Introduction to Talend Open Studio for Data Integration

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

Talend’s data integration solution helps companies deal with growing system complexities by addressing both ETL for analytics and ETL for operational integration needs and offering industrialization features.
Prerequisites to using Talend Open Studio for Data Integration

This chapter provides basic software and hardware information required and recommended to get started with your Talend Open Studio for Data Integration.

- Memory requirements on page 5
- Software requirements on page 5

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

- Installing Java on page 6
- Setting up the Java environment variable on Windows on page 6 or Setting up the Java environment variable on Linux on page 6
- Installing 7-Zip (Windows) on page 7

Memory requirements

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

<table>
<thead>
<tr>
<th>Memory usage</th>
<th>3GB minimum, 4 GB recommended</th>
</tr>
</thead>
<tbody>
<tr>
<td>Disk space</td>
<td>3GB</td>
</tr>
</tbody>
</table>

Software requirements

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

Required software

- Operating System for Talend Studio:

<table>
<thead>
<tr>
<th>Support type</th>
<th>Operating system (64 bits only)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Recommended</td>
<td>Ubuntu 18.04 LTS</td>
</tr>
<tr>
<td>Recommended</td>
<td>Microsoft Windows 10</td>
</tr>
<tr>
<td>Supported</td>
<td>Apple macOS 10.14/Mojave</td>
</tr>
<tr>
<td></td>
<td>Apple macOS 10.13/High Sierra</td>
</tr>
<tr>
<td></td>
<td>Apple macOS 10.12/Sierra</td>
</tr>
</tbody>
</table>

- You should install either Oracle Java 8 or 11 JRE, or OpenJDK 1.8 or 11.
- A properly installed and configured MySQL database, with a database named gettingstarted.
Optional software

- 7-Zip.

Installing Java

To use your Talend product, you need a JAVA environment installed on your computer.

This documentation is about Oracle JRE installation. For more information on OpenJDK, see http://openjdk.java.net/.

Procedure

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

Results

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

- Setting up the Java environment variable on Windows on page 6.
- Setting up the Java environment variable on Linux on page 6.

Setting up the Java environment variable on Windows

Prior to installing your Talend product, you need to set the JAVA_HOME and Path environment variables.

Procedure

1. Go to the Start Menu of your computer, right-click on Computer and select Properties.
2. In the Control Panel Home window, click Advanced system settings.
3. In the System Properties window, click Environment Variables....
4. Under System Variables, click New... to create a variable. Name the variable JAVA_HOME, enter the path to the Java 8 JRE, and click OK.
   
   Example of default JRE path: C:\Program Files\Java\jre1.8.0_77.
5. Under System Variables, select the Path variable and click Edit... to add the previously defined JAVA_HOME variable at the end of the Path environment variable, separated with semi colon.
   
   Example: <PathVariable>;%JAVA_HOME%\bin.

Setting up the Java environment variable on Linux

Prior to installing your Talend product, you have to set the JAVA_HOME and Path environment variables.
Prerequisites to using Talend Open Studio for Data Integration

Procedure

1. Find the JRE installation home directory.
   
   Example: `/usr/lib/jvm/jre1.8.0_65`

2. Export it in the `JAVA_HOME` environment variable.
   
   Example:
   ```
   export JAVA_HOME=/usr/lib/jvm/jre1.8.0_65
   export PATH=$JAVA_HOME/bin:$PATH
   ```

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

4. Log on again.

**Installing 7-Zip (Windows)**

Talend recommends to install 7-Zip and to use it to extract the installation files: [http://www.7-zip.org/download.html](http://www.7-zip.org/download.html).

Procedure

1. Download the 7-Zip installer corresponding to your Operating System.

2. Navigate to your local folder, locate and double-click the 7z exe file to install it.

Results

The download will start automatically.
Downloading and installing Talend Open Studio for Data Integration

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

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

Downloading Talend Open Studio for Data Integration

Talend Open Studio for Data Integration is a free open source product that you can download directly from Talend’s Website.

Procedure

2. When prompted, click Save File and then OK.

Results

The zip file of download Talend Open Studio for Data Integration is downloaded.

Installing Talend Studio

Unzip the archive previously downloaded to install Talend Studio.

You can do it either by using:

- 7Zip (Windows recommended): Extracting via 7-Zip (Windows recommended) on page 8.

Extracting via 7-Zip (Windows recommended)

For Windows, Talend recommends you to install 7-Zip and use it to extract files.

To install the Studio, follow the steps below:

Procedure

1. Navigate to your local folder, locate the ZIP file previously downloaded and move it to another location with a path as short as possible and without any space character.
   Example: C:/Talend/
2. Unzip it by right-clicking on the compressed file and selecting 7-Zip > Extract Here.
Extracting via Windows default unzipping tool

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

Procedure

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

Extracting via the Linux GUI unzipper

To install the Studio, follow the steps below:

Procedure

1. Navigate to your local folder, locate the previously downloaded ZIP file and move it to another location with a path as short as possible and without any space character.
   
   Example: home/user/talend/

2. Unzip it by right-clicking on the compressed file and selecting Extract Here.
Configuring and setting up your Talend product

This chapter provides basic information required to configure and set up your Talend Open Studio for Data Integration.

Launching the Studio for the first time

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

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

Procedure

1. Double-click the executable file corresponding to your operating system, for example:
   - TOS_*=win-x86_64.exe, for Windows.
   - TOS_*=linux-gtk-x86_64, for Linux.
   - TOS_*=macosx-cocoa.app, for Mac.
2. In the User License Agreement dialog box that opens, read and accept the terms of the end user license agreement to proceed.

Logging on to the Studio

To log on to the Talend Studio for the first time, do the following:

Procedure

1. In the Talend Studio login window, select Create a new project, specify the project name: getting_started and click Finish to create a new local project.
2. Depending on the product you are using, either of the following opens:
   - the Quick Tour. Play it to get more information on the User Interface of the Studio, and click Stop to end it.
   - the Welcome page. Follow the links to get more information about the Studio, and click Start Now! to close the page and continue opening the Studio.

Tip:

After your Studio successfully launches, you can also click the Videos link on the top of the Studio main window to watch a couple of short videos that help you get started with your Talend Studio. For some operating systems, you may need to install an MP4 decoder/player to play the videos.

Results

Now you have successfully logged on to the Talend Studio. Next you need to install additional packages required for the Talend Studio to work properly.
Installing additional packages

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

Before you begin

**Warning:** Make sure that the `-Dtalend.disable.internet` parameter is not present in the Studio .ini file or is set to false.

**Procedure**

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

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

   For more information, see the section about installing additional packages in the Talend Open Studio for Data Integration Installation and Upgrade Guide.

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

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

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

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

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

You will work with data about movies and directors and data about your customers as you learn how to filter data in order to separate movie entries with valid director information from those without.

Reading movies information from a CSV file

The examples provided in this chapter assume that:

- You have launched your Talend Studio and opened the Integration perspective.
- You have installed all the required third-party libraries and database drivers in your Talend Studio.
- You have properly installed and configured the MySQL database software, and created a database named gettingstarted.

In this scenario, you will learn:

- How to create a data integration Job. See Creating your first Job on page 12 for details.
- How to add and link components in a data integration Job. See Dropping and linking components on page 13 for details.
- How to create file metadata in the Repository. See Preparing the movies metadata on page 14 for details.
- How to configure and execute a data integration Job. See Configuring and executing your Job on page 18 for details.

If you want to replicate the example described in this document and use the exact input data, you can download tos_di_gettingstarted_source_files.zip from the Downloads tab of the online version of this page at https://help.talend.com, and then save the source files in your local directory C:\getting_started\input_data\.

Creating your first Job

This procedure describes how to create a Job folder named getting_started and a Job named movies in the folder.

Procedure

1. In the Repository tree view, right click the Job Designs node, and select Create folder from the contextual menu.
2. In the New Folder wizard, name your Job folder getting_started and click Finish to create your folder.
3. Right-click the getting_started folder and select Create Job from the contextual menu.
4. In the New Job wizard, give a name to the Job you are going to create and provide other useful information if needed.

In this example, enter movies in the the Name field.
Performing data integration tasks

In this step of the wizard, **Name** is the only mandatory field. The information you provide in the **Description** field will appear as hover text when you move your mouse pointer over the Job in the **Repository** tree view.

5. Click **Finish** to create your Job.

An empty Job is opened in the Studio.

**Dropping and linking components**

This example describes how to add and link components in the newly created Job, so that it will read a CSV file and display the data on the console.

**Procedure**

1. Drop a **tFileInputDelimited** and a **tLogRow** component from the **Palette** onto the design workspace.

   You can find the **tFileInputDelimited** component in the **Input** group of the **File** family and the **tLogRow** component in the **Logs & Errors** family in the **Palette**.

2. Click the **tFileInputDelimited** component so that an **o** icon appears, drag and drop the **o** icon onto the **tLogRow** component.

   The two components are now connected via a **Row > Main** connection.
Performing data integration tasks

Results

Now you have added the required components to the Job. In the next steps you will need to prepare the required metadata and configure the Job.

Preparing the movies metadata

This example describes how to set up the metadata of the source file movies.csv in the Repository. Repository metadata can be used across Jobs, allowing you to configure your Jobs quickly without having to define each parameter and schema manually.

Before you begin

• You have the source file movies.csv ready in the directory C:\getting_started \input_data\.

Procedure

1. In the Repository tree view, expand the Metadata node, right-click File delimited, and select Create file delimited from the contextual menu to open the New Delimited File wizard.
2. In the New Delimited File wizard, enter a name for the file metadata, movies in this example, and other useful information to better describe your file metadata, and then click Next to go to the next step and define the general properties of the file.
In this step of the wizard, **Name** is the only mandatory field. The information you provide in the **Description** field will appear as a tooltip when you move your mouse pointer over the file connection.

3. In the **File** field specify the path of the source file, or click **Browse** to browse to the file.
Performing data integration tasks

The file is loaded, and the File Viewer area displays an abstract of the file, allowing you to check the file consistency, the presence of header and more generally the file structure.

4. From the Format list, select your operating system, and click Next to parse the file.

5. On the Preview tab, select the Set heading row as column names check box to retrieve the file column names from the first row, and then click Refresh Preview.
Performing data integration tasks

6. If the file contains more than one heading row, which need to be skipped in file parsing, specify the number in this field and click Refresh Preview again.

7. Click Next to retrieve the file schema.

   The Description of the Schema table displays the generated file schema.

8. Name the schema movies_schema and check the file schema and edit it according to your actual needs.

   In this example, increase the length of the title and url columns.
Performing data integration tasks

9. Click **Finish** to validate the schema close the wizard.

The created file metadata is shown in the **Repository** tree view.

**Results**

You now have the movies file metadata ready for use. Next, you need to apply the created metadata to the component that reads the source file.

**Configuring and executing your Job**

This example describes how to configure the components using the metadata created in the previous procedure and run your Job.
Performing data integration tasks

**Procedure**

1. In the **Repository** tree view, double-click the Job **movies** to open it in the design workspace. You can skip this step if the Job is already open and active in the design workspace.

2. In the **Repository** tree view, expand **Metadata > File delimited**, and drag and drop the file connection **movies** or its schema **movies_schema** onto the **tFileInputDelimited** component in the design workspace. When asked whether to propagate the changes to the output component, click **Yes**.

   In the **Basic settings** tab of the **Component** view, you’ll find that all the parameters of the component have been automatically filled.

   ![Diagram of tFileInputDelimited and tLogRow components](image)

3. Double-click the **tLogRow** component to open its **Basic settings** tab view.

4. In the **Mode** area, select the **Vertical (each row is a key/value list)** option for better readability of long fields on the **Run** console.
Performing data integration tasks

5. Press F6 or click the Run button on the Run view to execute your Job.

Results

The Run console displays the movies information read from the source file.

Filtering the movies information

This scenario will extend the Job described in Reading movies information from a CSV file on page 12 to filter the data flow to get only those movies with valid director information.

This scenario demonstrates:

- How to duplicate a Job. See Duplicating the existing Job on page 24 for details.
- How to add a component by typing its name on a connection or on the design workspace. See Adding a mapping component on page 25 for details.
Performing data integration tasks

- How to drop a metadata item or its schema as a component on the design workspace. See Adding a lookup component on page 27 for details.
- How to perform basic processing to data flows using tMap. See Configuring mappings and executing the Job on page 29 for details.

Preparing directors file metadata

This procedure shows how to set up the metadata of the reference file directors.txt in the Repository. This metadata item will be used to add and set up the lookup input in this scenario.

Before you begin

- You have the file directors.txt ready in the directory C:\getting_started\input_data.

Procedure

1. In the Repository tree view, expand the Metadata node, right-click File delimited, and select Create file delimited from the contextual menu to open the New Delimited File wizard.
2. Enter a name for the file connection, directors in this example, and other useful information to better describe your file metadata, and then click Next to go to the next step and define the general properties of the file.

3. In the File field specify the path of the source file, or click Browse to browse to the file.
Performing data integration tasks

The file is loaded, and the File Viewer area displays an abstract of the file, allowing you to check the file consistency, the presence of header and more generally the file structure.

4. Select Windows from the Format list, and click Next to parse the file.
5. From the Field Separator list of the File Settings area, select Comma.
6. Click **Next** to retrieve the file schema.

The **Description of the Schema** table displays the generated file schema.

7. Name the schema `directors_schema` and rename the columns to `directorID` and `directorName` respectively, and change the data type of the `directorID` columns from `Integer` to `String`.
8. Click **Finish** to validate the schema close the wizard.

The created file metadata is shown in the **Repository** tree view.

![Metadata Tree View]

Results

You now have the directors file metadata ready for use when you set up the component to read the reference file.

**Duplicating the existing Job**

This procedure shows how to create a Job based on an existing Job.
Performing data integration tasks

Before you begin

- You have created and successfully executed the Job named movies as described in Reading movies information from a CSV file on page 12.

Procedure

1. In the Repository tree view, right-click the Job named movies and select Duplicate from the contextual menu.

2. In the Duplicate dialog box, enter a name for the new Job, filter_movies in this example, and click OK to validate the Job creation and close the dialog box.

The Job named filter_movies is created, which is a duplicate of the Job named movies.

Adding a mapping component

The procedure below shows how to add a mapping component by typing the component name directly on the existing connection.

Procedure

1. In the new Job named filter_movies, select the Row connection linking the tFileInputDelimited and tLogRow components, and type name of tMap or part of it.
Performing data integration tasks

When you start typing the component name, a list of components that match your input appears. You can select a component to view its description besides the component list.

2. Double-click tMap on the list to added it onto the connection.

The newly added tMap component is now connected with the input component, and a dialog box opens asking you to give a name to the new output connection.

3. Enter a name for the new output connection, Valid_movies in this example, and click OK. When asked whether you want to propagate the input schema to the target output component, click Yes.

Results

The tMap component is now added to the Job and connected with the two existing components via Row > Main connections.
Performing data integration tasks

Adding a lookup component

The procedure below shows how to add a lookup input component from the Repository, connect it to the tMap, and enable column trimming in the component.

Before you begin

• You have centralized the metadata for directors.txt in the Repository as described in Preparing directors file metadata on page 21.

Procedure

1. In the Repository tree view, expand Metadata > File delimited, drag and drop the file connection directors or its schema directors_schema onto the design workspace.

   The Components dialog box opens, showing a list of components you can add to the Job from this metadata item.

2. Select tFileInputDelimited and click OK.

   A tFileInputDelimited labelled directors is added to the design workspace, with its basic settings automatically filled.
3. Right-click the newly added `tFileInputDelimited` component, select Row > Main from the contextual menu, and click the `tMap` component.

The `tFileInputDelimited` is connected to the `tMap` via a lookup connection now.

4. In the Advanced settings tab of the new `tFileInputDelimited` component, and select the Trim all columns check box.
Some records of the reference input file `directors.txt` contains leading white spaces. This option allows you to remove such white spaces from the lookup input flow when the Job is executed.

### Results

You have now all the components in the Job needed for filtering the movies information. Next you’ll need to configure mappings in the tMap component to filter the main input flow against the lookup flow and output the desired information.

### Configuring mappings and executing the Job

The procedure below shows how to configure mappings and an inner join to output movies information with valid director IDs.

#### Procedure

1. Double-click the tMap component to open the map editor.

   The map editor shows three tables, named row1, row2 and Valid_movies in this example, corresponding respectively to the movies file schema, the directors file schema, and the schema of the output for valid movies information, and columns in the row1 table are already mapped to the columns in the Valid_movies table.
2. Select the `directorID` column in the `row1` table, and drop it onto the `directorID` column in the `row2` table to create a join between the two input data sets based on the director IDs.

3. Click the `tMap settings` button, then click `Value` field for `Join Model`, and then click the `[...]` button that appears to open the `Options` dialog box. In the dialog box, select `Inner Join` and click `OK` to define the join as an inner join.
Performing data integration tasks

With this setting, only the movie records with the director IDs matching with those in the reference file will be passed to the output.

4. In the **Schema editor** at the bottom of the map editor, select **directorID** column of the output schema, **Valid_movies** in this example, and click the **X** button to remove it.

5. Click the **[+]** button beneath the output table to add a new column, name it **directedBy**, set its length to **20**, and move it up so that it’s between the **title** and **releaseYear** columns.

6. Select the **directorName** column in the **row2** table, and drop it to the **Expression** field corresponding to the **directedBy** column in the output table.

   A new mapping is created between lookup table and the output table.
7. Click OK to validate the mappings and close the map editor, and click Yes when asked whether to propagate the changes.

The mapping configurations are saved and the output schema is synchronized to the output component tLogRow.

8. Press F6 or click the Run button on the Run view to execute your Job.

Results

Only movie records with valid director information are displayed on the Run console.

**Gathering rejected movies information and saving processing results to a database**

Based on the scenario described in Filtering the movies information on page 20, this scenario further extends the Job to gather movies data missing director information and writes both valid and invalid data to a MySQL database.

This scenario demonstrates:
• How to add a component by typing on the design workspace or dragging from an existing component. See Adding database output components to your Job on page 33 for details.
• How to configure mappings for rejected information in tMap. See Configuring mappings for rejected data on page 35 for details.
• How to configure database outputs. See Configuring MySQL database outputs on page 36 for details.

Adding database output components to your Job

In the example below we will create a new Job from the Job filter_movies and add two tMySqlOutput components. These components will be used to write the processed movies information to the specified database tables.

Before you begin
• You have created and successfully executed the Job filter_movies as described in Filtering the movies information on page 20.

Procedure
1. Create a new Job by duplicating the Job created in the previous scenario, and name the new Job write_movies_to_db, and then double-click the Job to open it in the design workspace.
2. Right-click the tLogRow component and select Delete from the contextual menu to delete it.
3. Click where the tLogRow was on the design workspace and type the name of tMySqlOutput or part of it, and then select and double-click tMySqlOutput on the list to add it onto the design workspace.

When you start typing the component name, a list of components that match your input appears. You can select a component to view its description besides the component list.

4. Right-click the tMap component, select Row > Valid_movies from the context menu, and click the tMySqlOutput to link it with the tMap.

The connection name Valid_movies corresponds to the name of the existing output table in tMap.
5. Click the **tMap** component, and drag and drop the o icon onto the design workspace.

A text field and a list of suggested components appear. You can select a component to view its description besides the component list.

6. In the text field, type the name of **tMySqlOutput**, select the component on the list, and press Enter to add another **tMySqlOutput** component onto the design workspace.

A dialog box appears, asking you to enter a name for the output connection.

7. In the dialog box, enter **Invalid_movies** and click **OK** to connect **tMap** to the second **tMySqlOutput** component.
Performing data integration tasks

Results

Now you have added and connected the database output components you need to write the processed movies information to a MySQL database. Next, you’ll need to configure new mappings in the tMap and database settings in the tMySqlOutput components.

Configuring mappings for rejected data

This procedure shows how to configure mappings to gather rejected information.

Procedure

1. Double-click the tMap component to open the map editor.

An second output table named Invalid_movies has been automatically created.

2. Drop the movieID and title columns from the row1 table to the Invalid_movies table.
3. Click the **tMap settings** button on the **Invalid_movies** table, click the **Value** field for **Catch lookup inner join reject**, and then click the [...] button that appears to open the **Options** dialog box. In the dialog box, select **true** and click **OK**.

With this setting, any records without director IDs or with director IDs that do not match with those in the reference file will be passed to this output.

4. Click **OK** to validate the mappings and close the map editor, and click **Yes** when asked whether to propagate the changes.

The mapping configurations are saved and the output schema is synchronized to the output component.

### Results

Now you have configured mappings for the rejected output. Next, you’ll need to configure the output components to write the output flows to database tables.

#### Configuring MySQL database outputs

This procedure shows how to configure database output components to write movies information to MySQL database tables.

### Procedure

1. Double-click the first **tMySqlOutput** component to open its **Basic settings** in the **Component** view.
2. Provide the connection details needed to access your database, including the host name or IP address, port number, database name, user name and password, in the relevant fields.

When entering your password, you need first to click the [...] button next to the Password field to open a dialog box, enter your password between double quotation marks in the text field, and then click OK.

3. In the Table field, enter the name of the target database table.

In this example, the table for valid movies information is **valid_movies**.

4. Select the Action on table and Action on data options according to your needs.

In this example, we want to remove the table first if it already exists and then create an empty one, and use the default option for the action on data.

5. In the Basic settings of the second tMySqlOutput component, use the same settings as in the first tMySqlOutput except the name for the target database table.

In this example, the table for invalid movies information is **invalid_movies**.

6. Press F6 or click the Run button on the Run view to execute your Job.

**Results**

The movies records with valid director information are saved to the database table named **valid_movies**, and those without valid director information are saved to the database table named **invalid_movies**.

**What’s next?**

You have seen how Talend Studio helps you manage your data using Talend Jobs. You have learned how to access your data via Talend Studio, filter and transform your data, and store the filtered and transformed data in a database. Along the way, you have learned how to centralize frequently used connections in the Repository and easily reuse these connections in your Jobs.

To learn more about Talend Studio, see:

- Talend Studio User Guide
- Talend components documentation
Performing data integration tasks

To ensure that your data is clean, you can try Talend Open Studio for Data Quality and Talend Data Preparation Free Desktop.

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