

Департамент образования и науки города Москвы
Государственное автономное образовательное учреждение
высшего образования города Москвы
«Московский городской педагогический университет»
Институт цифрового образования
Департамент информатики, управления и технологий

ДИСЦИПЛИНА:

«Инструменты для хранения и обработки больших данных»

пр_03

Тема:

«Архитектура хранилищ данных: традиционная и облачная»

Выполнила: Николаева С.Г., АДЭУ-201

Преподаватель: Босенко Тимур Муртазович

Москва

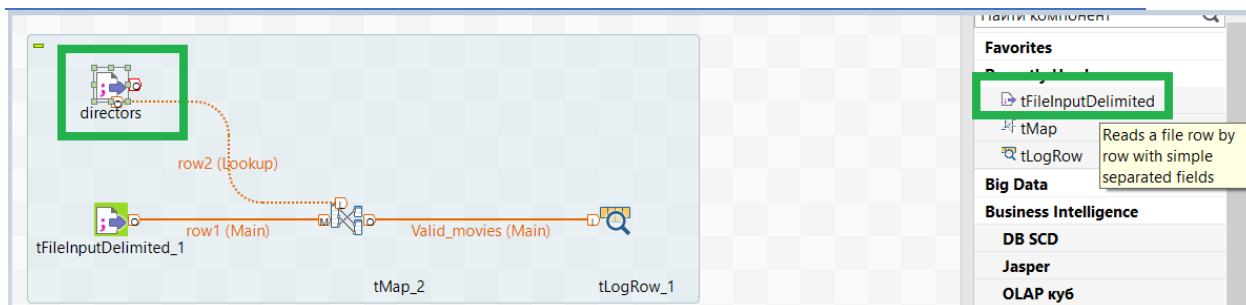
2023

Практика 3

В качестве практики вам необходимо выявить 8-10 подсистем в **ETL Pentaho DI и Talend**, написать отчет, в котором Вы приложите **print screen** компонента (ETL подсистемы) и напишите про его свойства. Результат сохраните в вашем **Git**.

Talend

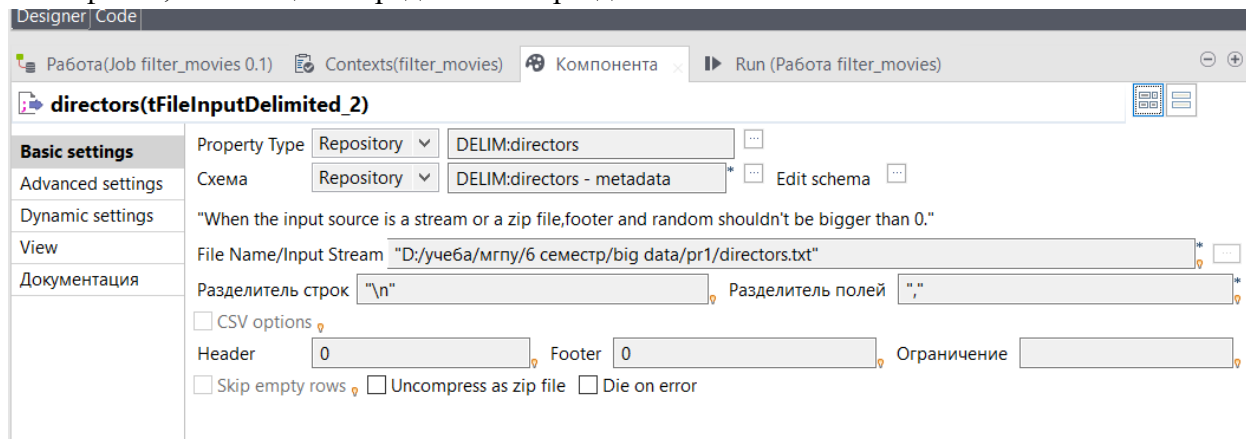
tFileInputDelimited



Читает заданный файл построчно с простыми разделенными полями.

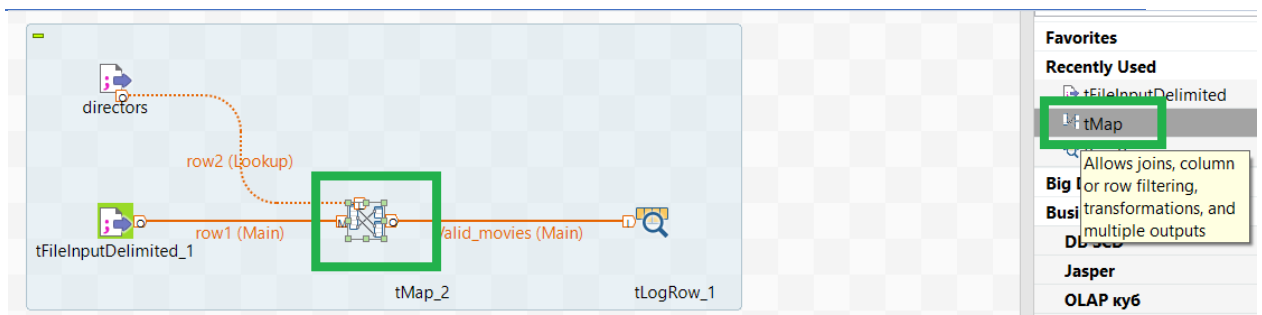
Открывает файл и считывает его строку за строкой, чтобы разделить их на поля, а затем отправляет поля, как определено в схеме, в следующий компонент задания через ссылку строки.

Используйте этот компонент для чтения файла и разделения полей, содержащихся в этом файле, с помощью определенного разделителя.



Имя файла/поток	Имя файла: имя и путь к файлу, который необходимо обработать. Поток: поток данных для обработки. Данные должны быть добавлены в поток, чтобы tFileInputDelimited мог получить эти данные через соответствующую репрезентативную переменную.
Разделитель строк	Введите разделитель, используемый для обозначения конца строки.
Разделитель полей	Введите символ, строку или регулярное выражение, чтобы разделить поля для передаваемых данных.
Параметры CSV	Установите этот флажок, чтобы включить параметры, характерные для CSV, такие как экранирующий символ и вложение текста.
Заголовок	Введите количество строк, которые необходимо пропустить в начале файла.
Нижний колонтитул	Количество строк, которые необходимо пропустить в конце файла.
Ограничение	Максимальное количество обрабатываемых строк. Если Limit = 0, ни одна строка не считывается и не обрабатывается.

tMap



Преобразует и направляет данные из одного или нескольких источников в одно или несколько мест назначения.

tMap — это расширенный компонент, который интегрируется в качестве плагина в Talend Studio.

Designer | Code

Работа(Job filter_movies 0.1) Contexts(filter_movies) Компонента Run (Работа filter_movies)

tMap_2

Map Editor: Mapping links display as: Auto

Сохранить на диск
Temp data directory path:

Предпросмотр:

row1	Column
	movieID
	title
	releaseYear
	url
	directorID

row2	Expr. key	Column
	row1.directorID	directorID
		directorName

Find:

Var

row1	Column
	movieID
	title
	releaseYear
	url
	directorID

row2	Expr. key	Column
	row1.directorID	directorID
		directorName

Find:

Var

Valid_movies

Выражение	Column
row1.movieID	movieID
row1.title	title
row2.directorName	directorBy
row1.releaseYear	releaseYear
row1.url	url

Schema editor Редактор выражений

Колонка	K.	Тип	N.	Date Pattern (Ctrl+...)	Длина	Precision	Default	Коммента...
movieID		Integer	<input checked="" type="checkbox"/>		4	0		
title		String	<input checked="" type="checkbox"/>		72	0		
releaseYear		Integer	<input checked="" type="checkbox"/>		4	0		
url		String	<input checked="" type="checkbox"/>		120	0		
directorID		String	<input checked="" type="checkbox"/>		3	0		

Колонка	K.	Тип	N.	Date Pattern (Ctrl+...)	Длина	Precision	Default	Коммента...
movieID		Integer	<input checked="" type="checkbox"/>		4	0		
title		String	<input checked="" type="checkbox"/>		72	0		
directorBy		String	<input checked="" type="checkbox"/>		20	0		
releaseYear		Integer	<input checked="" type="checkbox"/>		4	0		
url		String	<input checked="" type="checkbox"/>		120	0		

Применить Ок Отменить

Map editor

Позволяет определить свойства маршрутизации и преобразования tMap.

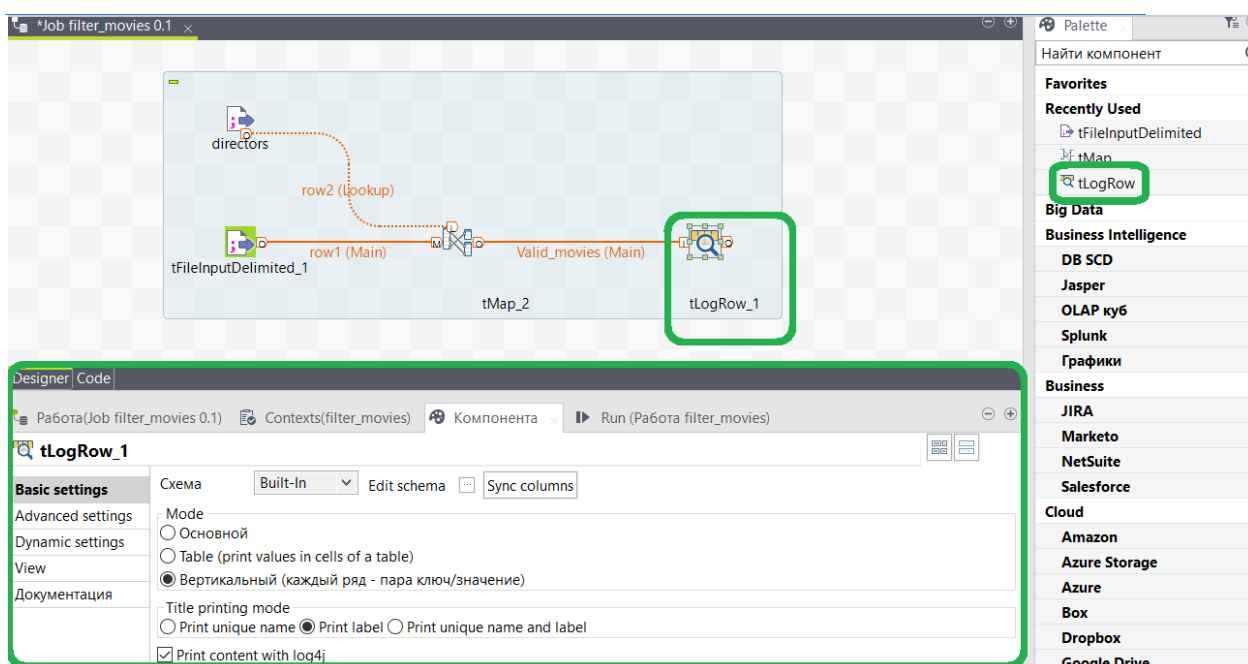
Die on error: установите этот флажок, если вы хотите завершить задание в случае ошибки. Этот флажок установлен по умолчанию.

Lookup in parallel: выберите этот флажок, чтобы максимизировать производительность преобразования данных в задании, которое обрабатывает несколько входных потоков поиска с большими объемами данных.

Enable Auto-Conversion of types: если ваши входные и выходные столбцы в сопоставлении имеют разные типы данных, установите этот флажок, чтобы включить автоматическое преобразование типов во время выполнения, чтобы избежать ошибок компиляции.

Mapping links display as	Auto: настройка по умолчанию кривые ссылки Curves: отображение в виде кривых Lines: сопоставление отображается в виде прямых линий. Эта последняя опция позволяет немного повысить производительность.
Temp data directory path	Введите путь, по которому вы хотите хранить временные данные, сгенерированные для загрузки поиска.
Preview	Предварительный просмотр — это мгновенный снимок данных Mapper. Он становится доступным, когда свойства Mapper заполнены данными. Синхронизация предварительного просмотра вступает в силу только после сохранения изменений.

tLogRow



Отображает данные или результаты в консоли «**Выполнение**» для отслеживания обработанных данных.

Schema and Edit Schema	<p>Схема — это описание строки, оно определяет поля, которые необходимо обработать и передать следующему компоненту.</p> <p>Схема этого компонента доступна только для чтения. Он описывает свойства данных журнала. Вы можете нажать кнопку [...] рядом с... Изменить схему, чтобы просмотреть предопределенную схему, которая содержит следующие поля:</p> <ul style="list-style-type: none"> moment : время, когда сообщение было перехвачено. pid: идентификатор процесса задания. root_pid: идентификатор корневого процесса.
-------------------------------	---

	<ul style="list-style-type: none"> • Father_pid: идентификатор родительского процесса. • project: название проекта. • job: название задания. • context: контекст, используемый для запуска задания. • priority: уровень приоритета сообщения. • type: тип сообщения. • origin: имя компонента, который запускает сообщение. • message: содержимое сообщения. • code: уровень кода ошибки.
Mode	<p>Выберите тип значений для измеренных данных: Absolute: регистрируется фактическое количество строк.</p> <p>Relative: регистрируется соотношение (%) количества строк. Когда выбран этот параметр, отображается список соединений, позволяющий выбрать эталонное соединение.</p>

tDBOutput

The screenshot displays the Talend Designer interface. At the top, a data flow diagram is visible, showing a sequence of components: 'directors', 'tFileInputDelimited_1', 'tMap_2', and 'tDBOutput_1'. The 'tDBOutput_1' component is highlighted with a green pentagon. Below the diagram, the 'Designer' tab is active, showing the configuration for 'tDBOutput_1(MySQL)'. The configuration is divided into several sections: 'Basic settings', 'Advanced settings', 'Dynamic settings', 'View', and 'Документация'. The 'Basic settings' section is expanded, showing the following fields: 'Database' (MySQL), 'Property Type' (Built-In), 'Версия БД' (Mysql 8), 'Использовать существующее соединение' (unchecked), 'Хост' (localhost), 'Порт' (3306), 'Database' (nikolaevadb), 'Имя пользователя' (root), 'Пароль' (*****), 'Таблица' (valid_movies), 'Action on table' (Create table if not exists), 'Действие над данными' (Вставить), 'Схема' (Built-In), 'Edit schema' (button), 'Sync columns' (button), 'Data source' (This option only applies when deploying and running in the Talend Runtime), 'Specify a data source alias' (unchecked), and 'Die on error' (unchecked).

Записывает, обновляет, вносит изменения или подавляет записи в базе данных.

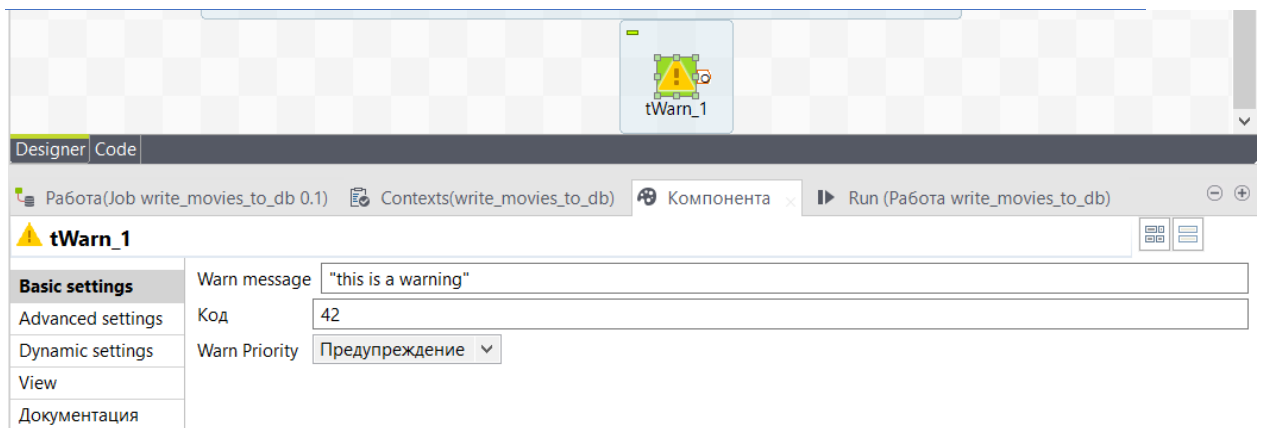
Этот компонент работает с различными базами данных в зависимости от вашего выбора.

Этот компонент служит точкой входа для следующих баз данных. Чтобы настроить этот компонент, выберите тип базы данных из списка **База данных** и нажмите кнопку **Применить** в представлении **основных параметров**.

- Access
- Amazon Mysql
- Amazon Oracle
- Amazon Redshift
- Greenplum
- IBM DB2
- Microsoft SQL Server
- MySQL
- Oracle

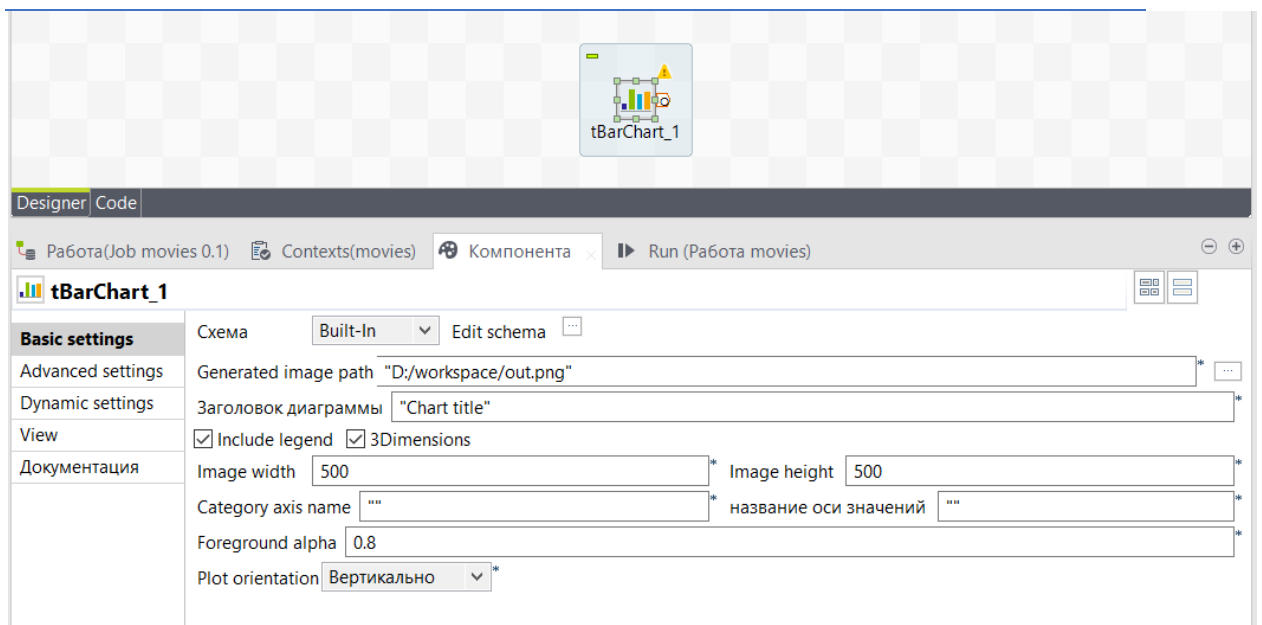
Возможны и другие варианты БД.

tWarn



Позволяет получить код и сообщение, связанные с поднятым исключением. Обычно он используется для обозначения завершения задания или исключений, которые не блокируют выполнение задания. Вы можете выбрать сообщение и код, который хотите отправить компоненту ловушки. Например, вы можете использовать tWarn, чтобы сигнализировать об окончании определенного конвейерного потока в вашем задании.

tBarChart



Создает гистограмму из входных данных для облегчения технического анализа.

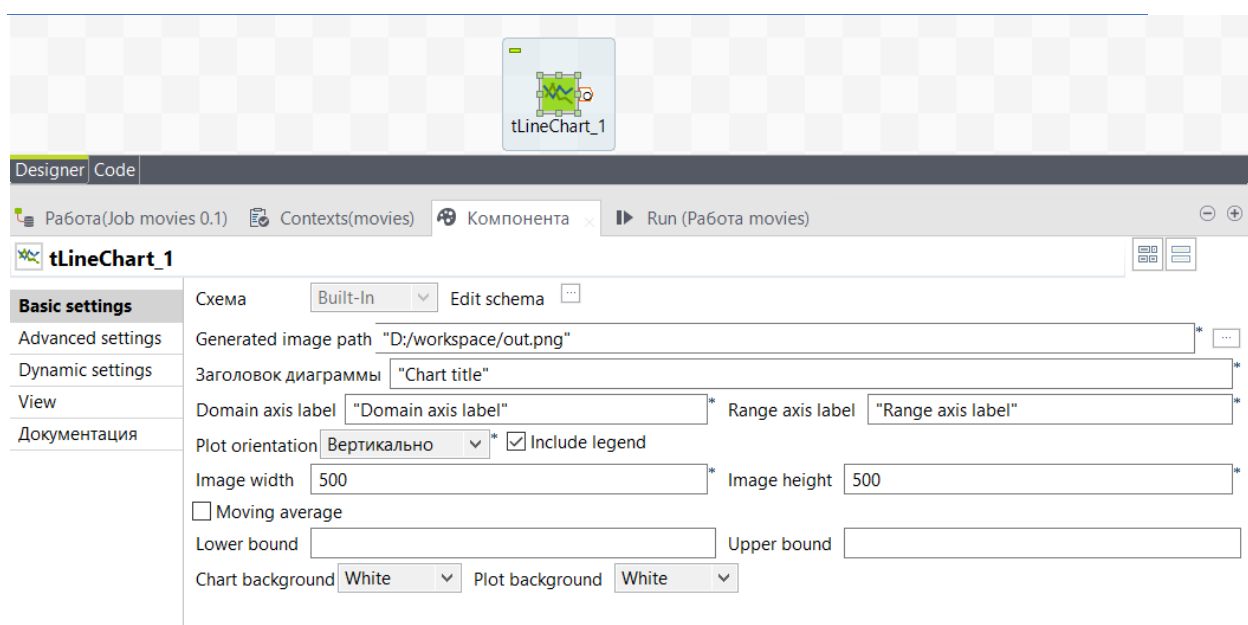
tBarChart считывает данные из входного потока и преобразует их в гистограмму в файле изображения PNG.

Схема и схема редактирования

Схема представляет собой описание строки. Он определяет количество полей (столбцов), которые необходимо обработать и передать следующему компоненту.

Синхронизировать столбцы	Нажмите, чтобы синхронизировать схему выходного файла со схемой входного файла. Функция синхронизации отображается только после того, как соединение Row связано с выходным компонентом.
Сгенерированный путь к изображению	Имя и путь к выходному файлу изображения.
Название диаграммы	Введите название гистограммы, которую необходимо сгенерировать.
Включить легенду	Установите этот флажок, если вы хотите, чтобы линейчатая диаграмма включала легенду, обозначающую все ряды разными цветами.
3 измерения	Установите этот флажок, чтобы создать изображение с 3D-эффектом. По умолчанию этот флажок установлен, и столбцы, представляющие серию каждой категории, будут располагаться друг над другом. Если этот флажок снят, будет создано 2D-изображение с полосами, расположенными одна за другой вдоль оси категорий.
Ширина изображения и высота изображения	Введите ширину и высоту файла изображения в пикселях.
Имя оси категорий и имя оси значений	Введите имя оси категорий и имя оси значений.
Альфа переднего плана	Введите целое число в диапазоне от 0 до 100, чтобы задать прозрачность изображения. Чем меньше введенное число, тем более прозрачным будет изображение.
Ориентация сюжета	Выберите ориентацию гистограммы: ВЕРТИКАЛЬНАЯ или ГОРИЗОНТАЛЬНАЯ.

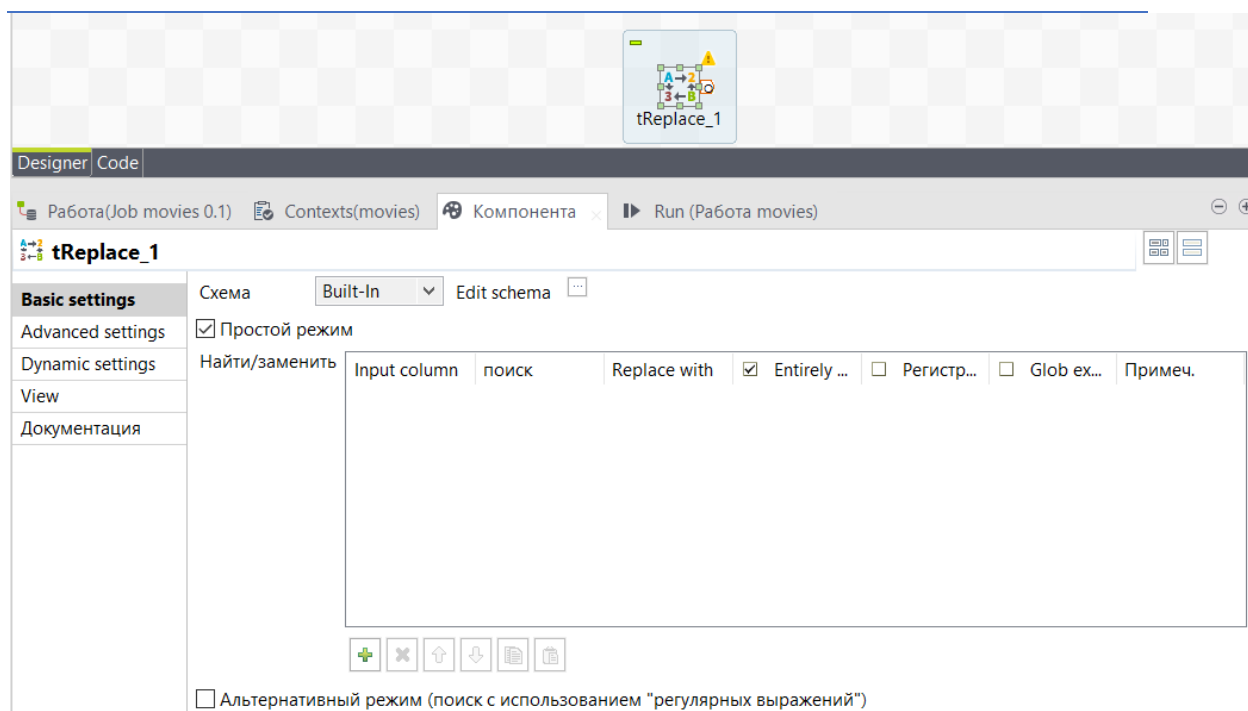
tLineChart



Считывает данные из входного потока и преобразует данные в линейную диаграмму в файле изображения PNG для упрощения технического анализа.

Остальные свойства аналогичны предыдущему компоненту.

tReplace




Очищает все файлы перед дальнейшей обработкой.

Выполняет операцию поиска и замены в определенных входных столбцах.

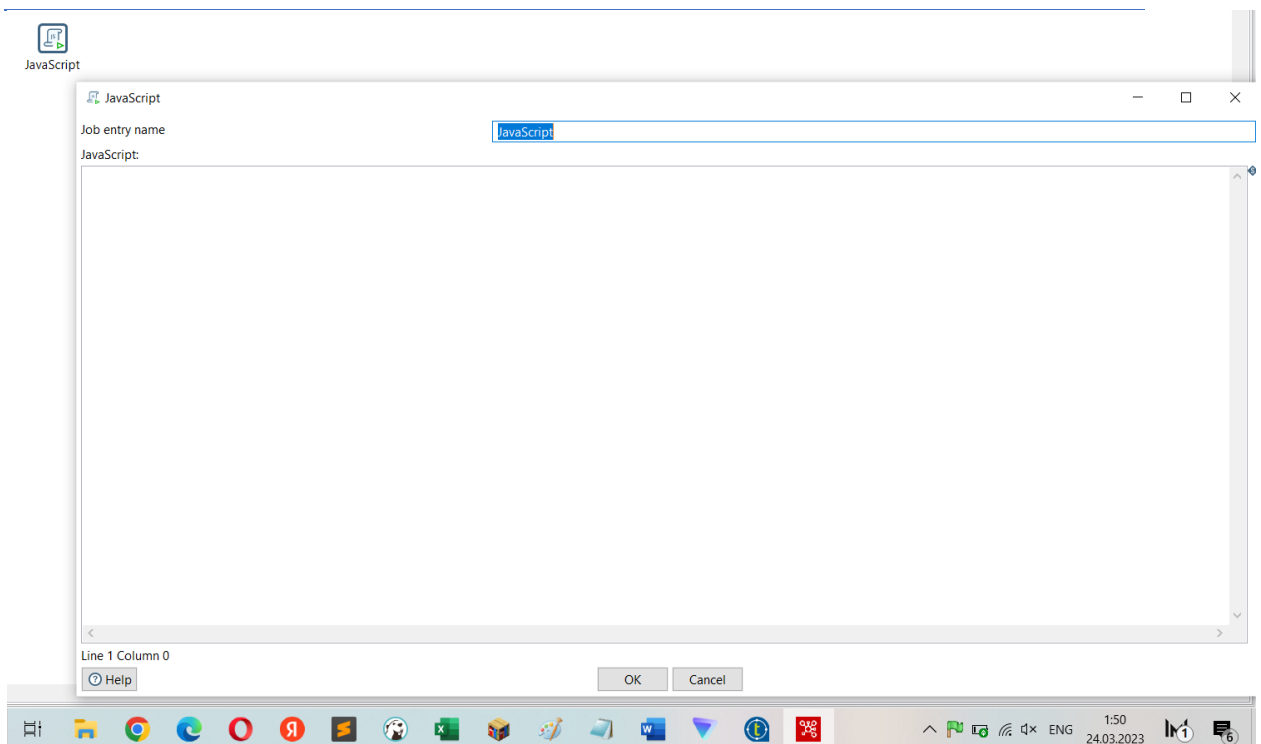
Схема и редактирование схемы

Схема представляет собой описание строки. Он определяет количество полей (столбцов), которые

	необходимо обработать и передать следующему компоненту.
Простой режим поиска/замены	Нажмите  кнопку, чтобы добавить столько условий, сколько необходимо. Условия выполняются одно за другим для каждой строки.
Использовать расширенный режим	Установите этот флажок, если операцию, которую вы хотите выполнить, нельзя выполнить в простом режиме. В текстовом поле введите необходимое регулярное выражение.

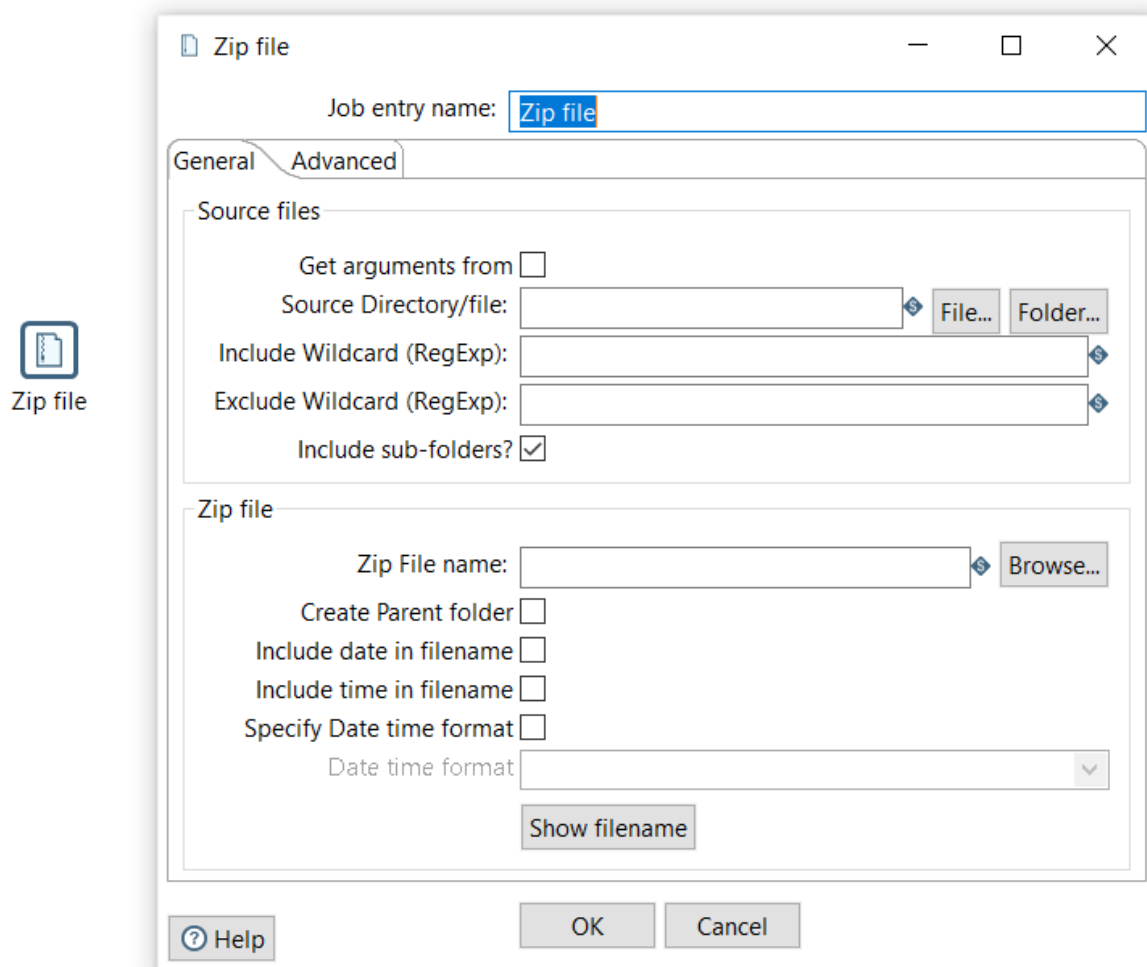
Pentaho DI

JavaScript



Дает возможность писать код на языке программирования JavaScript, чтобы преобразовать данные.

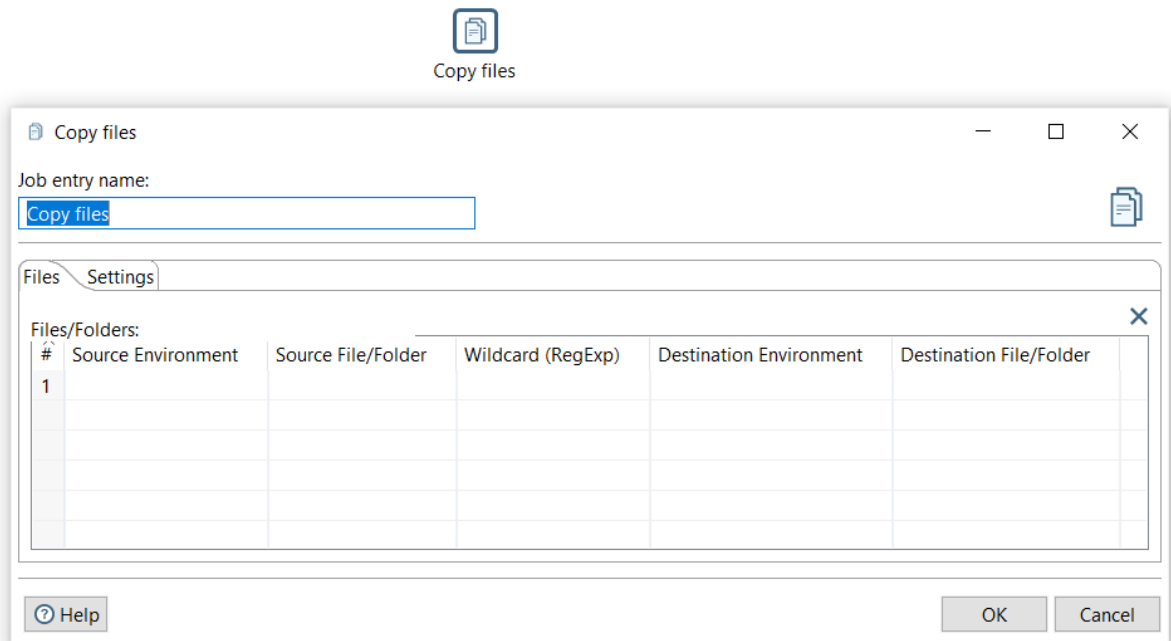
Zip File



На этом шаге создается стандартный ZIP-архив с использованием параметров, указанных в диалоговом окне.

Job entry name	Имя этой записи, как оно появляется в рабочем пространстве преобразования.
Source directory	Исходный каталог архивируемых файлов.
Include date in filename	Установите этот флажок, если вы хотите добавить дату в имя файла.
Include time in filename	Установите этот флажок, если вы хотите добавить время в имя файла.
Specify Date time format	Укажите формат даты или времени для включения в имя файла.

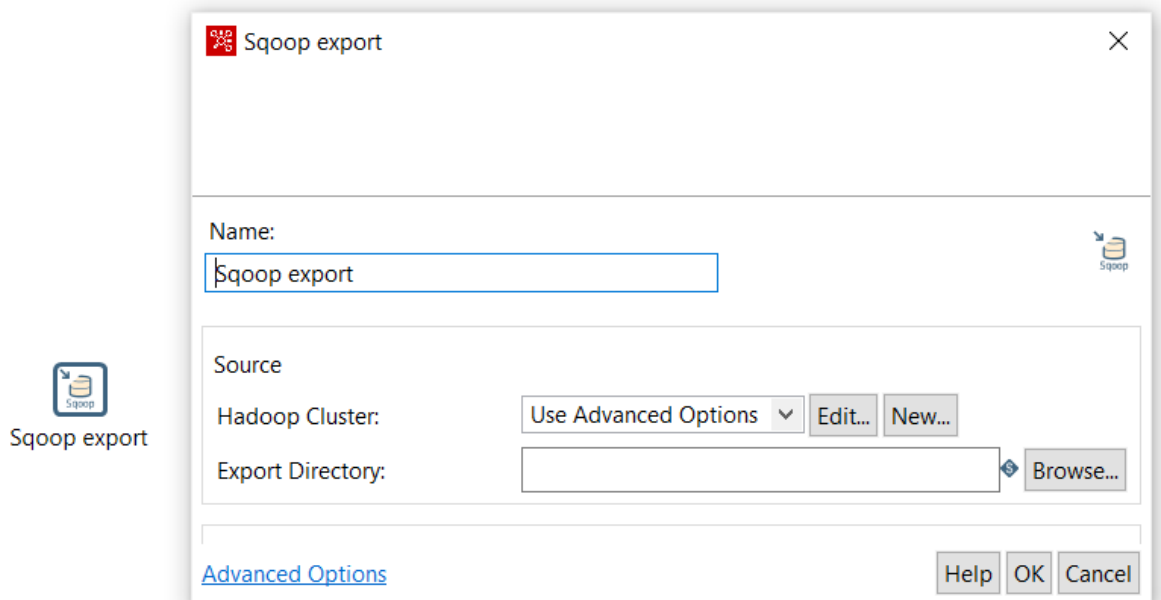
Copy files



Вы можете скопировать один или несколько файлов или папок с этой записью задания.

Job entry name	Имя этой записи, как оно появляется в рабочем пространстве преобразования.
Include subfolders	Установите этот флажок, если вы хотите включить вложенные папки. Примечание: Этот параметр будет работать только в том случае, если источником является папка.
Copy empty folders	Установите этот флажок, чтобы также копировать пустые папки
Replace existing files	Если целевой файл или папка существует, этот параметр заменит его, иначе PDI его проигнорирует.
File/Folder destination	Введите целевой файл или папку.

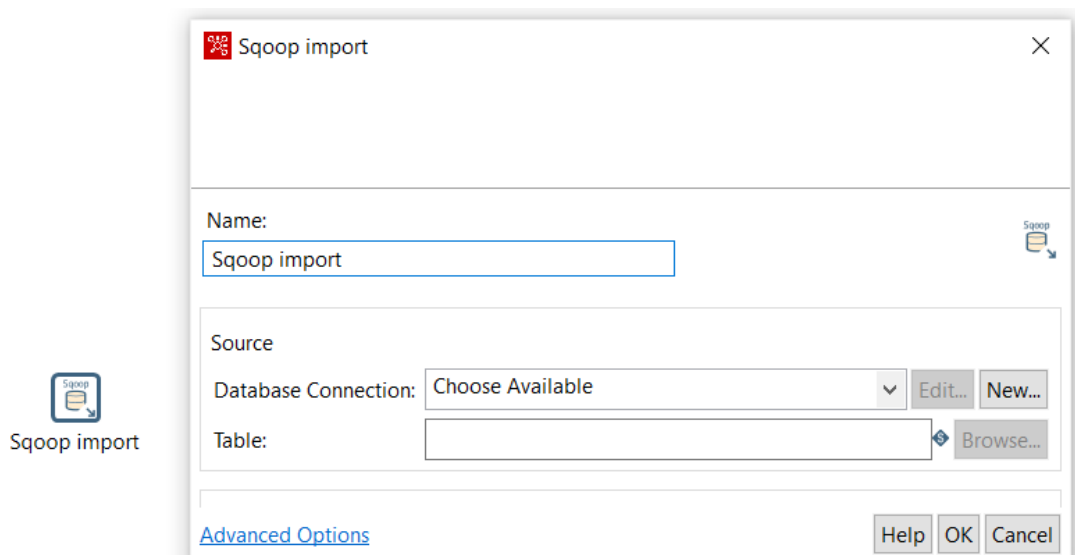
Sqoop export



Задание Sqoop Export позволяет экспортировать данные из Hadoop в СУБД с помощью Apache Sqoop. Эта задача имеет два режима настройки:

- Быстрый режим предоставляет минимальные параметры, необходимые для успешного экспорта Sqoop.
- Представление по умолчанию в расширенном режиме предоставляет параметры для лучшего управления экспортом Sqoop. Расширенный режим также имеет представление командной строки, которое позволяет повторно использовать существующую команду Sqoop из командной строки.

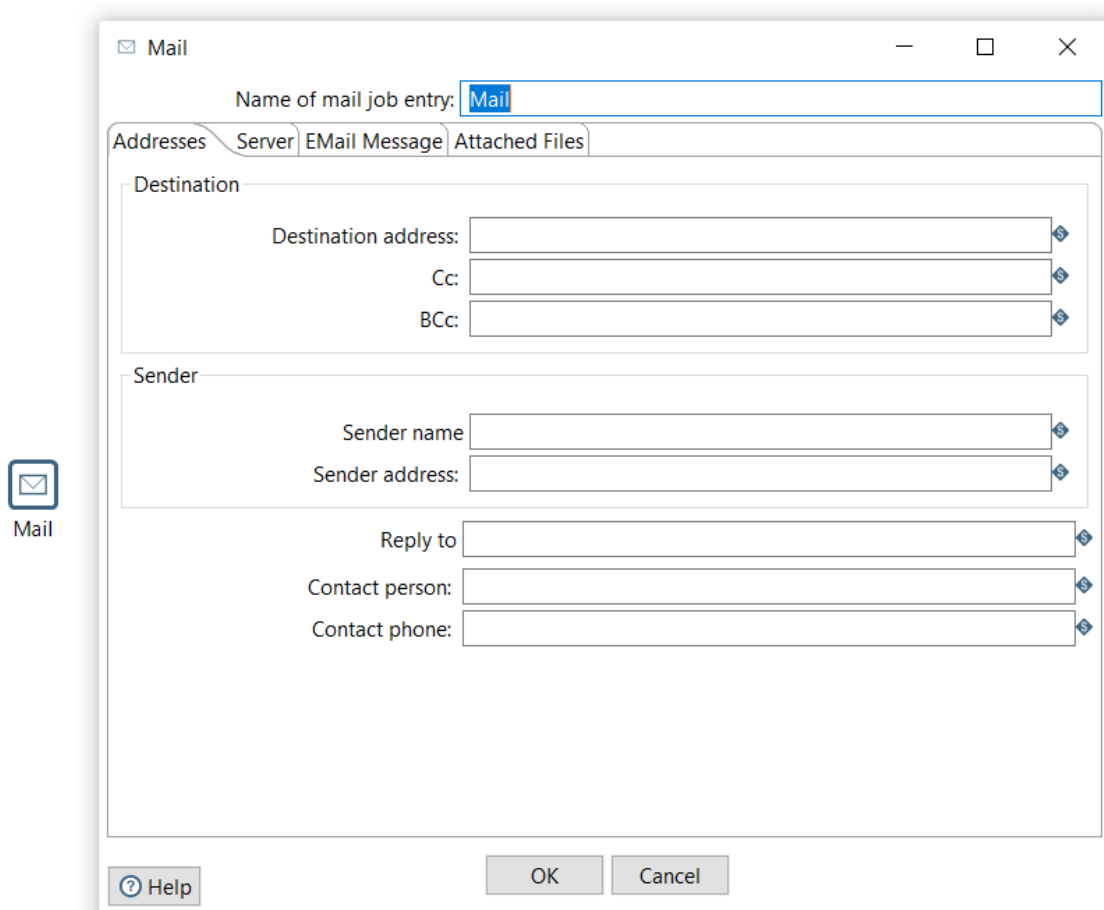
Sqoop import



Задание Sqoop Import позволяет импортировать данные из реляционной базы данных в распределенную файловую систему Hadoop (HDFS) с использованием Apache Sqoop. Это задание имеет два режима настройки:

- Быстрый режим предоставляет минимальные параметры, необходимые для успешного импорта Sqoop.
- Представление по умолчанию в расширенном режиме предоставляет параметры для лучшего контроля импорта Sqoop. Расширенный режим также имеет представление командной строки, которое позволяет вставить существующий аргумент командной строки Sqoop.

Mail



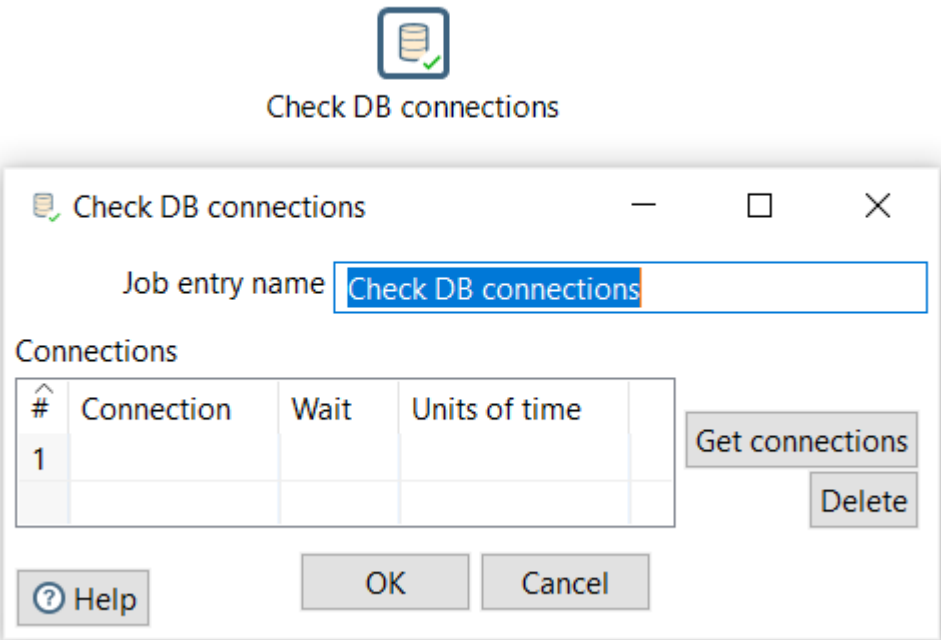
Используйте запись задания «Mail», чтобы отправить текстовое или HTML-сообщение электронной почтой с дополнительными вложенными файлами. Эта запись задания используется в конце после выполнения работы. Его можно использовать, чтобы объявить как о неудаче, так и об успешном выполнении задания. Например, не редкость в конце успешной загрузки отправить электронное письмо в список рассылки с уведомлением об успешной загрузке и включить файл журнала. Если есть ошибки, можно отправить электронное письмо, чтобы предупредить людей из списка рассылки.

Важно! При сбое задания во время выполнения работы сообщения по электронной почте не отправляются.

Для записи почтового задания требуется SMTP-сервер. Вы можете использовать аутентификацию и безопасность как часть соединения, но вы должны иметь учетные данные SMTP.

К сообщениям электронной почты можно прикреплять файлы, например журналы ошибок и обычные журналы. Кроме того, журналы можно заархивировать в единый архив для удобства.

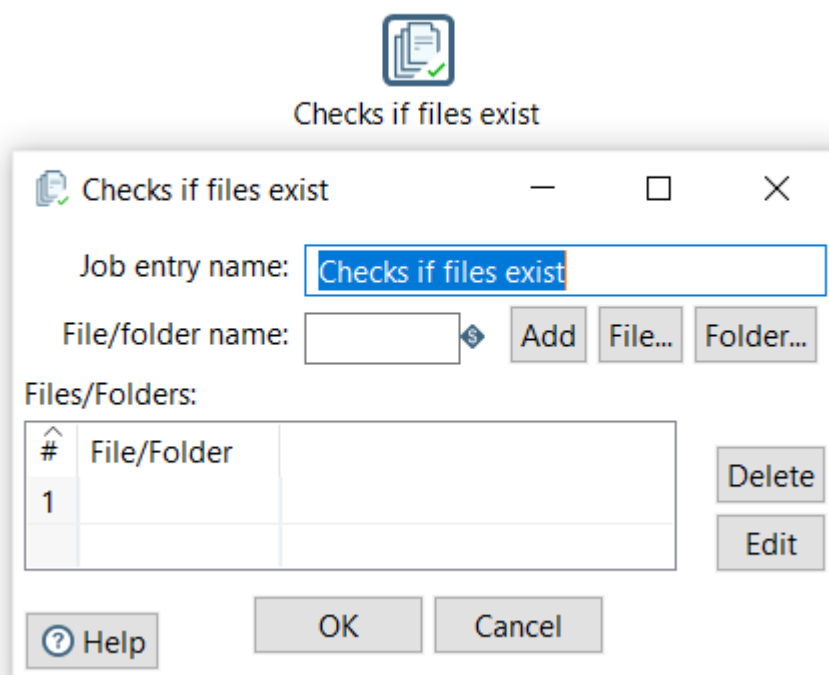
Check DB Connections



Эта запись задания позволяет проверить подключение к одной или нескольким базам данных.

Job entry name	Имя этой записи, как оно отобразится в преобразовании рабочего пространства.
Connection	Список подключений.
Wait	После открытия соединения подождите x (с, мин, час).
Units of Time	Укажите единицу измерения времени, в течение которого оставаться в соединении.
Get connections	Получить доступные подключения.

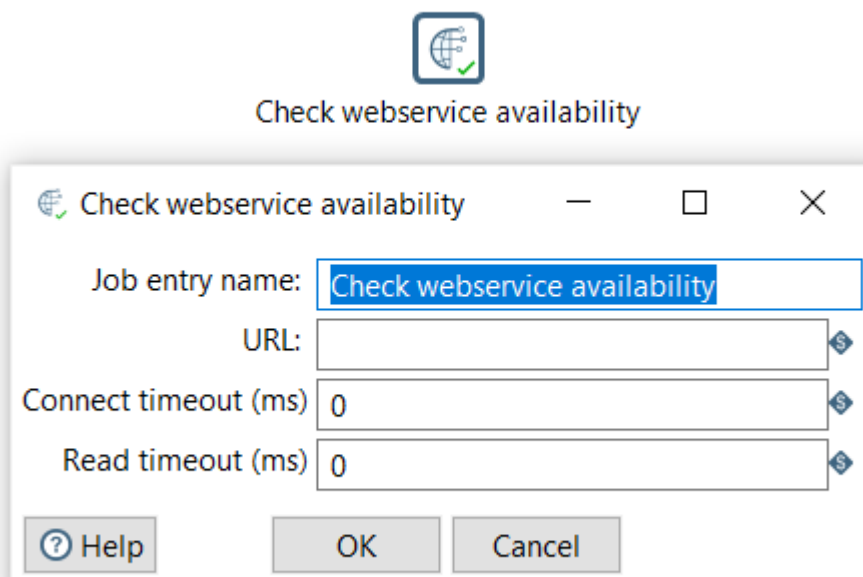
Check if files exist



Позволяет проверить существование файла.

Необходимо ввести имя файла, а также указать путь до самого файла.

Check Webservice Availability



Этот шаг проверяет, действителен ли данный URL-адрес, может ли он быть подключен и могут ли данные быть прочитаны.

Если он подключается в течение заданного тайм-аута и данные могут быть прочитаны, он возвращает «true», в противном случае — «false».

Дополнительную информацию о причине сбоя можно найти в журнале в виде записи журнала ошибок.

Job entry name	Имя записи о работе. Это имя должно быть уникальным в рамках одной работы.
URL field	Задаёт имя поля URL в потоке данных. URL-адрес проверяется для каждой строки, поступающей на этот шаг.
Connect timeout (ms)	Время ожидания соединения в миллисекундах. Значение зависит от качества обслуживания этого URL и опыта.
Read timeout (ms)	После подключения шаг пытается прочитать данные. Это значение задаёт время чтения в миллисекундах. Значение зависит от качества обслуживания этого URL и опыта.