# **Урок № 1 - Установка и настройка Pentaho Data Integration + работа с файлами в PDI**

1. Скачать Pentaho DI Community Edition по ссылке: <https://sourceforge.net/projects/pentaho/>

Потребует 1,7 GB свободного места.

Если ссылка изменилась с момента публикации урока, найти актуальную ссылку на загрузку по ключевым словам: «pentaho di community edition download».

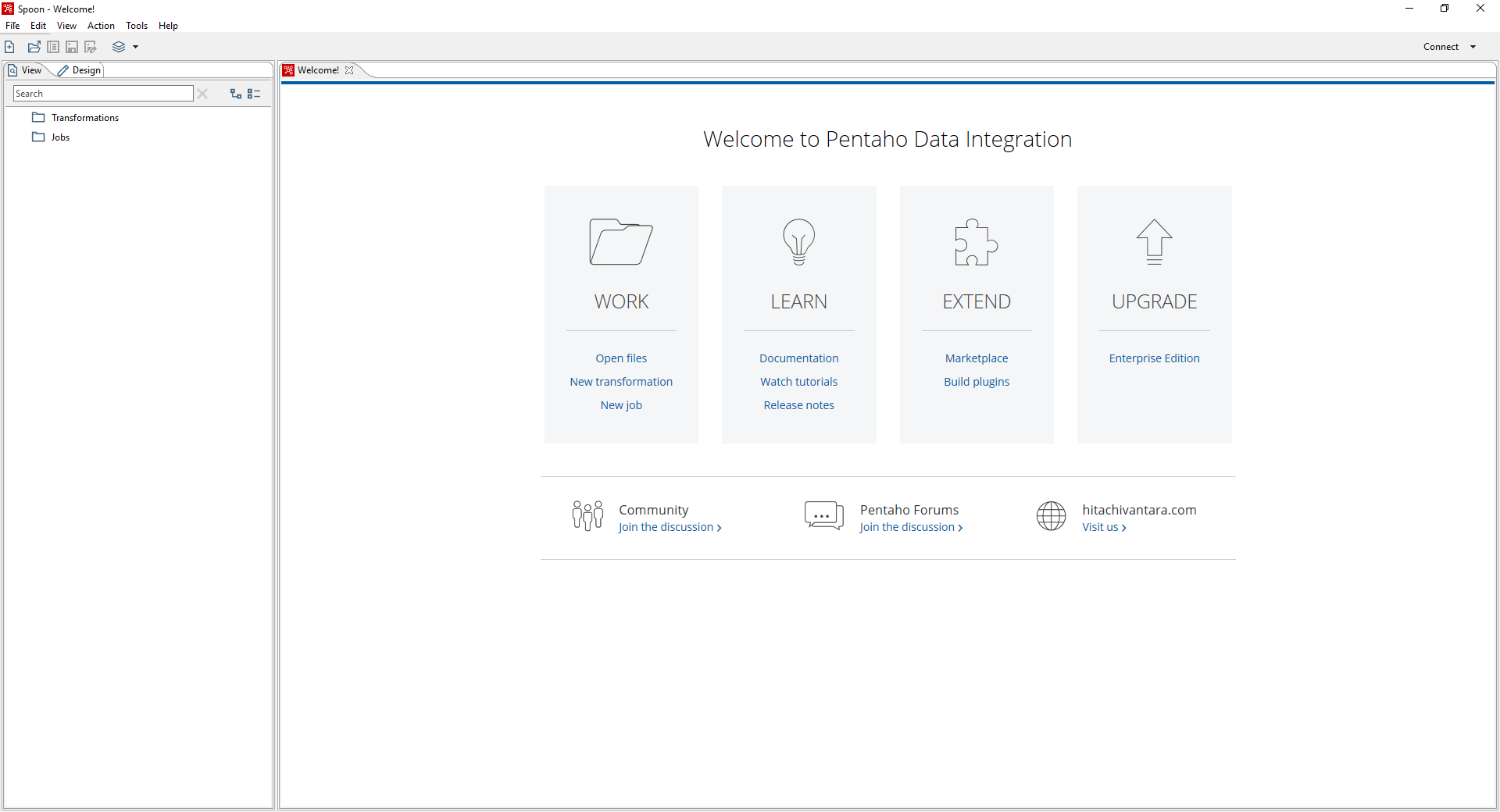
1. После окончания загрузки, распаковать архив в любое удобное место.
2. Установить Java (если она ещё не установлена) и выполнить пункт «Set environment variables» из инструкции:

Windows: <https://help.pentaho.com/Documentation/9.1/Setup/Prepare_your_Windows_environment_for_an_archive_install>

Linux: <https://help.pentaho.com/Documentation/9.1/Setup/Prepare_your_Linux_environment_for_an_archive_install>

Пункт «Install Java» поможет снять вопросы, возникающие при установке java.

1. Проверить корректность выполнения запустив файл data-integration\spoon.bat (для Windows) или data-integration/spoon.sh (для linux). Если все шаги выполнены корректно, то после запуска файла откроется главное окно программы:



Если в результате запуска файла происходит ошибка, то стоит проверить лог для определения причины ошибки. Логи расположены в папке data-integration\logs.

1. Ознакомиться с документацией

<https://help.pentaho.com/Documentation/8.0/Products/Data_Integration/Data_Integration_Perspective/010>

<https://help.pentaho.com/Documentation/9.1/Setup/Pentaho_Data_Integration_(PDI)_tutorial>

<https://help.pentaho.com/Documentation/9.1/Setup/PDI_job_tutorial>

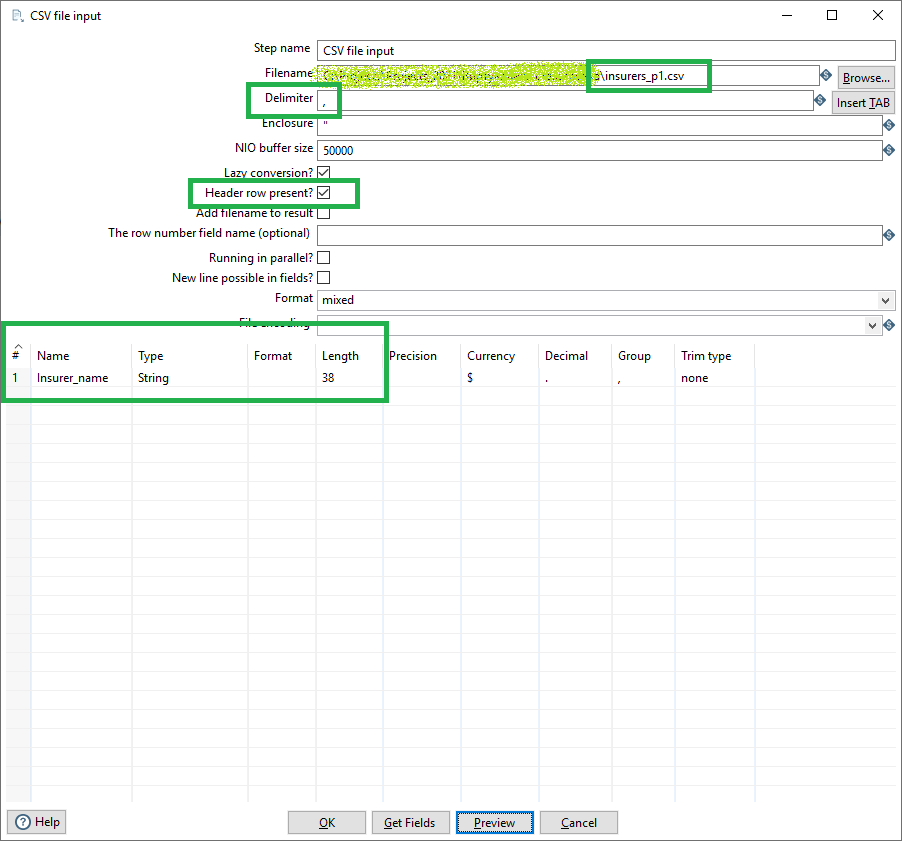
<https://help.pentaho.com/Documentation/9.1/Products/Learn_about_the_PDI_client>

# **Работа с файлами в PDI**

Предисловие: от заказчика поступили файлы с данными (5 шт). Два файла содержат наименования страховых компаний (insurers\_p1.csv и insurers\_p2.csv), перечень страховых продуктов (producttype.csv), перечень штатов в которых физически расположены компании (states.csv) и историю продаж услуг клиентам (fact.csv). Все эти данные в дальнейшем потребуется загрузить в базу данных. Но в начале требуется поработать с файлами.

Pentaho DI умеет работать с различными источниками и приёмниками данных. Самое простое – начать освоение с файлов.

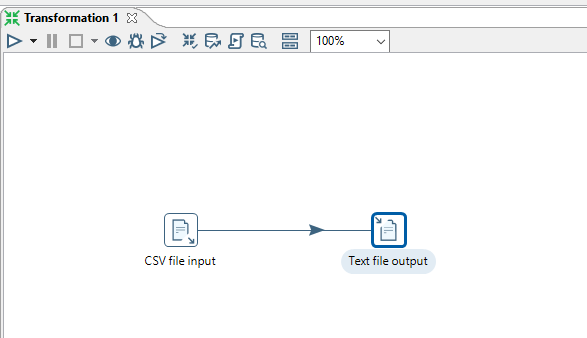
1. Ознакомиться с содержимым файлов insurers\_p1.csv и insurers\_p2.csv. Лучше всего это делать с помощью Notepad++ (<https://notepad-plus-plus.org/downloads/>) либо с помощью другого текстового редактора, но не через MS Excel – он отображает большинство csv файлов некорректно.
2. Создать трансформацию, объединяющую данные из двух файлов insurers\_p1.csv и insurers\_p2.csv в третий итоговый файл. Для этого потребуется два компонента: CSV file input и Text file output. Требуется открыть PDI, создать новую трансформацию и разместить на ней шаг «CSV file input». Далее, двойным щелчком открыть свойства этого шага и выбрать первый файл в поле filename. Далее, убедиться, что разделителем выбрана запятая «,», что стоит галка «Header row present?». После этого, нажать на кнопку «Get fields». При этом файл будет проанализирован и PDI увидит, что в файле содержится только один столбец с именем «Insurer\_name», тип столбца – String, длина 38 символов. Результат должен быть таким:



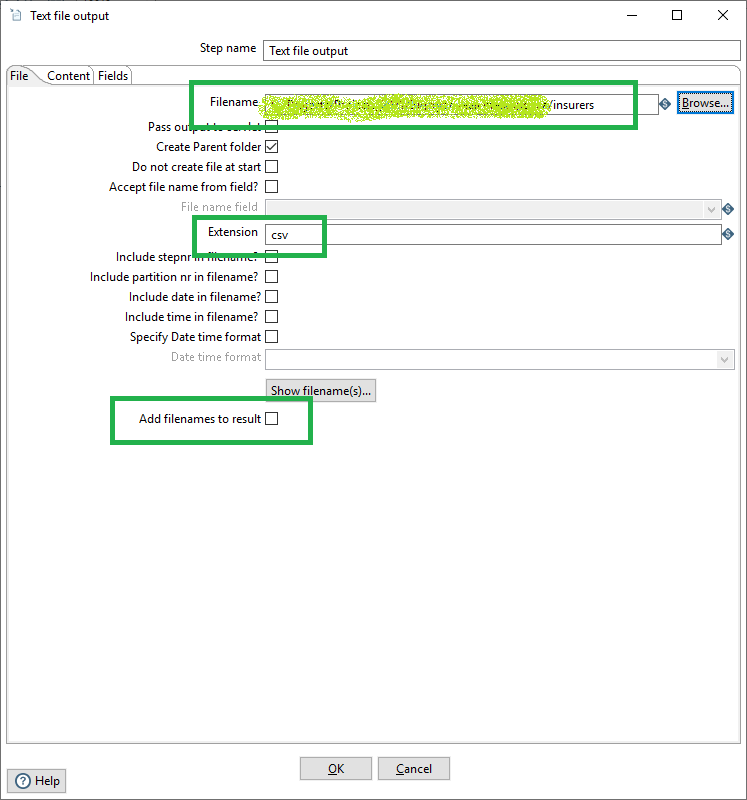
Можно нажать «Preview» чтобы просмотреть содержимое файла как его видит PDI.

Далее, нажать OK и сохранить трансформацию.

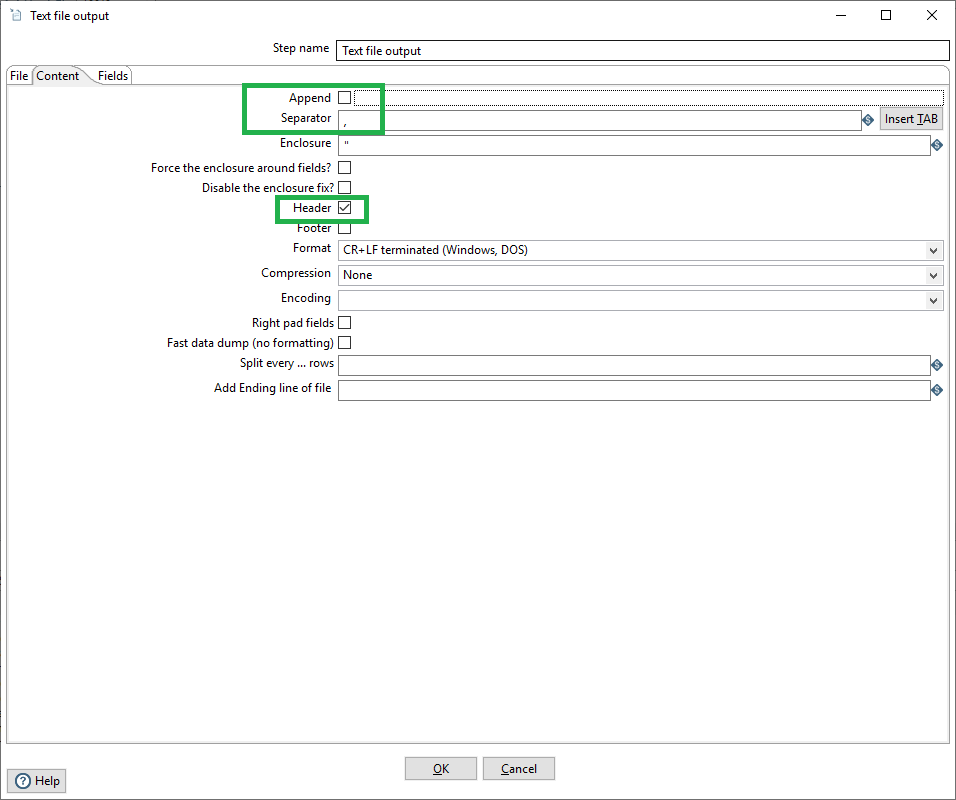
1. Открыть свойства шага «CSV file input», нажать «Help» и прочитать назначение галки «Header row present?».
2. Добавить в трансформацию шаг «Text file output» и соединить его с предыдущим шагом:



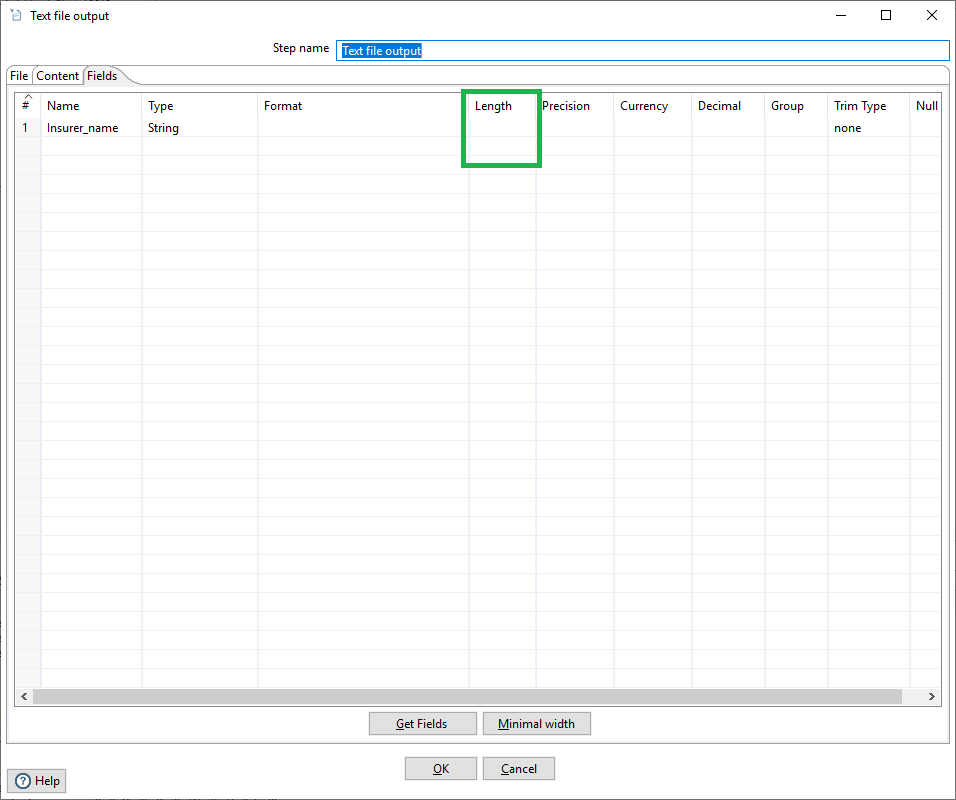
Открыть свойства «Text file output» и указать имя результирующего файла. Выбрать путём сохранения папку, содержащую исходные файлы. Указать что файл будет иметь расширение csv и убедиться, что галка «Add filenames to result» снята:



Перейти на вкладку «Content» и убедиться в отсутствии галки «Append», и заменить separator с двоеточия на запятую т поставить галку «Header»:



Перейти на вкладку «Fields» и нажать «Get fields» и убрать длину у поля «Insurer\_name» (Иначе строки в новом файле будут заполнены пробелами до указанной длины):



После этого нажать ОК и сохранить трансформацию.

1. Запустить трансформацию и убедиться, что в указанной папке появился файл с идентичный по содержанию файлу «insurers\_p1.csv».
2. Запустить трансформацию ещё раз, но в этот раз изменить исходный файл на «insurers\_p2.csv» и добавив галку «Append» в Text file output.
3. Убедиться, что теперь файл «insurers.csv» содержит 51 строку: 50 названий компаний и первая строка-заголовок «Insurer\_name»