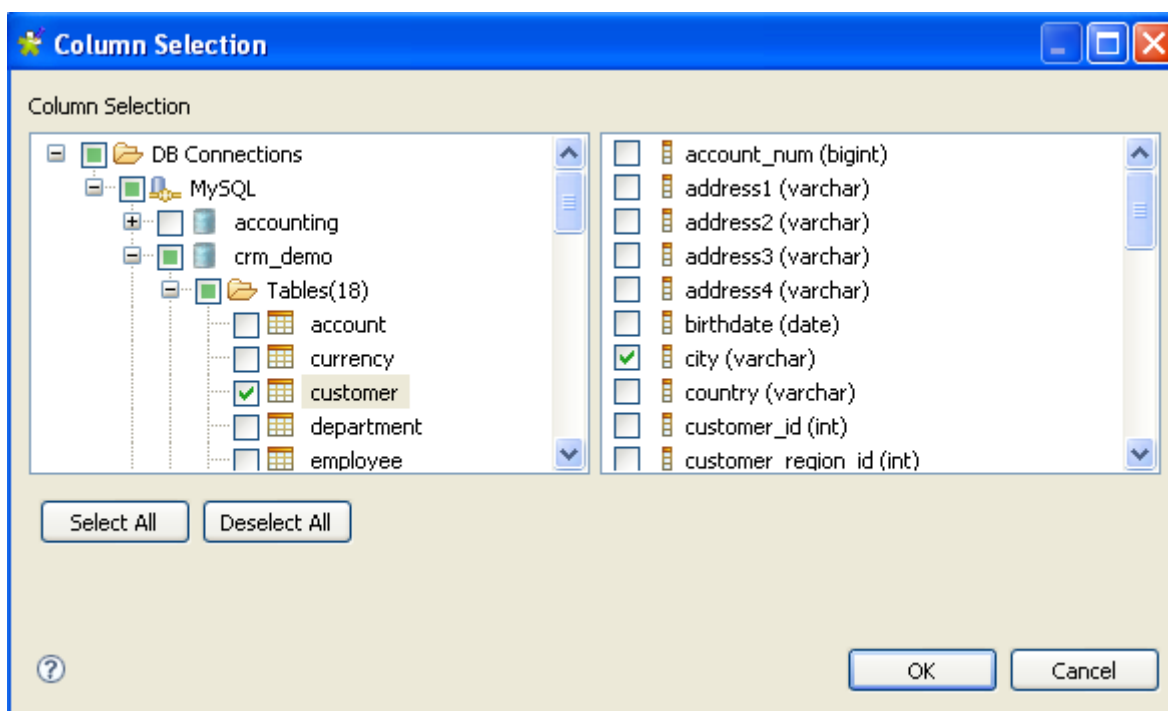
A folder for the newly created functional dependency analysis shows under **Analysis** in the **DQ Repository** tree view, and the Functional Dependency Analysis editor opens with the defined metadata.



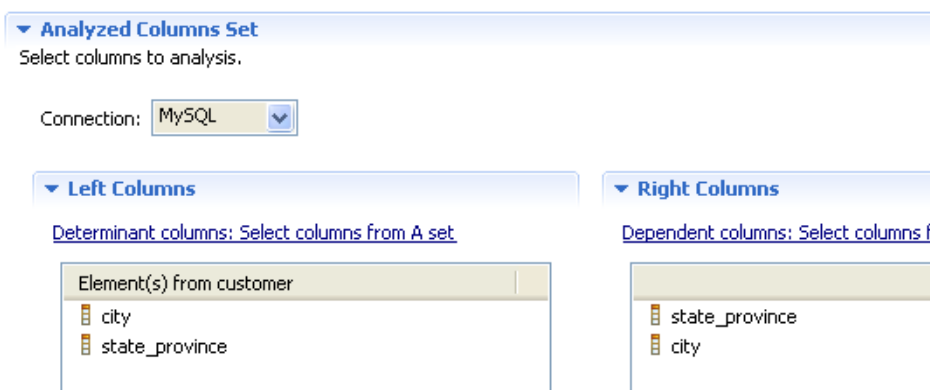
The screenshot shows a window titled '*functional-dependencies 0.1'. The main content area is titled 'Functional Dependency Analysis'. It contains two main sections:

- Analysis Metadata**: A section with the instruction 'Set the properties of analysis.' containing fields for Name (filled with 'functional-dependencies'), Purpose, Description, Author (filled with 'user@company.com'), and Status (a dropdown menu set to 'Draft').
- Analyzed Columns Set**: A section with the instruction 'Select columns to analysis.' containing a 'Connection' dropdown menu set to 'MySQL'. Below this are two panels: 'Left Columns' with the text 'Determinant columns: Select columns from A set' and 'Right Columns' with the text 'Dependent columns: Select col'.

- Click the **Analyzed Column Set** tab to open the corresponding view.
- Click **Select columns for A set** to open the **[Column Selection]** dialog box where you can select the first set of columns, or drag it directly from the **DQ Repository** tree view to the left column panel.



- In the [Column Selection] dialog box, expand **DB Connections** and the relevant database connection folder in succession and browse to the column(s) you want to analyze.
- Select the column(s)' check boxes and click **OK** to close the dialog box. The selected column(s) display(s) in the **Left Columns** panel of the **Analyzed Columns Set** view.



- Do the same to select the second set of columns or drag it to the right column panel.



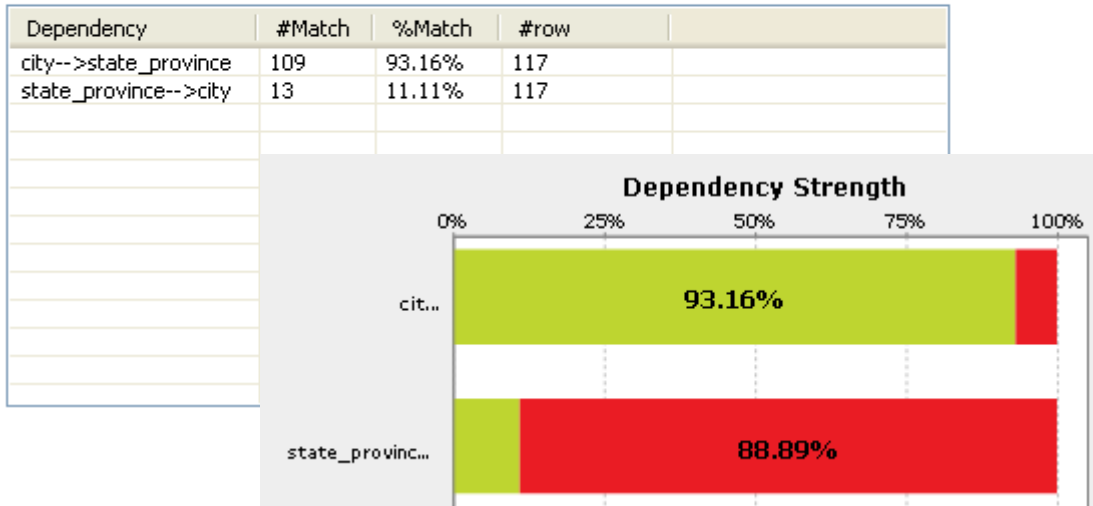
You can change your database connection by selecting another database from the **Connection** list. If the columns listed in the **Analyzed Columns Set** view do not exist in the new database connection you want to set, you will receive a warning message that enables you to continue or cancel the operation.

- Press **F6** to execute the current analysis. A progress information pop-up opens to confirm that the operation is in progress. The results of column functional dependency analysis display in the **Analysis Results** view.
- Click the **Analysis Results** tab at the bottom of the editor to display the corresponding view where you can see column functional dependency analysis results.

Analysis Result

▶ Analysis Summary

▼ Analysis Results



The above functional dependency analysis evaluated the records present in the *city* column and those present in the *state_province* column against each other to see if state names match to the listed city names and vice versa.

This functional dependency analysis returns functional dependency strength for each determinant column. It indicate “dirty” records by red color.



The presence of null values in either of the two analyzed columns will lessen the “dependency strength”. The system does not ignore null values, but rather calculates them as values that violates the functional dependency.

2.10 Adding a task to an item

In **Talend Open Profiler**, you can add tasks to different items either:

- in the **DQ Repository** tree view on catalogs, schemas, tables, columns and created analyses, or
- on columns, or patterns and indicators set on columns directly in the current Column Analysis editor.

For example, you can add a general task to any item in a database connection via the **Metadata** node in the **DQ Repository** tree view. You can add a more specific task to the same item defined in the context of an analysis through the **Analyses** node. And finally, you can add a task to a column in an analysis context (also to a pattern or an indicator set on this column) directly in the current Column Analysis editor.

The procedure to add a task to any of these items is exactly the same.

For examples on how to add a task to an item, see *How to add a task to a column in a database connection on page 70*, *How to add a task to an item in a specific analysis context on page 71*, *How to add a task to an indicator in a column analysis on page 72*.

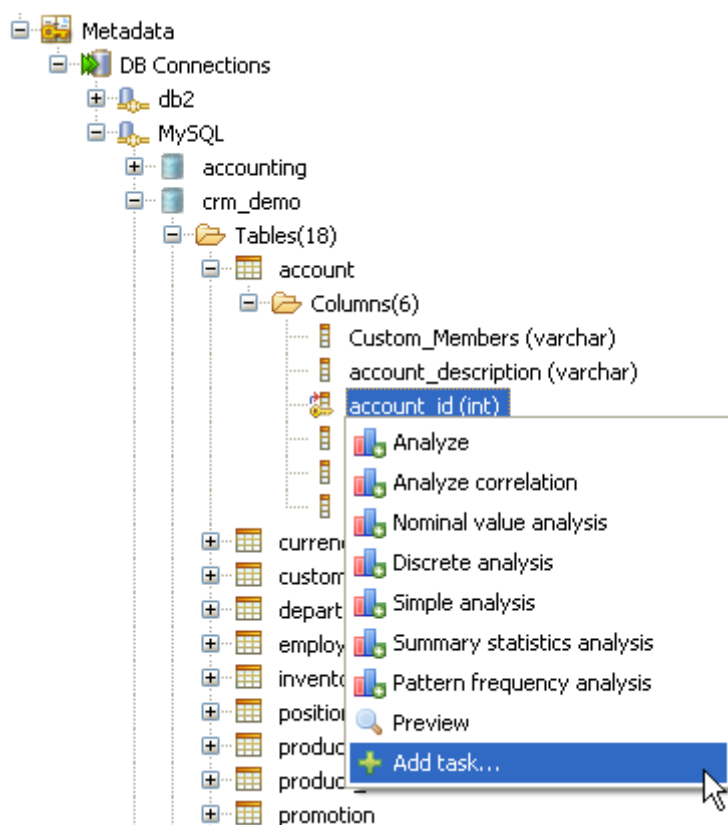
Adding tasks to such items will list these tasks in the **Tasks** list accessible through the **Window - Show view...** combination. Later, you can open the editor corresponding to the relevant item by double clicking the appropriate task in the **Tasks** list.

2.10.1 How to add a task to a column in a database connection

Prerequisite(s): **Talend Open Profiler** main window is open. You have created at least one database connection.

To add a task to a column in a database connection:

- In the **DQ Repository** tree view, expand the **Metadata** and the **DB Connections** folders in succession.
- Navigate to the column you want to add a task to, `account_id` in this example.
- Right-click the `account_id` and select **Add task...** from the drop-down list.



The **[Properties]** dialog box opens showing the metadata of the selected column.

- In the **Description** field, enter a short description for the task you want to carry on the selected item.
- On the **Priority** list, select the priority level and then click **OK** to close the dialog box. The created task is added to the **Tasks** list.

5 items						
	!	Description	Resource	Path	Locat...	Type
<input checked="" type="checkbox"/>		check for null values?	account_id	TOP_DEFAULT_PRJ/TD...	D:/TO...	Task
<input checked="" type="checkbox"/>		pattern task	Address	TOP_DEFAULT_PRJ/TD...	D:/TO...	Task
<input type="checkbox"/>		chek this report?	catalog_Analysis	TOP_DEFAULT_PRJ/TD...	D:/TO...	Task
<input type="checkbox"/>		test this indicator?	Address	TOP_DEFAULT_PRJ/TD...	D:/TO...	Task
<input type="checkbox"/>		test this pattern?	product_id	TOP_DEFAULT_PRJ/TD...	D:/TO...	Task



- You can double-click a task to open the editor where this task has been set.
- You can select the task check box once the task is completed in order to be able to delete it.
- You can filter the task view according to your needs using the options in a drop-down list accessible through the drop-down arrow on the top-right corner of the **Tasks** panel.

For information on how to display the **Tasks** list, see *How to delete a completed task on page 73*.

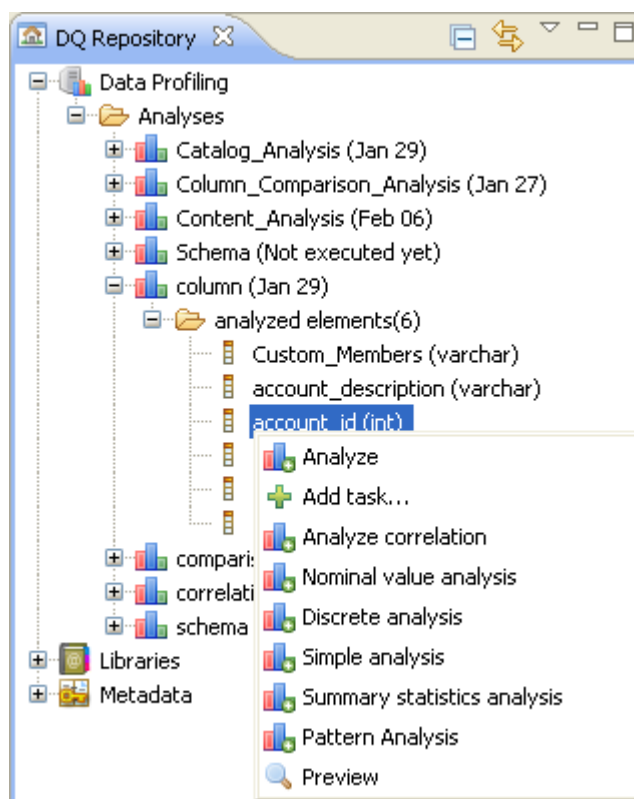
2.10.2 How to add a task to an item in a specific analysis context

The below procedure gives an example of adding a task to a column in an analysis context. You can follow the same steps to add tasks to other elements in the created analyses.

To add a task to an item in a specific analysis context:

Prerequisite(s): **Talend Open Profiler** main window is open. The appropriate analysis has been created.

- In the **DQ Repository** tree view, expand the **Analyses** node and in an already created analysis, navigate to the item you want to add a task to, the *account_id* column in this example.
- Right-click *account id* and select **Add task...** from the drop-down list.



Continue following the same steps as in *How to add a task to a column in a database connection* to add a task to `account_id` in the selected analysis context.

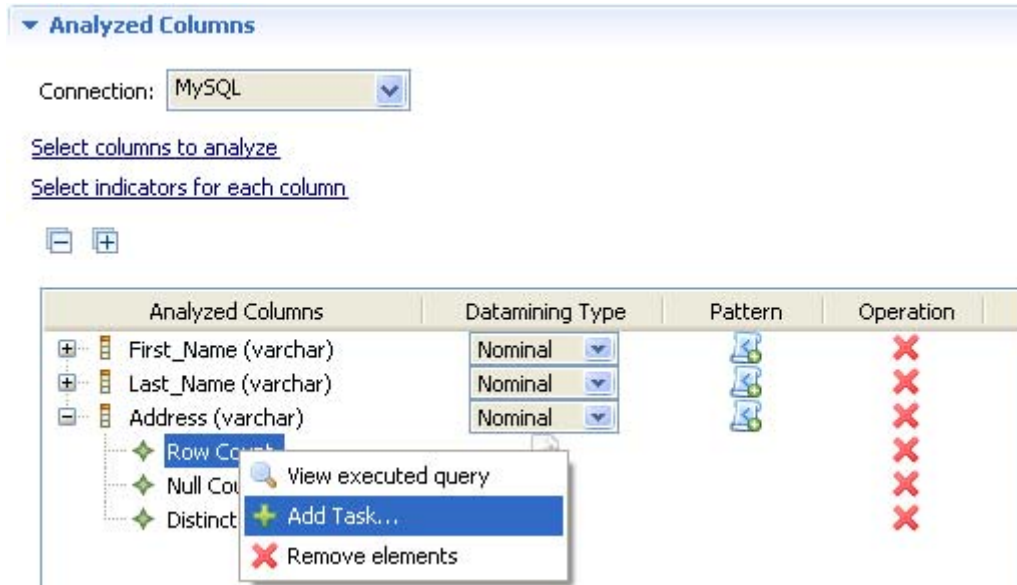
2.10.3 How to add a task to an indicator in a column analysis

You can add a task to indicators set on columns directly in the open Column Analysis editor of **Talend Open Profiler**. This can be used as a reminder to modify the indicator or to flag a problem that needs to be solved later, for example.

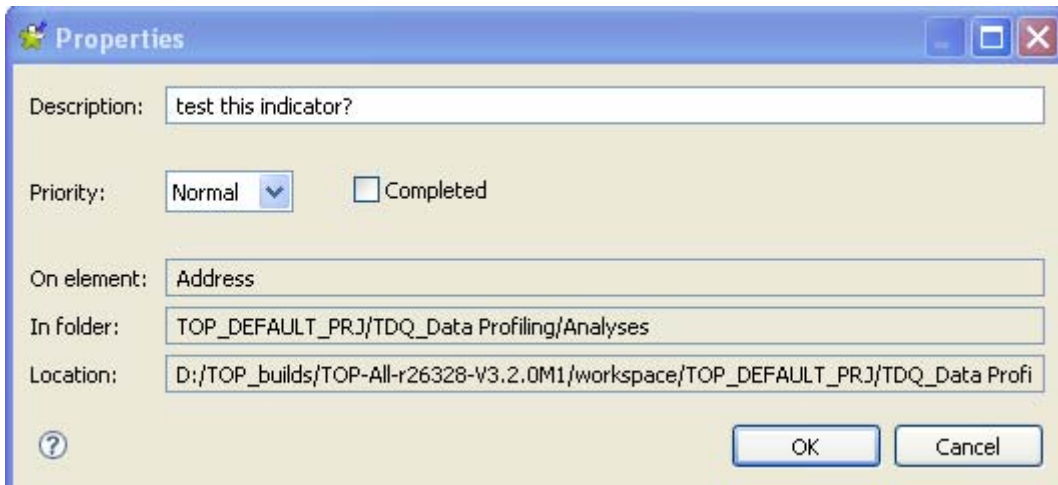
Prerequisite(s): **Talend Open Profiler** main window is open. An analysis of a set of columns is open in the Column Analysis editor. At least one indicator is set for the columns to be analyzed.

To add a task to an indicator:

- In the open Column Analysis editor, click **Analyzed columns** to open the relevant view.
- In the **Analyzed Columns** list, right-click the indicator name and select **Add task...** from the drop-down list.



The **[Properties]** dialog box opens showing the metadata of the selected indicator.



- In the **Description** field, enter a short description for the task you want to attach to the selected indicator.
- On the **Priority** list, select the priority level and then click **OK** to close the dialog box. The created task is added to the **Tasks** list.

For information about how to display the **Tasks** list, see *How to delete a completed task on page 73*.

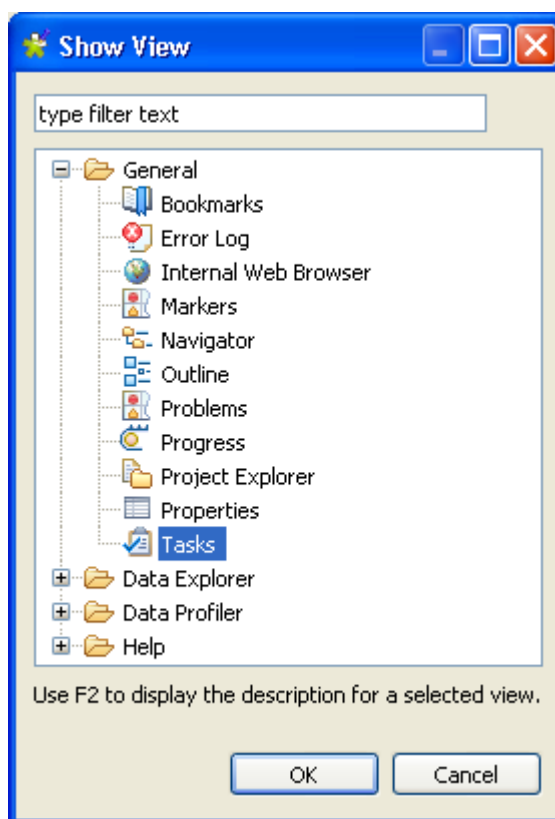
2.10.4 How to delete a completed task

When tasks' goals are met, you can delete these task from the **Tasks** list after labeling them as completed.

Prerequisite(s): **Talend Open Profiler** main window is open. At least, one task is added to an item.

To delete a completed task:

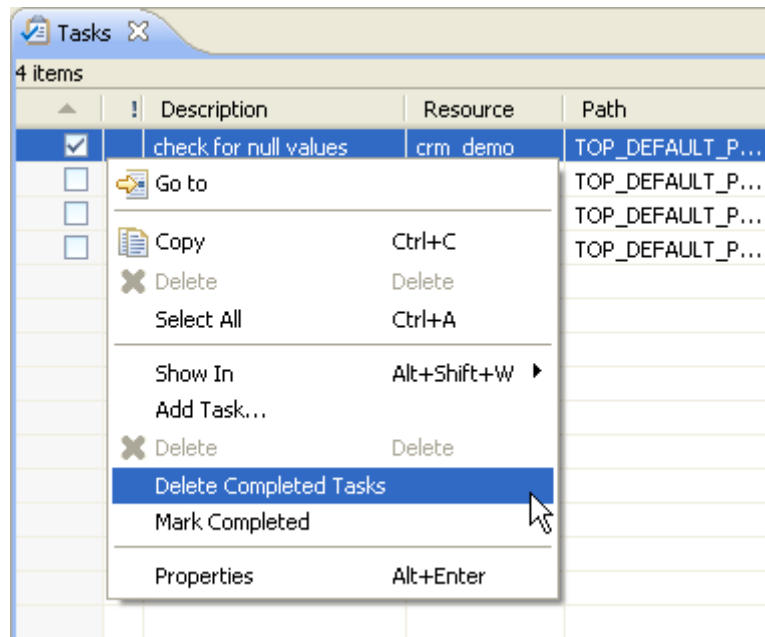
- On the menu bar of **Talend Open Profiler**, select **Window - Show view...** . The **[Show View]** dialog box displays.



- Select **Tasks** from the **General** list and click **OK** to close the dialog box. The **Tasks** panel opens in **Talend Open Profiler** listing the added task(s).

	!	Description	Resource	Path	Locat...	Type
<input checked="" type="checkbox"/>		check for null values?	account_id	TOP_DEFAULT_PRJ/TD...	D:/TO...	Task
<input checked="" type="checkbox"/>		pattern task	Address	TOP_DEFAULT_PRJ/TD...	D:/TO...	Task
<input type="checkbox"/>		chek this report?	catalog_Analysis	TOP_DEFAULT_PRJ/TD...	D:/TO...	Task
<input type="checkbox"/>		test this indicator?	Address	TOP_DEFAULT_PRJ/TD...	D:/TO...	Task

- When one or more tasks are completed or you do not need them anymore, select the check boxes next to each of the tasks' description lines and right-click anywhere in the **Tasks** list.



- In the drop-down list, select **Delete Completed Tasks**. A confirmation message displays to validate the operation.
- Click **OK** to close the confirmation message. All tasks marked as completed are deleted from the **Tasks** list.

2.11 Generic procedures for all types of analyses

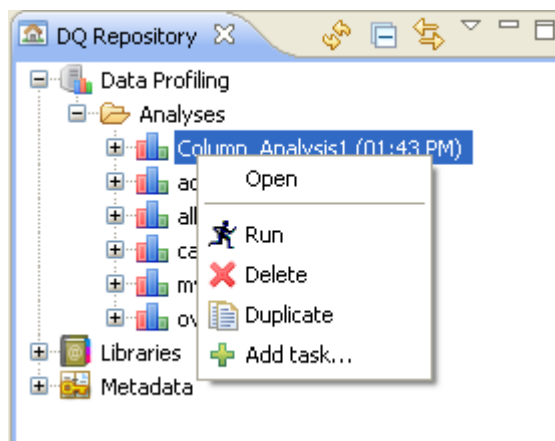
The below are generic procedure with similar functionalities implemented for all types of analyses.

2.11.1 How to open an analysis

Prerequisite(s): **Talend Open Profiler** main window is open. At least one analysis type has been created.

To open an analysis:

- In the **DQ Repository** tree view, expand the **Data Profiling** and **Analyses** folders in succession.
- Either, double-click the analysis you want to open, or
- Right-click the analysis you want to open and select **Open** in the drop-down list.



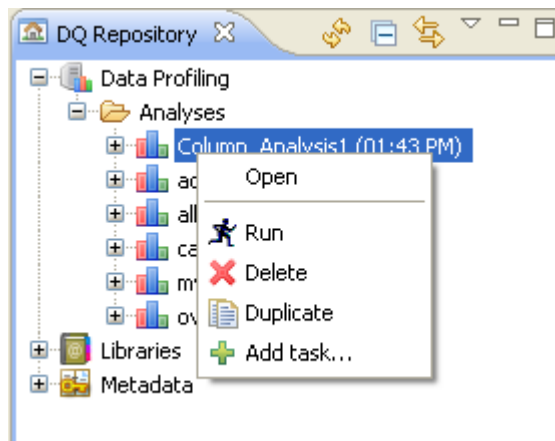
The corresponding analysis editor for the selected analysis displays.

2.11.2 How to delete an analysis

Prerequisite(s): **Talend Open Profiler** main window is open. At least one analysis type has been created.

To delete an analysis:

- In the **DQ Repository** tree view, expand the **Data Profiling** and **Analyses** folders in succession.
- Right-click the analysis you want to delete and select **Delete** from the drop-down list.



A confirmation pop-up appears prompting you to confirm the deletion operation or to cancel it. Click **OK** to close the pop-up and delete the analysis from the **Analyses** folder in the tree view.

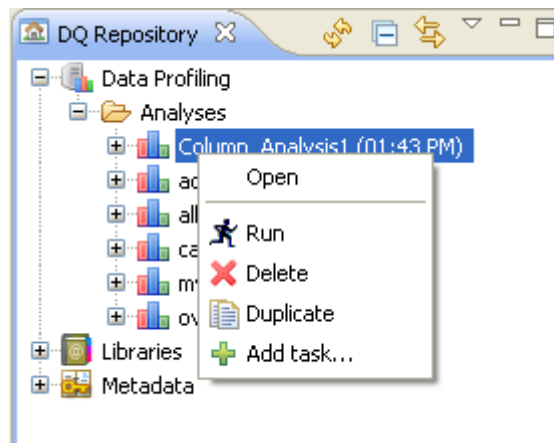
2.11.3 How to execute an analysis

Prerequisite(s): **Talend Open Profiler** main window is open. At least one analysis type has been created.

To execute an analysis:

- In the **DQ Repository** tree view, expand the **Data Profiling** and **Analyses** folders in succession.

- Right-click the analysis you want to execute and select **Run** from the drop-down list.



A progress bar appears to convey the progress of the analysis execution.

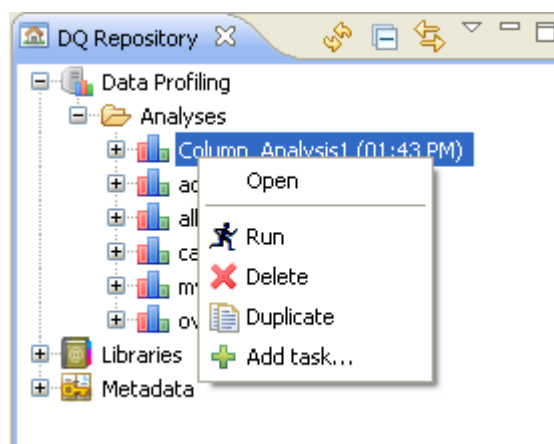
2.11.4 How to duplicate an analysis

To avoid creating an analysis from scratch, for example, you can duplicate an existing one in the **Analyses** folder and work around its metadata to have a new analysis.

Prerequisite(s): **Talend Open Profiler** main window is open. At least one analysis type has been created.

To duplicate an analysis:

- In the **DQ Repository** tree view, expand the **Data profiling** and the **Analyses** folders in succession.
- Right-click the analysis you want to duplicate and select **Duplicate...** from the drop-down list.



The duplicated analysis shows in the analysis list in the **DQ Repository** tree view. You can now open the duplicated analysis and modify its metadata as needed.

2.11.5 How to add a task to an analysis

You can add a task to an analysis to indicate a problem that needs to be solved later, for example.

For more information, see *Adding a task to an item on page 69*.



CHAPTER 3

Advanced analysis procedures

This chapter provides information about how to use data quality rules when analyzing tables and how to use patterns and indicators when analyzing a set of columns.

3.1 Managing data quality rules

Talend Open Profiler allows you to set up data quality rules based on WHERE clauses and add them as indicators to table analyses. You can as well define expected thresholds on the data quality indicator's value. The range defined is used for measuring the quality of the data in the selected table.

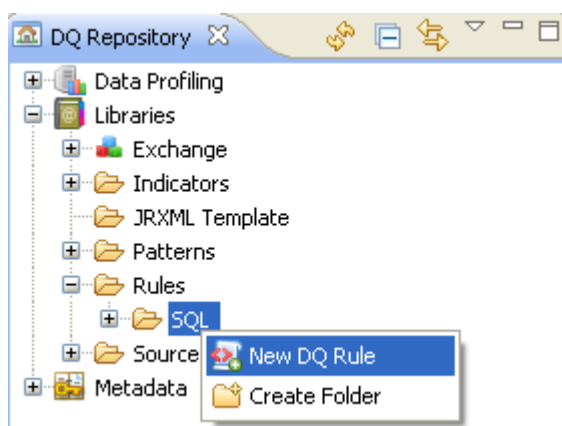
For more information about using data quality rules as indicators on a table analysis, see *Managing table analyses* on page 60.

3.1.1 How to create a DQ rule

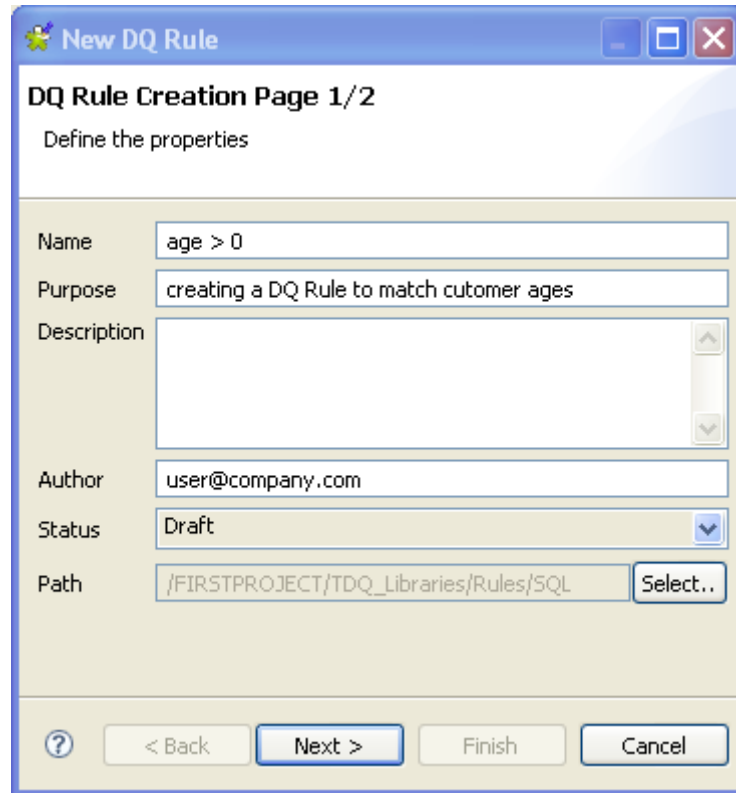
Prerequisite(s): **Talend Open Profiler** main window is open.

To create a new DQ rule:

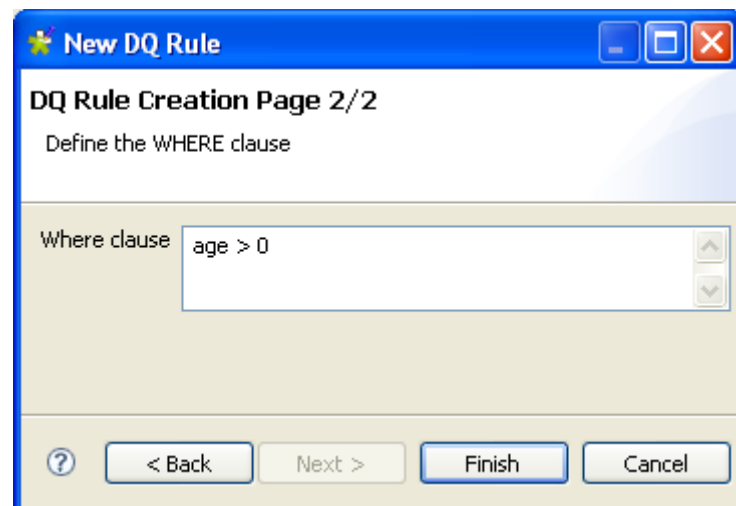
- In the **DQ Repository** tree view, expand the **Libraries** folder and right-click **DQ Rules**.



- From the drop-down list, select **New DQ Rule** to open the [New DQ Rule] wizard.

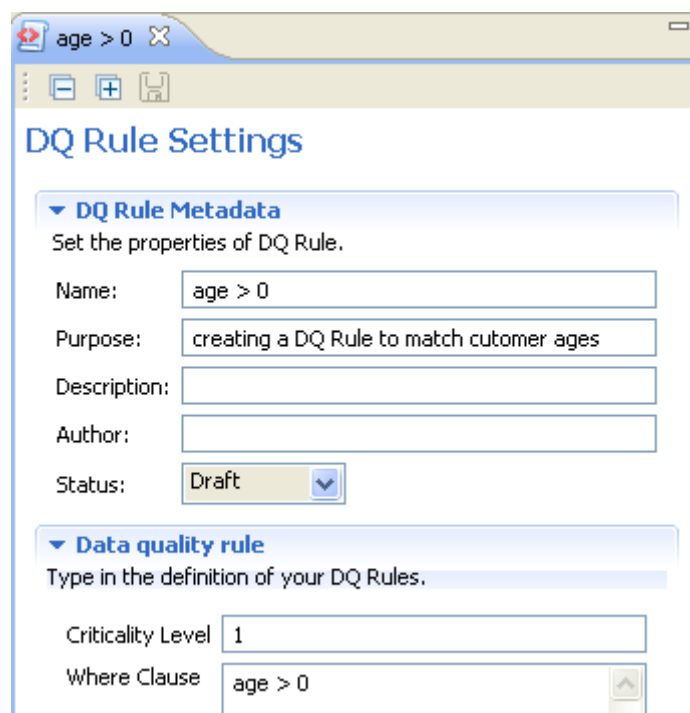


- In the **Name** field, enter a name for this new DQ rule.
- If needed, set other metadata (purpose, description and author name) in the corresponding fields and click **Next** to open a new view on the wizard.

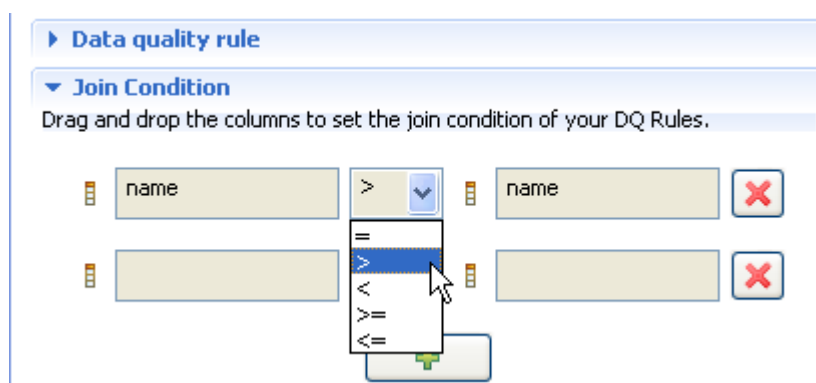


- In the **Where clause** field, enter the WHERE clause to be used in the analysis. The DQ rule must be surrounded by single quotes.
- Click **Finish** to close the [New DQ Rule] wizard.

A sub folder for this new DQ rule shows under the **Rules** folder in the **DQ Repository** tree view, and the DQ Rule editor opens with the defined metadata.



- In the DQ Rule editor, click **Data quality rule** and modify the WHERE clause or add a new one.
- If needed, set a value in the **Criticality Level** field that acts as an indicator to measure the importance of the DQ Rule. This value is saved in the database and can be used later in the [Data Quality Portal](#).
- In the DQ Rule editor, click **Join Condition** to open the corresponding view and then click the plus button to add as many join conditions as you want on selected columns.
- Drop the first selected column in the first box and drop the second in the second box.



- Select the desired sign from the join operator box and save your modifications.

You can now drop this newly created DQ rule onto a table in the Table Analysis editor. The table should contain at least one of the columns used in the DQ Rule. When you run the analysis, the join to the second column is done automatically.

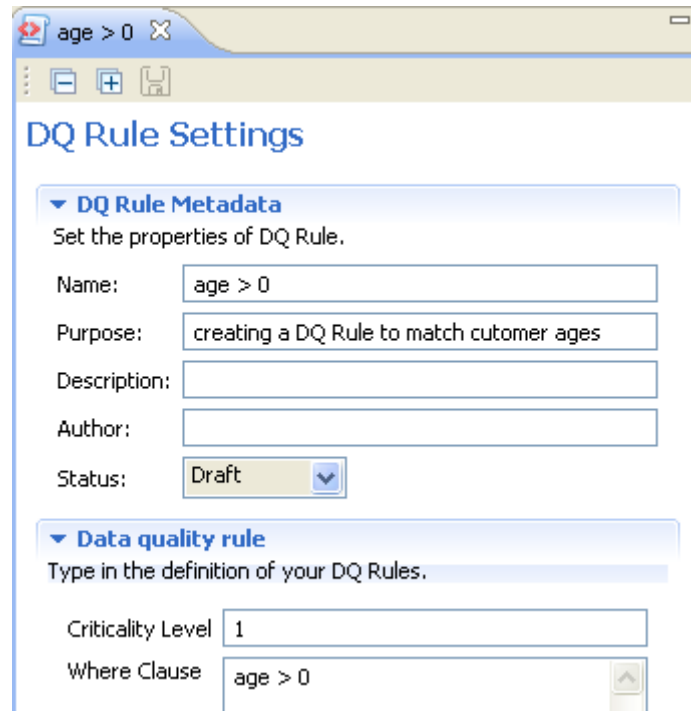
3.1.2 How to open a DQ rule

Prerequisite(s): [Talend Open Profiler](#) main window is open.

To open a DQ rule from the Repository tree view:

- In the **DQ Repository** tree view, expand the **Libraries** and the **DQ Rules** folders in succession.
- Right-click the DQ rule you want to open and select **Open** from the drop-down list.

The DQ Rule editor opens displaying the DQ rule metadata.




In the DQ Rule editor, you can click **Data quality rule** and modify the WHERE clause or add a new one.

3.2 Managing patterns

Patterns are sets of strings against which you can match the content of the columns to analyze.

Talend Open Profiler allows you to manage SQL patterns and regular expressions, including those for Java, and use them in the analyzed columns. Management processes for both types of patterns are the same.

 *Regular expressions do not work with all types of databases. An **UDF Preferences** option in **Preferences > Data Profiler** enables you through an **Add** button to define your own patterns in specific databases and use them later in your analyses. For more information about defining regular expressions, see *How to declare a regular expression in a specific database* on page 83.*

3.2.1 How to declare a regular expression in a specific database

The regular expression function is not built into all different databases environments. This is why you need, when using some databases, to create a User-Defined Function (UDF) to extend the functionality of the database server. For example, the following databases natively support regular expressions: MySQL, PostgreSQL, Oracle 10g, Ingres, etc., while Microsoft SQL server does not.

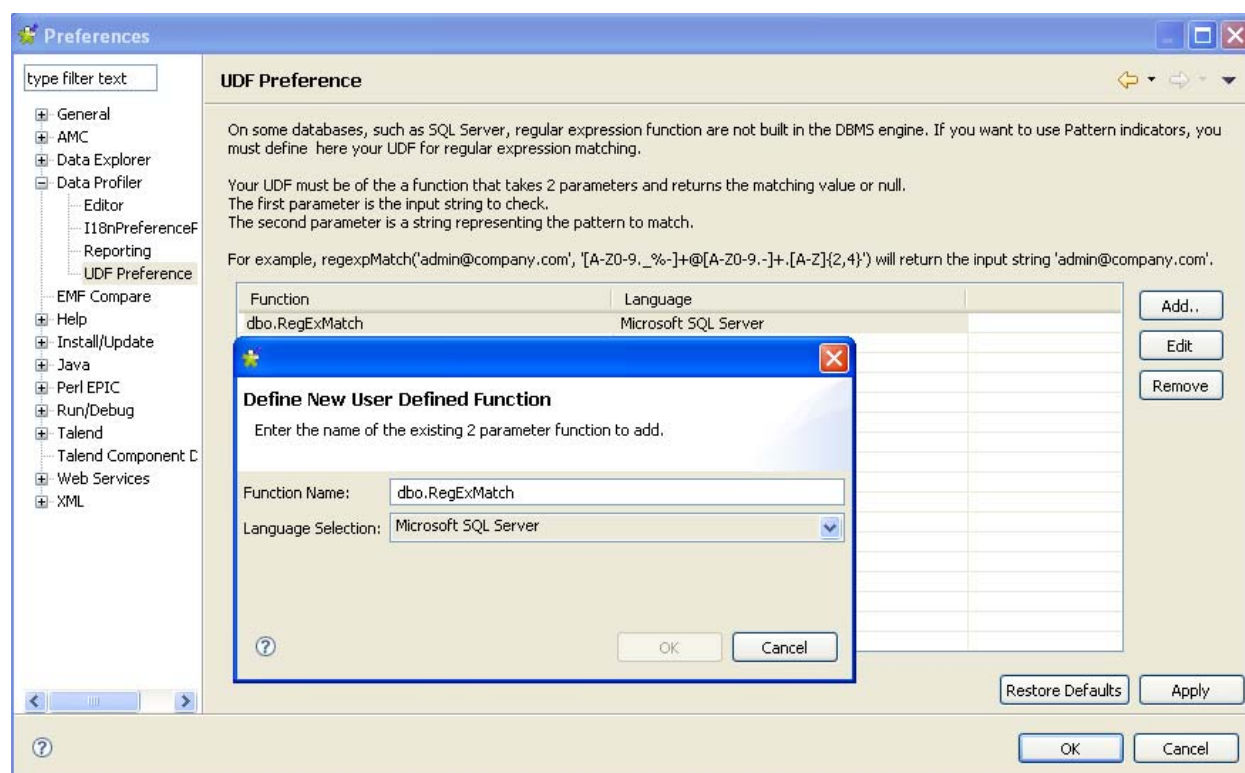
Appendix C gives a detailed example on how to create a user-defined regular expression function in an SQL server.

After you create the regular expression function, you should use **Talend Open Profiler** to declare that function in a specific database before being able to use regular expressions on analyzed columns.

Prerequisite(s): **Talend Open Profiler** main window is open.

To declare a User-Defined Function in a database:

- On the menu bar, select **Window > Preferences** to display the **[Preferences]** dialog box.
- Expand **Data Profiler** and select **UDF Preferences** from the list.



- Click the **Add** button to open another dialog box where you can define the regular pattern function you want to use in a specific database and then select the target database from the **Language Selection** list.
- Click **OK** to close the dialog box. The defined function shows in the **Function** list of the **UDF Preferences** page.
- In the **[Preferences]** dialog box, click **Apply** to validate your changes and then click **OK** to close the dialog box.

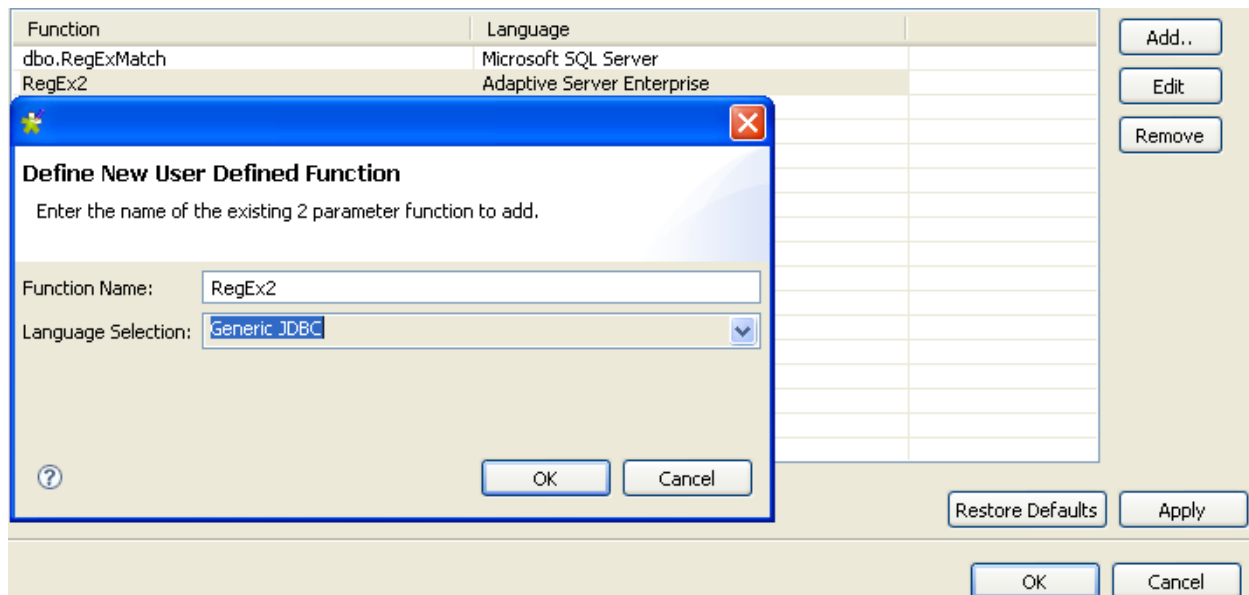
3.2.2 How to edit or delete the User-Defined Function

Talend Open Profiler enables you to edit or delete the User-Defined Function you create for a specific database.

Prerequisite(s): **Talend Open Profiler** main window is open. You have created at least one regular expression function.

To edit a User-Defined Function created for a specific database:

- On the menu bar, select **Window > Preferences** to display the **[Preferences]** dialog box.
- Expand **Data Profiler** and select **UDF Preferences** from the list.
- From the **Function** list, select the function you want to edit.
- Click the **Edit** button to open another dialog box where you can modify function name or assigned database in the corresponding lists.



- Click **OK** to close the dialog box. The function is modified accordingly in the **Function** list of the **UDF Preferences** page.
- In the **[Preferences]** dialog box, click **Apply** to validate your changes and then click **OK** to close the dialog box.

To delete a regular expression function created for a specific database:

- On the menu bar, select **Window > Preferences** to display the **[Preferences]** dialog box.
- Expand **Data Profiler** and select **UDF Preferences** from the list.
- From the **Function** list, select the function you want to delete.
- Click the **Remove** button to delete the selected function from the **Function** list.
- In the **[Preferences]** dialog box, click **Apply** to validate your changes and then click **OK** to close the dialog box.

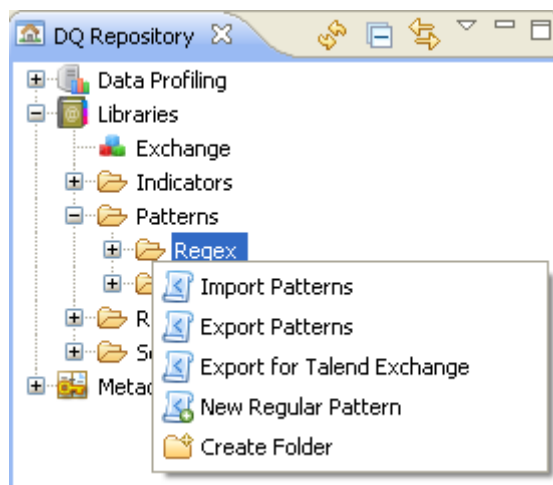
3.2.3 How to create a new pattern

Talend Open Profiler enables you to create regular patterns, including those for Java, that you can use later on analyzed columns.

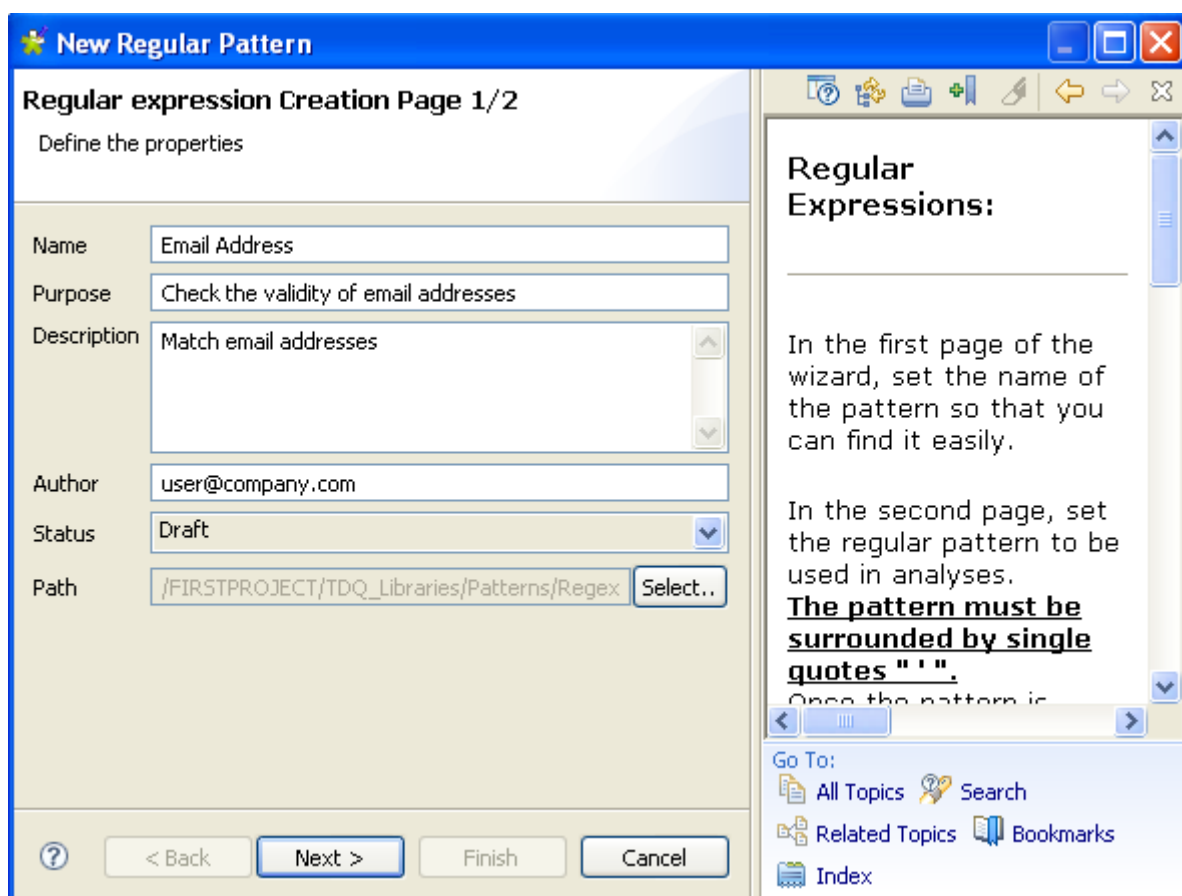
Prerequisite(s): **Talend Open Profiler** main window is open.

To create a new regular pattern:

- In the **DQ Repository** tree view, expand the **Libraries** and **Pattern** folders in succession and right-click **Regex** or **SQL**.



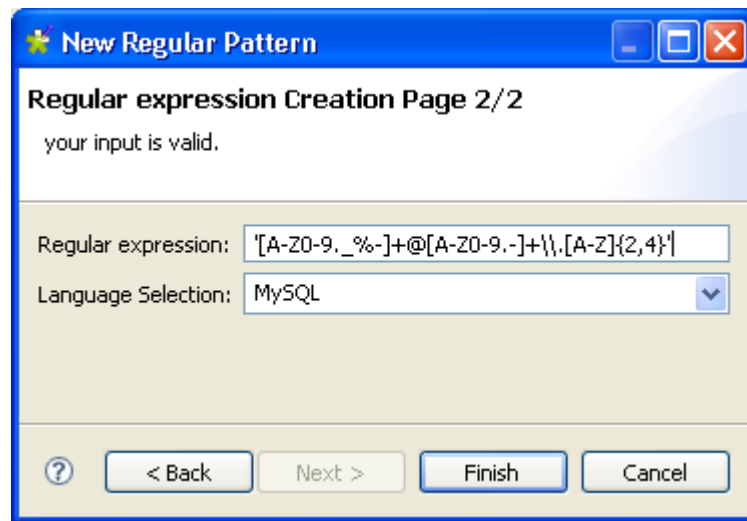
- From the drop-down list, select **New regular pattern** to open the [Create a new regular pattern] wizard.



When you open the [New Regular Pattern] wizard, a help panel automatically opens with the wizard. This help panel guides you through the steps of creating new regular patterns.

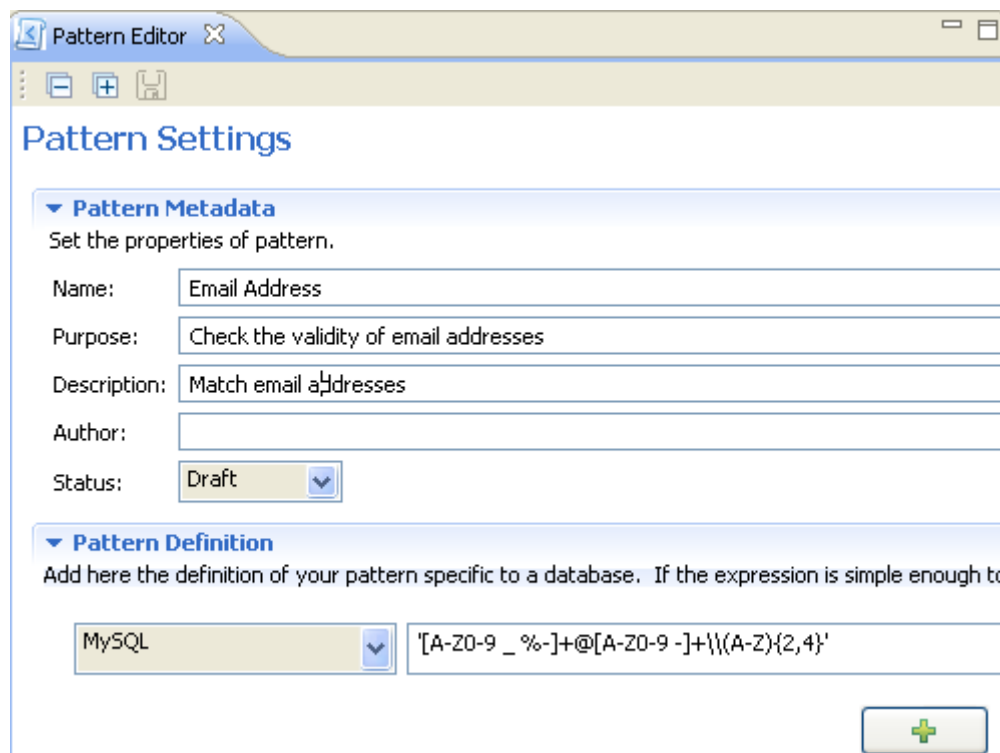
- In the **Name** field, enter a name for this new regular pattern.

- If needed, set other metadata (purpose, description and author name) in the corresponding fields and click **Next** to open a new dialog box.



- In the **Regular expression** field, enter the regular pattern to be used in the analysis. The pattern must be surrounded by single quotes.
- From the **Language Selection** list, select the relevant language.
- Click **Finish** to close the dialog box.

A sub folder for this new regular pattern shows under the **Patterns** folder in the **DQ Repository** tree view, and the Pattern editor opens with the defined metadata.



In the Pattern editor, you can click **Pattern Definition** and add patterns specific to the available databases through the plus button.

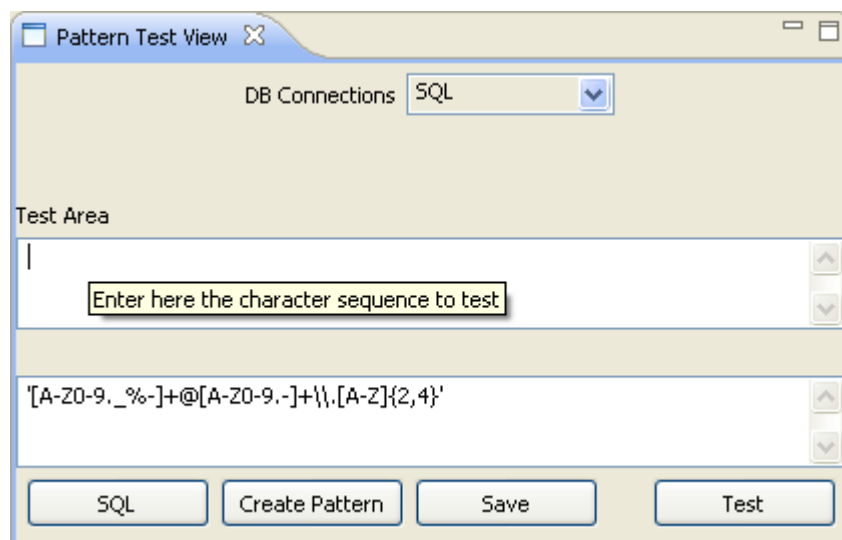


-If the regular expression is simple enough to be used in all databases, select ALL_DATABASE_TYPE from the list.

Sub folders labeled with the specified database types display below the created pattern name in the **DQ Repository** tree view. By clicking the pattern sub folder, you can display its expression in the **Detail View** below the tree view.

Once the pattern is created, you can drag it onto a column in the open editor.

If you click the **Test** button to the right in the **Pattern Definition** view, you can open the **Pattern Test View** panel where you can test character sequences against the created/selected pattern.



To test a character sequence:

- In the **Test Area**, enter the character sequence and click **Test**. An icon will display top right to indicate if the entered sequence character is accepted or not.
- Click **Save** to save the regular expression of the edited pattern.



You can create/modify patterns directly from the **Pattern Test View** via the **Create Pattern** button.

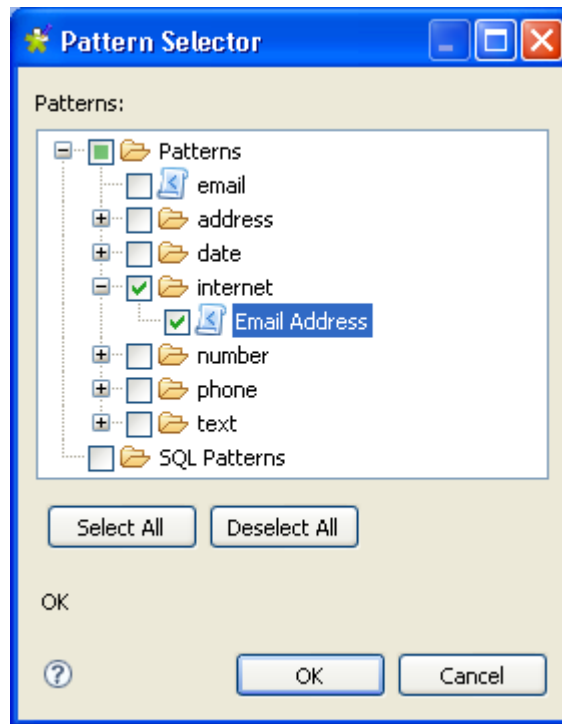
3.2.4 How to add patterns to analyzed columns

You can add one or more patterns to each analyzed column. Two types of patterns exist: regular patterns and SQL patterns. The process to add a pattern is the same for both types.

Prerequisite(s): **Talend Open Profiler** main window is open. An analysis of a set of columns is open in the Column Analysis editor.

To add a regular pattern to a column:

- Click **Analyze Columns** to display the analyzed columns view.
- Click the icon to display the [**Pattern Selector**] dialog box.



- Expand the **Patterns** folder and check the check box(es) of the pattern(s) you want to add to the current column analysis.
- Click **OK** to close the dialog box. The added pattern shows under the analyzed column in the **Analyzed Column** list.



You can add a regular pattern to a column simply by a drag&drop operation of a pattern from the **DQ Repository** tree view onto the analyzed column.



If the database you are using does not support regular expressions, you need first to define the needed patterns in this specific database through the [Preferences] dialog box before being able to add any of the specified patterns to the column analysis. For more information about defining patterns, see [How to declare a regular expression in a specific database on page 83](#).

3.2.5 How to analyze a set of columns with pattern indicators

When you add one or more patterns to an analyzed column, you check all existing data in the column against the specified pattern(s). After the execution of the column analysis, **Talend Open Profiler** provides you with the possibility to switch to the data explorer and access a list of all valid/invalid data in the analyzed column.

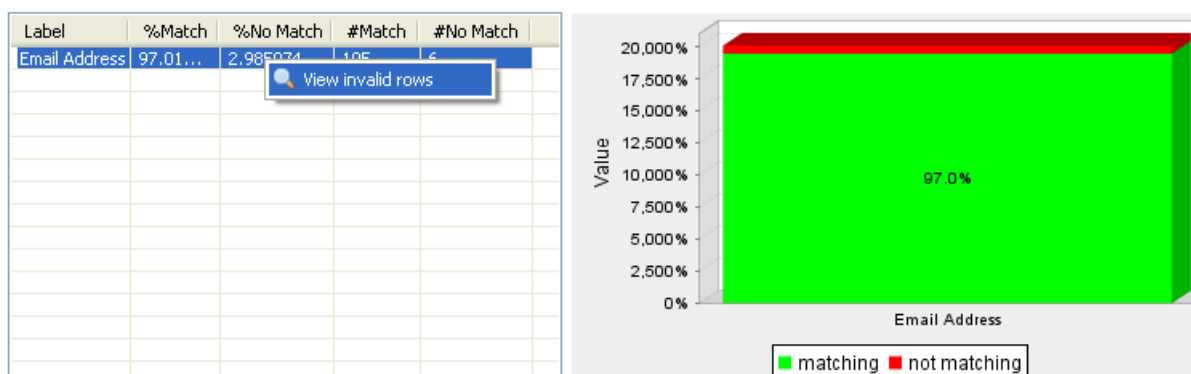
Example:

- Add the **Email Address** pattern to an analyzed column and execute the column analysis. For more information, see [How to analyze a set of columns on page 33](#) and [How to add patterns to analyzed columns on page 88](#).
- In the Column Analysis editor, click the **Analysis Results** tab at the bottom to open the corresponding view.

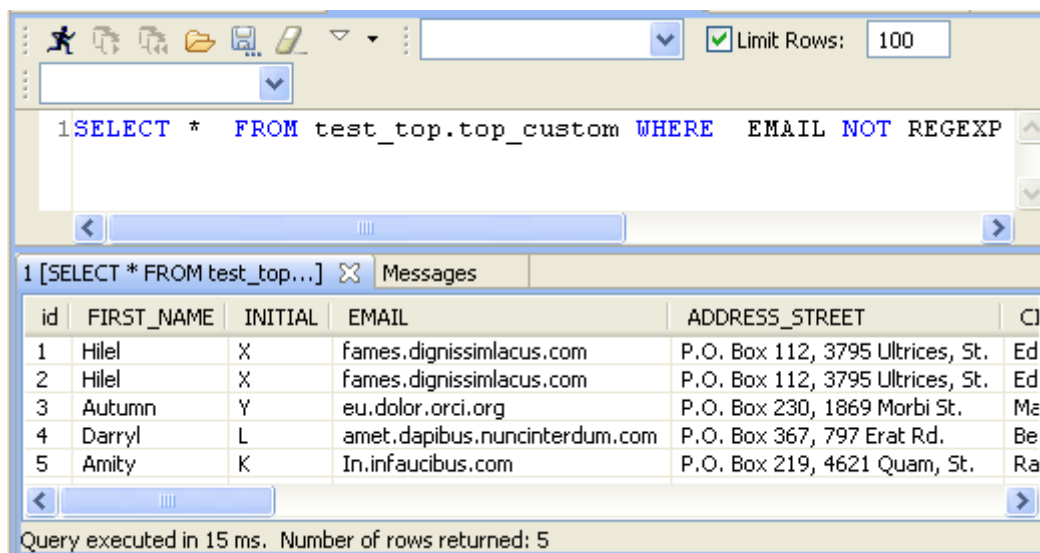
Analysis Results

Column: EMAIL

Pattern Matching



In the **Analysis Results** view, right-click the pattern line and select **View invalid rows**. The **[SQL Editor]** opens in the data explorer. A list of the analyzed data with all invalid email addresses displays in the **Messages** area in the **[SQL Editor]**.



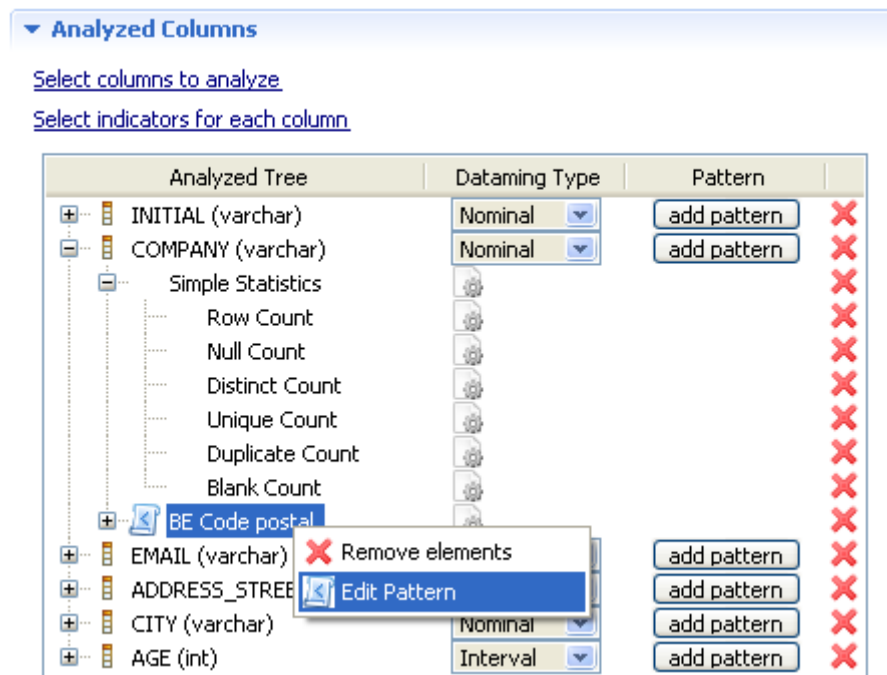
For more information about the data explorer Graphical User Interface, see *Data Explorer management GUI on page 125*

3.2.6 How to edit a pattern in the analyzed column

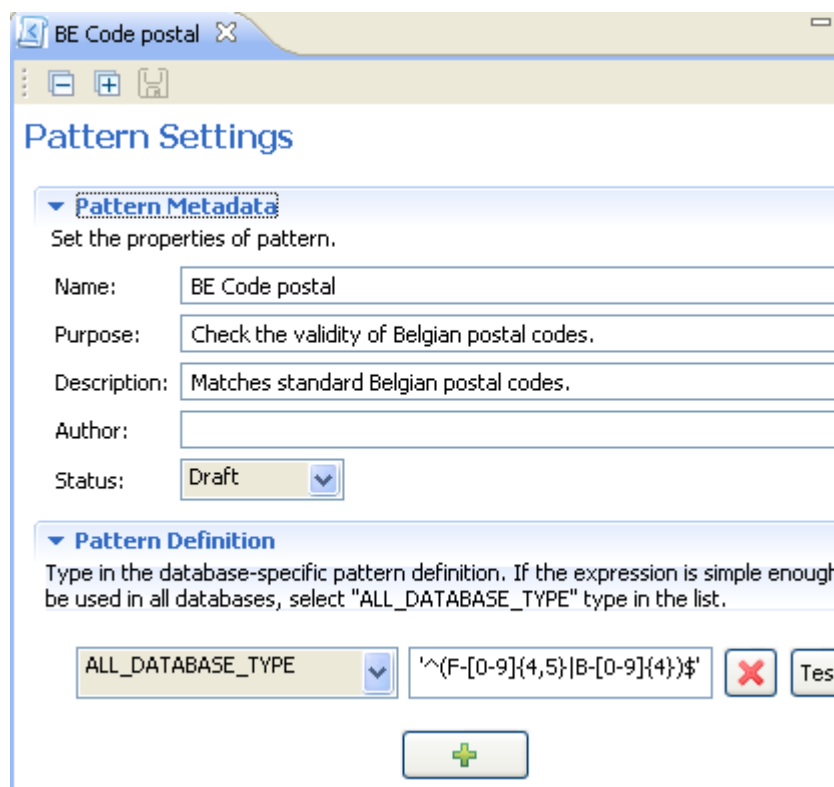
Prerequisite(s): **Talend Open Profiler** main window is open. An analysis of a set of columns is open in the Column Analysis editor.

To edit a pattern added to an analyzed column:

- Click **Analyze Columns** to display the analyzed columns view.
- Right-click the pattern you want to edit and select **Edit pattern** from the drop-down list.



The Pattern editor opens displaying the selected pattern metadata.



- In the Pattern editor, click **Pattern Definition** to edit the pattern definition, or change the selected database, or add other patterns specific to available databases through the plus button.
- On the toolbar, click the save icon to save your changes.



- If the regular pattern is simple enough to be used in all databases, select ALL_DATABASE_TYPE in the list.
- When you edit a pattern through the Column Analysis editor, you modify the pattern listed in the **DQ Repository** tree view. Make sure that your modifications are suitable for all other analyses that may be using the pattern modified.

3.2.7 How to edit a pattern

You can open the editor of any pattern to check its settings and/or edit its definition and metadata in order to adapt it to a specific database type, if needed.

Prerequisite(s): Talend Open Profiler main window is open.

To open/edit a pattern from the Repository tree view:

- In the **DQ Repository** tree view, expand the **Libraries** and the **Patterns** folders in succession.
- Browse through the pattern lists to reach the pattern you want to open/edit, right-click its name and select **Open** from the drop-down list.


The Pattern editor opens displaying the pattern settings.

The screenshot shows the 'Pattern Settings' dialog box for a pattern named 'UK Phone Number'. The dialog is divided into two main sections: 'Pattern Metadata' and 'Pattern Definition'.
Under 'Pattern Metadata', the following fields are visible:
- Name: UK Phone Number
- Purpose: Check the validity of UK phone numbers
- Description: Matches UK mobile phone number, with optional +44 national code. Allows optional br
- Author: (empty field)
- Status: Draft (dropdown menu)
Under 'Pattern Definition', there is a text area for the pattern definition and a dropdown menu to select the database type. The current database type is 'MySQL'. The pattern definition text is: '^(\{+44[[:space:]]?7[[:digit:]]{3}\|\{?07[[:digit:]]{3}\}\})?([:spa'. A plus sign button is located at the bottom right of the dialog.

- Modify the pattern metadata, if needed, and then click **Pattern Definition** to display the relevant view. In this view, you can: edit pattern definition, change the selected database and add other patterns specific to available databases through the plus button.
- Click the save icon on top of the editor to save your changes.



- If the regular pattern is simple enough to be used in all databases, select ALL_DATABASE_TYPE in the list.

 When you edit a pattern, you modify the pattern listed in the DQ Repository tree view. Make sure that your modifications are suitable for all analyses that may be using the modified pattern.

3.2.8 How to delete a pattern

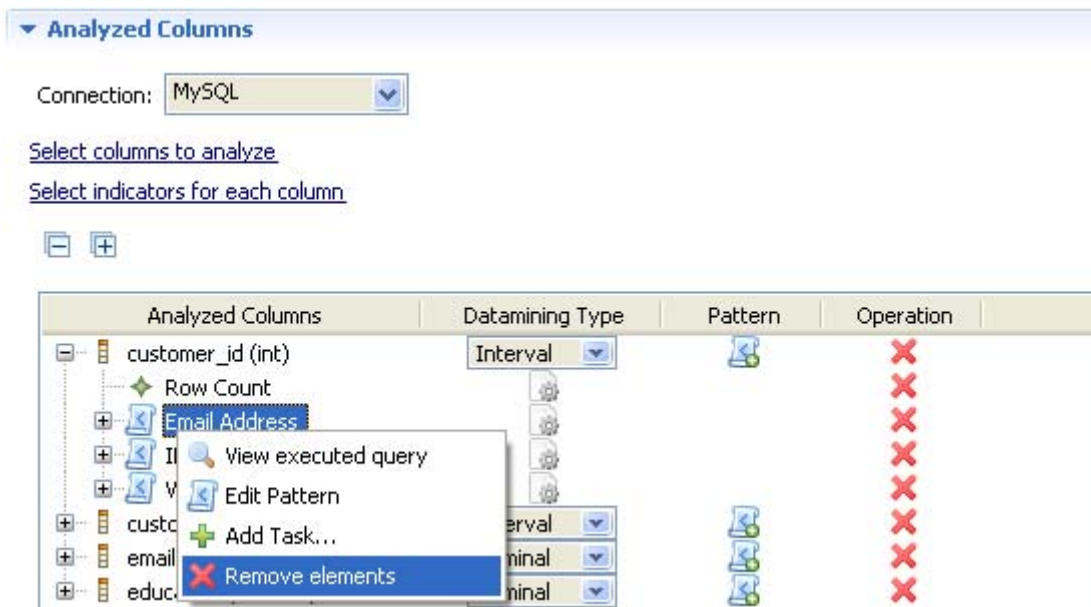
You can delete patterns directly from the **Analyzed Columns** view or from the **DQ Repository** tree view.

How to delete a pattern from the analyzed column:

Prerequisite(s): **Talend Open Profiler** main window is open. An analysis of a set of columns is open in the Column Analysis editor.

To delete a pattern from an analyzed column:

- Click **Analyze Columns** to display the analyzed columns view.
- Right-click the pattern you want to delete and select **Remove Elements** from the drop-down list.



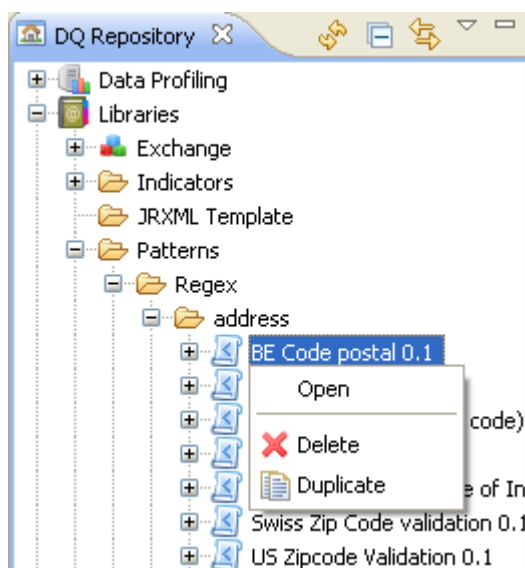
The selected pattern disappears from the **Analyzed Column** list.

How to delete a pattern from the DQ Repository

Prerequisite(s): **Talend Open Profiler** main window is open.

To delete a pattern from the DQ Repository:

- In the **DQ Repository** tree view, expand the **Libraries** and the **Patterns** folders in a succession.
- Right-click the pattern you want to delete and select **Delete** from the drop-down list.



A confirmation pop-up appears prompting you to confirm the deletion operation or to cancel it.

Click **OK** to delete the pattern from the **Patterns** folder in tree view.

3.2.9 How to duplicate a pattern

To avoid creating a pattern from scratch, for example, you can duplicate an existing one in the pattern list and work around its metadata to have a new pattern and use it later in data profiling analyses.

Prerequisite(s): **Talend Open Profiler** main window is open.

To duplicate a pattern from the Repository tree view:

- In the **DQ Repository** tree view, expand the **Libraries** and the **Patterns** folders in succession.
- Browse through the patterns lists to reach the pattern you want to duplicate, right-click its name and select **Duplicate...** from the drop-down list.

The duplicated pattern shows under the corresponding pattern folder in the **DQ Repository** tree view. You can now double-click the duplicated pattern to modify its metadata as needed.

3.2.10 How to import patterns from a csv file

In **Talend Open Profiler** you can import patterns stored locally in a csv file. The rules for laying out the content of the csv file is as the following:

- Column 1: Label: the label of the pattern (must not be empty),
- Column 2: Purpose: the purpose of the pattern (can be empty),
- Column 3: Description: the description of the pattern (can be empty),
- Column 4: Author: the author of the regular expression (can be empty),
- Column 5: Relative Path: the relative path to the root folder (can be empty),
- Column 6: All DB Regular: the regular expression applicable to all databases (can be empty),

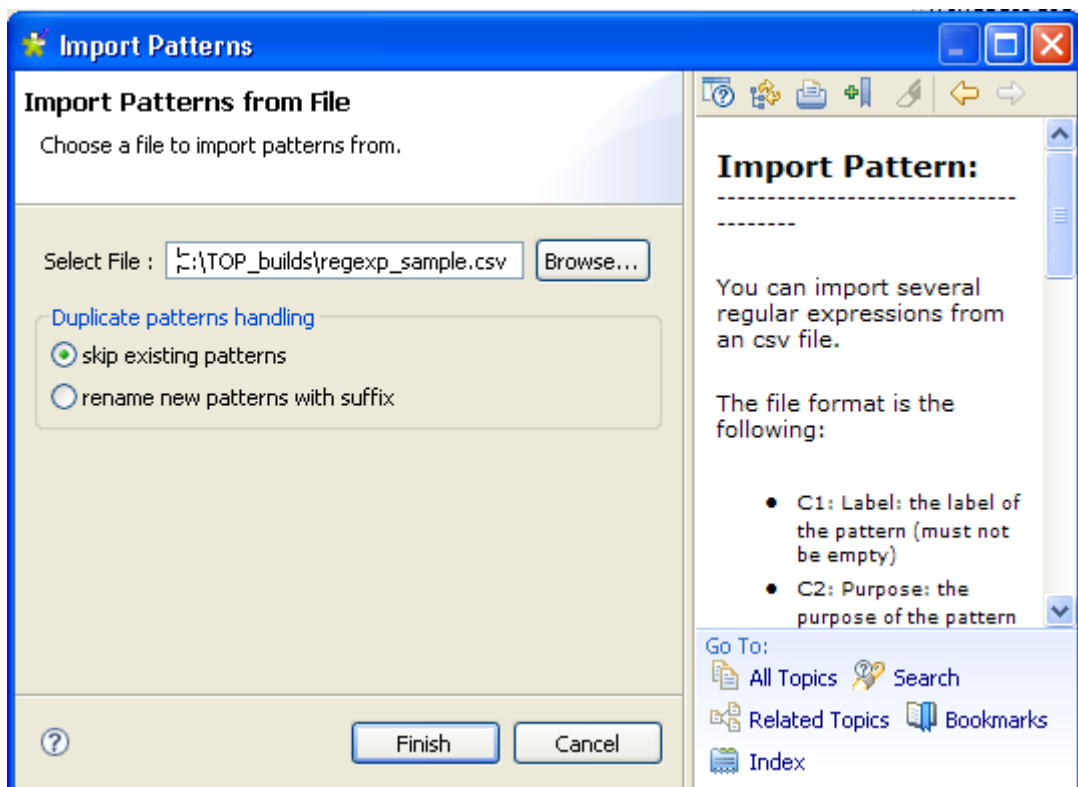
- Column 7: DB2 Regexp: the regular expression applicable to DB2 databases (can be empty),
- Column 8: MySQL Regexp: the regular expression applicable to MySQL databases (can be empty),
- Column 9: Oracle Regexp: the regular expression applicable to Oracle databases (can be empty),
- Column 10: PostgreSQL Regexp: the regular expression applicable to PostgreSQL databases (can be empty),
- Column 11: SQL Server Regexp: the regular expression applicable to SQL Server databases (can be empty).

Prerequisite(s): **Talend Open Profiler** main window is open. The csv file is stored locally.

To import regular patterns from a csv file:

- In the **DQ Repository** tree view, expand the **Libraries** folder.
- Right-click **Patterns** and select **Import patterns**.

The [**Import Patterns**] wizard opens.



When you open the [**Import patterns**] wizard, a help panel automatically opens with the wizard. This help panel guides you through the steps of importing patterns from a csv file.

- Browse to the csv file holding the regular expressions.
- In the **Duplicate patterns handling** panel, either click **skip existing patterns** to only import patterns that do not already exist in the pattern list in the **DQ Repository** tree view, or
- Click **rename new patterns with suffix** to identify all the imported patterns with a suffix.

- Click **Finish** to close the wizard.

All imported patterns are listed under the **Patterns** folder in the **DQ Repository** tree view.



A warning icon next to the name of the imported pattern in the tree view identifies that it is not correct. You must open the pattern and try to figure out what is wrong. Usually, problems come from missing quotes at the beginning and end of the expressions. Check your regular expressions and ensure that they are encapsulated in single quotes.

3.2.11 How to import patterns from Talend Exchange

You can import regular expressions and SQL patterns from **Talend Exchange** to your current version of **Talend Open Profiler** and use them later on your analyzed columns.

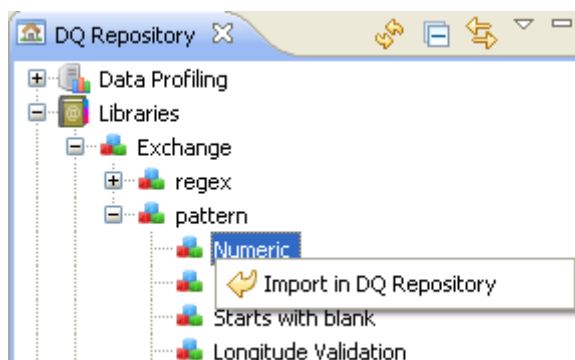
However, make sure that the **Talend Exchange** extension you want to import from is compatible with your current Studio version.

Compatibility means that **Talend Exchange** extension has the same two first sequences of the unique identifier of your current version of **Talend Open Profiler**. For example, if your current version of **Talend Open Profiler** is 3.2.0, compatible extensions could be 3.2.1, 3.2.0M1, 3.2.0M2, 3.2.0RC1 etc.

Prerequisite(s): **Talend Open Profiler** main window is open.

To import patterns from Talend Exchange:

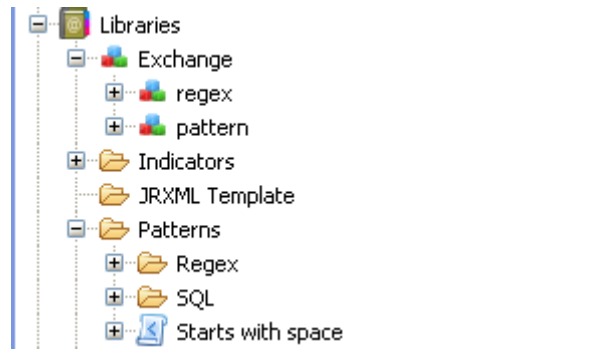
- In the **DQ Repository** tree view, expand **Libraries** and **Exchange** in succession.
- Under **Exchange**, expand **pattern** and right-click the name of the indicator you want to import and then select **Import in DQ Repository**.



A message displays to confirm the operation.

- Click **OK** in the confirmation message to close it.

The imported pattern from **Talend Exchange** is listed under the **Patterns** folder in the **DQ Repository** tree view.



3.2.12 How to export patterns

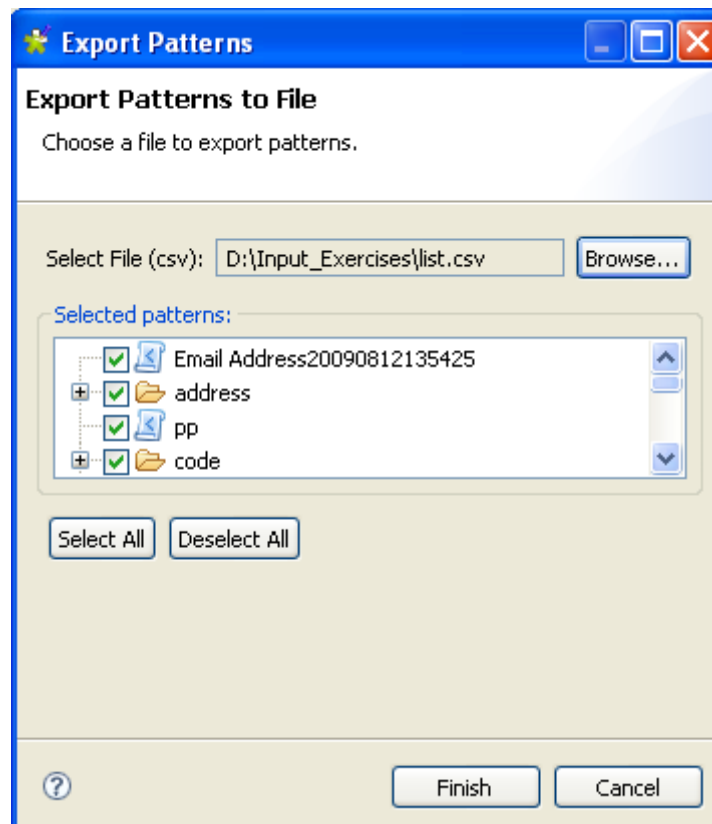
In **Talend Open Profiler** you can export patterns and store them locally in a csv file. For more information about the content lay out of the csv file, see *How to import patterns from a csv file on page 94*.

Prerequisite(s): **Talend Open Profiler** main window is open.

To export regular patterns to a csv file:

- In the **DQ Repository** tree view, expand the **Libraries** and **Patterns** folders in succession and right-click the pattern folder you want to export.
- From the drop-down list, select **Export Patterns**.

The [**Export Patterns**] wizard opens with the check boxes of all patterns selected by default.



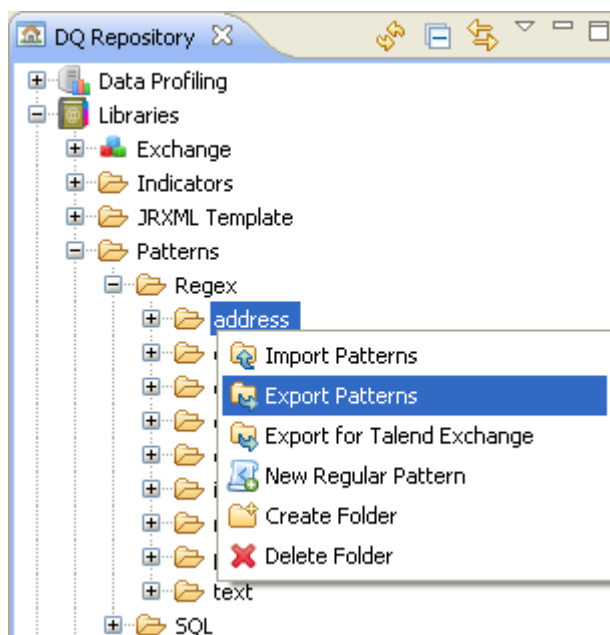
- Browse to the csv file where to save the regular expressions.

- If needed, clear the check boxes of the pattern families or patterns you do not want to export to the csv file.
- Click **Finish** to close the wizard.

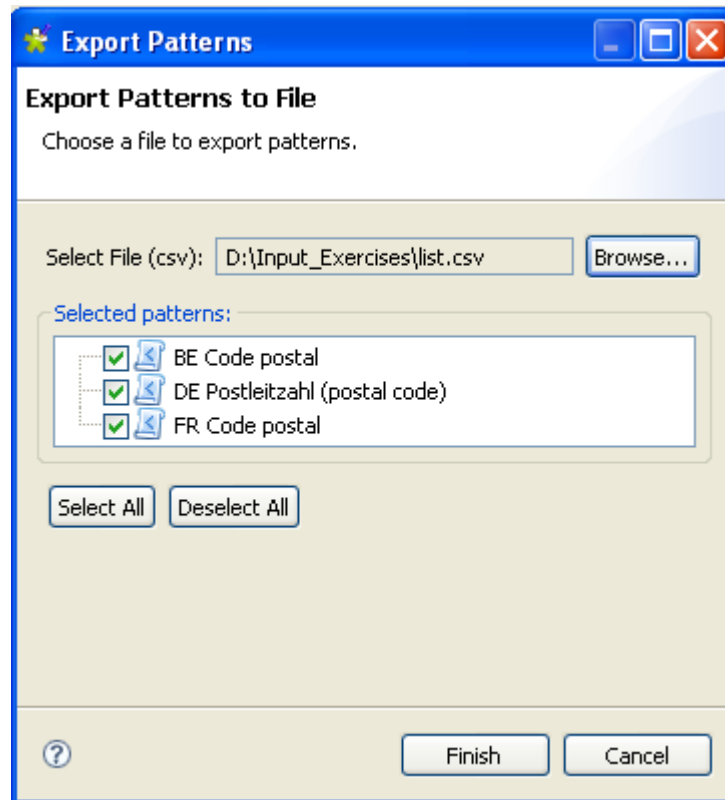
All exported patterns are saved in the defined csv file.

To export a single regular pattern family to a csv file directly from the DQ Repository tree view:

- In the **DQ Repository** tree view, expand the **Libraries** and **Patterns** in succession and right-click the pattern family you want to export.



- From the drop-down list, select **Export Patterns**.
The [**Export Patterns**] wizard opens with the check boxes of all patterns belonging to the family selected by default.



- If needed, clear the check boxes of the patterns you do not want to export to the csv file.
- Click **Finish** to close the wizard.

All exported patterns are saved in the defined csv file.

3.2.13 How to export patterns to Talend Exchange

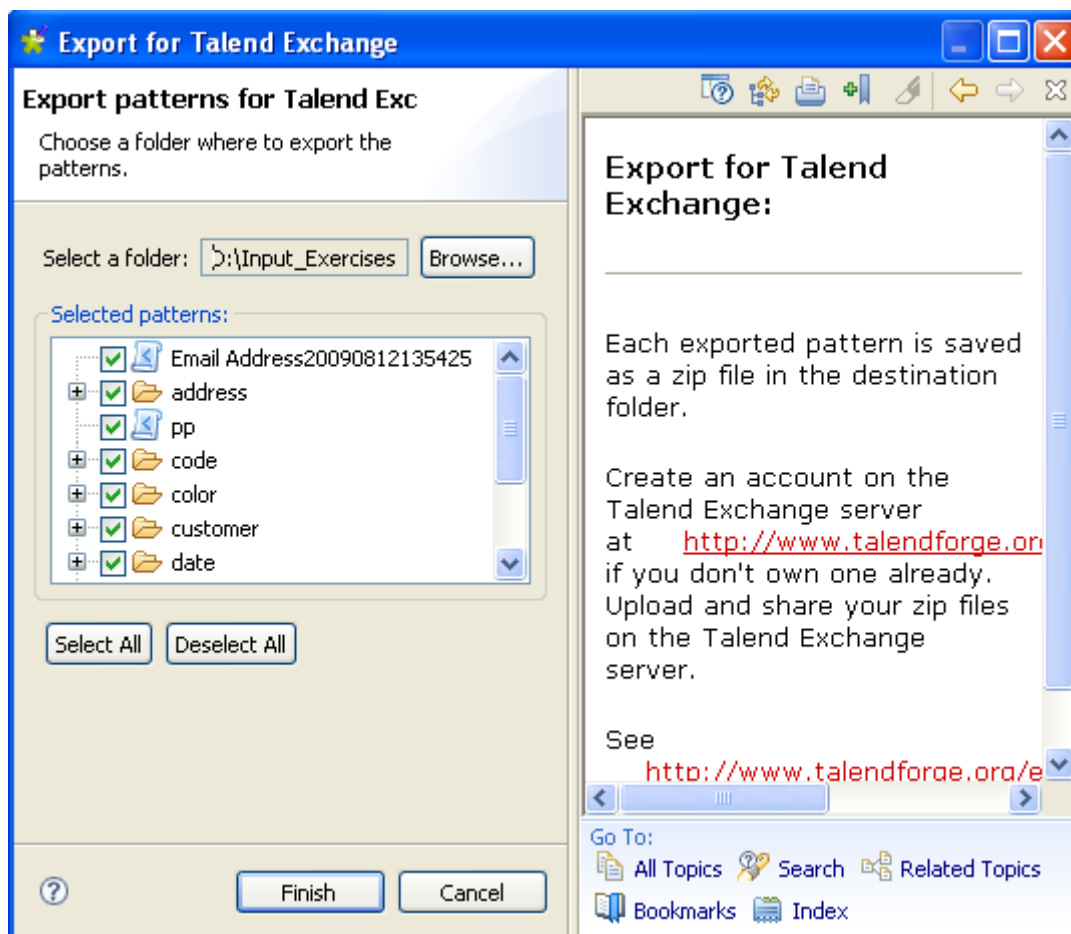
You can export patterns from your current version of **Talend Open Profiler** to **Talend Exchange** where you can share them with other users.

Prerequisite(s): **Talend Open Profiler** main window is open.

To export patterns to Talend Exchange:

- In the **DQ Repository** tree view, expand **Libraries** and **Patterns** in succession.
- Right-click the pattern folder you want to export to **Talend Exchange** and select **Export for Talend Exchange**.

The [**Export for Talend Exchange**] wizard displays with a help panel to the right.

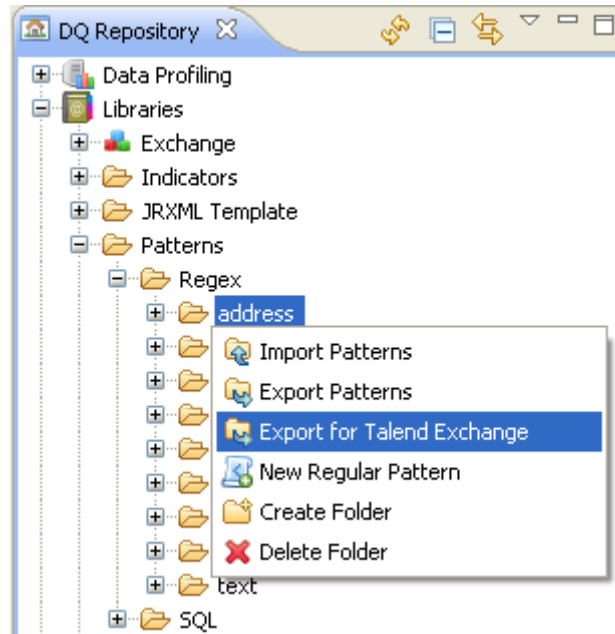


- Browse to the folder where to save patterns.
- If needed, clear the check boxes of the pattern families or patterns you do not want to export to the specified folder.
- Click **Finish** to close the wizard.

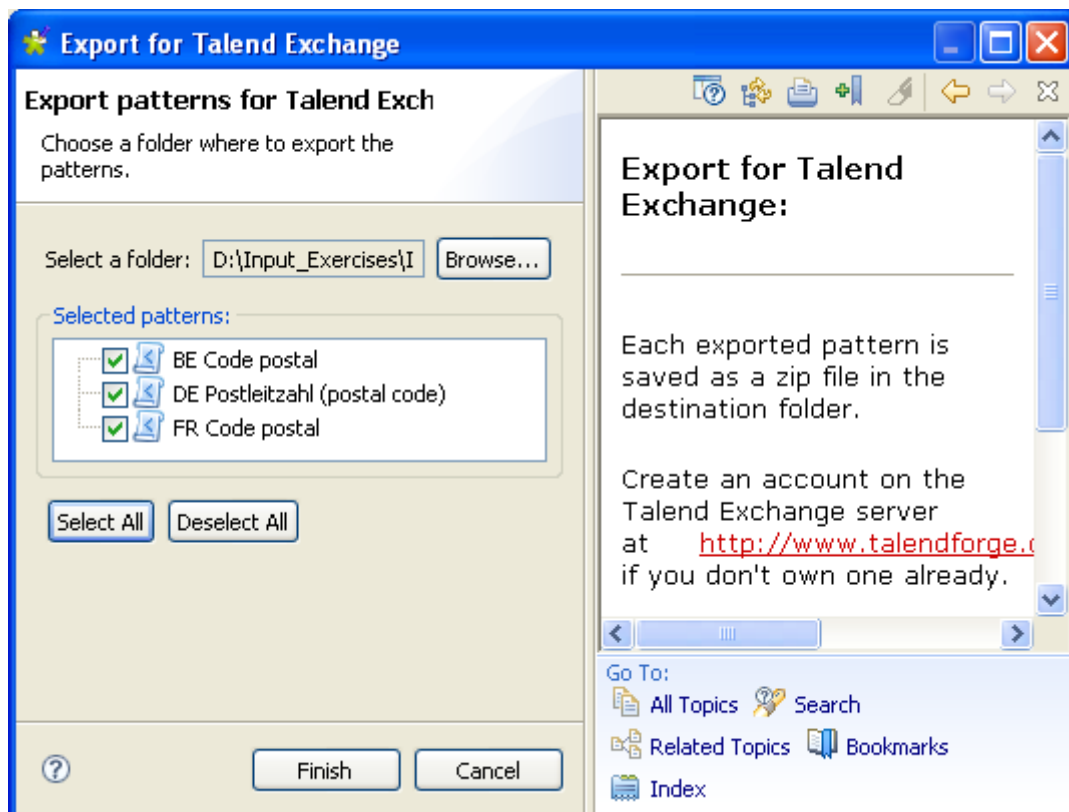
A distinct csv file is created for each exported pattern. Each csv file is compressed as a zip. All these zip files are saved in the defined folder. You need now to upload them to **Talend Exchange** at http://www.talendforge.org/exchange/top/help_guest.php.

To export a single pattern family to Talend Exchange:

- In the **DQ Repository** tree view, expand **Libraries** and **Patterns** in succession and right-click the pattern family you want to export.



- From the drop-down list, select **Export for Talend Exchange**. The [**Export for Talend Exchange**] wizard opens with the check boxes of all patterns belonging to the family selected by default. A help panel also shows to the right of the wizard.



- If needed, clear the check boxes of the patterns you do not want to export to the folder.
- Click **Finish** to close the wizard.

A distinct csv file is created for each exported pattern. Each csv file is compressed as zip. All these zip files are saved in the defined folder. You need now to upload them to **Talend Exchange** at http://www.talendforge.org/exchange/top/help_guest.php.

3.3 Managing indicators

Indicators are the results achieved through the implementation of different patterns that are used to define the content, structure and quality of your data. Indicators can represent the results of highly complex operations related to data- matching and different other data-related operations.

Talend Open Profiler lists all system or user-defined indicators under the **Indicators** node in the **DQ Repository** tree view.

Talend Open Profiler allows you to manage these indicators and define them for columns of database tables that need to be analyzed or monitored.

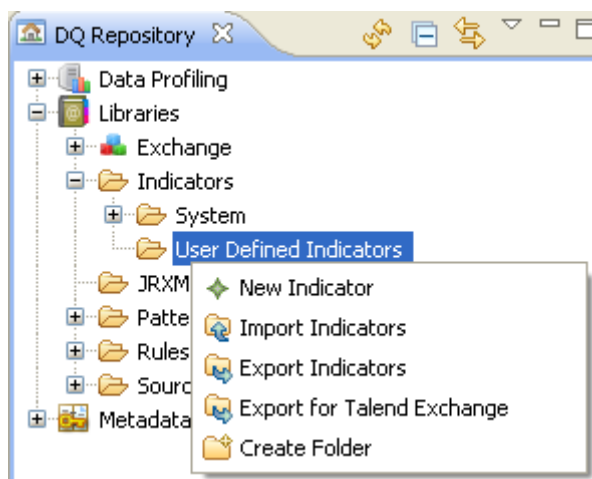
3.3.1 How to create a user-defined indicator

Talend Open Profiler enables you to create your own personalized indicators which you can later manage the same way you manage any system indicator.

Prerequisite(s): **Talend Open Profiler** main window is open.

To create a user-defined indicator:

- In the **DQ Repository** tree view, expand the **Libraries** and **Indicators** folders in succession and then right-click **User Defined Indicators**.



- Select **New Indicator** from the drop-down list to open the **[New Indicator]** wizard.

- In the **Name** field, enter a name for the indicator you want to create.
- If needed, set other metadata (purpose, description and author name) in the corresponding fields and click **Next** to open a new view on the wizard.

- From the **Language Selection** list, select the database that will support the created indicator.
- In the **SQL Template** field, enter the SQL template statement corresponding to the indicator you want to create and then click **Finish** to close the wizard.

The indicator editor opens displaying the created user defined indicator metadata.

Indicator Settings

Indicator Metadata

Name: Simple_Count

Purpose:

Description:

Author:

Status: Draft

Indicator Definition

This section is for indicator definition.

ALL_DATABASE_TYPE 'SELECT COUNT(*) FROM <%= __TABLE_NAME__ %> <%= __W

Indicator Category

This section is for indicator category.

User Defined Count

- In the editor, click **Indicator Definition** to display the corresponding view. In this view, you can: edit indicator definition, change the selected database and add other indicators specific to available databases through the plus button.
- Click **Indicator Category** to display the corresponding view. In this view, you can select from the list a category for the created indicator. The selected category will determine the type of chart that will represent the results of the executed analysis that uses the created indicator.



The by-default indicator category is **User Defined Match**.



To delete the user-defined indicator you created, expand **Indicators** and **User Defined indicators** in succession in the tree view, right-click the name of the created indicator and then select **Delete**.

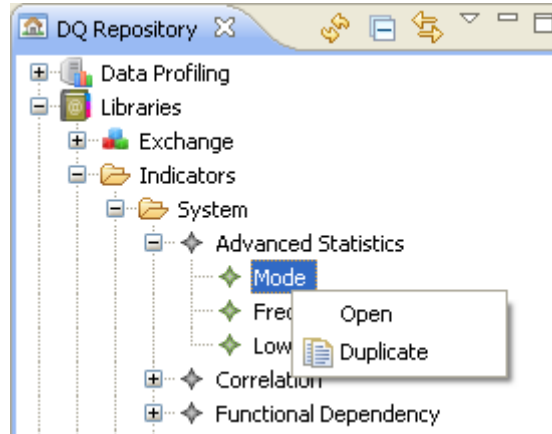
3.3.2 How to edit the definition of an indicator

You can open the editor of any system or user-defined indicator to check its settings and/or edit its definition and metadata in order to adapt it to a specific database type, if needed.

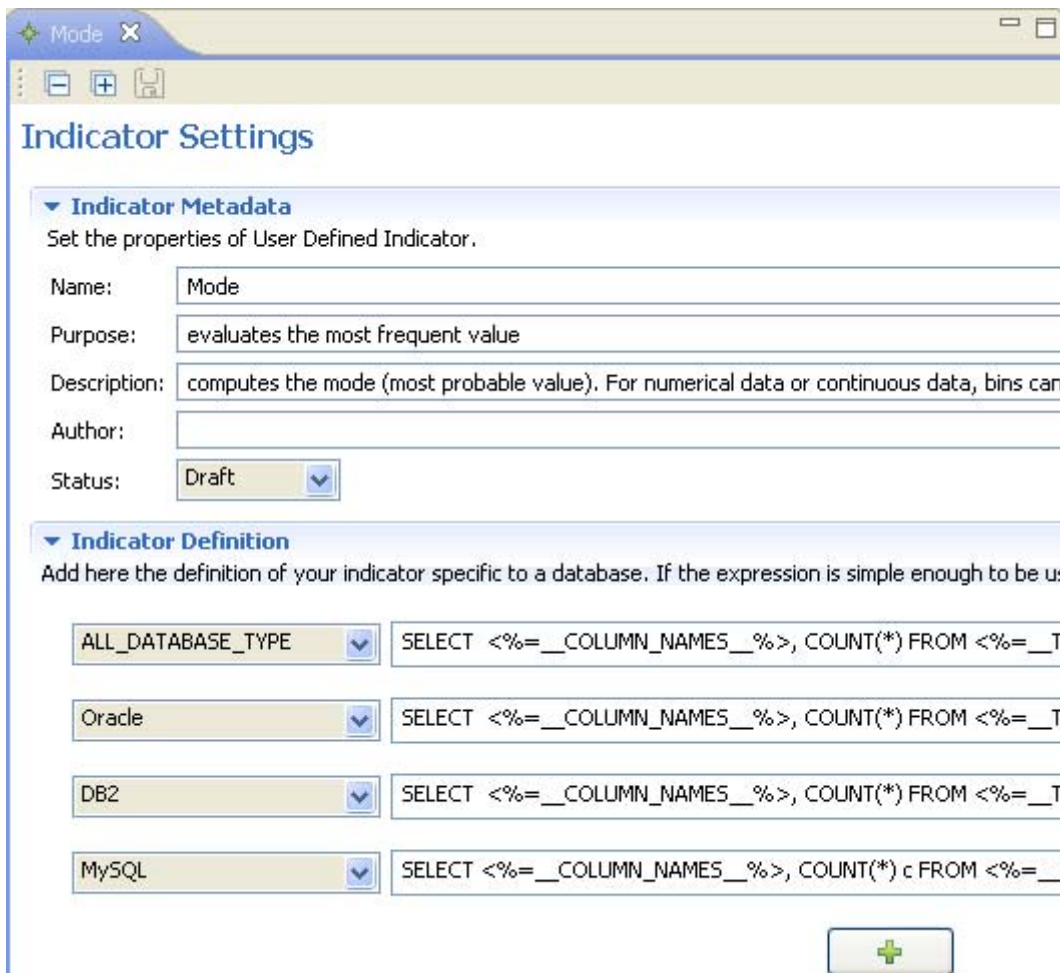
Prerequisite(s): **Talend Open Profiler** main window is open. At least, one user-defined indicator is created.

To edit the definition of a system or user-defined indicator:

- In the **DQ Repository** tree view, expand the **Libraries** and **Indicators** folders in succession and browse through the indicators lists to reach the indicator you want to modify the definition of.
- Right-click the indicator name and select **Open** from the drop-down list.



The indicator editor opens displaying the selected indicator settings.



- Modify the indicator metadata, if needed, and then click **Indicator Definition** to display the relevant view. In this view, you can: edit indicator definition, change the selected database and add other indicators specific to available databases through the plus button.
- Click the save icon on top of the editor to save your changes.



-If the indicator is simple enough to be used in all databases, select ALL_DATABASE_TYPE in the list.



When you edit an indicator, you modify the indicator listed in the DQ Repository tree view. Make sure that your modifications are suitable for all analyses that may be using the modified indicator.

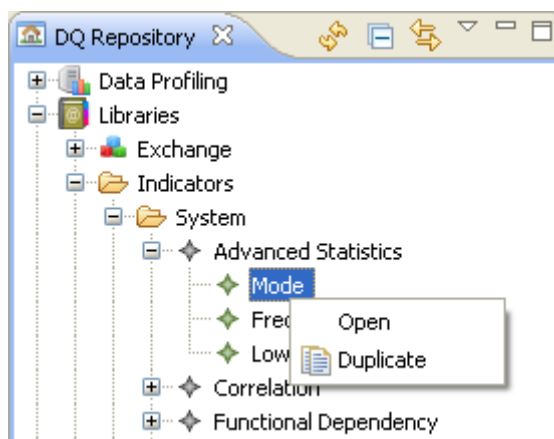
3.3.3 How to duplicate an indicator

To avoid creating an indicator from scratch, for example, you can duplicate an existing one in the indicator list and work around its metadata to have a new indicator and use it later in data profiling analyses.

Prerequisite(s): Talend Open Profiler main window is open.

To duplicate an indicator from the Repository tree view:

- In the **DQ Repository** tree view, expand the **Libraries** and the **Indicators** folders in succession.
- Browse through the indicators lists to reach the indicator you want to duplicate, right-click its name and select **Duplicate...** from the drop-down list.



The duplicated indicator shows under the User Defined Indicators folder in the **DQ Repository** tree view. You can now double-click the duplicated indicator to modify its metadata as needed.

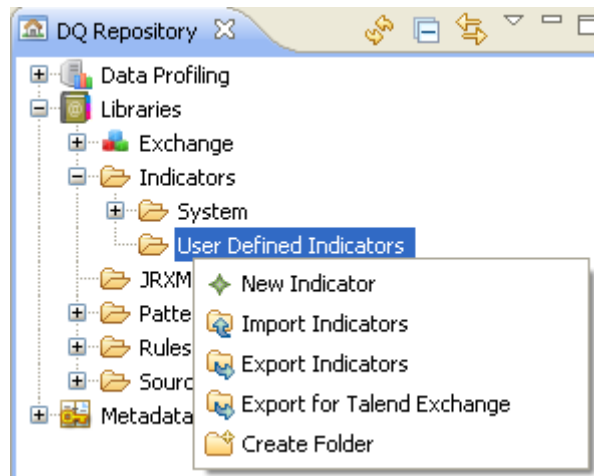
3.3.4 How to export user-defined indicators to a csv file

In **Talend Open Profiler** you can export user-defined indicators and store them locally in a csv file.

Prerequisite(s): **Talend Open Profiler** main window is open. At least, one user-defined indicator is created.

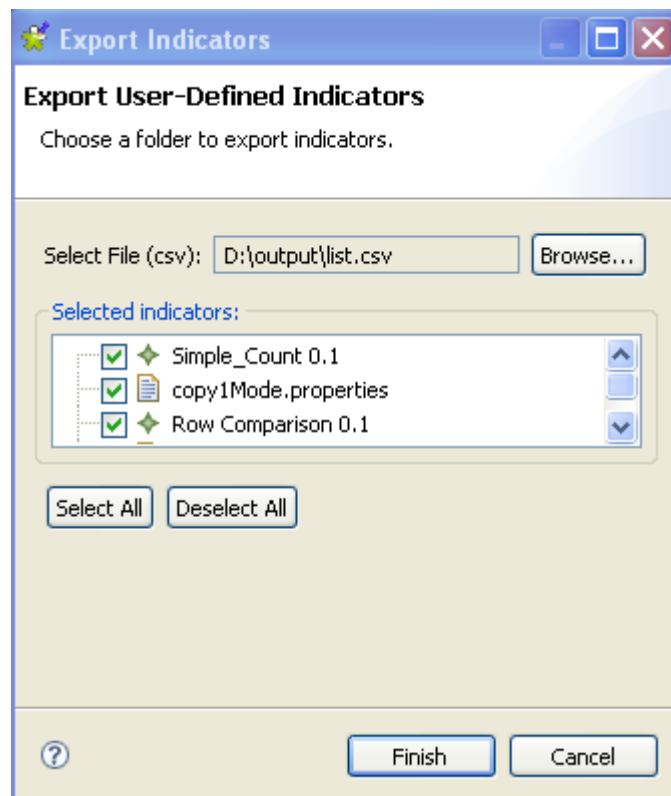
To export user-defined indicators to a csv file:

- In the **DQ Repository** tree view, expand the **Libraries** and **Patterns** folders in succession and right-click **User Defined Indicators**.



- From the drop-down list, select **Export Indicators**.

The [**Export Patterns**] wizard opens with the check boxes of all indicators selected by default.



- Browse to the csv file where to save the indicators.
- If needed, clear the check boxes of the indicators you do not want to export to the csv file.
- Click **Finish** to close the wizard.

All exported user-defined indicators are saved in the defined csv file.

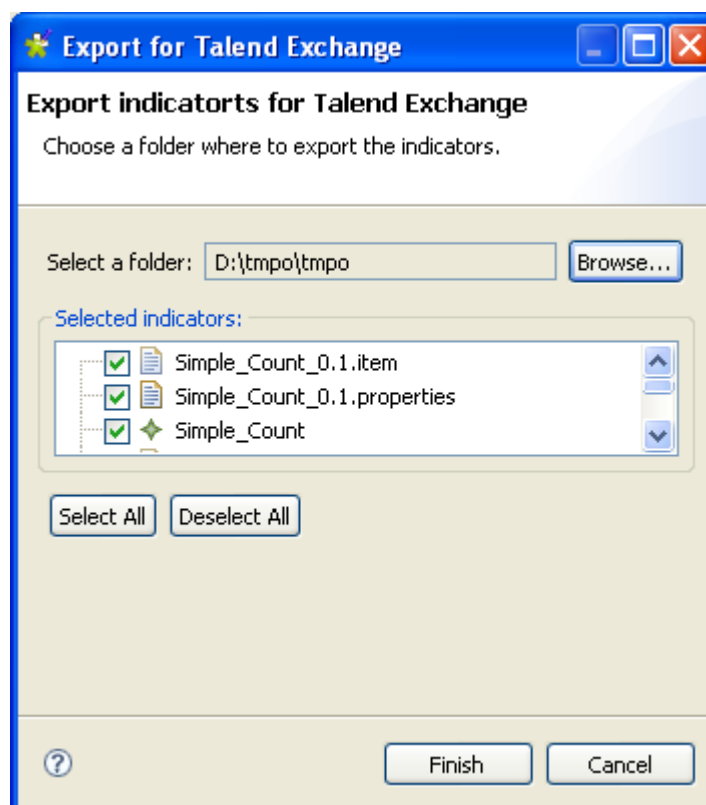
3.3.5 How to export user-defined indicators to Talend Exchange

You can export user-defined indicators from your current version of **Talend Open Profiler** to **Talend Exchange** where you can share them with other users.

Prerequisite(s): **Talend Open Profiler** main window is open. At least, one user-defined indicator is created.

To export indicators to Talend Exchange:

- In the **DQ Repository** tree view, expand **Libraries** and **Patterns** in succession.
- Right-click the **User Defined Indicator** folder and select **Export for Talend Exchange**. The **[Export for Talend Exchange]** wizard displays.



- Browse to the folder where to save indicators.
- If needed, clear the check boxes of the indicators you do not want to export to the specified folder.
- Click **Finish** to close the wizard.

A distinct csv file is created for each exported indicator. Each csv file is compressed as a zip. All these zip files are saved in the defined folder. You need now to upload them to **Talend Exchange** at http://www.talendforge.org/exchange/top/help_guest.php.

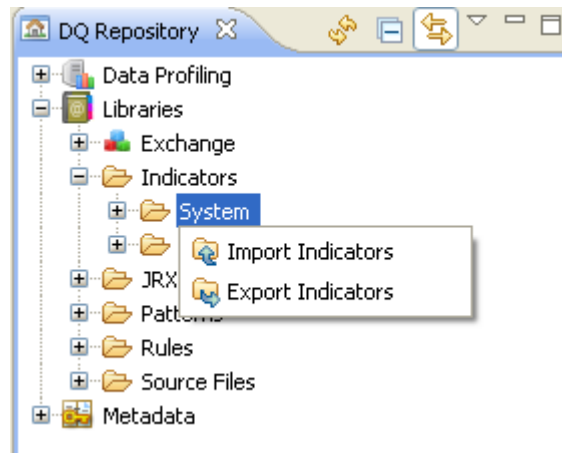
3.3.6 How to export system indicator to a definition file

In **Talend Open Profiler** you can export system indicators and store them locally in a definition file.

Prerequisite(s): **Talend Open Profiler** main window is open.

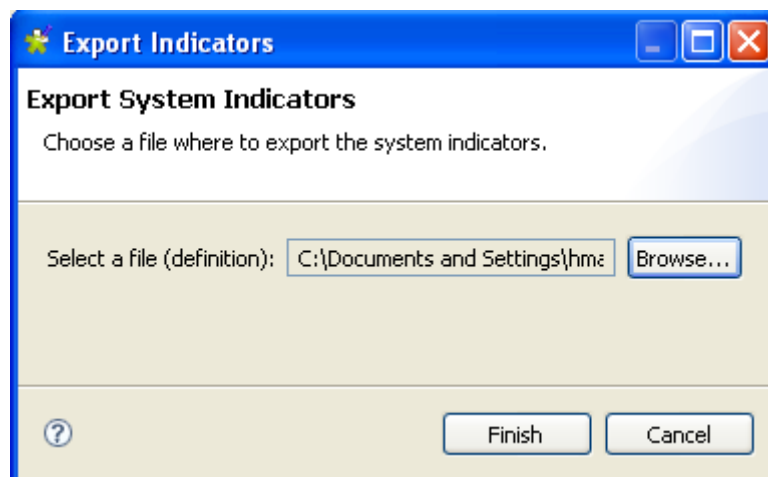
To export system indicators to a definition file:

- In the **DQ Repository** tree view, expand the **Libraries** and **Patterns** folders in succession and right-click **System**.



- From the drop-down list, select **Export Indicators**.

The [**Export Indicators**] wizard opens.



- Browse to the definition file where to save the indicators.
- Click **Finish** to close the wizard.

All exported system indicators are saved in the defined definition file.

3.3.7 How to import user-defined indicators from a csv file

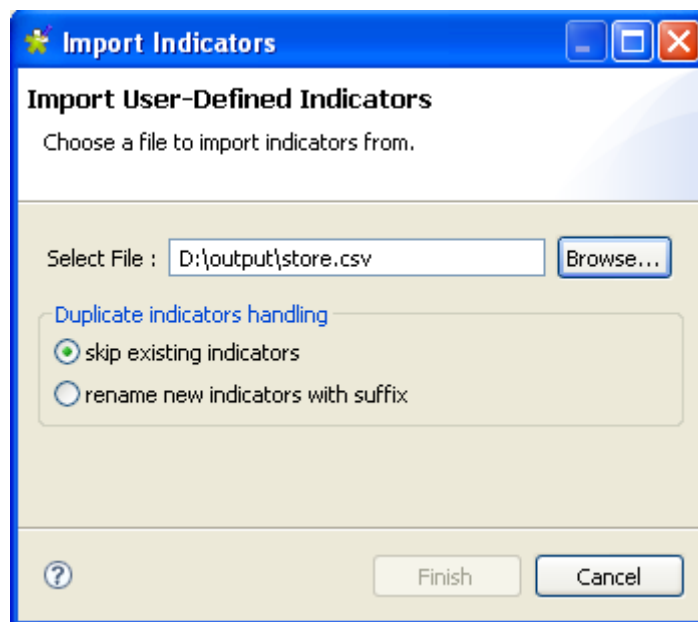
In **Talend Open Profiler** you can import indicators stored locally in a csv file to use them later on your column analyses.

Prerequisite(s): **Talend Open Profiler** main window is open. The csv file is stored locally.

To import user-defined indicators from a csv file:

- In the **DQ Repository** tree view, expand **Libraries** and **Indicators** in succession.
- Right-click **User Defined Indicators** and select **Import Indicators**.


The [**Import Indicators**] wizard opens.



- Browse to the csv file holding the user-defined indicators.
- In the **Duplicate indicators handling** panel, either click **skip existing indicators** to only import indicators that do not already exist in the indicators list in the **DQ Repository** tree view, or
- Click **rename new indicators with suffix** to identify all the imported indicators with a suffix.
- Click **Finish** to close the wizard.

All imported indicators are listed under the **User Defined Indicators** folder in the **DQ Repository** tree view.



A warning icon  next to the name of the imported user-defined indicator in the tree view identifies that it is not correct. You must open the indicator and try to figure out what is wrong.

3.3.8 How to import system indicators from a definition file

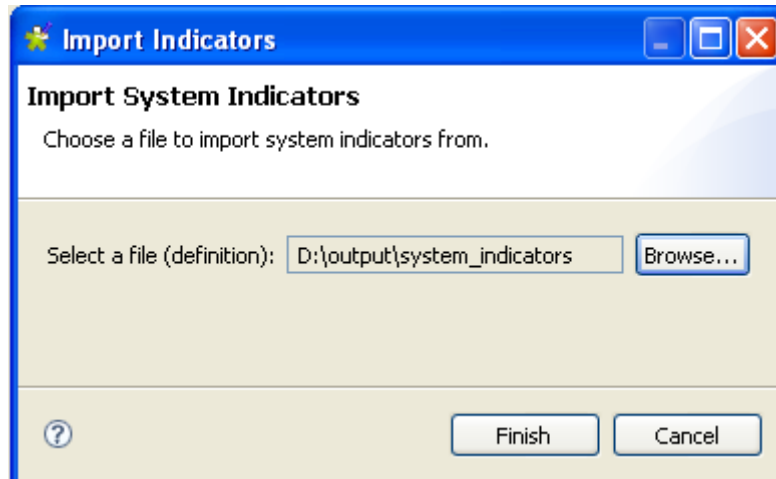
In **Talend Open Profiler** you can import system indicators stored locally in a definition file to use them later on your column analyses.

Prerequisite(s): **Talend Open Profiler** main window is open. The definition file is stored locally.

To import system indicators from a definition file:

- In the **DQ Repository** tree view, expand the **Libraries** and **Indicators** folders in succession.
- Right-click **System** and select **Import Indicators**.

The [**Import Indicators**] wizard opens.



- Browse to the definition file holding the system indicators.
- Click **Finish** to close the wizard.

All imported indicators are listed under the **System** folder in the **DQ Repository** tree view.

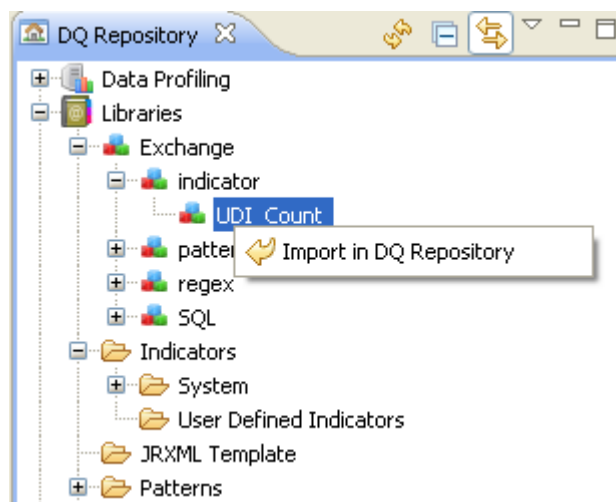
3.3.9 How to import indicators from Talend Exchange

You can import user-defined indicators created by other users and stored in **Talend Exchange** into your current version of **Talend Open Profiler** to use them later, as needed, on your column analysis.

Prerequisite(s): **Talend Open Profiler** main window is open.

To import user-defined indicators from Talend Exchange:

- In the **DQ Repository** tree view, expand **Libraries** and **Exchange** in succession.
- Under **Exchange**, expand **indicator** and right-click the indicator name you want to import and then select **Import in DQ Repository**.



A message displays to confirm the operation.

- Click **OK** in the confirmation message to close it.

The imported indicator from **Talend Exchange** is listed under the **User Defined Indicators** folder in the **DQ Repository** tree view.

3.3.10 How to set indicators for the columns to analyze

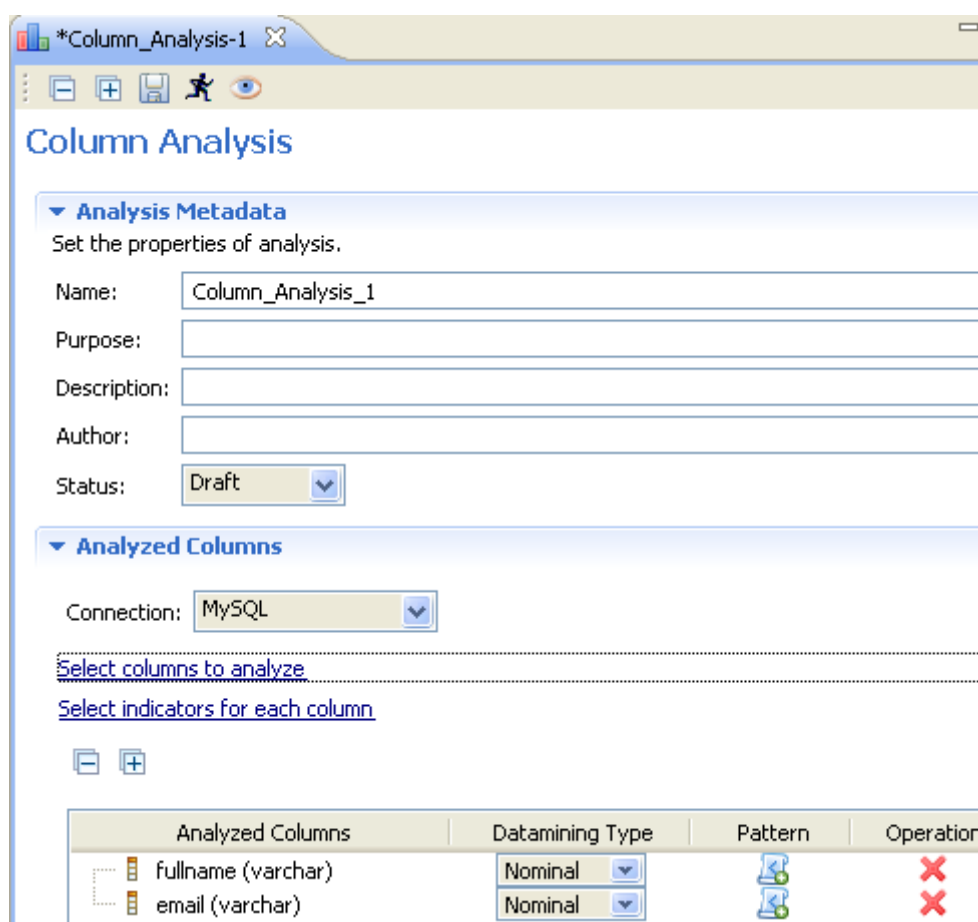
You can define indicators for columns of database tables that need to be analyzed or monitored.

Prerequisite(s): **Talend Open Profiler** main window is open. An analysis of a set of columns is open in the Column Analysis editor.

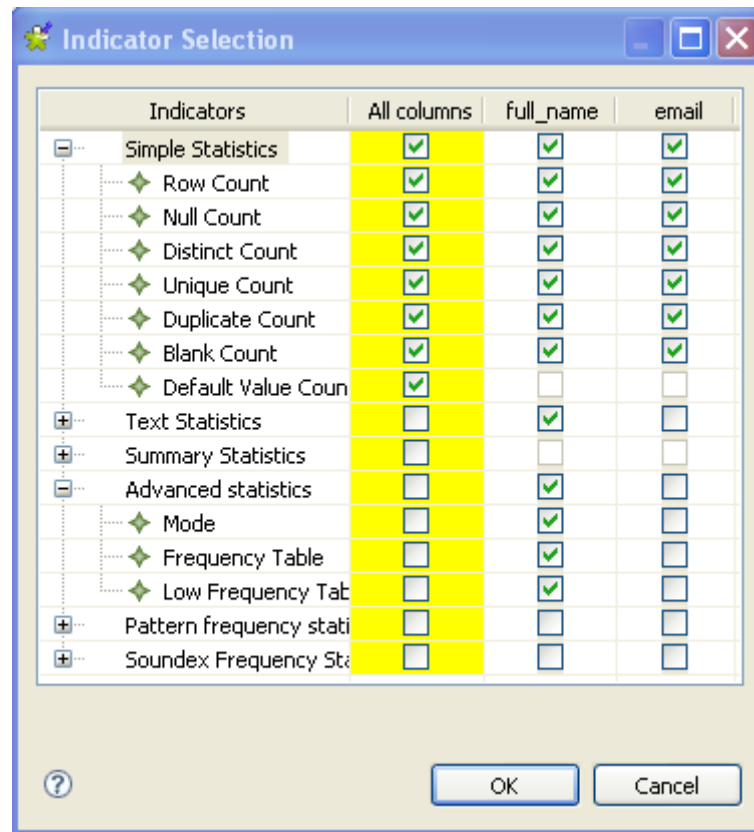
For more information about analyzing columns, see *How to analyze a set of columns on page 33*.

To set indicators for a column:

- In the Column Analysis editor, click **Analyzed Columns** to open the analyzed columns view.



- Click **Select indicators for each column** to open the **[Indicator Selection]** dialog box.



When you open the **[Indicator Selection]** dialog box, a help panel automatically opens with the wizard. This help panel guides you through the steps for setting indicators.

- Set indicator parameters for the analyzed columns as needed and click **OK** to close the dialog box. For more information about indicator types, see *Core features of Talend Open Profiler on page 3*, *Indicators on page 4*.
- Click **OK** to close the dialog box.

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

3.3.11 How to set options for indicators


Prerequisite(s): **Talend Open Profiler** main window is open. An analysis of a set of columns is open in the Column Analysis editor and a set of indicators is already defined for the analyzed columns.

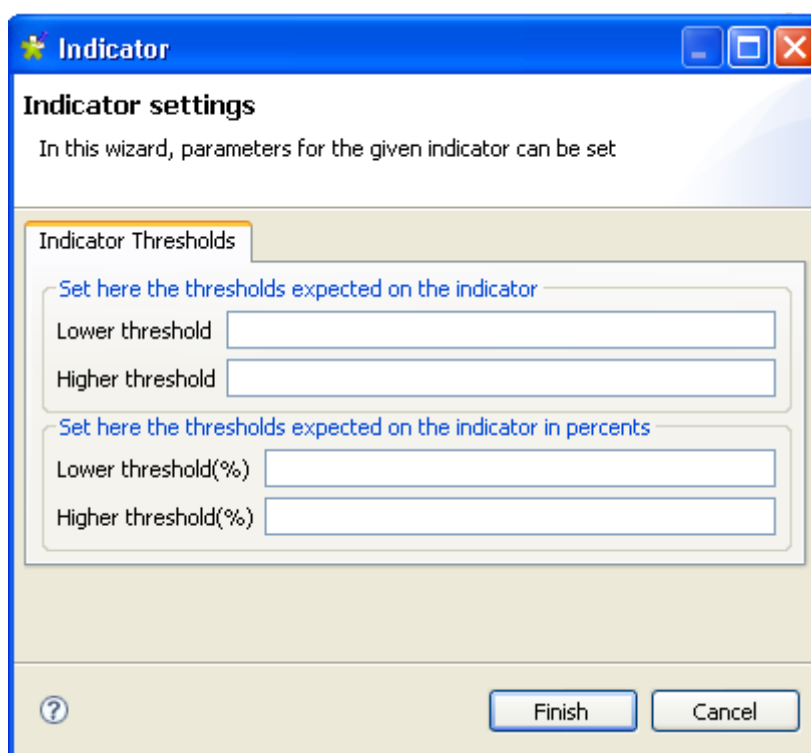
For more information about analyzing columns, see *How to analyze a set of columns on page 33*.


For more information about setting indicators, see *How to set indicators for the columns to analyze on page 112*.

To set options for indicators:

- In the Column Analysis editor, click **Analyzed Columns** to open the analyzed columns view.

- Click the option icon  next to the defined indicator to open the dialog box where you can set options for the given indicator.



 Indicators settings dialog boxes differ according to the parameters specific for each indicator. For more information about different indicator parameters, see *Indicators parameters on page 114*.

- Set the parameters for the given indicator.
- Click **Finish** to close the dialog box.

3.3.12 Indicators parameters

This section describes indicator parameters displayed in the different **Indicators Settings** dialog boxes.

Data Thresholds

Possible value	Description
Lower threshold	Data smaller than this value should not exist
Higher threshold	Data greater than this value should not exist

Indicator Thresholds

Possible value	Description
Lower threshold	Lower threshold of matching indicator values
Higher threshold	Higher threshold of matching indicator values

Possible value	Description
Lower threshold(%)	Lower threshold of matching indicator values in percentage relative to the total row count
Higher threshold(%)	Higher threshold of matching indicator values in percentage relative to the total row count

Bins Designer

Possible value	Description
Minimal value	Beginning of the first bin
Maximal value	End of the last bin
Number of bins	Number of bins

Text Length

Possible value	Description
Count nulls	When selected, null data are counted as zero length text field
Count blanks	When selected, blank texts (e.g. " ") are counted as zero length text fields

Text Parameters

Possible value	Description
Ignore case	When checked, comparison of text data is not case sensitive

Frequency Table Parameters

Possible value	Description
Number of results shown	Number of displayed results



APPENDIX A

Talend Open Profiler management GUI

This appendix describes the Graphical User Interfaces (GUI) of **Talend Open Profiler**.

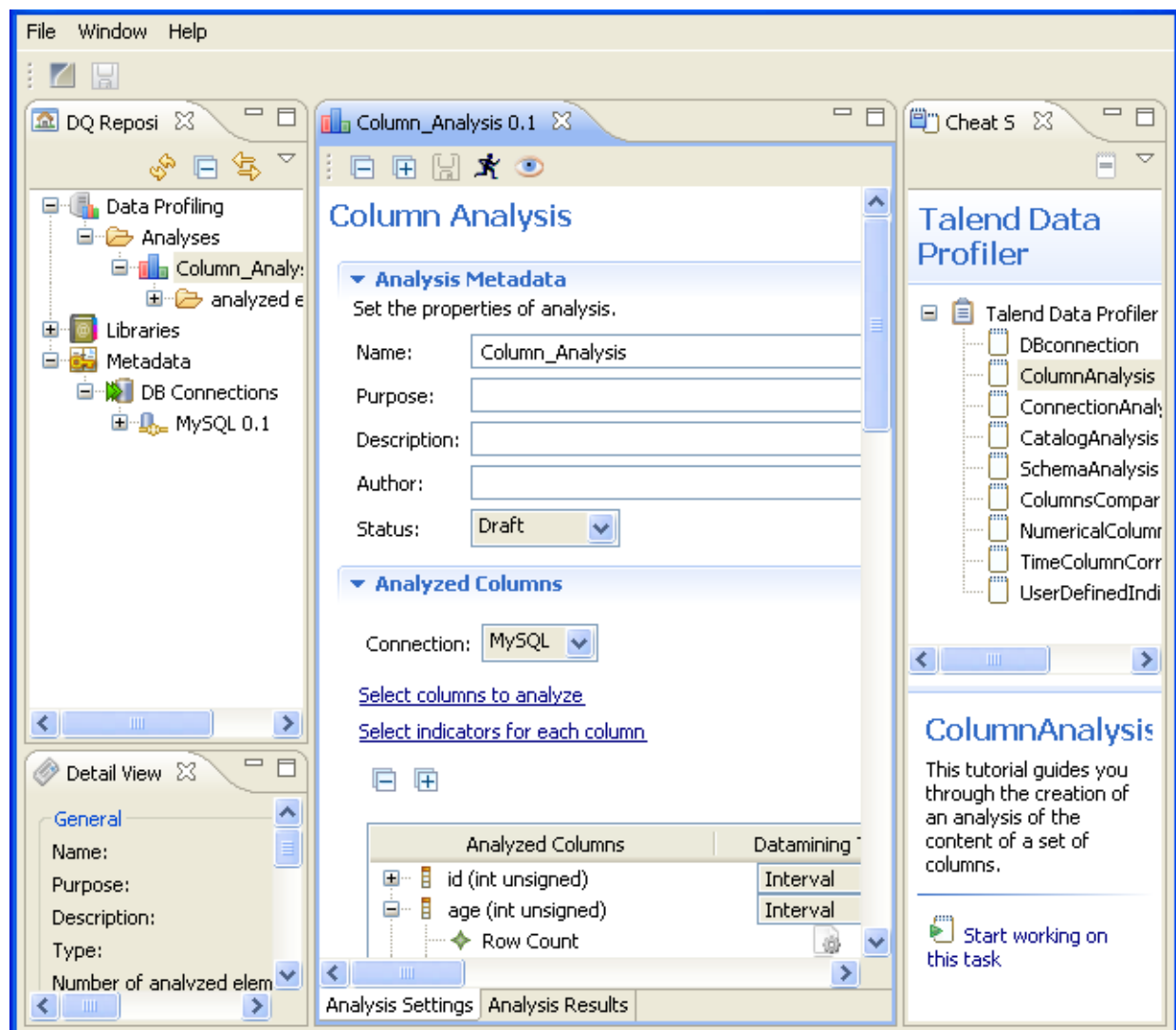
A.1 Main window of Talend Open Profiler

Talend Open Profiler main window is the interface from which you manage data profiling. For more information about managing data profiling, see *Data profiling management procedures on page 7*.

The **Talend Open Profiler** main window is divided into:

- the menu bar,
- the toolbar,
- the tree view area,
- a detail view
- the workspace,
- a tab panel (specific to Column Analysis editors),
- a cheat sheet view.

The figure below illustrates **Talend Open Profiler** main window and its possible views.



The following sections give detailed information about each of the above views.

A.2 Menu bar of Talend Open Profiler

The menu bar headers and submenus help you perform operations on your enterprise data.

Table 1 describes menus and menu items available to you.

Table 1—Management menus



Menu	Menu item	Description
File	Close	Closes the current open editor in the workspace
	Close All	Closes all open editors in the workspace
	Save	Saves any changes done in the current open editor
	Save All	Saves any changes done in all open editor
	Exit	Closes Talend Open Profiler main window
	Open File	Opens a file
Window	Perspective	Data profiler: Opens the data profiler GUI
		Data Explorer: Opens the data explorer GUI
	Preferences	Opens the [Preferences] window which enables you to set your preferences
	Reset Perspective...	Resets the current perspective to its default view after confirmation
Show View...	Opens the [Show View] dialog box which enables you to display different views on Talend Open Profiler	
Help	Welcome	Opens a welcoming page which has links to Talend Open Profiler documentation and Talend practical sites
	Help Contents	Opens the Eclipse help system documentation
	About Talend Open Profiler	Displays: <ul style="list-style-type: none"> -the software version you are using -detailed information on your software configuration that may be useful if there is a problem -detailed information about plug-in(s) -detailed information about Talend Open Profiler features
	Software Updates	Find and Install...: Opens the [Install/Update] wizard that helps searching for updates for the currently installed features, or searching for new features to install
		Manage Configuration...: Opens the [Product Configuration] window where you can manage Talend Open Profiler configuration
	Key Assist...	Opens a list of all short-cut keys
	View bookmarks	Opens a bookmarks panel that holds few useful links. These links enable you to easily access specific information related to the usage of Talend Open Profiler and/or its database management system

A.3 Toolbar of Talend Open Profiler

The toolbar contains icons that provide you with quick access to the commonly used operations you can perform from **Talend Open Profiler** main window.

Table 2 describes the toolbar icons and their functions.

Table 2—Management toolbar

Icon	Function
	Switches to Data Explorer
	Saves modifications

A.4 Tree view of Talend Open Profiler

The **DQ Repository** tree view area shows folders for data profiling analyses, patterns and metadata.

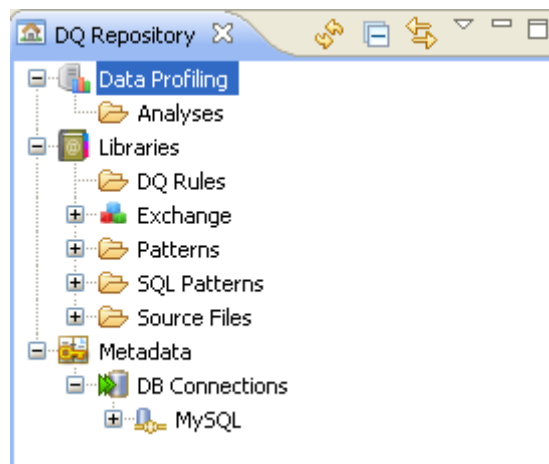
When expanding the **Data profiling** folder in the tree view list, you display the created analyses (either executed or not executed yet).

When expanding the **Libraries** folder in the tree view list, you display the list of the pre-defined patterns and SQL patterns. Imported patterns and patterns created by you will also show under the **Patterns** folder.

Under **Libraries** as well, you have all created DQ rules and all imported patterns from **Talend Exchange**.

When expanding the **Metadata** folder in the tree view list, you display the list of all created DB connections.

The figure below shows an example of an expanded **DQ Repository** tree view.

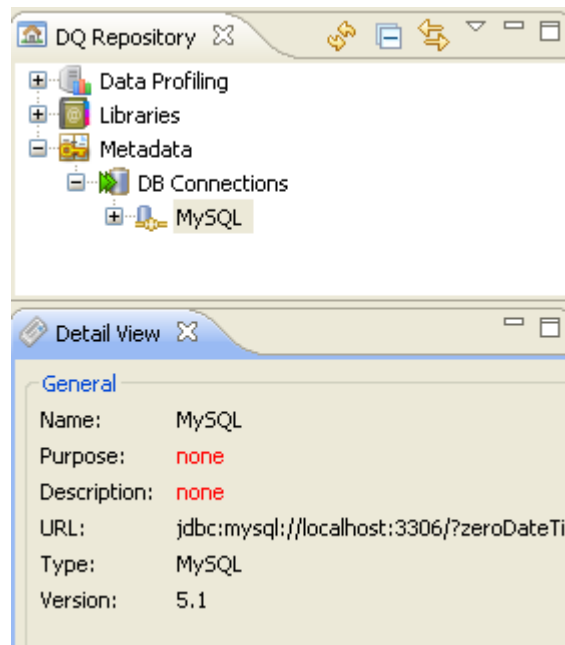


You can use the local toolbar icons to manage the display of the **DQ Repository** tree view.

A.5 Detailed View of Talend Open Profiler

This view shows below the **DQ Repository** tree view area. It displays detailed information about the selected element in the tree view area.

The figure below shows an example of the detailed view of the selected DB connection.



You can use the local toolbar icons to manage the display of Detail View.

A.6 Design workspace of Talend Open Profiler

This area contains:

- nothing if no analysis, pattern or DB connection is open,
- the parameter values of the open analysis, pattern or DB connection.

When you open a column analysis, a pattern or a DB connection through the tree view area, the relevant editor opens in **Talend Open Profiler** workspace.

You can use the local toolbar icons to manage the display of the workspace.

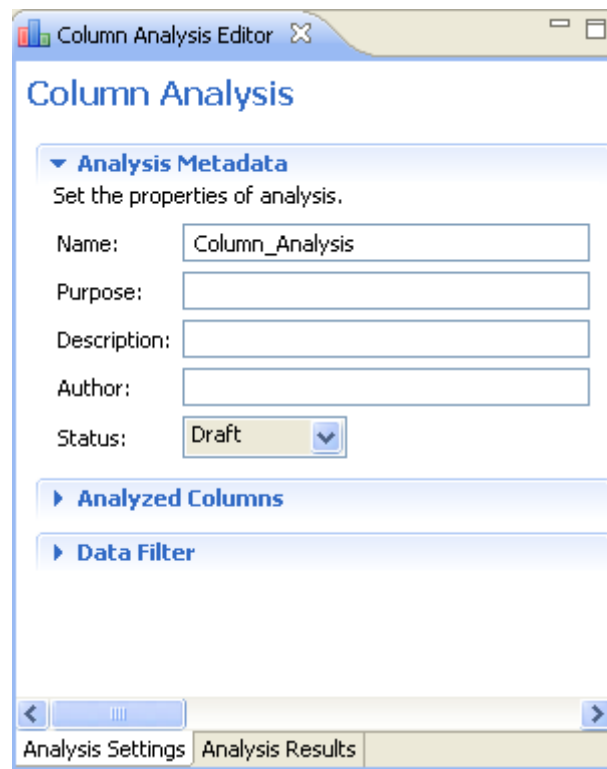
A.7 Tab panel of the column analysis editor

This management tab panel is located at the bottom of Column Analysis editors. It contains a pair of tabs:

- **Analysis Settings**,
- **Analysis Results**.

The **Analysis Settings** tab displays the settings for the current analysis in the Column Analysis editor.

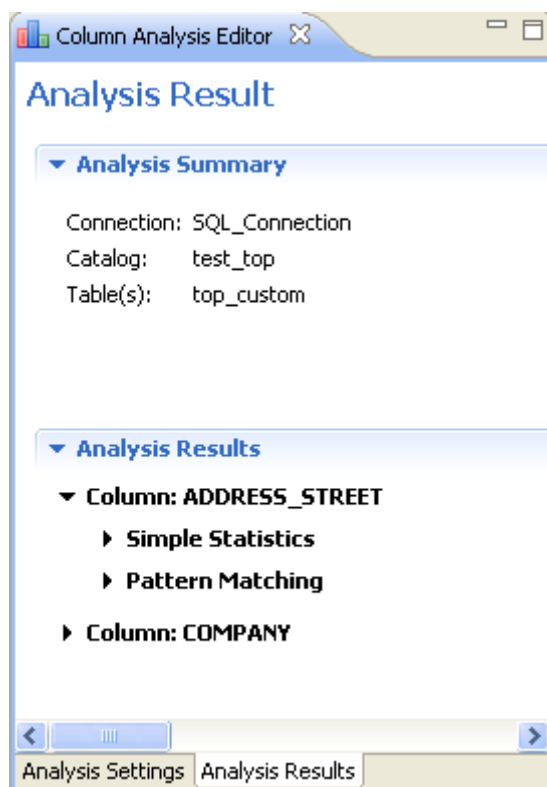
The figure below is an example of the parameters of a column analysis.



The **Analysis Results** tab displays.

- a summary of the executed analysis in the **Analysis Summary** view in which it specifies the connection, the database and the table names for the current analysis,
- the results of the executed analysis, graphics and tables, in the **Analysis Results** view.

The figure below is an example of a column analysis results.



In the **Analysis Results** view, you can:

- click the arrow located next to a column name to display the types of analyses done on that column,
- select a type of analysis to display the corresponding generated graphics and tables.

A.8 How to select a task from Talend Open Profiler management GUI

You have several ways to select a task from the **Talend Open Profiler** main window. You can, for example, use:

- a menu - submenu combination, or
- a toolbar icon, or
- a right-click list, or
- shortcut keys.

Example 1: To show a view in the **Talend Open Profiler** main window, either:

- use the **Window - Show View...** menu - submenu combination, or
- use the **Alt+Shift+Q, Q** shortcut key.

Example 2: To execute an analysis, do one of the followings:

- use the run icon on the toolbar, or
- right-click the analysis you want to execute and select **Run** from the contextual menu, or

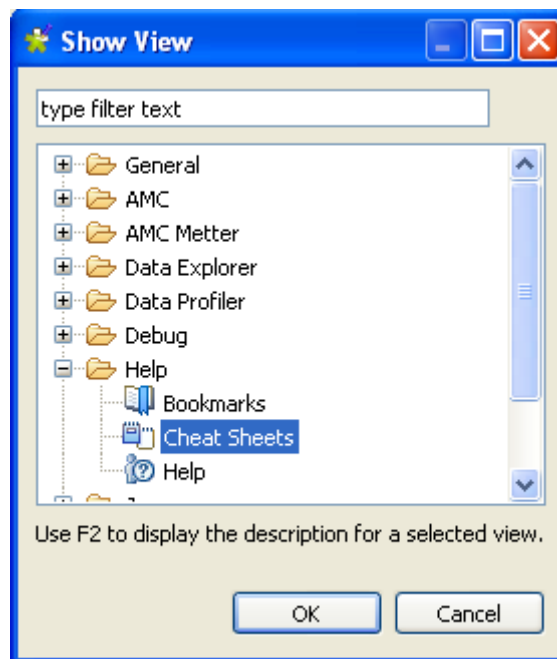
- click the **Run** button at the bottom of the editor, or
- use the **F6** shortcut key.

A.9 Cheat Sheets of Talend Open Profiler

The Cheat Sheets is a quick reference that guides you through all common tasks in **Talend Open Profiler**.

To display the Cheat Sheets:

- Either, press the **Alt+Shift+Q, H** shortcut key, or
- Select **Window - Show View** from the menu bar.



The [**Show View**] dialog box opens.

- Expand the **Help** folder and select **Cheat Sheets**.
- Click **OK** to close the dialog box.

The Cheat Sheets opens in the **Talend Open Profiler** main window. Use the local toolbar icons to manage the display of the Cheat Sheets.



APPENDIX B

Data Explorer management GUI

The Data Explorer embedded in **Talend Open Profiler** allows you to query and browse databases. This appendix introduces the Graphical User Interfaces (GUI) of the Data Explorer which is based on the SQL Explorer for which you can find documentation at <http://www.sqlexplorer.org/>.

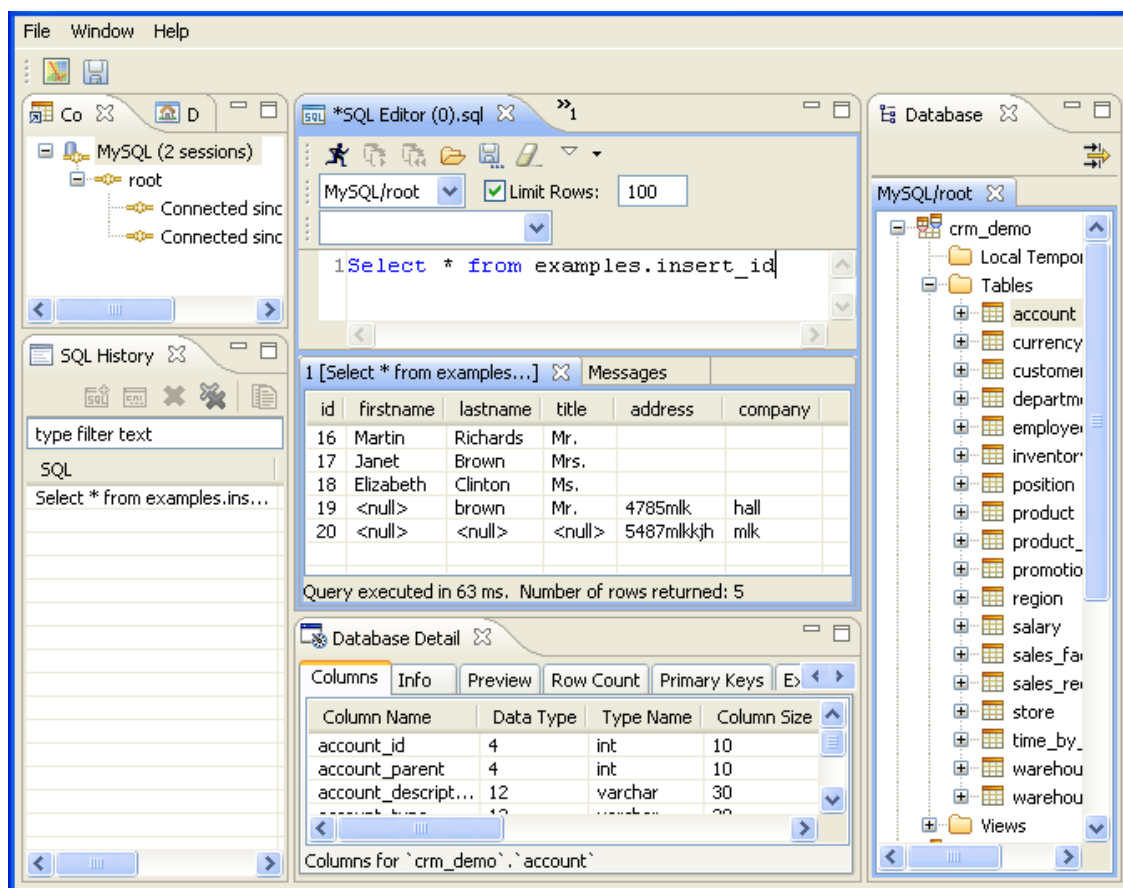
B.1 Main window of the Data Explorer

The main window of the Data Explorer is the interface from which you manage your database.

The Data Explorer main window is divided into:

- the menu bar,
- the toolbar,
- Connections view,
- SQL History view
- SQL editor view,
- Database Detail view,
- Database Structure view.

The figure below illustrates an example of Data Explorer main window and its components.



The following sections give detailed information about each of the above components.

B.2 Menu bar of the Data Explorer

The menu bar headers and submenus help you perform operations on your enterprise data.

Table 1 *Management menus on page 119* of Appendix A describes menus and menu items available to you.

B.3 Toolbar of the Data Explorer

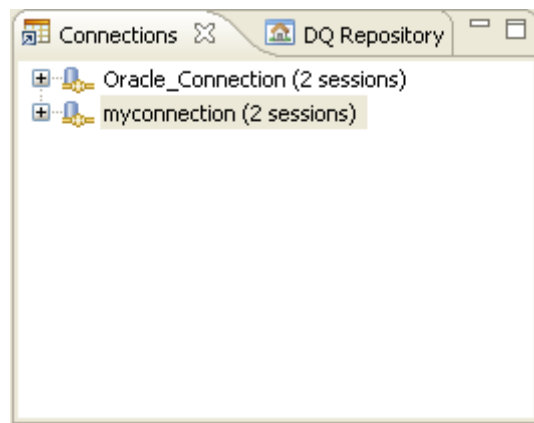
The toolbar contains icons that provide you with quick access to the commonly used operations you can perform from the Data Explorer main window.

Table 2 *Management toolbar on page 120* of Appendix A describes the toolbar icons and their functions.

B.4 Connections view

The Connections view shows all the connection profiles that you have set up.

The figure below shows an example of the Connections view.



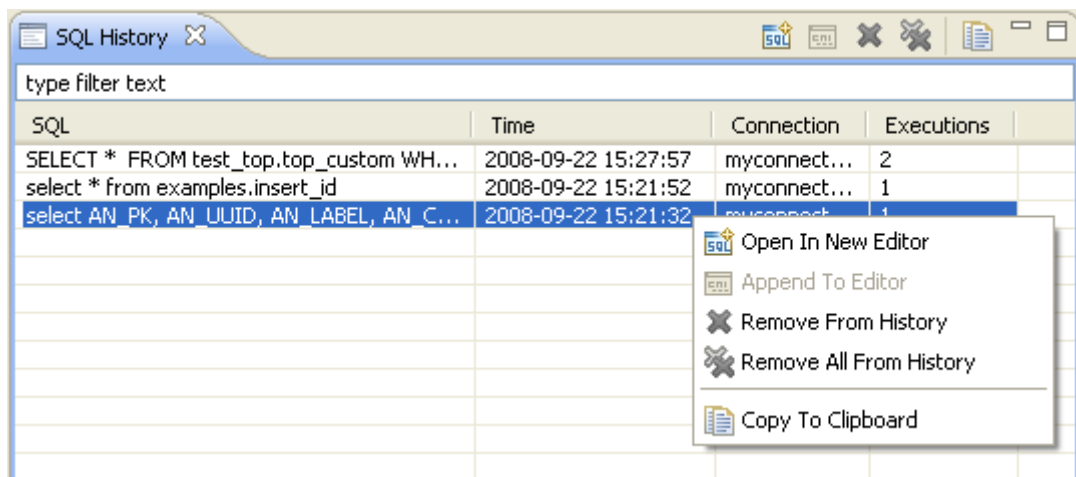
You can use the local toolbar icons to manage the display of the Connections view.

B.5 SQL History view

This view shows below the Connections view area. Every statement that was successfully executed is logged in the SQL History view.

The view shows the statement, the date and time when the statement was last executed, which connection was used and how many times the statement has been executed. The SQL statements can be filtered, sorted, removed and opened in or appended to the **[SQL Editor]**.

The figure below shows an example of the SQL History view.



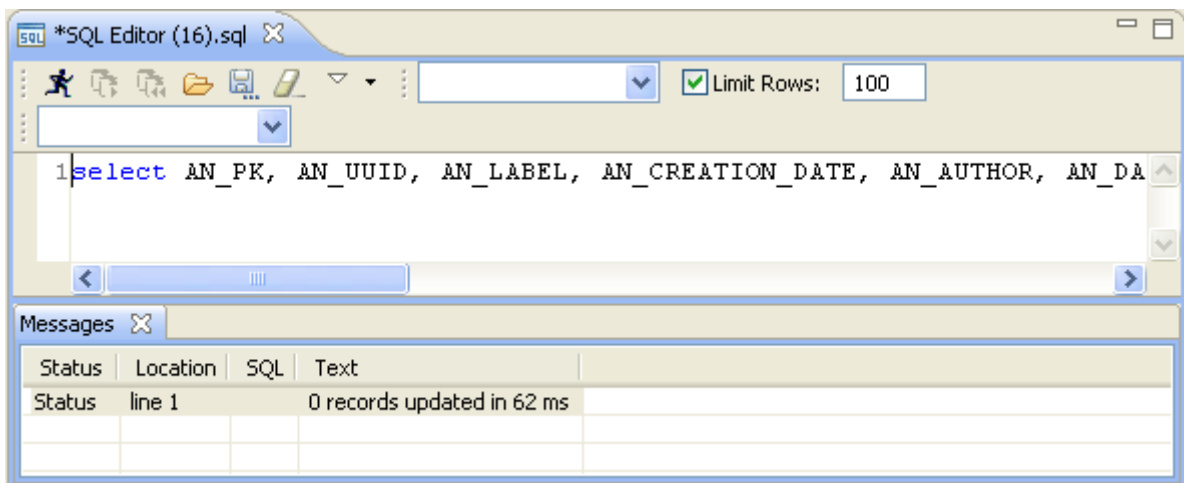
You can use the local toolbar icons to manage the display of SQL History View.

B.6 SQL Editor view

This area contains nothing if no **[SQL Editor]** is open. The **[SQL Editor]** provides the following features:

- executing queries using the CTRL-ENTER combination,
- Basic syntax coloring
- Basic Content Assist
- Overriding result limit
- Word wrapping (if enabled in preferences)
- Session/Catalog/Schema switching
- Loading/Saving SQL scripts
- Commit/Rollback buttons (if session is not in autocommit mode)
- Display of query execution time of last run query

The figure below shows an example of the **[SQL Editor]** view.

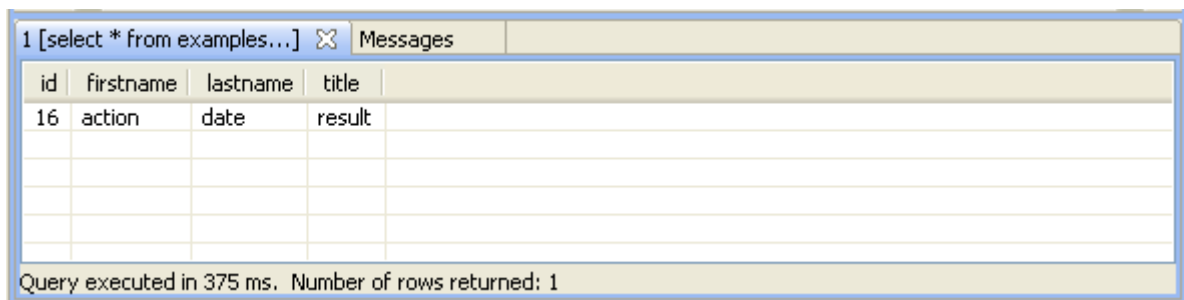


The lower part of the [SQL Editor] view, the Messages area, detail information about your data exploring actions. When you execute a query in the SQL query editor, the Messages area displays the query results.



You can save all the queries you execute in the Data Explorer under **Libraries - Source Files** in the DQ Repository tree view of [Talend Open Profiler](#).

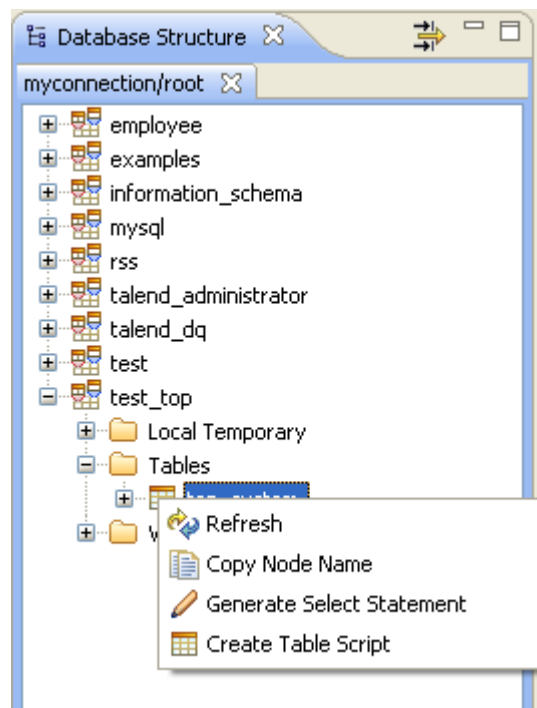
The figure below shows an example of the Messages area.



B.7 Database Structure view

Using the **Database Structure** view, you can explore multiple databases simultaneously.

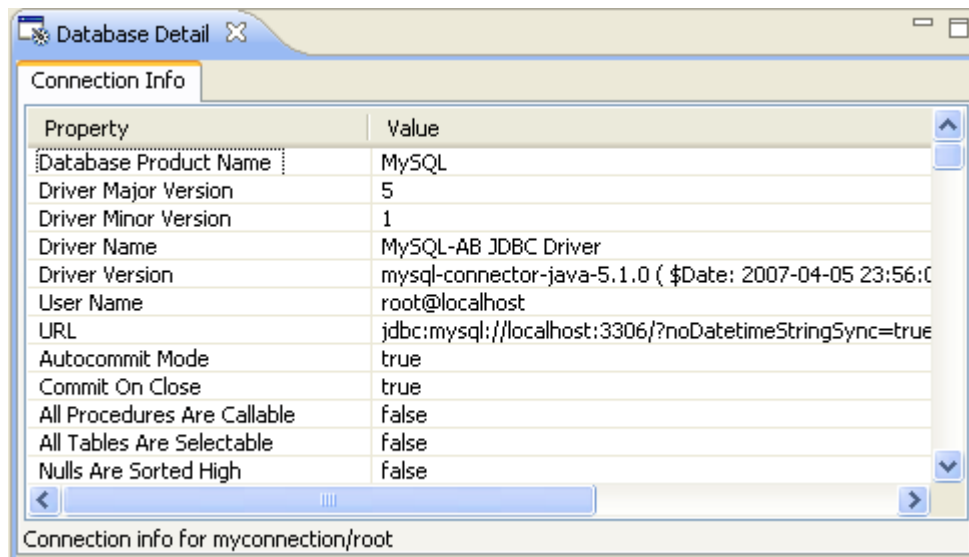
When you select a node in the Database Structure view, the corresponding detail is shown in the Database Detail view. For more information, see *Database Detail view on page 130*. If the detail view is not active, double clicking the node will bring the detail view to the front.

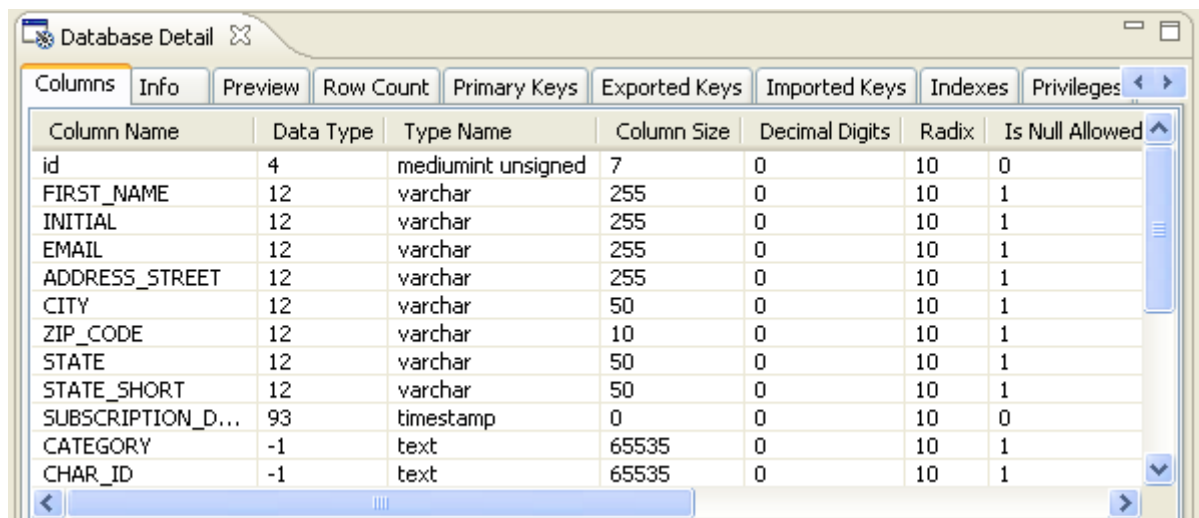


B.8 Database Detail view

Database Detail view shows detailed information for whatever node you select in the database structure view. What is displayed will depend on the type of database that you are using.

The figures below show two examples of a database detailed view.





The screenshot shows a 'Database Detail' window with a tabbed interface. The 'Columns' tab is selected, displaying a table with the following columns:

Column Name	Data Type	Type Name	Column Size	Decimal Digits	Radix	Is Null Allowed
id	4	mediumint unsigned	7	0	10	0
FIRST_NAME	12	varchar	255	0	10	1
INITIAL	12	varchar	255	0	10	1
EMAIL	12	varchar	255	0	10	1
ADDRESS_STREET	12	varchar	255	0	10	1
CITY	12	varchar	50	0	10	1
ZIP_CODE	12	varchar	10	0	10	1
STATE	12	varchar	50	0	10	1
STATE_SHORT	12	varchar	50	0	10	1
SUBSCRIPTION_D...	93	timestamp	0	0	10	0
CATEGORY	-1	text	65535	0	10	1
CHAR_ID	-1	text	65535	0	10	1

A	
Adding	
tasks	69
Advanced analysis	79
Analysis	
adding a task	77
deleting	76
duplicating	77
executing	76
opening	75
Analyzing	
columns	33
database content	21
C	
Comparing	
catalog and schema	15
column lists	18
table lists	17
Core features	3
Correlation analysis	46
Creating	
analysis with DQRule	60
catalog analysis	26
column analysis	33
column comparison analysis	42
content analysis	21
functional dependency analysis	65
new database connection	9, 12
nominal correlation analysis	55
numerical correlation analysis	46
schema analysis	30
time correlation analysis	51
D	
Data profiling	1
main concepts	2
problems addressed	2
Data profiling tool	2
Datamining types	41
interval	42
nominal	42
other	42
unstructured text	42
Declaring UDF	83
Deleting	
database connection	14
DQ rules	80
creating	80

opening	82
E	
Editing/Deleting UDF	84
Editor display	8
Exchange	
exporting patterns	99
exporting to Talend Exchange	108
importing indicators	111
importing patterns	96
G	
GUI	
cheat sheet	124, 129
detailed view	120, 127
main window	118, 126
menu bar	119, 126
select a task	123
tab panel	121
toolbar	119, 127
tree view	120, 127
workspace	121, 128
I	
Indicators	4
adding a task	72
advanced statistics	5
deleting a task	73
duplicating	106
editing	104
exporting	106, 108
pattern finder	5
simple statistics	4
soundex finder	5
summary statistics	5
text statistics	4
user defined	102
Indicators parameters	114
M	
Management procedures	7
Managing	
DQ rules	80
indicators	102
patterns	83
Metadata	3
O	
Opening	

database connection	13
P	
Pattern indicators	89
Patterns	3
adding	88
creating	85
deleting	93
duplicating	94
editing	90
exporting	97
importing	94, 109, 110
opening	92
Profiling	2
S	
Setting indicator parameters	113
Setting indicators	112
Synchronizing	
catalog and schema	19
columns	21
tables	20
Synchronizing metadata	19
T	
Talend Open Profiler	
managers	3
Tasks	69