Contents

Copyleft.............................................................................................................................................................................3

Introduction to Talend Open Studio for Big Data................................................................. 5
  Functional architecture of Talend Open Studio for Big Data.....................................................5

Prerequisites to using Talend Open Studio for Big Data.....................................................6
  Memory requirements................................................................................................................6
  Software requirements..............................................................................................................6
  Installing Java......................................................................................................................................................7
  Setting up the Java environment variable on Windows.......................................................7
  Setting up the Java environment variable on Linux............................................................8
  Installing 7-Zip (Windows)......................................................................................................8

Downloading and installing Talend Open Studio for Big Data................................. 9
  Downloading Talend Open Studio for Big Data................................................................9
  Installing Talend Open Studio for Big Data.......................................................................9

Configuring and setting up your Talend product.........................................................11
  Launching the Studio for the first time..................................................................................11
  Logging in to the Studio.......................................................................................................11
  Installing additional packages............................................................................................12
  Setting up Hadoop connection manually........................................................................12
  Setting up connection to HDFS........................................................................................14
  Uploading files to HDFS......................................................................................................17
  Preparing file metadata........................................................................................................20

Performing data integration tasks for Big Data.................................................... 25
  Joining movie and director information............................................................................25
  What’s next?........................................................................................................................................25
Adapted for 7.1.1. Supersedes previous releases.

Publication date: October 15, 2019

The content of this document is correct at the time of publication.

However, more recent updates may be available in the online version that can be found on Talend Help Center.

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.
Introduction to Talend Open Studio for Big Data

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

Built on top of Talend’s data integration solution, the big data solution is a powerful tool that enables users to access, transform, move and synchronize big data by leveraging the Apache Hadoop Big Data Platform and makes the Hadoop platform ever so easy to use.

Functional architecture of Talend Open Studio for Big Data

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

The following chart illustrates the main architectural functional blocks.

The different types of functional blocks are:

- From Talend Studio, you design and launch Big Data Jobs that leverage a Hadoop cluster to handle large data sets. Once launched, these Jobs are sent to, deployed on and executed on this Hadoop cluster.
- The Oozie workflow scheduler system is integrated within the Studio through which you can deploy, schedule, and execute Big Data Jobs on a Hadoop cluster and monitor the execution status and results of these Jobs.
- A Hadoop cluster independent of the Talend system to handle large data sets.
Prerequisites to using Talend Open Studio for Big Data

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

- Memory requirements on page 6
- Software requirements on page 6

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

- Installing Java on page 7
- Setting up the Java environment variable on Windows on page 7 or Setting up the Java environment variable on Linux on page 8
- Installing 7-Zip (Windows) on page 8

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>
<tr>
<td></td>
<td>Apple OS X 10.11/El Capitan</td>
</tr>
</tbody>
</table>

- Oracle Java 8 JRE or OpenJDK 1.8. See Installing Java on page 7.
- A properly installed and configured Hadoop cluster.
Ensure that the client machine on which the Talend Studio is installed can recognize the host names of the nodes of the Hadoop cluster to be used. For this purpose, add the IP address/hostname mapping entries for the services of that Hadoop cluster in the hosts file of the client machine.

For example, if the host name of the Hadoop Namenode server is talend-cdh550.weave.local, and its IP address is 192.168.x.x, the mapping entry reads 192.168.x.x talend-cdh550.weave.local.

**Optional software**
- 7-Zip. See Installing 7-Zip (Windows) on page 8.

**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/](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 7.
- Setting up the Java environment variable on Linux on page 8.

**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.

**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:**
   ```bash
   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 Big Data

Talend Open Studio for Big Data 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 Big Data

Talend Open Studio for Big Data 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 Big Data is downloaded.

Installing Talend Open Studio for Big Data

Unzip the archive previously downloaded to install Talend Open Studio for Big Data.

You can do it either by using:


Extracting via 7-Zip (Windows recommended)

For Windows, Talend recommends you to install 7-Zip and use it to extract files. For more information, see Installing 7-Zip (Windows) on page 8.

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 Big Data.

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 in to the Studio

To log in 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 in 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.

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 Big Data 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.

Setting up Hadoop connection manually

Setting up the connection to a given Hadoop distribution in the Repository allows you to avoid configuring that connection each time when you need to use the same Hadoop distribution.

Before you begin

- Ensure that the client machine on which the Talend Studio is installed can recognize the host names of the nodes of the Hadoop cluster to be used. For this purpose, add the IP address/hostname mapping entries for the services of that Hadoop cluster in the hosts file of the client machine.

  For example, if the host name of the Hadoop Namenode server is talend-cdh550.weave.local, and its IP address is 192.168.x.x, the mapping entry reads 192.168.x.x talend-cdh550.weave.local.

- The Hadoop cluster to be used has been properly configured and is running.

The Cloudera Hadoop cluster to be used in this example is of the CDH V5.5 in the Yarn mode and applies the default configuration of the distribution without enabling the Kerberos security. For further information about the default configuration of the CDH V5.5 distribution, see Deploy CDH 5 on a cluster and Default ports used in CDH5.

Procedure

1. In the Repository tree view of your studio, expand Metadata and then right-click Hadoop cluster.

2. Select Create Hadoop cluster from the contextual menu to open the Hadoop cluster connection wizard.
3. Fill in generic information about this connection, such as **Name** and **Description** and click **Next** to open the **Hadoop configuration import wizard** that helps you import the ready-for-use configuration if any.

4. Select the **Enter manually Hadoop services** check box to manually enter the configuration information for the Hadoop connection being created.

5. Click **Finish** to close this import wizard.

6. From the **Distribution** list, select **Cloudera** and then from the **Version** list, select **Cloudera CDH5.5 (YARN mode)**.

7. In the **Namenode URI** field, enter the URI pointing to the machine used as the NameNode service of the Cloudera Hadoop cluster to be used.

   The NameNode is the master node of a Hadoop system. For example, assume that you have chosen a machine called machine1 as the NameNode, then the location to be entered is `hdfs://machine1:portnumber`.

   On the cluster side, the related property is specified in the configuration file called `core-site.xml`. If you do not know what URI is to be entered, check the `fs.defaultFS` property in the `core-site.xml` file of your cluster.

8. In the **Resource manager** field and the **Resource manager scheduler** field, enter the URIs pointing to these two services, respectively.

   On the cluster side, these two services share the same host machine but use different default portnumbers. For example, if the machine hosting them is resourcemanager.company.com,
the location of the Resource manager is resourcemanager.company.com:8032 and the location of the Resource manager scheduler is resourcemanager.company.com:8030.

If you do not know the name of the hosting machine of these services, check the yarn.resourcemanager.hostname property in the configuration file called yarn-site.xml of your cluster.

9. In the **Job history** field, enter the location of the JobHistory service. This service allows the metrics information of the current Job to be stored in the JobHistory server.

   The related property is specified in the configuration file called mapred-site.xml of your cluster. For the value you need to put in this field, check the mapreduce.jobhistory.address property in this mapred-site.xml file.

10. In the **Staging directory** field, enter this directory defined in your Hadoop cluster for temporary files created by running programs.

   The related property is specified in the mapred-site.xml file of your cluster. For further information, check the yarn.app.mapreduce.am.staging-dir property in this mapred-site.xml file.

11. Select the **Use datanode hostname** check box to allow the Studio to access each Datanode of your cluster via their host names.

   This actually sets the dfs.client.use.datanode.hostname property of your cluster to true.

12. In the **User name** field, enter the user authentication name you want the Studio to use to connect to your Hadoop cluster.

13. Since the Hadoop cluster to be connected to is using the default configuration, leave the other fields or check boxes in this wizard as they are because they are used to define any custom Hadoop configuration.

14. Click the **Check services** button to verify that the Studio can connect to the NameNode and the ResourceManager services you have specified.

   A dialog box pops up to indicate the checking process and the connection status.

   If the connection fails, you can click **Error log** at the end of each progress bar to diagnose the connection issues.

15. Once this check indicates that the connection is successful, click **Finish** to validate your changes and close the wizard.

**Results**

The new connection, called my_cdh in this example, is displayed under the Hadoop cluster folder in the Repository tree view.

You can then continue to create the child connections to different Hadoop elements such as HDFS or Hive based on this connection.

**Setting up connection to HDFS**

A connection to HDFS in the Repository allows you to reuse this connection in related Jobs.
**Before you begin**

- The connection to the Hadoop cluster hosting the HDFS system to be used has been set up from the Hadoop cluster node in the Repository.

  For further information about how to create this connection, see Setting up Hadoop connection manually on page 12.

- The Hadoop cluster to be used has been properly configured and is running and you have the proper access permission to that distribution and its HDFS.

- Ensure that the client machine on which the Talend Studio is installed can recognize the host names of the nodes of the Hadoop cluster to be used. For this purpose, add the IP address/hostname mapping entries for the services of that Hadoop cluster in the hosts file of the client machine.

  For example, if the host name of the Hadoop Namenode server is talend-cdh550.weave.local, and its IP address is 192.168.x.x, the mapping entry reads 192.168.x.x talend-cdh550.weave.local.

**Procedure**

1. Expand the Hadoop cluster node under Metadata in the Repository tree view, right click the Hadoop connection to be used and select Create HDFS from the contextual menu.

2. In the connection wizard that opens up, fill in the generic properties of the connection you need create, such as Name, Purpose and Description.

   ![New HDFS Connection on repository - Step 1/2](image)

   - **Name**: cdh_hdfs
   - **Purpose**: Reuse file schema
   - **Description**: This is the file schema set up to be managed and reused.
   - **Author**: ychen@talend.com

3. Click Next when completed. The second step requires you to fill in the HDFS connection data.
The **User name** property is automatically pre-filled with the value inherited from the Hadoop connection you selected in the previous steps.

The **Row separator** and the **Field separator** properties are using the default values.

4. Select the **Set heading row as column names** check box to use the data in the heading rows of the HDFS file to be used to define the column names of this file.

   The **Header** check box is then automatically selected and the **Header** field is filled with 1. This means that the first row of the file will be ignored as data body but used as column names of the file.

5. Click **Check** to verify your connection.

   A message pops up to indicate whether the connection is successful.

6. Click **Finish** to validate these changes.

**Results**

The new HDFS connection is now available under the **Hadoop cluster** node in the **Repository** tree view. You can then use it to define and centralize the schemas of the files stored in the connected HDFS system in order to reuse these schemas in a Talend Job.
Configuring and setting up your Talend product

Uploading files to HDFS

Uploading a file to HDFS allows the Big Data Jobs to read and process it.

In this procedure, you will create a Job that writes data in the HDFS system of the Cloudera Hadoop cluster to which the connection has been set up in the Repository as explained in Setting up Hadoop connection manually on page 12. This data is needed for the use case described in Performing data integration tasks for Big Data on page 25. For the files needed for the use case, download tos_bd_gettingstarted_source_files.zip from the Downloads tab of the online version of this page at https://help.talend.com.

Before you begin

• The connection to the Hadoop cluster to be used and the connection to the HDFS system of this cluster have been set up from the Hadoop cluster node in the Repository.
  
  If you have not done so, see Setting up Hadoop connection manually on page 12 and then Setting up connection to HDFS on page 14 to create these connections.

• The Hadoop cluster to be used has been properly configured and is running and you have the proper access permission to that distribution and the HDFS folder to be used.

• Ensure that the client machine on which the Talend Jobs are executed can recognize the host names of the nodes of the Hadoop cluster to be used. For this purpose, add the IP address/hostname mapping entries for the services of that Hadoop cluster in the hosts file of the client machine.

  For example, if the host name of the Hadoop Namenode server is talend-cdh550.weave.local, and its IP address is 192.168.x.x, the mapping entry reads 192.168.x.x talend-cdh550.weave.local.

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.

   For example, enter write_to_hdfs in the Name field.

   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.

6. Expand the Hadoop cluster node under Metadata in the Repository tree view.

7. Expand the Hadoop connection you have created and then the HDFS folder under it. In this example, it is the my_cdh Hadoop connection.

8. Drop the HDFS connection from the HDFS folder into the workspace of the Job you are creating. This connection is cdh_hdfs in this example.
9. Select tHDFSPut and click OK to validate your choice.

This Components window is closed and a tHDFSPut component is automatically placed in the workspace of the current Job, with this component having been labelled using the name of the HDFS connection mentioned in the previous step.

10. Double-click tHDFSPut to open its Component view.
The connection to the HDFS system to be used has been automatically configured by using the configuration of the HDFS connection you have set up and stored in the Repository. The related parameters in this tab therefore becomes read-only. These parameters are: Distribution, Version, NameNode URI, Use Datanode Hostname, User kerberos authentication and Username.

11. In the Local directory field, enter the path, or browse to the folder in which the files to be copied to HDFS are stored.

The files about movies and their directors are stored in this directory.

12. In the HDFS directory field, enter the path, or browse to the target directory in HDFS to store the files.

This directory is created on the fly if it does not exist.

13. From the Overwrite file drop-down list, select always to overwrite the files if they already exist in the target directory in HDFS.

14. In the Files table, add one row by clicking the [+ ] button in order to define the criteria to select the files to be copied.

15. In the Filemask column, enter an asterisk (*) within the double quotation marks to make tHDFSPut select all the files stored in the folder you specified in the Local directory field.

16. Leave the New name column empty, that is to say, keep the default double quotation marks as is, so as to make the name of the files unchanged after being uploaded.

17. Press F6 to run the Job.

The Run view is opened automatically. It shows the progress of this Job.
Results

When the Job is done, the files you uploaded can be found in HDFS in the directory you have specified.

Preparing file metadata

In the Repository, setting up the metadata of a file stored in HDFS allows you to directly reuse its schema in a related Big Data component without having to define each related parameter manually.

Since the movies.csv file you need to process has been stored in the HDFS system being used, you can retrieve its schema to set up its metadata in the Repository.
The schema of the `directors.txt` file can also be retrieved, but is intentionally ignored in the retrieval procedure explained below, because in this scenario, this `directors.txt` file is used to demonstrate how to manually define a schema in a Job.

**Before you begin**

- You have launched your Talend Studio and opened the **Integration** perspective.
- The source files `movies.csv` and `directors.txt` have been uploaded into HDFS as explained in Uploading files to HDFS on page 17.
- The connection to the Hadoop cluster to be used and the connection to the HDFS system of this cluster have been set up from the **Hadoop cluster** node in the **Repository**.

  If you have not done so, see Setting up Hadoop connection manually on page 12 and then Setting up connection to HDFS on page 14 to create these connections.
- The Hadoop cluster to be used has been properly configured and is running and you have the proper access permission to that distribution and the HDFS folder to be used.
- Ensure that the client machine on which the Talend Studio is installed can recognize the host names of the nodes of the Hadoop cluster to be used. For this purpose, add the IP address/hostname mapping entries for the services of that Hadoop cluster in the `hosts` file of the client machine.

  For example, if the host name of the Hadoop Namenode server is `talend-cdh550.weave.local`, and its IP address is `192.168.x.x`, the mapping entry reads `192.168.x.x talend-cdh550.weave.local`.

**Procedure**

1. Expand the **Hadoop cluster** node under **Metadata** in the **Repository** tree view.
2. Expand the Hadoop connection you have created and then the **HDFS** folder under it.

   In this example, it is the **my_cdh** Hadoop connection.
3. Right click the HDFS connection in this **HDFS** folder and from the contextual menu, select **Retrieve schema**.

   In this scenario, this HDFS connection is named **cdh_hdfs**.

   A **Schema** wizard is displayed, allowing you to browse to files in HDFS.
4. Expand the file tree to show the **movies.csv** file, from which you need to retrieve the schema, and select it.

   In this scenario, the **movies.csv** file is stored in the following directory: `/user/ychen/input_data`.

5. Click **Next** to display the retrieved schema in the wizard.

   The schema of the movie data is displayed in the wizard and the first row of the data is automatically used as the column names.
If the first row of the data you are using is not used this way, you need to review how you set the Header configuration when you were creating the HDFS connection as explained in Setting up connection to HDFS on page 14.

6. Click Finish to validate these changes.

Results

You can now see the file metadata under the HDFS connection you are using in the Repository tree view.
Hadoop Cluster
- my_cdh 0.1
- HDFS(1)
  - cdh_hdfs 0.1
    - movies
      - Columns(5)
        - directorID
        - movieID
        - releaseYear
        - title
        - url
Performing data integration tasks for Big Data

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

You will work with data about movies and directors and data about your customers as you learn how to:

- upload data stored in a local file system to the HDFS file system of the company’s Hadoop cluster.
- join the director data to the movie data to produce a new dataset and store this dataset in the HDFS system too.

Joining movie and director information

This scenario demonstrates:

1. How to create a Talend Job. See Creating the Job on page 25 for details.
2. How to drop and link the components to be used in a Job. See Dropping and linking components on page 26 for details.
3. How to configure the input components using the related metadata from the Repository. See Configuring the input data for Pig on page 27 for details.
4. How to configure the transformation to join the input data. See Configuring the data transformation for Pig on page 30 for details.
5. How to write the transformed data to HDFS. See Writing the output on page 32 for details.

Creating the Job

A Talend Job allows you to access and use the Talend components to design technical processes to read, transform or write data.

Before you begin

- You have launched your Talend Studio and opened the Integration perspective.

Procedure

1. Right-click the getting_started folder and select Create Job from the contextual menu.
2. In the New Job wizard, give a name to the Job you are going to create and provide other useful information if needed.
   
   For example, enter aggregate_movie_director in the Name field.
   
   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.
3. Click Finish to create your Job.
   
   An empty Job is opened in the Studio.

Results

The component Palette is now available in the Studio. You can start to design the Job by leveraging this Palette and the Metadata node in the Repository.
**Dropping and linking components**

The Pig components to be used are orchestrated in the Job workspace to compose a Pig process for data transformation.

**Before you begin**

- You have launched your Talend Studio and opened the **Integration** perspective.
- An empty Job has been created as described in **Creating the Job** on page 25 and is open in the workspace.

**Procedure**

1. In the Job, enter the name of the component to be used and select this component from the list that appears. In this scenario, the components are two **tPigLoad** components, a **tPigMap** component and two **tPigStoreResult** components.
   - The two **tPigLoad** components are used to load the movie data and the director data, respectively, from HDFS into the data flow of the current Job.
   - The **tPigMap** component is used to transform the input data.
   - The **tPigStoreResult** components write the results into given directories in HDFS.
2. Double-click the label of one of the **tPigLoad** component to make this label editable and then enter **movie** to change the label of this **tPigLoad**.
3. Do the same to label another **tPigLoad** component to **director**.
4. Right click the **tPigLoad** component that is labelled **movie**, then from the contextual menu, select **Row > Pig combine** and click **tPigMap** to connect this **tPigLoad** to the **tPigMap** component. This is the main link through which the movie data is sent to **tPigMap**.
5. Do the same to connect the **director** **tPigLoad** component to **tPigMap** using the **Row > Pig combine** link. This is the **Lookup** link through which the director data is sent to **tPigMap** as lookup data.
6. Do the same to connect the **tPigMap** component to **tPigStoreResult** using the **Row > Pig combine** link, then in the pop-up wizard, name this link to **out1** and click **OK** to validate this change.
7. Repeat these operations to connect the **tPigMap** component to another **tPigStoreResult** component using the **Row > Pig combine** link and name it to **reject**.

**Results**

Now the whole Job looks as follows in the workspace:
Configuring the input data for Pig

Two tPigLoad components are configured to load data from HDFS into the Job.

Before you begin

- The source files, movies.csv and directors.txt have been uploaded into HDFS as explained in Uploading files to HDFS on page 17.
- The metadata of the movie.csv file has been set up in the HDFS folder under the Hadoop cluster node in the Repository.
  
  If you have not done so, see Preparing file metadata on page 20 to create the metadata.

Procedure

1. Expand the Hadoop cluster node under the Metadata node in the Repository and then the my_cdh Hadoop connection node and its child node to display the movies schema metadata node you have set up under the HDFS folder as explained in Preparing file metadata on page 20.
2. Drop this schema metadata node onto the movie tPigLoad component in the workspace of the Job.
3. Double-click the movie tPigLoad component to open its Component view.
   
   This tPigLoad has automatically reused the HDFS configuration and the movie metadata from the Repository to define the related parameters in its Basic settings view.
4. From the **Load function** drop-down list, select **PigStorage** to use the PigStorage function, a built-in function from Pig, to load the movie data as a structured text file. For further information about the PigStorage function of Pig, see [PigStorage](#).

5. From the Hadoop connection node called *my_cdh* in the **Repository**, drop the *cdh_hdfs* HDFS connection node under the **HDFS** folder onto the **tPigLoad** component labelled **director** in the workspace of the Job.

   This applies the configuration of the HDFS connection you have created in the **Repository** on the HDFS-related settings in the current **tPigLoad** component.

6. Double-click the **director** **tPigLoad** component to open its **Component** view.

   This **tPigLoad** has automatically reused the HDFS configuration from the **Repository** to define the related parameters in its **Basic settings** view.
7. Click the [...] button next to **Edit schema** to open the schema editor.

8. Click the [+ ] button twice to add two rows and in the **Column** column, rename them to **ID** and **Name**, respectively.

9. Click **OK** to validate these changes and accept the propagation prompted by the pop-up dialog box.

10. From the **Load function** drop-down list, select **PigStorage** to use the PigStorage function.

11. In the **Input file URI** field, enter the directory where the data about the director data is stored. As is explained in *Uploading files to HDFS* on page 17, this data has been written in /user/ychen/input_data/directors.txt.

12. Click the **Field separator** field to open the **Edit parameter using repository** dialog box to update the field separator.
You need to change this field separator because this **tPigLoad** is reusing the default separator, a semicolon (;), you have defined for the HDFS metadata while the director data is actually using a comma (,) as separator.

13. Select **Change to built-in property** and click **OK** to validate your choice.

The **Field separator** field becomes editable.

14. Enter a comma within double quotation marks.

**Results**

The **tPigLoad** components are now configured to load the movie data and the director data to the Job.

**Configuring the data transformation for Pig**

The tPigMap component is configured to join the movie data and the director data.

Once the movie data and the director data are loaded into the Job, you need to configure the tPigMap component to join them to produce the output you expect.
Procedure

1. Double-click tPigMap to open its Map Editor view.

2. Drop the movieID column, the title column, the releaseYear column and the url column from the left side onto each of the output flow table.

On the input side (left side) of the Map Editor, each of the two tables represents one of the input flow, the upper one for the main flow and the lower one for the lookup flow.

On the output side (right side), the two tables represent the output flows that you named to out1 and reject when you linked tPigMap to tPigStoreResult in Dropping and linking components on page 26.

3. On the input side, drop the directorID column from the main flow table to the Expr.key column of the ID row in the lookup flow table.

This way, the join key between the main flow and the lookup flow is defined.

4. Drop the directorID column from the main flow table to the reject table on the output side and drop the Name column from the lookup flow table to the out1 table.

The configuration in the previous two steps describes how the columns of the input data are mapped to the columns of the output data flow.

From the Schema editor view in the lower part of the editor, you can see the schemas on both sides have been automatically completed.

5. On the out1 output flow table, click the button to display the editing field for the filter expression.
6. Enter `row1.directorId` is not null

This allows tPigMap to output only the movie records in each of which the `directorID` field is not empty. A record with an empty `directorID` field is filtered out.

7. On the reject output flow table, click the button to open the settings panel.

8. In the **Catch Output Reject** row, select **true** to output the records with empty `directorID` fields in the reject flow.

9. Click **Apply**, then click **OK** to validate these changes and accept the propagation prompted by the pop-up dialog box.

Results

The transformation is now configured to complete the movie data with the names of their directors and write the movie records that do not contain any director data into a separate data flow.

Writing the output

Two tPigStoreResult components are configured to write the expected movie data and the rejected movie data to different directories in HDFS.

Before you begin

- Ensure that the client machine on which the Talend Jobs are executed can recognize the host names of the nodes of the Hadoop cluster to be used. For this purpose, add the IP address/hostname mapping entries for the services of that Hadoop cluster in the `hosts` file of the client machine.

  For example, if the host name of the Hadoop Namenode server is `talend-cdh550.weave.local`, and its IP address is `192.168.x.x`, the mapping entry reads `192.168.x.x talend-cdh550.weave.local`.

- The Hadoop cluster to be used has been properly configured and is running.

Procedure

1. Double-click the tPigStoreResult which receives the `out1` link.

   Its **Basic settings** view is opened in the lower part of the Studio.
2. In the **Result file** field, enter the directory you need to write the result in. In this scenario, it is `/user/ychen/output_data/out`, which receives the records that contain the names of the movie directors.

3. Select **Remove result directory if exists** check box.

4. In the **Store function** list, select **PigStorage** to write the records in human-readable UTF-8 format.

5. In the **Field separator** field, enter `;` within double quotation marks.

6. Repeat the same operations to configure the **tPigStoreResult** that receives the `reject` link, but set the directory, in the **Result file** field, to `/user/ychen/output_data/reject`.

7. Press **F6** to run the Job.

**Results**

The **Run** view is automatically opened in the lower part of the Studio and shows the execution progress of this Job.

Once done, you can check, for example in the web console of your HDFS system, that the output has been written in HDFS.

---

**What's next?**

You have seen how Talend Studio helps you manage your big data using Talend Jobs. You have learned how to access and move your data to a given Hadoop cluster via Talend Jobs, filter and transform your data, and store the filtered and transformed data in the HDFS system of the Hadoop cluster. Along the way, you have learned how to centralize frequently used Hadoop 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

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.