



Talend Open Studio for Big Data

Getting Started Guide

6.0.0

Adapted for v6.0.0. Supersedes previous releases.

Publication date: July 2, 2015

Copyright

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

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

Notices

Talend is a trademark of Talend, Inc.

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

License Agreement

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

This product includes software developed at AOP Alliance (Java/J2EE AOP standards), ASM, Amazon, AntLR, Apache ActiveMQ, Apache Ant, Apache Avro, Apache Axiom, Apache Axis, Apache Axis 2, Apache Batik, Apache CXF, Apache Cassandra, Apache Chemistry, Apache Common Http Client, Apache Common Http Core, Apache Commons, Apache Commons Bcel, Apache Commons JXPath, Apache Commons Lang, Apache DataFu, Apache Derby Database Engine and Embedded JDBC Driver, Apache Geronimo, Apache HCatalog, Apache Hadoop, Apache Hbase, Apache Hive, Apache HttpClient, Apache HttpComponents Client, Apache JAMES, Apache Log4j, Apache Lucene Core, Apache Neethi, Apache Oozie, Apache POI, Apache Parquet, Apache Pig, Apache PiggyBank, Apache ServiceMix, Apache Squirrel, Apache Thrift, Apache Tomcat, Apache Velocity, Apache WSS4J, Apache WebServices Common Utilities, Apache Xml-RPC, Apache Zookeeper, Box Java SDK (V2), CSV Tools, Cloudera HTrace, ConcurrentLinkedHashMap for Java, Couchbase Client, DataNucleus, DataStax Java Driver for Apache Cassandra, Ehcache, Ezmorph, Ganymed SSH-2 for Java, Google APIs Client Library for Java, Google Gson, Groovy, Guava: Google Core Libraries for Java, H2 Embedded Database and JDBC Driver, Hector: A high level Java client for Apache Cassandra, Hibernate BeanValidation API, Hibernate Validator, HighScale Lib, HsqlDB, Ini4j, JClouds, JDO-API, JLine, JSON, JSR 305: Annotations for Software Defect Detection in Java, JUnit, Jackson Java JSON-processor, Java API for RESTful Services, Java Agent for Memory Measurements, Jaxb, Jaxen, JetS3T, Jettison, Jetty, Joda-Time, Json Simple, LZ4: Extremely Fast Compression algorithm, LightCouch, MetaStuff, Metrics API, Metrics Reporter Config, Microsoft Azure SDK for Java, Mondrian, MongoDB Java Driver, Netty, Ning Compression codec for LZ4 encoding, OpenSAML, Paracel JDBC Driver, Parboiled, PostgreSQL JDBC Driver, Protocol Buffers - Google's data interchange format, Resty: A simple HTTP REST client for Java, Riak Client, Rocoto, SDSU Java Library, SL4J: Simple Logging Facade for Java, SQLite JDBC Driver, Scala Lang, Simple API for CSS, Snappy for Java a fast compressor/decompressor, SpyMemCached, SshJ, StAX API, StAXON - JSON via StAX, Super SCV, The Castor Project, The Legion of the Bouncy Castle, Twitter4J, Uuid, W3C, Windows Azure Storage libraries for Java, Woden, Woodstox: High-performance XML processor, Xalan-J, Xerces2, XmlBeans, XmlSchema Core, Xmlsec - Apache Santuario, YAML parser and emitter for Java, Zip4J, atinject, dropbox-sdk-java: Java library for the Dropbox Core API, google-guice. Licensed under their respective license.

Table of Contents

Preface	v
1. General information	v
1.1. Purpose	v
1.2. Audience	v
1.3. Typographical conventions	v
2. Feedback and Support	v
Chapter 1. Getting Started with Talend Studio	1
1.1. Launching Talend Studio	2
1.1.1. How to launch the Studio for the first time	2
1.1.2. How to connect to TalendForge	3
1.1.3. How to access a Repository	5
1.1.4. How to set up a project	6
1.2. Working with different workspace directories	7
1.2.1. How to create a new workspace directory	7
1.2.2. How to connect to a different workspace directory	8
1.3. Working with projects	10
1.3.1. How to create a project	10
1.3.2. How to import the demo project	12
1.3.3. How to import projects	14
1.3.4. How to open a project	16
1.3.5. How to delete a project	17
1.3.6. How to export a project	18
Chapter 2. Getting started with Talend Big Data using the demo project	21
2.1. Introduction to the Big Data demo project	22
2.1.1. Hortonworks_Sandbox_Samples	22
2.1.2. NoSQL_Examples	24
2.2. Setting up the environment for the demo Jobs to work	24
2.2.1. Installing Hortonworks Sandbox	24
2.2.2. Understanding context variables used in the demo project	25
Chapter 3. Working in <i>Talend Studio</i> - basic Job examples	29
3.1. Getting started with a basic Job	30
3.1.1. Creating a Job	30
3.1.2. Adding components to the Job	32
3.1.3. Connecting the components together	35
3.1.4. Configuring the components	36
3.1.5. Executing the Job	38

Preface

1. General information

1.1. Purpose

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

Information presented in this document applies to *Talend Open Studio for Big Data* **6.0.0**.

1.2. Audience



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



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

1.3. Typographical conventions

This guide uses the following typographical conventions:

- text in **bold**: window and dialog box buttons and fields, keyboard keys, menus, and menu and options,
- text in **[bold]**: window, wizard, and dialog box titles,
- text in *courier*: system parameters typed in by the user,
- text in *italics*: file, schema, column, row, and variable names,
- text in *italics*: file, schema, column, row, and variable names,
- The  icon indicates an item that provides additional information about an important point. It is also used to add comments related to a table or a figure,
- The  icon indicates a message that gives information about the execution requirements or recommendation type. It is also used to refer to situations or information the end-user needs to be aware of or pay special attention to.
- Any command is highlighted with a grey background or code typeface.

2. Feedback and Support

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

<http://talendforge.org/forum>



Chapter 1. Getting Started with Talend Studio

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

1.1. Launching Talend Studio

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

1.1.1. How to launch the Studio for the first time

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

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



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

2. In the **[User License Agreement]** dialog box that opens, read and accept the terms of the end user license agreement to proceed.
3. In the *Talend Studio* login window, select an option to define your project that will hold all Jobs and Business models designed in the Studio.



This login window appears only when the Studio is started for the first time. When you launch the Studio again, the normal login window opens, which provides one more option, a connection list box, for subscription-based users to select a repository connection when launching the Studio.

If you plan to use the same repository connection and / or project at your next Studio launch, you can skip the login window to speed up Studio launch by clearing the **Always ask me at startup** check box. Then, if you want to see the login window again, go to the menu **Window > Preferences** to open the **[Preferences]** window, select **Talend**, and select the **Always show project dialog at startup** check box.

- Select **Create a new project**, specify a project name and click **Finish** to create a new project. For more information, see [How to create a project](#).
- Select **Import a demo project** and click **Finish** to import a demo project that includes numerous samples of ready-to-use Jobs. This Demo project can help you understand the functionalities of different *Talend* components. For more information, see [How to import the demo project](#).
- Select **Import an existing project** and click **Finish** to import an existing projects. For more information, see [How to import projects](#).

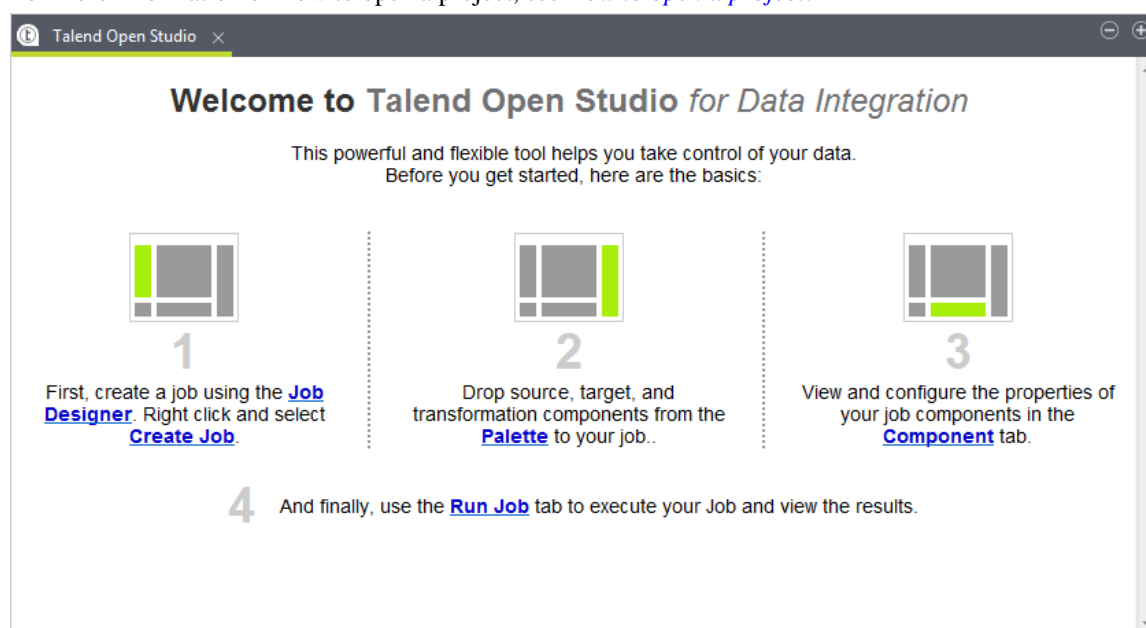
- If you want to modify the default repository connection, click **Manage Connections** to set up your connection before setting up a project. For further information about connecting to a repository, see [How to access a Repository](#).

As the purpose of this procedure is to create a new project, select **Create a new project**, fill in a project name in the text field, and click **Finish**.

The **[Welcome]** window opens. From this window you have direct links to Demo projects, user documentation, tutorials, **Talend** forum, **Talend** on-demand training and **Talend** latest news.

4. Click **Start now!** to open *Talend Studio* main window, which displays a welcome page that provides useful tips for beginners on how to get started with the Studio. Clicking an underlined link brings you to the corresponding tab view or opens the corresponding dialog box.

For more information on how to open a project, see [How to open a project](#).



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

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

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

1.1.2. How to connect to TalendForge

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

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

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

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

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

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



Be assured that any personal information you may provide to **Talend** will never be transmitted to third parties nor used for any purpose other than joining and logging in to the **Talend** Community and being informed of **Talend** latest updates.

Connect to TalendForge

Connect your Studio to TalendForge, the Talend Online Community.

- Download **new components and connectors** from Talend Exchange.
- Access the most recent **Documentation and Tech articles** from Talend social knowledgebase.
- See the latest messages in the Talend **Discussions Forums**.

talend FORGE

user ✓

user@comapny.com ✓

..... ✓

..... ✓

United States ▼

I agree to the [TalendForge Terms of Use](#)

I want to help to improve Talend by sharing anonymous usage statistics

CREATE ACCOUNT

Connect to Existing Account | Skip this Step

If you already have created an account at <http://www.talendforge.org>, click **Connect to Existing Account**, fill in your user name and password, and click **CONNECT TO MY ACCOUNT** to sign in the **Talend Community** and continue launching the Studio.

Connect your Studio to TalendForge, the Talend Online Community.

- Download **new components and connectors** from Talend Exchange.
- Access the most recent **Documentation and Tech articles** from Talend social knowledgebase.
- See the latest messages in the Talend **Discussions Forums**.



✔

✔

I want to help to improve Talend by sharing anonymous usage statistics

CONNECT TO MY ACCOUNT

[Create a New Account](#)

[Skip this Step](#)



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

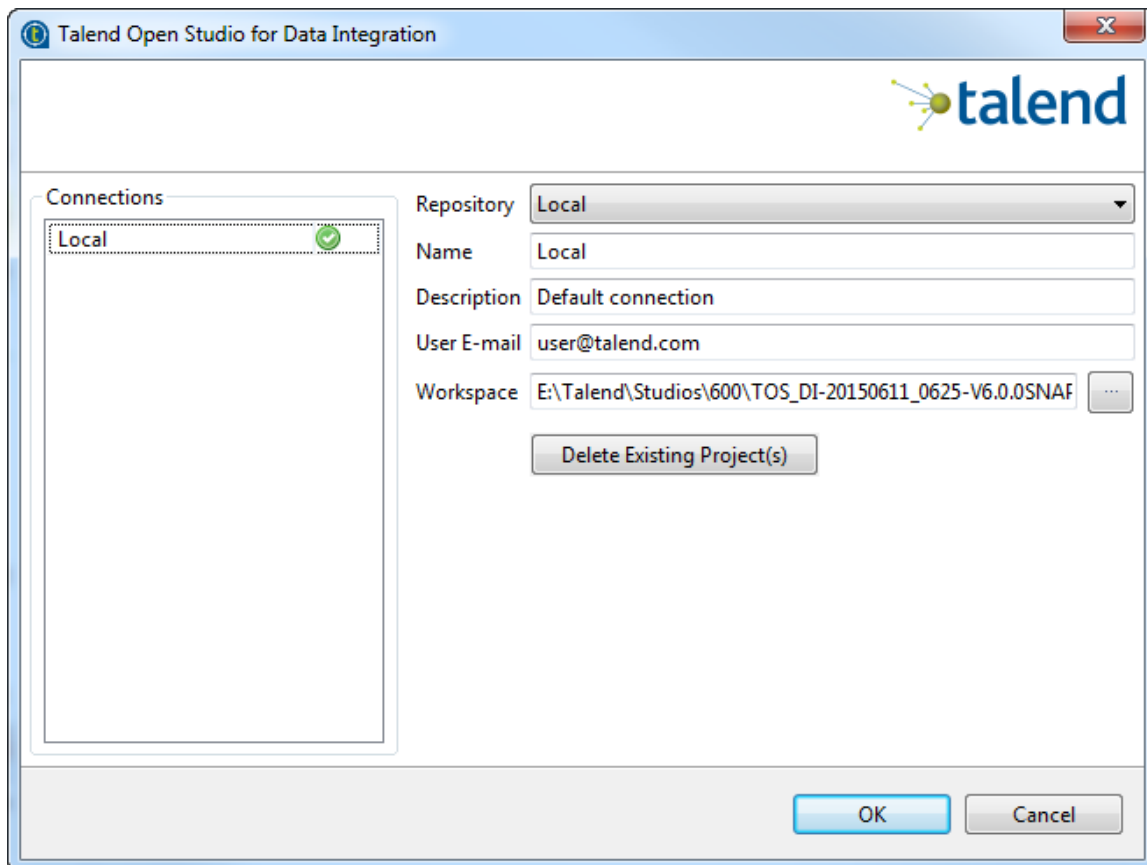
1.1.3. How to access a Repository

When launching *Talend Studio*, you can connect to a local repository where you store the data for your projects, including Jobs and business models, metadata, routines, etc. You can also connect to a remote repository where you store the same type of data to work collaboratively on projects.

1.1.3.1. How to connect to a local repository

To set a connection to a local repository, do the following:

1. On the login window of *Talend Studio*, click the **Manage Connections** button to open the repository connection setup dialog box.



Depending on the Studio product you are using, the product information displayed in your Studio may differ slightly from what is shown above.

2. If needed, type in a name and a description for your connection in the relevant fields.
3. In the **User E-mail** field, type in the email address that will be used as your user login. This field is compulsory to be able to use *Talend Studio*.

Be aware that the email entered is never used for purposes other than logging in.

4. By default, the **Workspace** field shows the path to the current workspace directory which contains all of the folders belonging to the project created. To change the workspace directory, type in the name of an existing directory or click the [...] button next to the **Workspace** field and browse to your preferred workspace directory. Upon changing your workspace directory, unless it is the first startup, you need to restart your *Talend Studio* by clicking the **Restart** button back on the login window for your change to take effect.

For more information about workspace directories, see [Working with different workspace directories](#).

5. Click **OK** to validate your changes and return to the login window.

1.1.4. How to set up a project

To open *Talend Studio*, you must first set up a project.

You can set up a project by:

- creating a new project. For more information, see [How to create a project](#).
- importing one or more projects you already created in other sessions of *Talend Studio*. For more information, see [How to import projects](#).

- importing the Demo project. For more information, see [How to import the demo project](#).

1.2. Working with different workspace directories

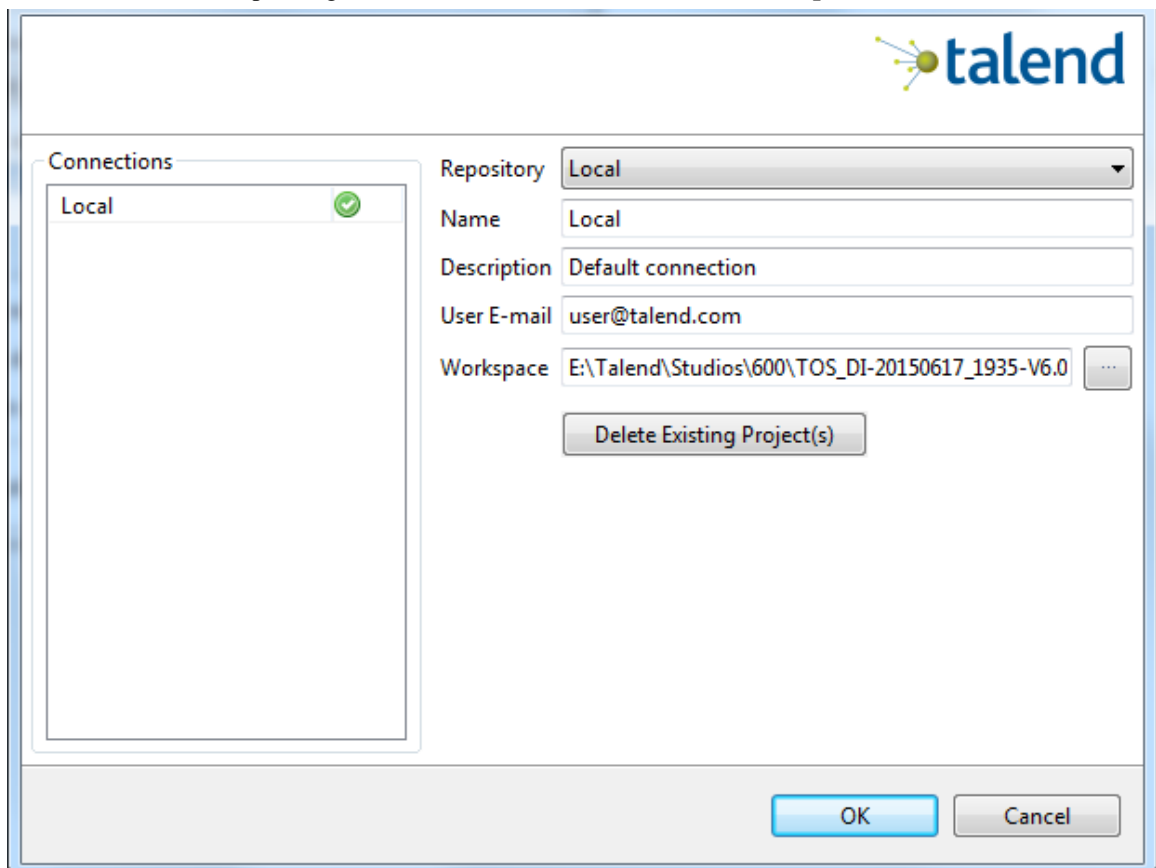
Talend Studio makes it possible to create many workspace directories and connect to a workspace different from the one you are currently working on, if necessary.

This flexibility enables you to store these directories wherever you want and give the same project name to two or more different projects as long as you store the projects in different directories.

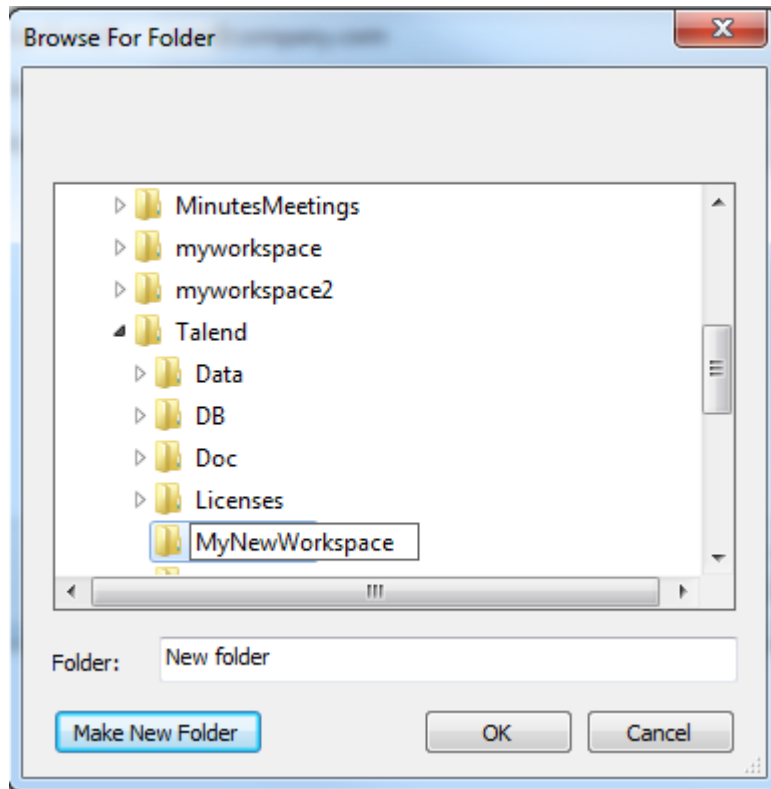
1.2.1. How to create a new workspace directory

Talend Studio is delivered with a default workspace directory. However, you can create as many new directories as you want and store your project folders in them according to your preferences.

1. If you have already started the Studio, select **File > Switch Project or Workspace** from the menu bar to restart the Studio.
2. On the login window, click **Manage Connections** to open the connection setup dialog box.
3. On the connection setup dialog box, click the [...] button next to the **Workspace** field.



4. In the **[Browse For Folder]** dialog box, browse to the parent directory under which you want to create a new workspace directory, click **Make New Folder**, and enter the name of your new workspace directory. Then click **OK** to validate directory creation and close the dialog box.

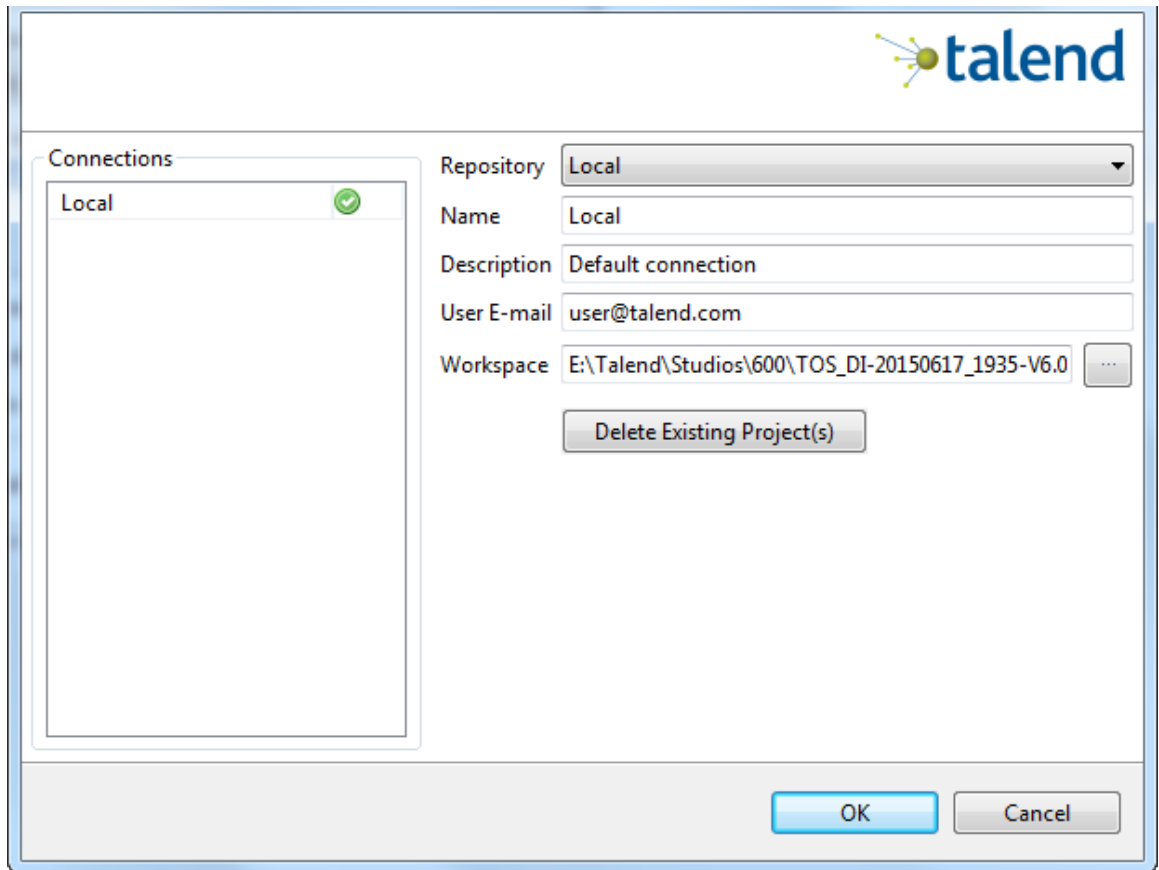


5. Click **OK** to validate your connection setup and go back to the login window.
6. Back on the login window, click the **Restart** button to restart your *Talend Studio* for the change to take effect.

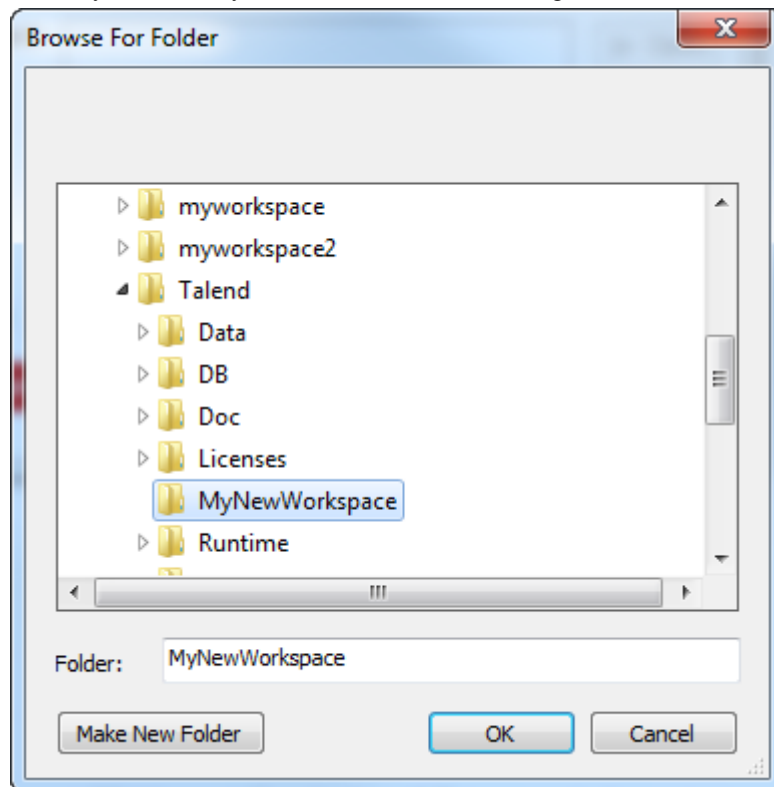
1.2.2. How to connect to a different workspace directory

In *Talend Studio*, you can select the workspace directory you want to store your project folders in according to your preferences.

1. If you have already started the Studio, select **File > Switch Project or Workspace** from the menu bar to restart the Studio.
2. On the login window, click the **Manage Connections** button to open the connection setup dialog box.
3. On the connection setup dialog box, click the [...] button next to the **Workspace** field.



4. In the **[Browse For Folder]** dialog box, browse to your preferred folder to use as the new workspace directory, and click **OK** to validate your directory selection and close the dialog box.



5. Click **OK** to validate your connection setup and go back to the login window.

6. Back on the login window, click the **Restart** button to restart your *Talend Studio* for the change to take effect.

1.3. Working with projects

In *Talend Studio*, the highest physical structure for storing all different types of data integration Jobs, metadata, routines, etc. is the "project".

From the login window of the Studio, you can:

- create a local project.

When you launch the Studio for the first time, there are no default projects listed. You need to create a project that will hold all data integration Jobs and business models you design in the current instance of the Studio.

You can create as many projects as you need to store your data of different instances of your Studio.

When creating a new project, a tree folder is automatically created in the workspace directory on your repository server. This will correspond to the **Repository** tree view displayed on the main window of the Studio.

For more information, see [How to create a project](#).

- import the Demo project to discover the features of *Talend Studio* based on samples of different ready-to-use Jobs. When you import the Demo project, it is automatically installed in the workspace directory of the current session of the Studio.

For more information, see [How to import the demo project](#).

- import projects you have already created with previous releases of *Talend Studio* into your current *Talend Studio* workspace directory.

For more information, see [How to import projects](#).

- open a project you created or imported in the Studio.

For more information, see [How to open a project](#).

- delete local projects that you already created or imported and that you do not need any longer.

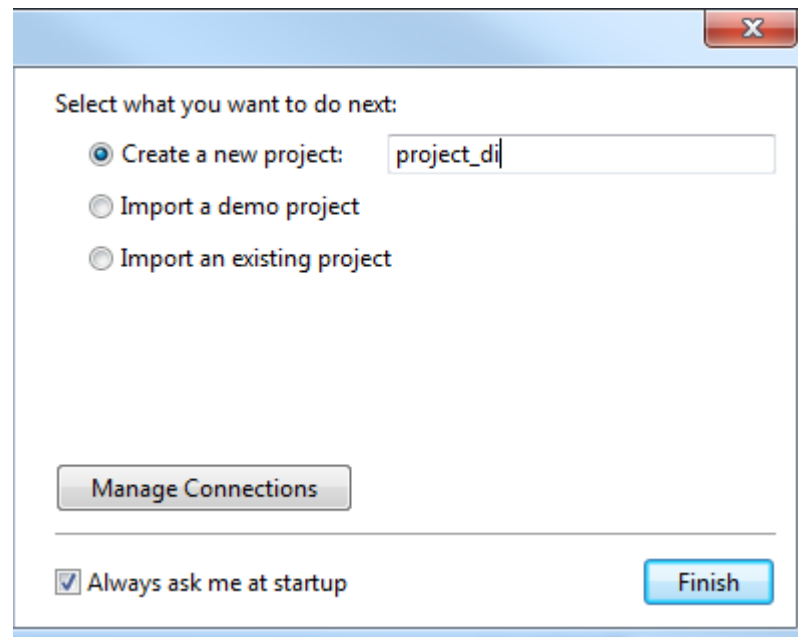
For more information, see [How to delete a project](#).

Once you launch *Talend Studio*, you can export the resources of one or more of the created projects in the current instance of the Studio. For more information, see [How to export a project](#).

1.3.1. How to create a project

To create a project at the initial startup of the Studio, do the following:

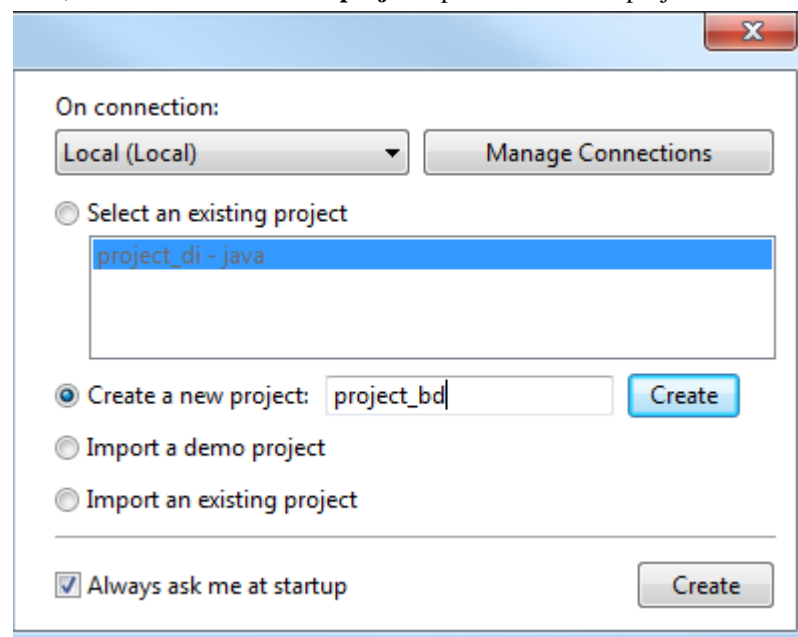
1. Launch *Talend Studio*.
2. On the login window, select the **Create a new project** option and enter a project name in the field.



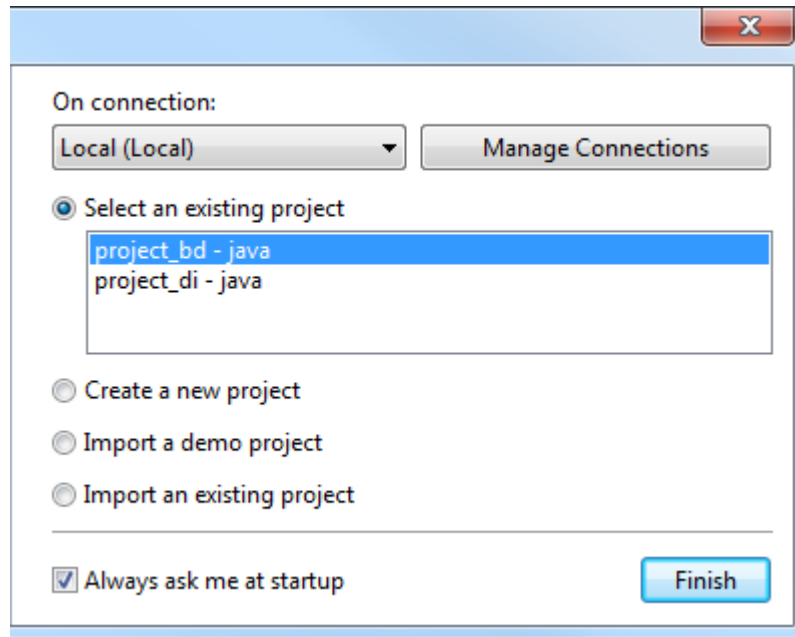
3. Click **Finish** to create the project and open it in the Studio.

To create a new project after the initial startup of the Studio, do the following:

1. On the login window, select the **Create a new project** option and enter a project name in the field.



2. Click **Create** to create the project. The newly created project is displayed on the list of existing projects.



3. Select the project on the list and click **Finish** to open the project in the Studio.


Later, if you want to switch between projects, on the Studio menu bar, use the combination **File > Switch Project or Workspace**.

1.3.2. How to import the demo project

You can import one or more demo projects that include numerous samples of ready to use Jobs into your *Talend Studio* to help you understand the functionalities of different **Talend** components.

To import a demo project, proceed as follows:

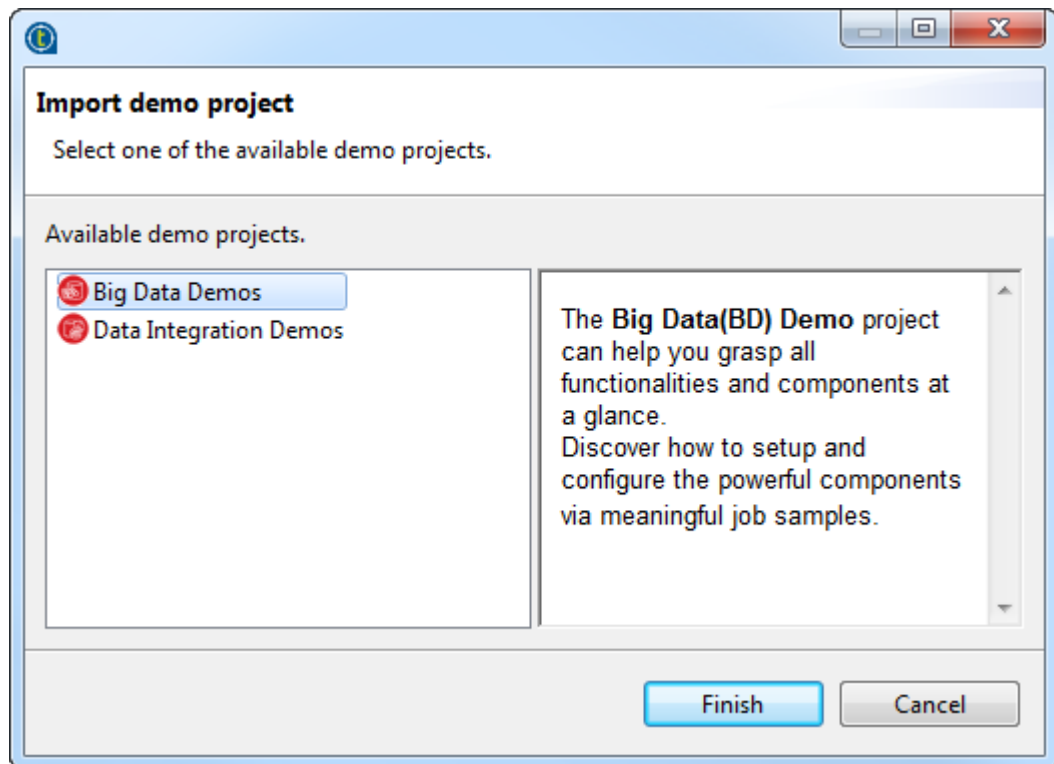
1. When launching your *Talend Studio*, select the **Import a demo project** option on the Studio login window and click **Select**, or click the **Demos** link on the welcome window, to open the **[Import demo project]** dialog box.

After launching the Studio, click  button on the toolbar, or select **Help > Welcome** from the Studio menu bar to open the welcome window and then click the **Demos** link, to open the **[Import demo project]** dialog box.

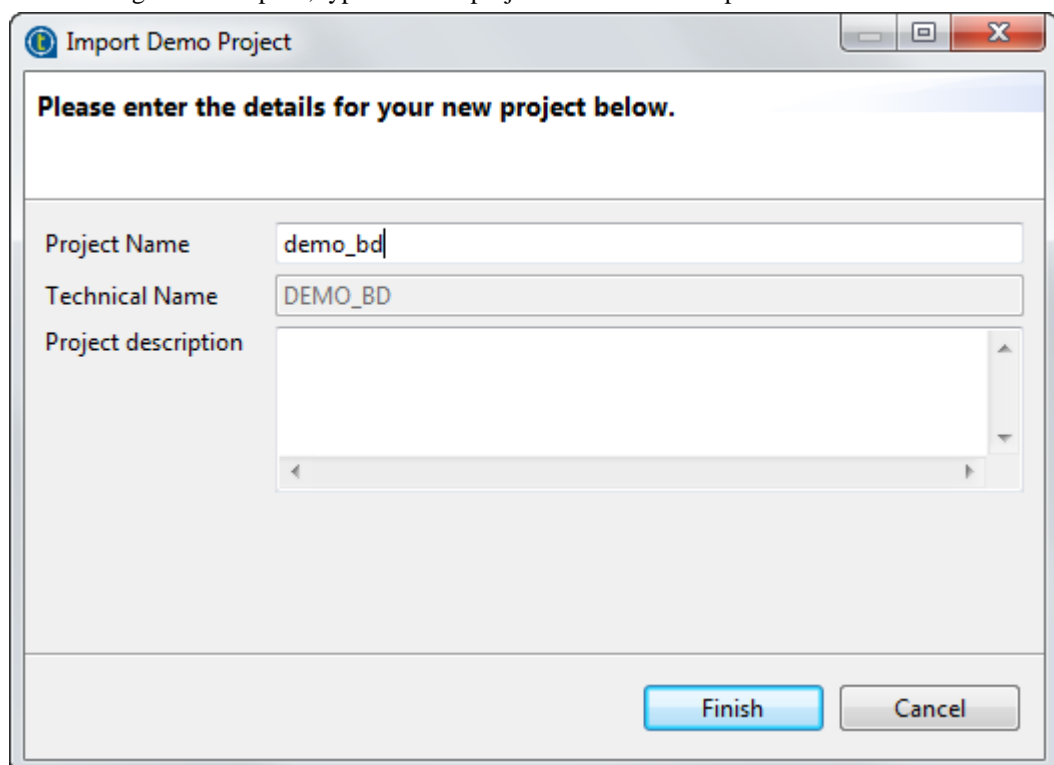
2. In the **[Import Demo Project]** dialog box, select the demo project you want to import and view the description on the right panel.



The demo projects available in the dialog box may vary depending on the product you are using.



3. Click **Finish** to close the dialog box.
4. In the new dialog box that opens, type in a new project name and description information if needed.



5. Click **Finish** to create the project.

All the samples of the demo project are imported into the newly created project, and the name of the new project is displayed in the **Project** list on the login screen.

6. To open the imported demo project in *Talend Studio*, back on the login window, select it from the **Project** list and then click **Finish**.

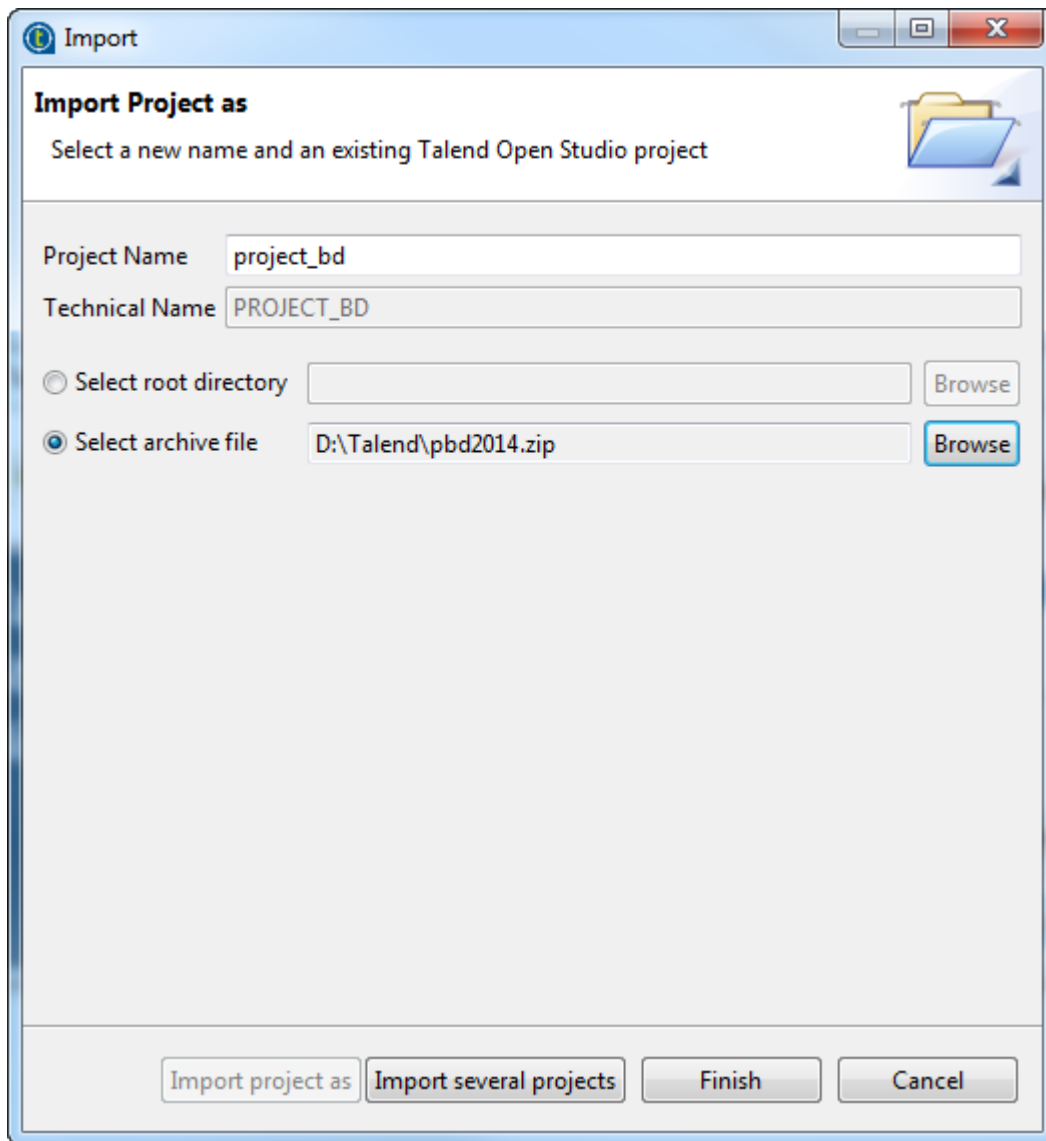
The Job samples in the open demo project are automatically imported into your workspace directory and made available in the **Repository** tree view under the **Job Designs** folder.

1.3.3. How to import projects

In *Talend Studio*, you can import one or more projects you already created with previous releases of the Studio.

To import a single project, do the following:

1. From the Studio login window, select **Import an existing project** then click **Select** to open the **[Import]** wizard.

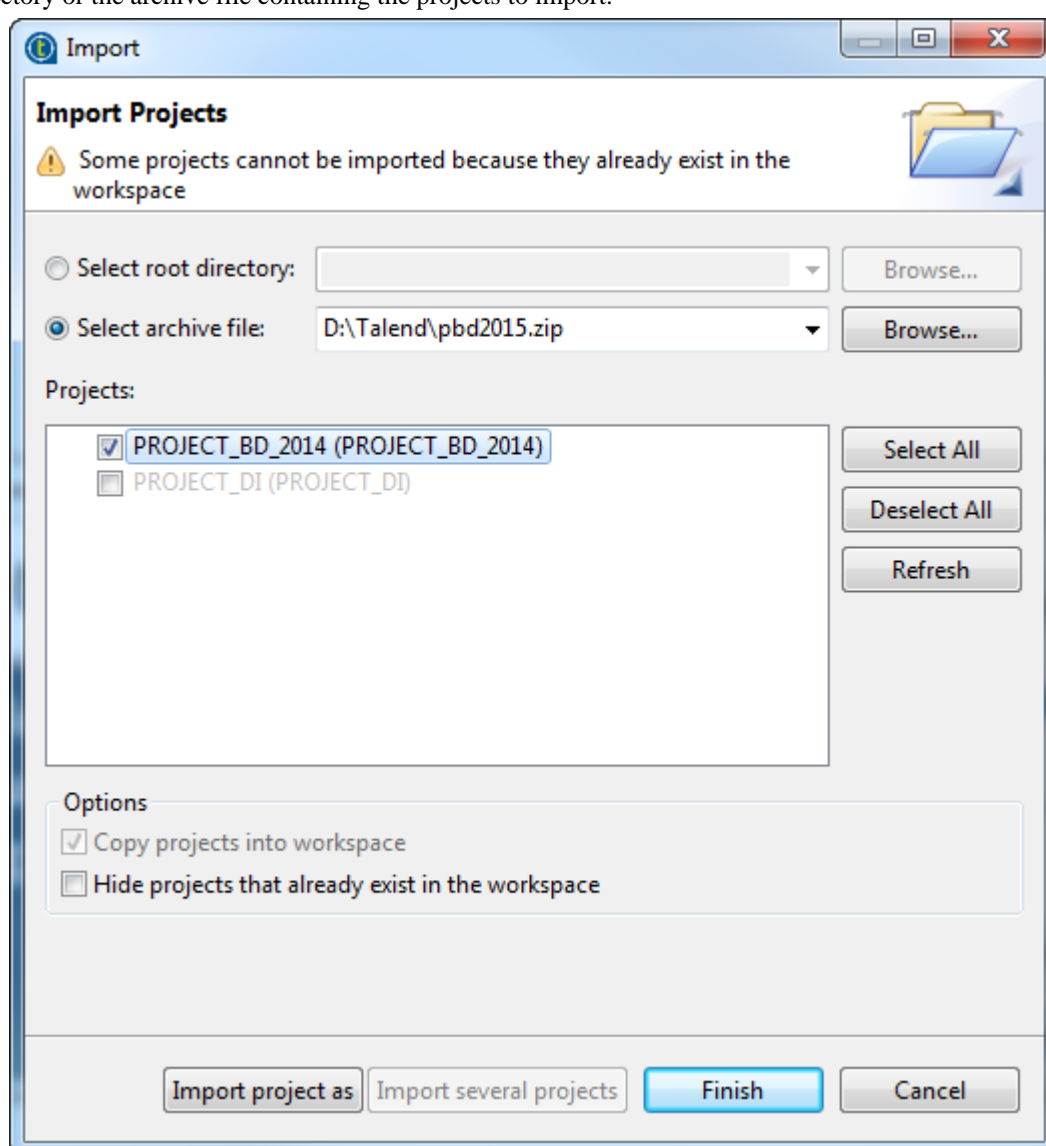


2. Click the **Import project as** button and enter a name for your new project in the **Project Name** field.
3. Click **Select root directory** or **Select archive file** depending on the source you want to import from.

4. Click **Browse...** to select the workspace directory/archive file of the specific project folder. By default, the workspace in selection is the current release's one. Browse up to reach the previous release workspace directory or the archive file containing the projects to import.
5. Click **Finish** to validate the operation and return to the login window.

To import several projects simultaneously, do the following:

1. From the Studio login window, select **Import an existing project** then click **Select** to open the **[Import]** wizard.
2. Click **Import several projects**.
3. Click **Select root directory** or **Select archive file** depending on the source you want to import from.
4. Click **Browse...** to select the workspace directory/archive file of the specific project folder. By default, the workspace in selection is the current release's one. Browse up to reach the previous release workspace directory or the archive file containing the projects to import.



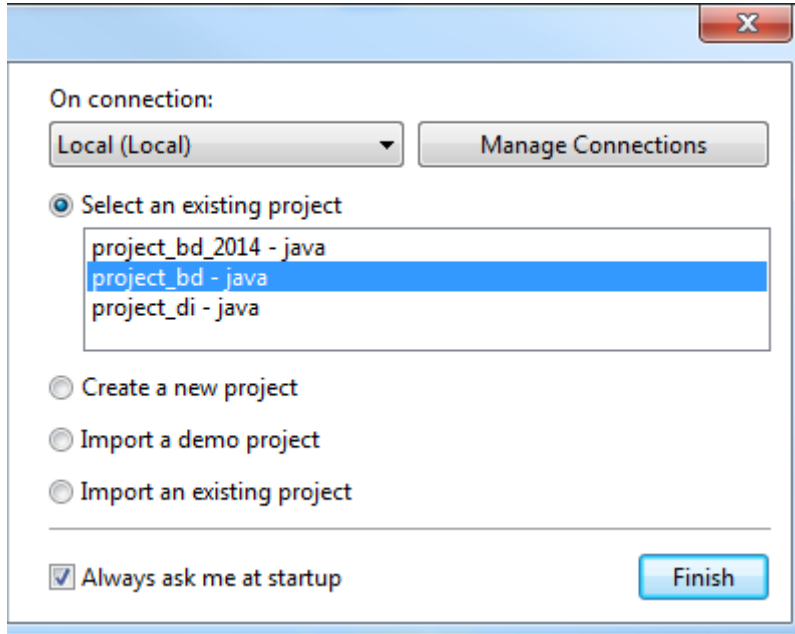
5. Select the **Copy projects into workspace** check box to make a copy of the imported project instead of moving it. This option is available only when you import several projects from a root directory.



If you want to remove the original project folders from the *Talend Studio* workspace directory you import from, clear this check box. But we strongly recommend you to keep it selected for backup purposes.

6. Select the **Hide projects that already exist in the workspace** check box to hide existing projects from the **Projects** list. This option is available only when you import several projects.
7. From the **Projects** list, select the projects to import and click **Finish** to validate the operation.

Upon successful project import, the names of the imported projects are displayed on the **Project** list of the login window.



You can now select the imported project you want to open in *Talend Studio* and click **Finish** to launch the Studio.



A generation initialization window might come up when launching the application. Wait until the initialization is complete.

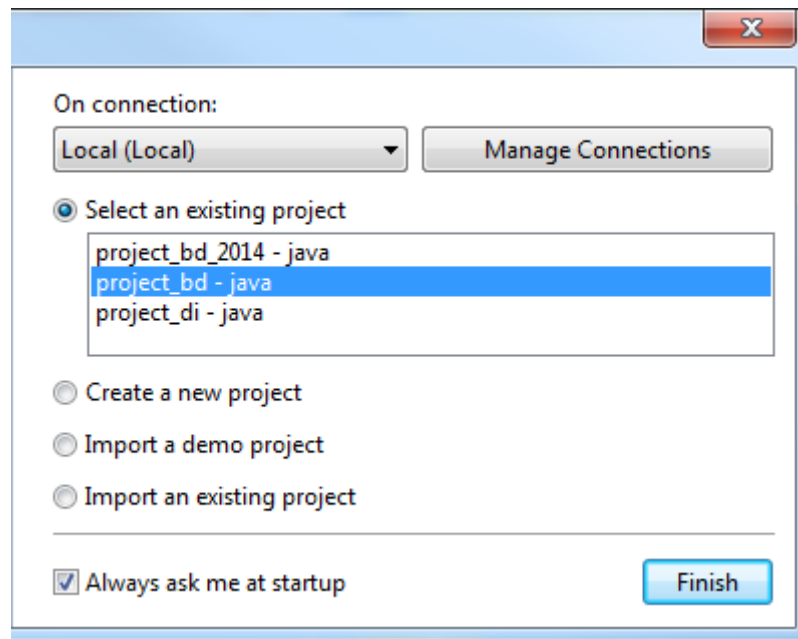
1.3.4. How to open a project



*When you launch Talend Studio for the first time, no project names are displayed on the **Project** list. First you need to create a project or import a Demo project in order to populate the **Project** list with the corresponding project names that you can then open in the Studio.*

To open a project in *Talend Studio*:

On the Studio login screen, select the project of interest from the project list and click **Finish**.



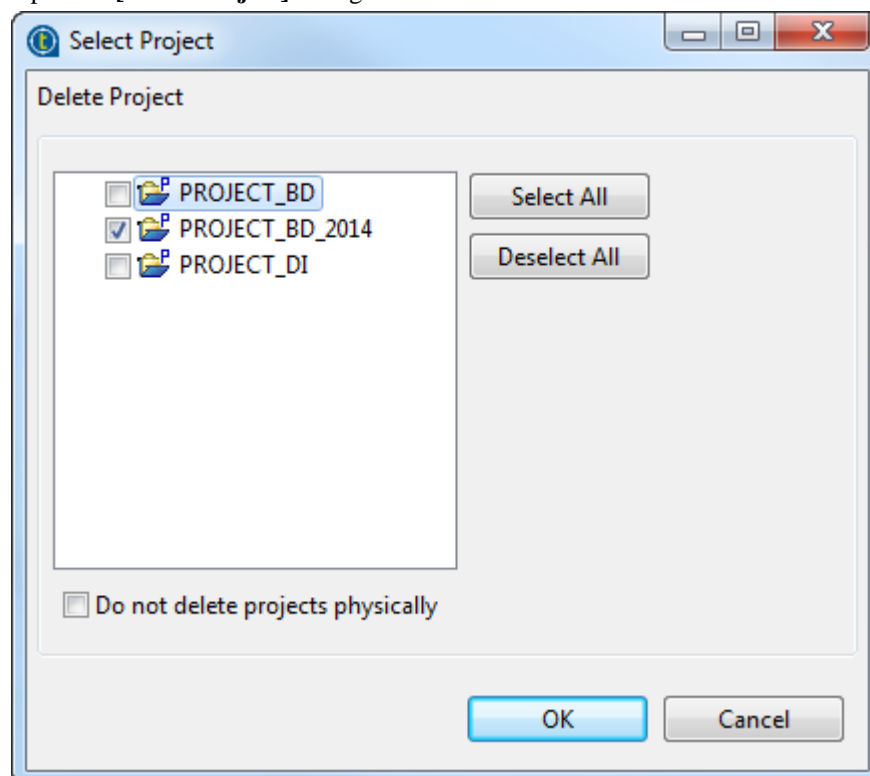
A progress bar appears. Wait until the task is complete and the *Talend Studio* main window opens.



When you open a project imported from a previous version of the Studio, an information window pops up to list a short description of the successful migration tasks.

1.3.5. How to delete a project

1. On the login screen, click **Manage Connections**, then on the dialog box that opens click **Delete Existing Project(s)** to open the [Select Project] dialog box.



2. Select the check box(es) of the project(s) you want to delete.

3. Click **OK** to validate the deletion.

The project list on the login window is refreshed accordingly.



*Be careful, this action is irreversible. When you click **OK**, there is no way to recuperate the deleted project(s).*

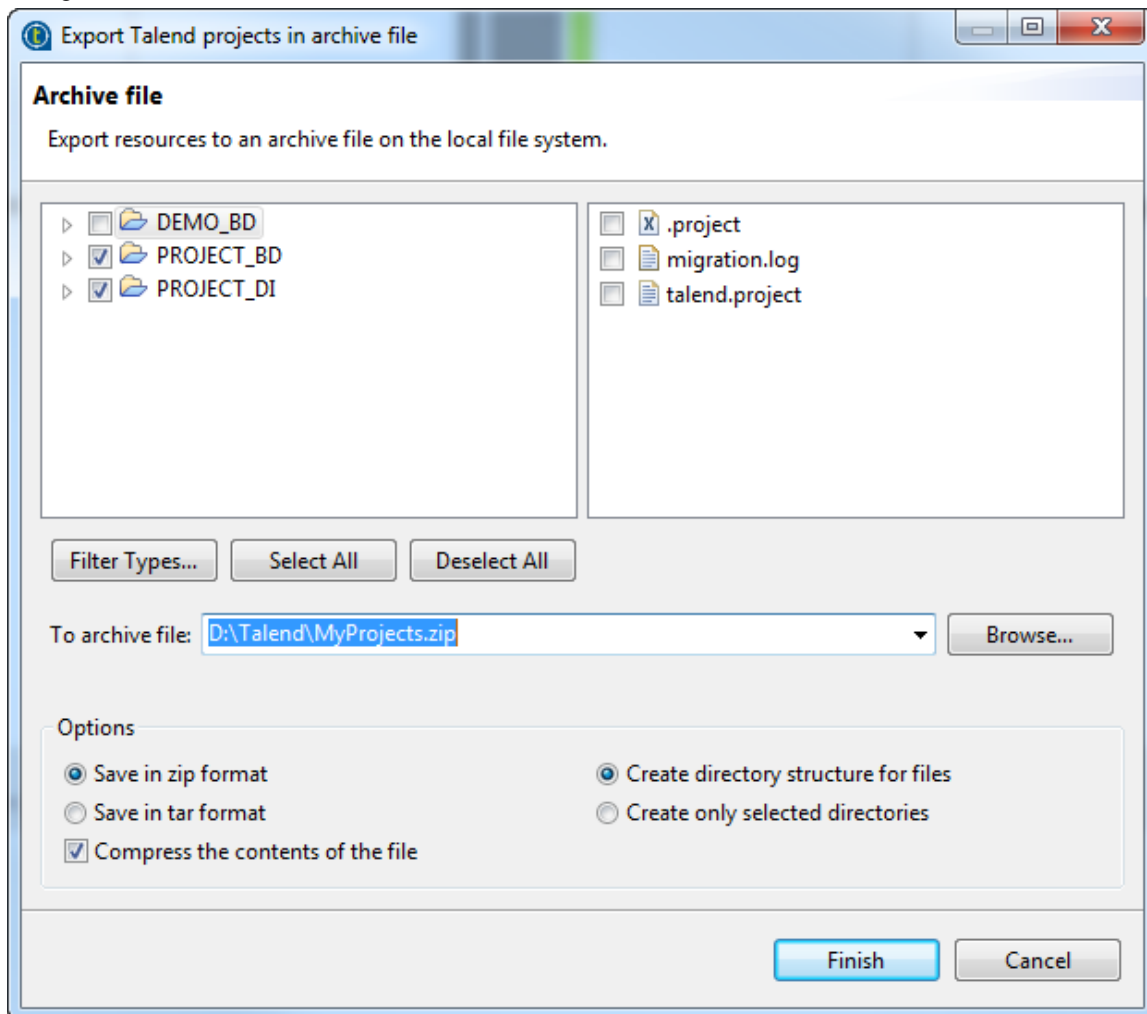


If you select the **Do not delete projects physically** check box, you can delete the selected project(s) only from the project list and still have it/them in the *workspace* directory of *Talend Studio*. Thus, you can recuperate the deleted project(s) any time using the **Import existing project(s) as local** option on the **Project** list from the login window.

1.3.6. How to export a project

Talend Studio allows you to export projects created or imported in the current instance of *Talend Studio*.

1. On the toolbar of the Studio main window, click  to open the **[Export Talend projects in archive file]** dialog box.



2. Select the check boxes of the projects you want to export. You can select only parts of the project through the **Filter Types...** link, if need be (for advanced users).
3. In the **To archive file** field, type in the name of or browse to the archive file where you want to export the selected projects.
4. In the **Option** area, select the compression format and the structure type you prefer.

5. Click **Finish** to validate the changes.

The archived file that holds the exported projects is created in the defined place.



Chapter 2. Getting started with Talend Big Data using the demo project

This chapter provides short descriptions about the sample Jobs included in the demo project and introduces the necessary preparations required to run the sample Jobs on a Hadoop platform. For how to import a demo project, see the section on importing a demo project of *Talend Studio User Guide*.

Before you start working in the studio, you need to be familiar with its Graphical User Interface (GUI). For more information, see the appendix describing GUI elements of *Talend Studio User Guide*.

2.1. Introduction to the Big Data demo project

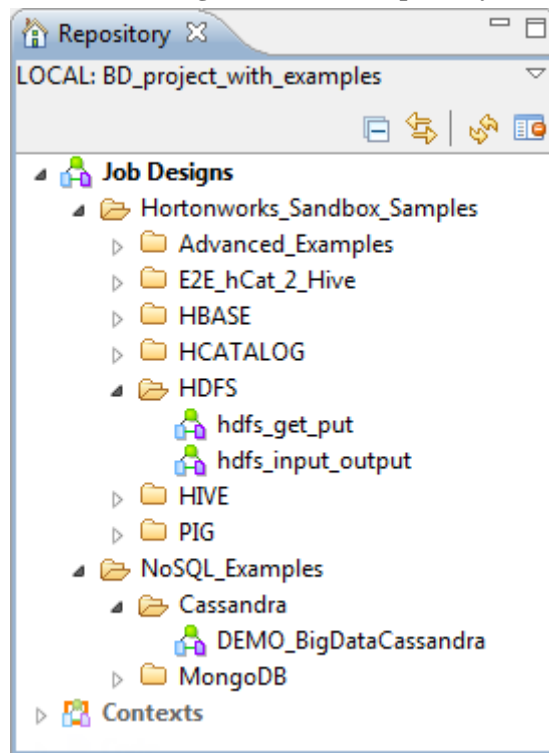
Talend provides a Big Data demo project that includes a number of easy-to-use sample Jobs. You can import this demo project into your **Talend** studio to help familiarize yourself with **Talend** studio and understand the various features and functions of **Talend** components.



Due to license incompatibility, some third-party Java libraries and database drivers (.jar files) required by **Talend** components used in some Jobs of the demo project could not be shipped with your *Talend Studio*. You need to download and install those .jar files, known as external modules, before you can run Jobs that involve such components. Fortunately, your *Talend Studio* provides a wizard that let you install external modules quickly and easily. This wizard automatically appears when you run a Job and the studio detects that one or more required external modules are missing. It also appears when you click **Install** on the top of the **Basic settings** or **Advanced settings** view of a component for which one or more required external modules are missing.

For more information on installing third-party modules, see the sections on identifying and installing external modules of the *Installation and Upgrade Guide*.

With the Big Data demo project imported and opened in your **Talend** studio, all the sample Jobs included in it are available in different folders under the **Job Designs** node of the **Repository** tree view.



The following sections briefly describe the Jobs contained in each sub-folder of the main folders.

2.1.1. Hortonworks_Sandbox_Samples

The main folder *Hortonworks_Sandbox_Samples* gathers standard **Talend** Jobs that are intended to demonstrate how handle data on a Hadoop platform.

Folder	Sub-folder	Description
Advanced_Examples		The Advanced_Examples folder has some use cases, including an example of processing Apache Weblogs using the Talend 's Apache Weblog, HCatalog and Pig components, an example of computing US Government Spending data using a Hive query, and an example of extracting data from any MySQL database and loading all the data from the tables dynamically.

Folder	Sub-folder	Description
		If there are multiple steps to achieve a use case they are named <i>Step_1</i> , <i>Step_2</i> and so on.
	ApacheWebLog	<p>This folder has a classic Weblog file process that shows loading an Apache web log into HCatalog and HDFS and filtering out specific codes. There are two examples computing counts of unique IP addresses or web codes. These examples use Pig Scripts and HCatalog load.</p> <p>There are 6 steps in this example, run each step in the order listed in the Job names.</p> <p>For more details of this example, see the chapter on Big Data Job examples of your Studio User Guide, which guides you step by step through the creation and configuration of the example Jobs.</p>
	Gov_Spending_Analysis	<p>This is an example of a two-step process that loads some sample US Government spending data into HCatalog and then in step two it uses a Hive query to get the total spending amount per Government agency. There is an extra Data Integration Job that takes a file from the http://usaspending.gov/data web site and prepares it for the input to the Job that loads the data to HCatalog. You will need to replace the tFixedFlowInput component with the input file.</p> <p>There are 2 steps in this example, run each step in the order listed in the Job names.</p>
	<i>RDBMS_Migration_SQOOP</i>	<p>This is a two step process that will read data from any MySQL schema and load it to HDFS. The database can be any MySQL5.5 or newer. The schema needs to have as many tables as you desire. Set the database and the schema in the context variables labeled <i>SQOOP_SCENARIO_CONTEXT</i> and the first Job will dynamically read the schema and create two files with list of tables. One file consists of tables with Primary Keys to be partitioned in HCatalog or Hive if used and the second one consists of the same tables without Primary Keys. The second step uses the two files to then load all the data from MySQL tables in the schema to HDFS. There will be a file per table.</p> <p>Keep in mind when running this process not to select a schema with a large amount of volume if you are using the Sandbox single node VM, as it has not a lot of power. For more information on using the proposed Sandbox single node VM, see Installing Hortonworks Sandbox.</p>
E2E_hCat_2_Hive		This folder contains a very simple process that loads some sample data to HCatalog in the first step and then in the second step just shows how you can use the Hive components to access and process the data.
HBASE		This folder contains simple examples of how to load data to HBase and read data from it.
HCATALOG		There are two examples for HCatalog: the first one puts a file directly on the HDFS and then loads the meta store with the information into HCatalog. The second example loads data streaming directly into HCatalog in the defined partitions.
HDFS		The examples in this folder show the basic HDFS operations like Get, Put, and Streaming loads.
HIVE		This folder contains three examples: the first Job shows how to use the Hive components to complete basic operations on Hive like creating a database, creating a table and loading data to the table. The next two Jobs shows first how to load two tables to Hive, which are then used in the second step, an example of how you can do ELT with Hive.
PIG		This folder contains many different examples of how Pig components can be used to perform many different key functions such as aggregations and filtering and an example of how the Pig Code works.

2.1.2. NoSQL_Examples

The main folder *NoSQL_Examples* gathers Jobs that are intended to demonstrate how to handle data with NoSQL databases.

Folder	Description
Cassandra	This is another example of how to do the basic write and read to the Cassandra Database to then start using the Cassandra NoSQL database right away.
MongoDB	This folder has an example of how to use MongoDB to easily and quickly search open text unstructured data for blog entries with key words.

2.2. Setting up the environment for the demo Jobs to work

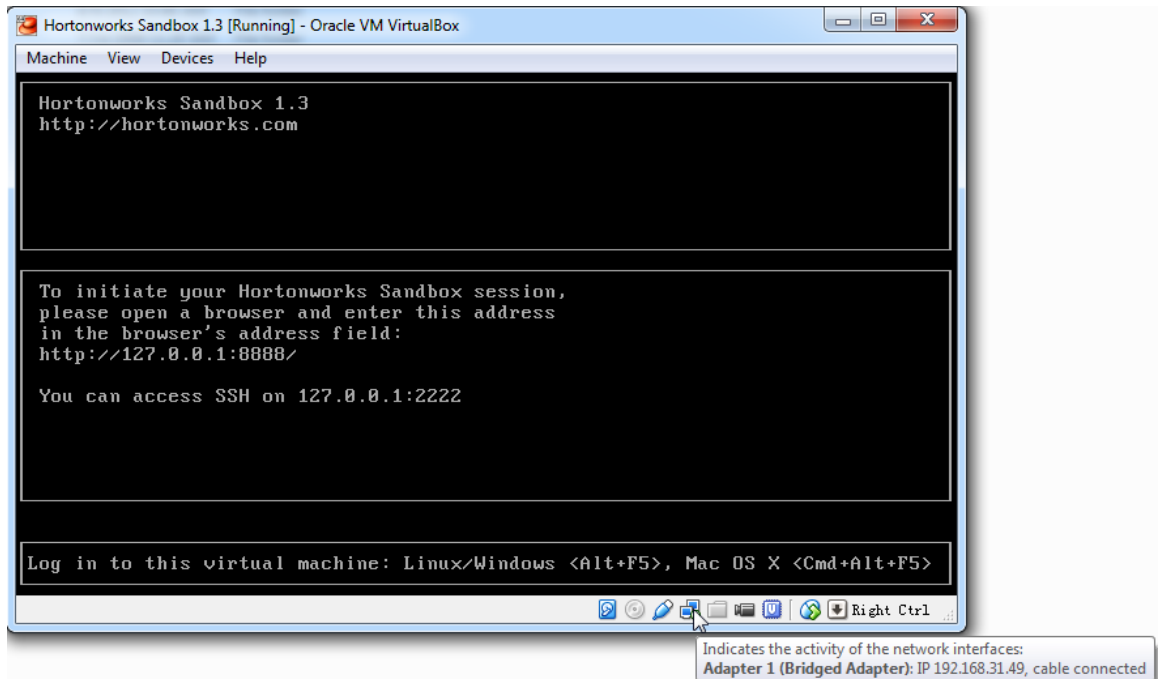
The Big Data demo project is intended to give you some easy and practical examples of many of the basic functionality of **Talend's** Big Data solution. To run the demo Jobs included in the demo project, you need to have your Hadoop platform up and running, and configure the context variables defined in the demo project or configure the relevant components directly if you do not want to use the proposed Hortonworks Sandbox virtual appliance.

2.2.1. Installing Hortonworks Sandbox

For ease of use, one of the methods to get a Hadoop platform up quickly is to use a Virtual Appliance from one of the top Hadoop Distribution Vendors. Hortonworks provides a Virtual Appliance/Machine or a VM called the Sandbox that is fast and easy to set up. Using context variables, the samples Jobs within the *Hortonworks_Sandbox_Samples* folder of the demo project have been configured to work on Hortonworks Sandbox VM.

Below is a brief procedure of setting up the single-node VM with Hortonworks Sandbox on Oracle VirtualBox, which is recommended by Hortonworks. For details, see the documentation of the relevant vendors.

1. Download the recommended version of Oracle VirtualBox from <https://www.virtualbox.org/> and the Sandbox image for VirtualBox from <http://hortonworks.com/products/hortonworks-sandbox/>.
2. Install and set up Oracle VirtualBox by following Oracle VirtualBox documentation.
3. Install the Hortonworks Sandbox virtual appliance on Oracle VirtualBox by following Hortonworks Sandbox instructions.
4. In the [**Oracle VM VirtualBox Manager**] window, click **Network**, select the **Adapter 1** tab, select **Bridged Adapter** from the **Attached to** list box, and then select your working physical network adapter from the **Name** list box.
5. Start the Hortonworks Sandbox virtual appliance to get the Hadoop platform up and running, and check that the IP address assigned to the Sandbox virtual machine is pingable.



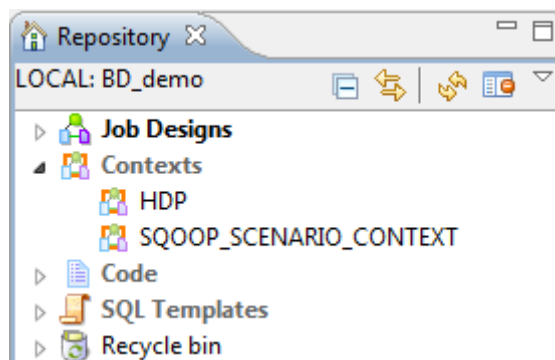
Then, before launching the demo Jobs, add an IP-domain mapping entry in your *hosts* file to resolve the host name *sandbox*, which is defined as the value of a couple of context variables in this demo project, rather than using an IP address of the Sandbox virtual machine, as this will minimize the changes you will need to make to the configured context variables.

For more information about context variables used in the demo project, see [Understanding context variables used in the demo project](#).

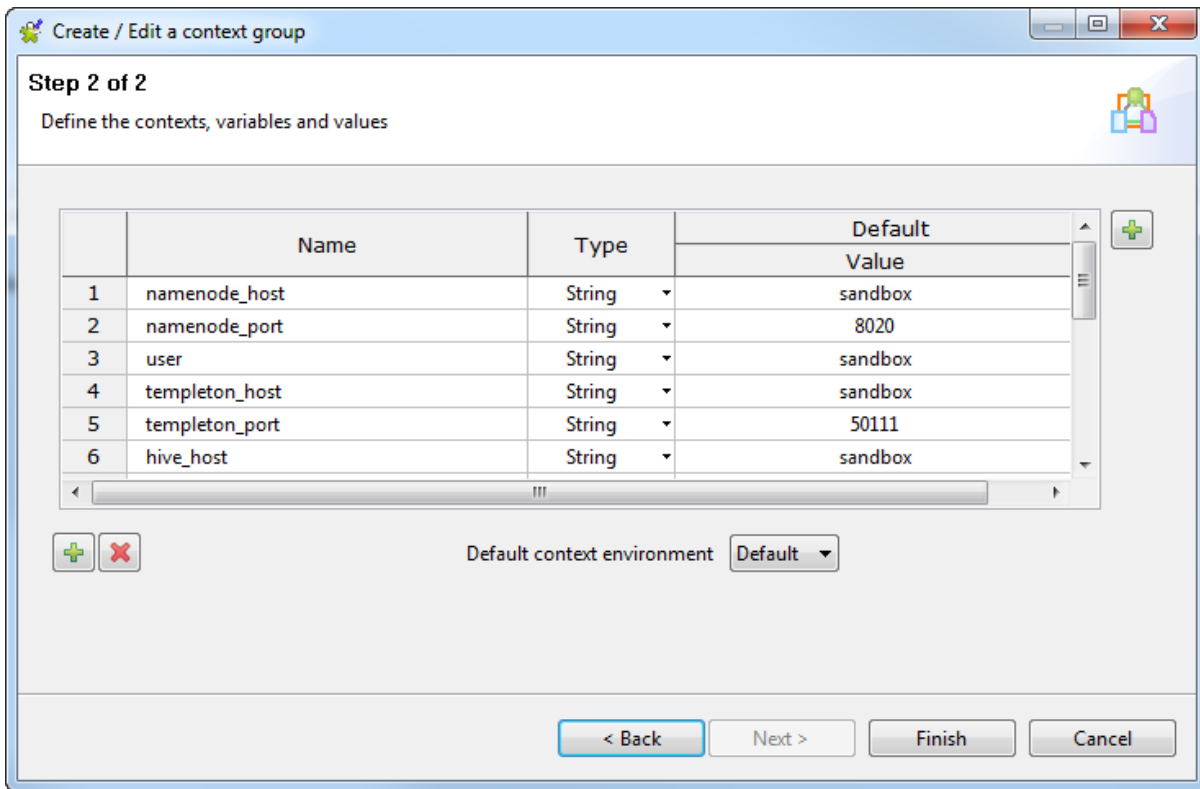
2.2.2. Understanding context variables used in the demo project

In *Talend Studio*, you can define context variables in the Repository once in a project and use them in many Jobs, typically to help define connections and other things that are common across many different Jobs and processes. The advantage for this is obvious. For example, if you define the namenode IP address in a context variable and you have 50 Jobs that use that variable, to change the IP address of the namenode you simply need to update the context variable. Then, the studio will inform you of all the Jobs impacted by this update and change them for you.

Repository-stored context variables are grouped under the **Contexts** node in the **Repository** tree view. In the Big Data demo project, two groups of context variables have been defined in the **Repository**: *HDP* and *SQOOP_SCENARIO_CONTEXT*.



To view or edit the settings of the context variables of a group, double-click the group name in the **Repository** tree view to open the **[Create / Edit a context group]** wizard and go to Step 2.




The context variables in the *HDP* group are used in all the demo examples in the *Hortonworks_Sandbox_Samples* folder. If you want, you can change the values of these variables. For example, if you want to use the IP address for the Sandbox Platform VM rather than the host name *sandbox*, you can change the value of the host name variables to the IP address. If you change any of the default configurations on the Sandbox VM, you need to change the context settings accordingly; otherwise the demo examples may not work as intended.

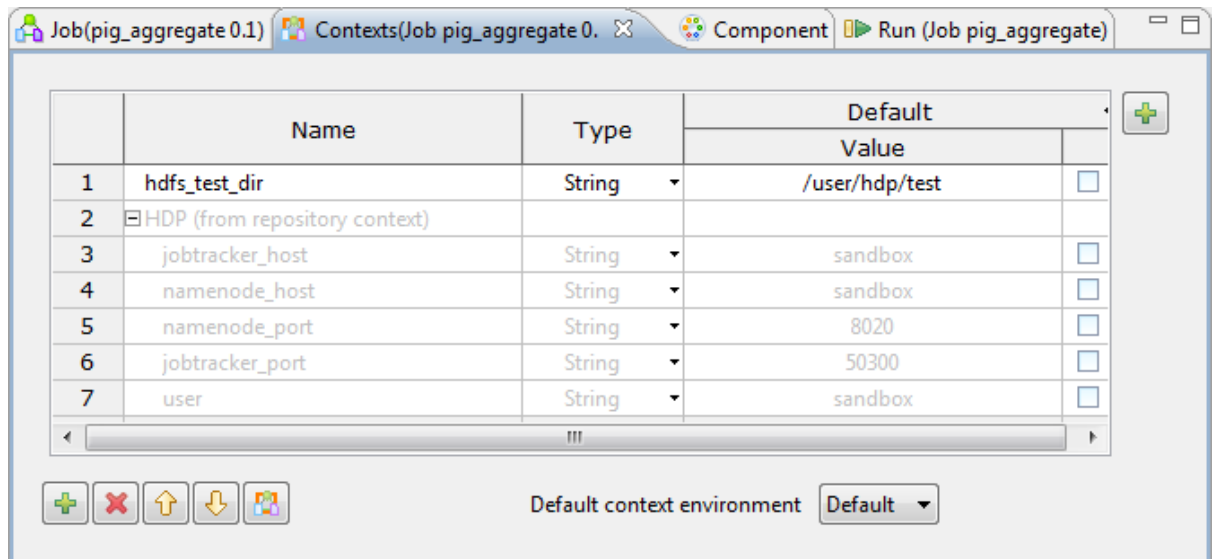
Variable name	Description	Default value
namenode_host	Namenode host name	sandbox
namenode_port	Namenode port	8020
user	User name to connect to the Hadoop system	sandbox
templeton_host	HCatalog server host name	sandbox
templeton_port	HCatalog server port	50111
hive_host	Hive metastore host name	sandbox
hive_port	Hive metastore port	9083
jobtracker_host	Jobtracker host name	sandbox
jobtracker_port	Jobtracker port	50300
mysql_host	Host of the Sandbox for the Hive metastore	sandbox
mysql_port	Port of the Hive metastore	3306
mysql_user	User name to connect to the Hive metastore	hep
mysql_passed	Password to connect to the Hive metastore	hep
mysql_testes	Name of the test database for the Hive metastore	testes
hbase_host	HBase host name	sandbox
hbase_port	HBase port	2181

The context variables in the *SQOOP_SCENARIO_CONTEXT* group are used for the *RDBMS_Migration_SQOOP* demo examples only. You will need to go through the following context variables and update your information for the Sandbox VM on your local MySQL connections if you want to use the *RDBMS_Migration_SQOOP* demo.

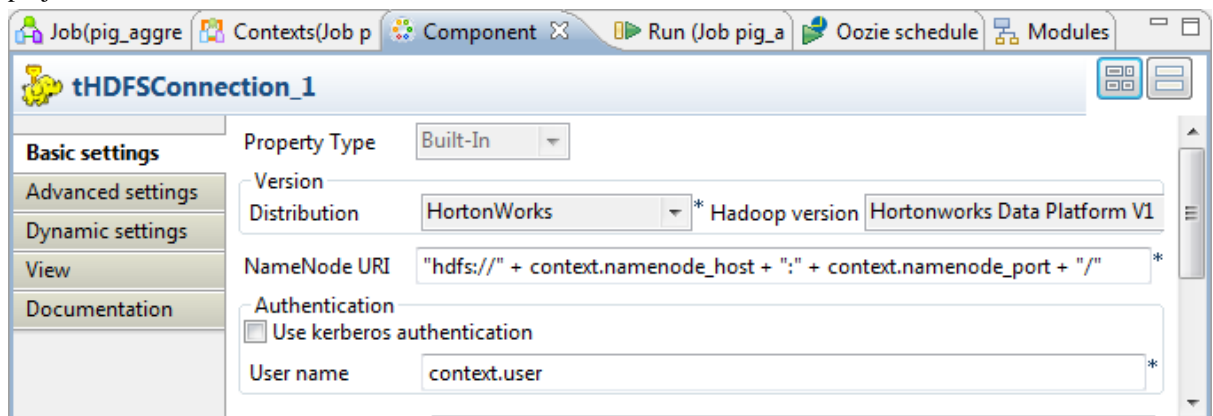
Variable name	Description	Default value
KEY_LOGS_DIRECTORY	A directory holding table files on your local machine that the studio has full access to	C:/Talend/BigData/
MYSQL_DBNAME_TO_MIGRATE	Name of your own MySQL database to migrate to HDFS	dstar_crm
MYSQL_HOST_or_IP	Host name or IP address of the MySQL database	192.168.56.1
MYSQL_PORT	Port of the MySQL database	3306
MYSQL_USERNAME	User name to connect to the MySQL database	tisadmin
MYSQL_PWD	Password to connect to the MySQL database	
HDFS_LOCATION_TARGET	Target location on the Sandbox HDFS where you want to load the data	/user/hdp/sqoop/

To use Repository-stored context variables in a Job, you need to import them into the Job first by clicking the  button in the **Contexts** view. You can also define context variables in the **Contexts** view of a Job. These variables are built-in variables that work only for that Job.

The **Contexts** view shows the built-in context variables defined in the Job and the Repository-stored context variables imported into the Job.



Once defined, variables are referenced in the configurations of components. The following example shows how context variables are used in the configurations of the **tHDFSConnection** component in a Pig Job of the demo project.



Once these are setup to reflect how you have configured the HortonWorks Sandbox, the examples will run with little intervention. You can see how many of the core functions work allowing you to have good samples to implement your Big Data projects.

For more information on defining and using context variables, see the section on using contexts and variables of the *Talend Studio User Guide*.

For how to run a Job from the **Run** console, see the section on how to run a Job of the *Talend Studio User Guide*.

For how to run a Job from the **Oozie scheduler** view, see the *Talend Studio User Guide*.



Chapter 3. Working in *Talend Studio* - basic Job examples

This chapter provides basic Job examples to help users get started with *Talend Studio*.

3.1. Getting started with a basic Job

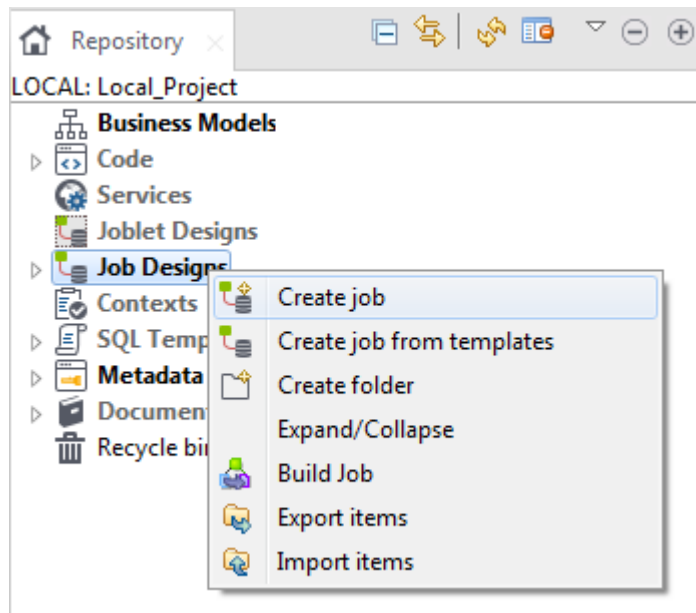
This section provides a continuous example that will help you create, add components to, configure, and execute a simple Job. This Job will be named *A_Basic_Job* and will read a text file, display its content on the **Run** console, and then write the data into another text file.

3.1.1. Creating a Job

Talend Studio enables you to create a Job by dropping different technical components from the **Palette** onto the design workspace and then connecting these components together.

To create the example Job described in this section, proceed as follows:

1. In the **Repository** tree view of the **Integration** perspective, right-click the **Job Designs** node and select **Create job** from the contextual menu.



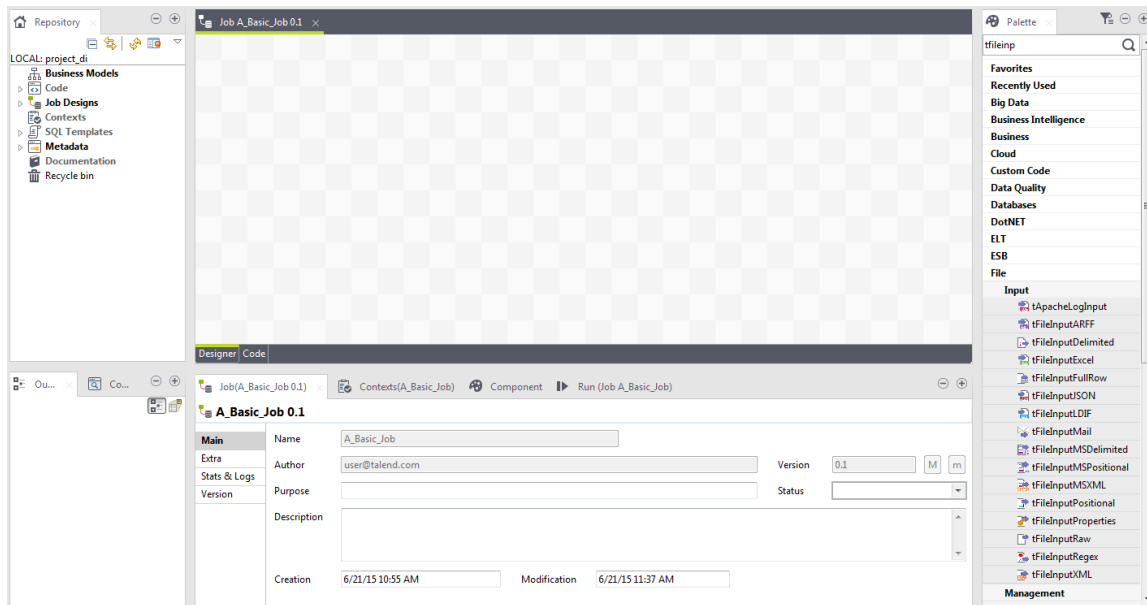
The [**New Job**] wizard opens to help you define the main properties of the new Job.

- Fill the Job properties as shown in the previous screenshot.

The fields correspond to the following properties:

Field	Description
Name	the name of the new Job. Note that a message comes up if you enter prohibited characters.
Purpose	Job purpose or any useful information regarding the Job use.
Description	Job description containing any information that helps you describe what the Job does and how it does it.
Author	a read-only field that shows by default the current user login.
Locker	a read-only field that shows by default the login of the user who owns the lock on the current Job. This field is empty when you are creating a Job and has data only when you are editing the properties of an existing Job.
Version	a read-only field. You can manually increment the version using the M and m buttons.
Status	a list to select from the status of the Job you are creating.
Path	a list to select from the folder in which the Job will be created.

- An empty design workspace opens up showing the name of the Job as a tab label.



The Job you created is now listed under the **Job Designs** node in the **Repository** tree view.

You can open one or more of the created Jobs by simply double-clicking the Job label in the **Repository** tree view.

Related topics:

- Classify the Jobs you created by creating folders. For more information, see your *Talend Studio* User Guide.
- Create a data integration Job. For more information, see your *Talend Studio* User Guide.
- Customize the workspace. For more information, see your *Talend Studio* User Guide.

3.1.2. Adding components to the Job

Now that the Job is created, components have to be added to the design workspace, a **tFileInputDelimited**, a **tLogRow**, and a **tFileOutputDelimited** in this example.

There are several ways to add a component onto the design workspace. You can:

- find your component on the **Palette** by typing the search keyword(s) in the search field of the **Palette** and drop it onto the design workspace.
- add a component by directly typing your search keyword(s) on the design workspace.
- add an output component by dragging from an input component already existing on the design workspace.
- drag and drop a centralized metadata item from the **Metadata** node onto the design workspace, and then select the component of interest from the **Components** dialog box.

This section describes the first three methods. For details about how to drop a component from the **Metadata** node, see your *Talend Studio* User Guide.

3.1.2.1. Dropping the first component from the Palette

The first component of this example will be added from the **Palette**. This component defines the first task executed by the Job. In this example, as you first want to read a text file, you will use the **tFileInputDelimited** component.

For more information regarding components and their functions, see *Talend Open Studio Components Reference Guide*.

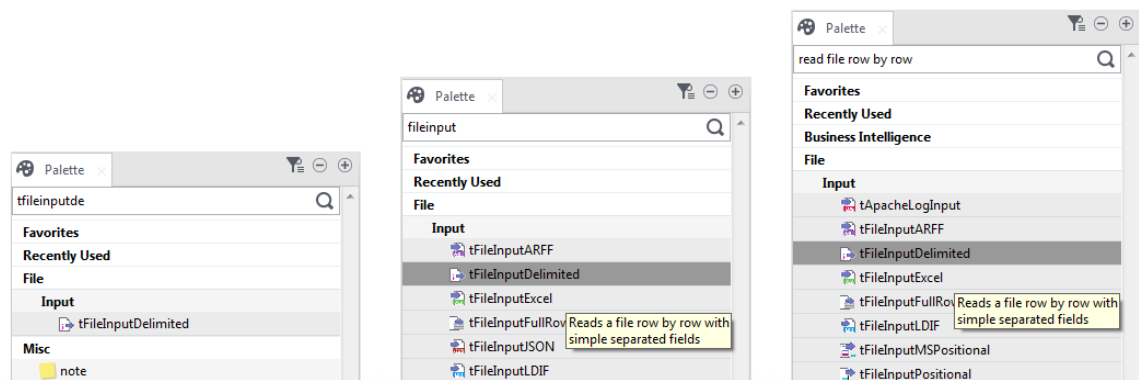
To drop a component from the **Palette**, proceed as follows:

1. Enter the search keyword(s) in the search field of the **Palette** and press **Enter** to validate your search.

The keyword(s) can be the partial or full name of the component, or a phrase describing its functionality if you don't know its name, for example, *tfileinputde*, *fileinput*, or *read file row by row*.

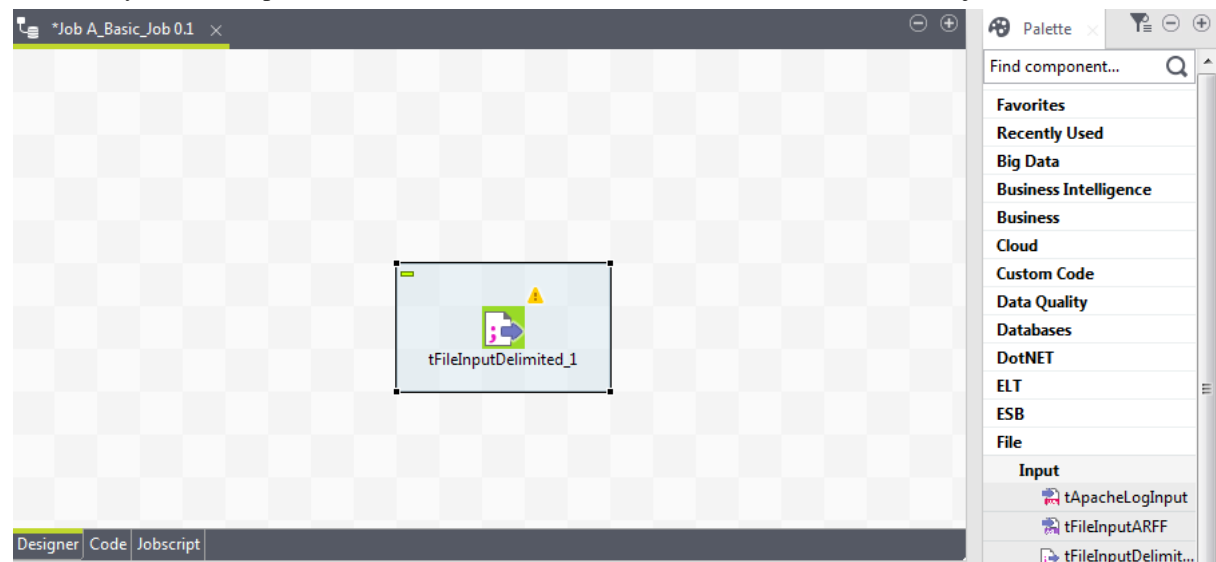


To use a descriptive phrase as keywords for a fuzzy search, make sure the **Also search from Help when performing a component searching** check box is selected on the **Preferences > Palette Settings** view. For more information, see your *Talend Studio User Guide*.



2. Select the component you want to use and click on the design workspace where you want to drop the component.

Each newly-added component is shown in a blue box to show that it as an individual Subjob.



3.1.2.2. Adding the second component by typing on the design workspace

The second component of our Job will be added by typing its name directly on the workspace, instead of dropping it from the **Palette** or from the **Metadata** node.

Prerequisite: Make sure you have selected the **Enable Component Creation Assistant** check box in the Studio preferences. For more information, see your *Talend Studio* User Guide.

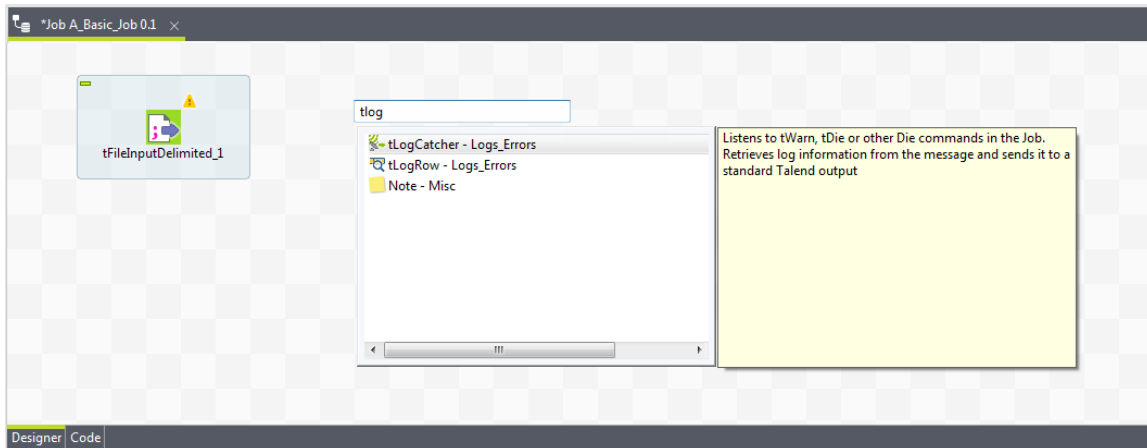
To add a component directly on the workspace, proceed as follows:

1. Click where you want to add the component on the design workspace, and type your keywords, which can be the full or partial name of the component, or a phrase describing its functionality if you don't know its name. In our example, start typing *tlog*.



To use a descriptive phrase as keywords for a fuzzy search, make sure the **Also search from Help when performing a component searching** check box is selected on the **Preferences > Palette Settings** view. For more information, see your *Talend Studio* User Guide.

A list box appears below the text field displaying all the matching components in alphabetical order.



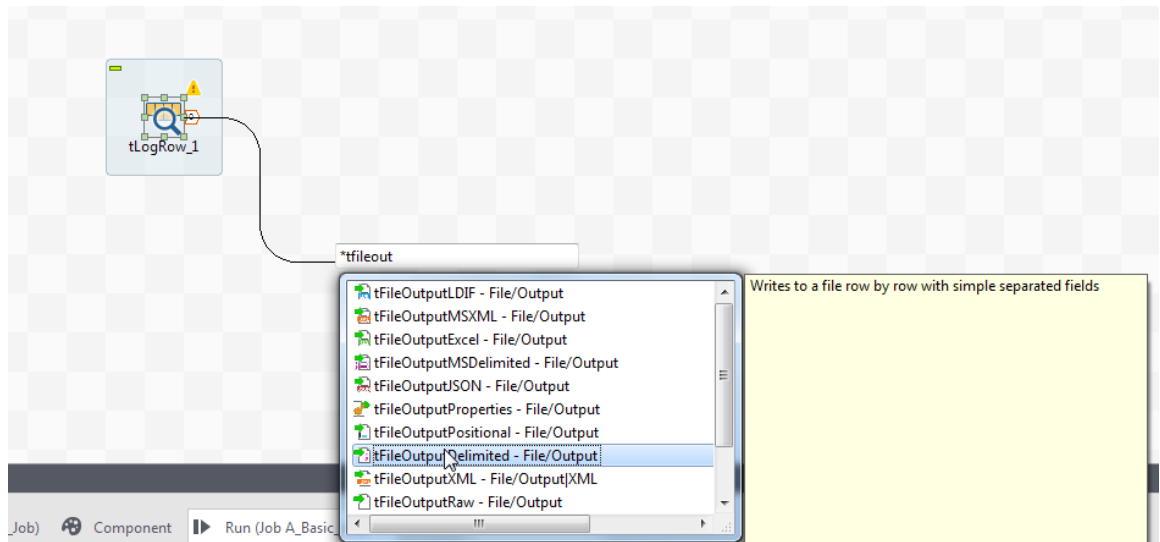
2. Double-click the desired component to add it on the workspace, **tLogRow** in our example.

3.1.2.3. Adding an output component by dragging from an input one

Now you will add the third component, a **tFileOutputDelimited**, to write the data read from the source file into another text file. We will add the component by dragging from the **tLogRow** component, which serves as an input component to the new one to be added.

1. Click the **tLogRow** component to show the **o** icon docked to it.
2. Drag and drop the **o** icon where you want to add a new component.

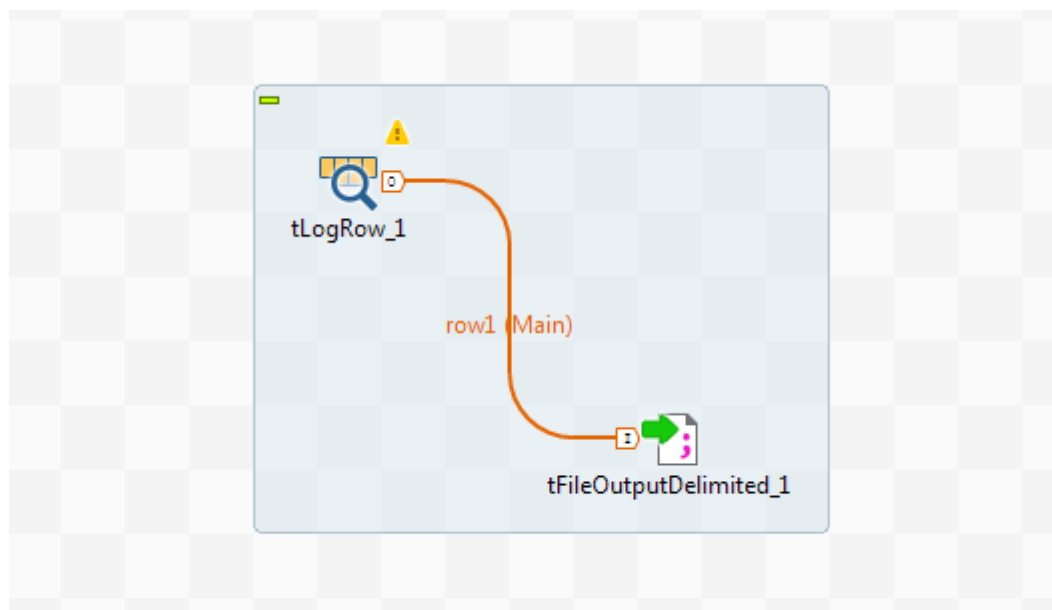
A text field and a component list appear. The component list shows all the components that can be connected with the input component.



- To narrow the search, type in the text field the name of the component you want to add or part of it, or a phrase describing the component's functionality if you don't know its name, and then double-click the component of interest, **tFileOutputDelimited** in this example, on the component list to add it onto the design workspace. The new component is automatically connected with the input component **tLogRow**, using a **Row > Main** connection.



To use a descriptive phrase as keywords for a fuzzy search, make sure the **Also search from Help when performing a component searching** check box is selected on the **Preferences > Palette Settings** view. For more information, see your *Talend Studio User Guide*.



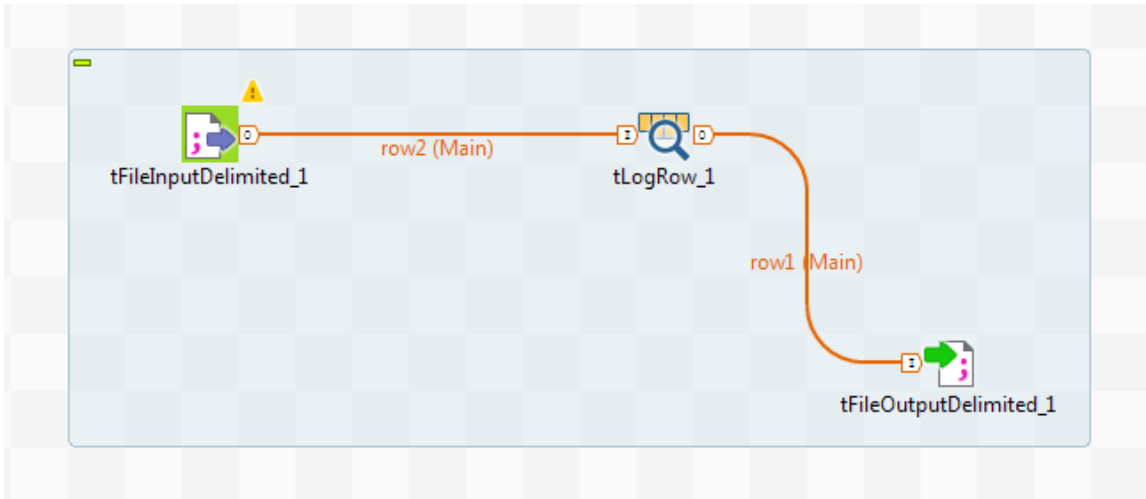
3.1.3. Connecting the components together

Now that the components have been added on the workspace, they have to be connected together. Components connected together form a subjob. Jobs are composed of one or several subjobs carrying out various processes.

In this example, as the **tLogRow** and **tFileOutputDelimited** components are already connected, you only need to connect the **tFileInputDelimited** to the **tLogRow** component.

To connect the components together, proceed as follows:

1. Right-click the source component, **tFileInputDelimited** in this example.
2. In the contextual menu that opens, select the type of connection you want to use to link the components, **Row > Main** in this example.
3. Click the target component to create the link, **tLogRow** in this example.



Note that a black crossed circle is displayed if the target component is not compatible with the link.



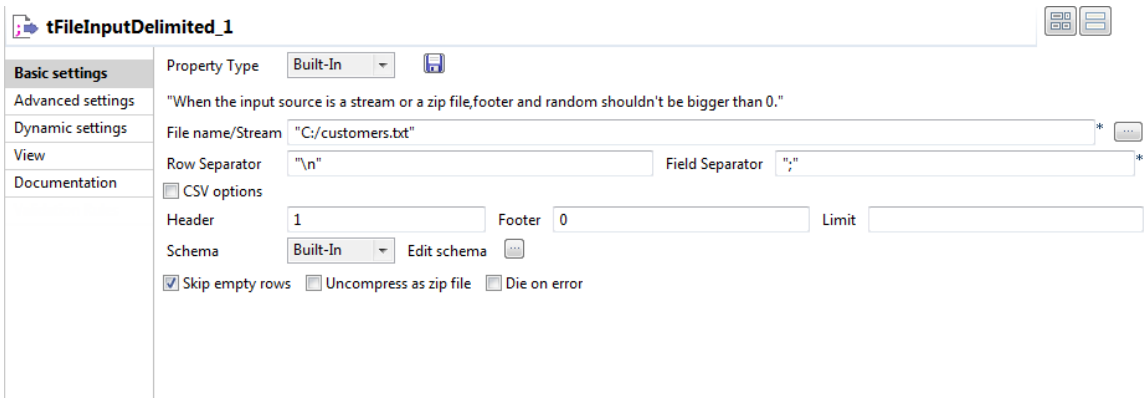
According to the nature and the role of the components you want to link together, several types of link are available. Only the authorized connections are listed in the contextual menu.

3.1.4. Configuring the components

Now that the components are linked, their properties should be defined.

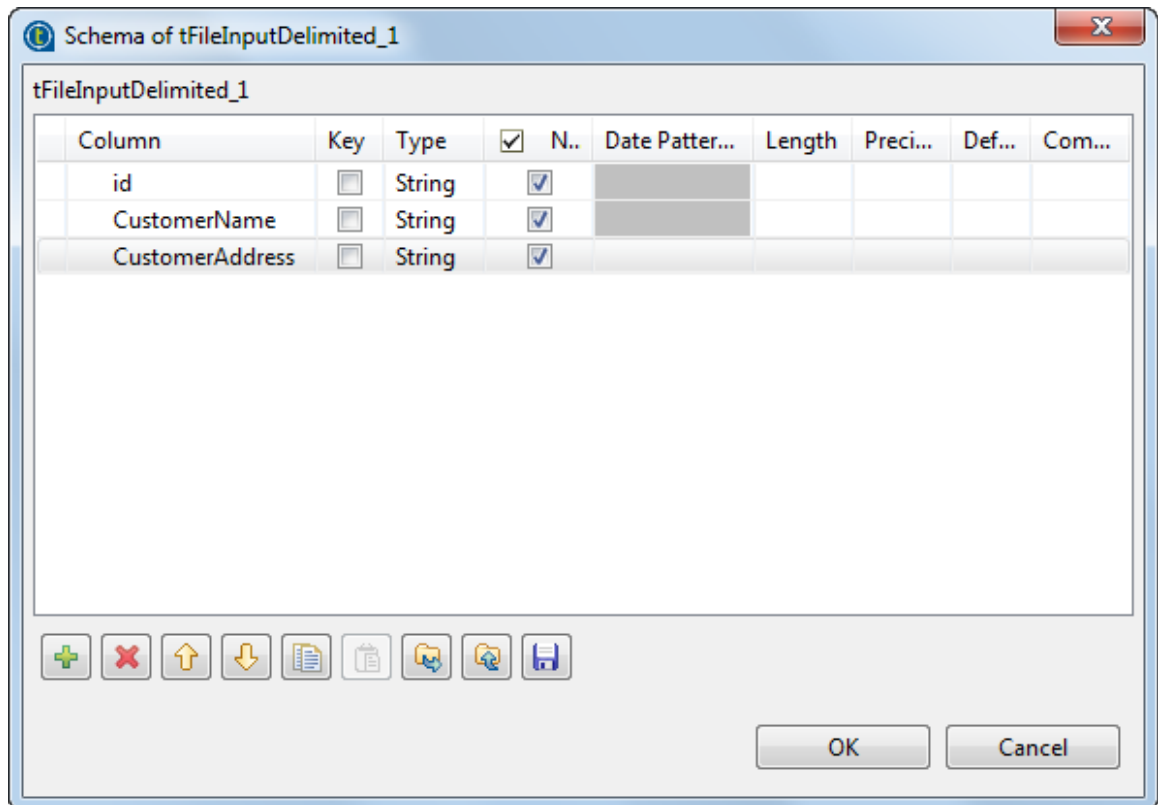
Configuring the tFileInputDelimited component

1. Double-click the **tFileInputDelimited** component to open its **Basic settings** view.



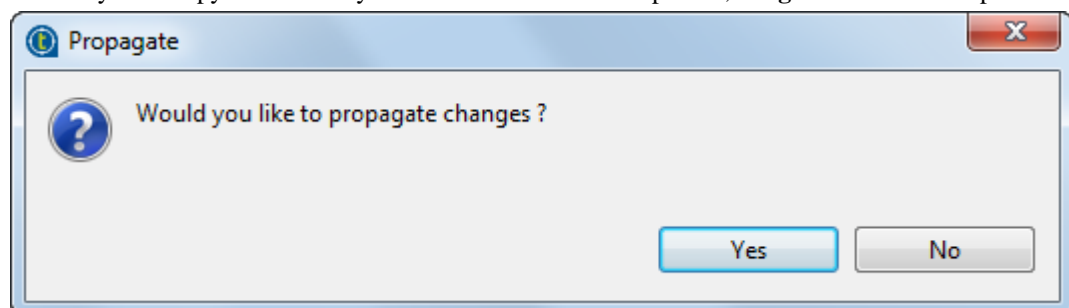
2. Click the [...] button next to the **File Name/Stream** field.

3. Browse your system or enter the path to the input file, *customers.txt* in this example.
4. In the **Header** field, enter *1*.
5. Click the [...] button next to **Edit schema**.
6. In the Schema Editor that opens, click three times the [+] button to add three columns.
7. Name the three columns *id*, *CustomerName* and *CustomerAddress* respectively and click **OK** to close the editor.



8. In the pop-up that opens, click **OK** accept the propagation of the changes.

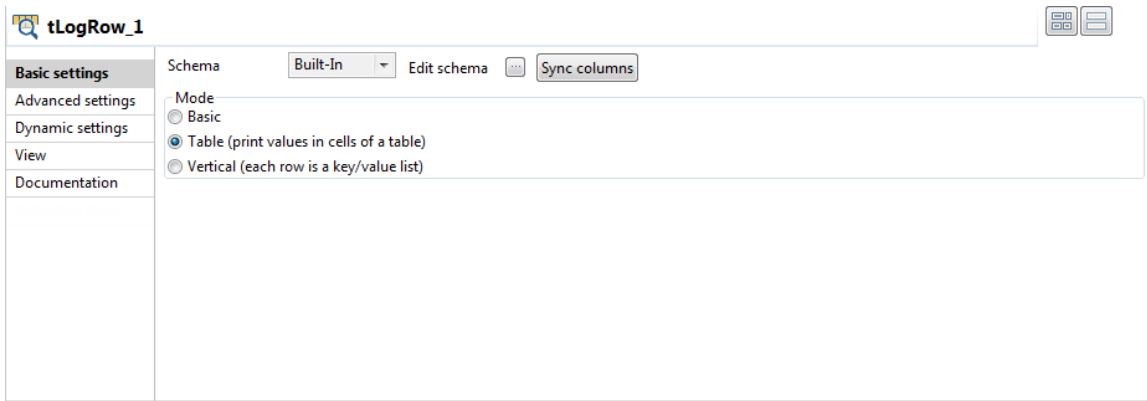
This allows you to copy the schema you created to the next component, **tLogRow** in this example.



Configuring the tLogRow component

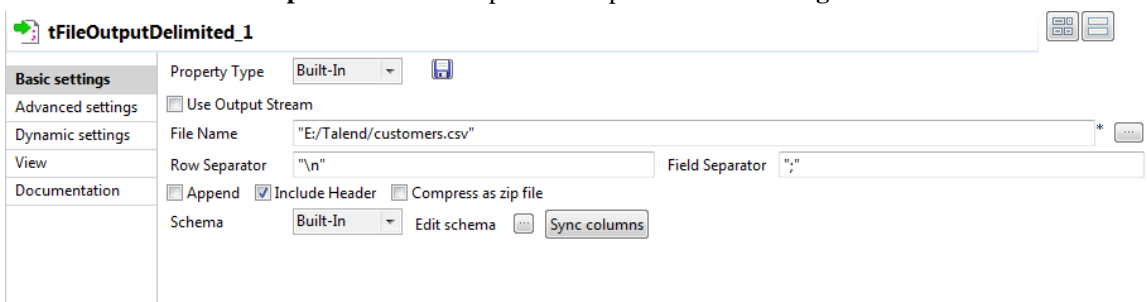
1. Double-click the **tLogRow** component to open its **Basic settings** view.
2. In the **Mode** area, select **Table (print values in cells of a table)**.

By doing so, the contents of the *customers.txt* file will be printed in a table and therefore more readable.



Configuring the tFileOutputDelimited component

1. Double-click the **tFileOutputDelimited** component to open its **Basic settings** view.



2. Click the [...] button next to the **File Name** field.
3. Browse your system or enter the path to the output file, *customers.csv* in this example.
4. Select the **Include Header** check box.
5. If needed, click the **Sync columns** button to retrieve the schema from the input component.

3.1.5. Executing the Job

Now that components are configured, the Job can be executed.

To do so, proceed as follows:

1. Press **Ctrl+S** to save the Job.
2. Go to **Run** tab, and click on **Run** to execute the Job.

The file is read row by row and the extracted fields are displayed on the **Run** console and written to the specified output file.

Job A_Basic_Job

Execution

Run Kill Clear

Starting job A_Basic_Job at 11:37 21/06/2015.

```
[statistics] connecting to socket on port 3648
[statistics] connected
```

tLogRow_1		
id	CustomerName	CustomerAddress
1	Griffith Paving and Sealcoat	talend@apres91
2	Bill's Dive Shop	511 Maple Ave. Apt. 1B
3	Childress Child Day Care	662 Lyons Circle
4	Facelift Kitchen and Bath	220 Vine Ave.
5	Terrinni & Son Auto and Truck	770 Exmoor Rd.
6	Kermit the Pet Shop	1860 Parkside Ln.
7	Tub's Furniture Store	807 Old Trail Rd.
8	Toggle & Myerson Ltd	618 Sheriden rd.
9	Childress Child Day Care	788 Tennyson Ave.
10	Elle Hypnosis and Therapy Cent	2032 Northbrook Ct.

```
[statistics] disconnected
Job A_Basic_Job ended at 11:37 21/06/2015. [exit code=0]
```

Line limit Wrap

