

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

Практическая работа 2

Тема:

Работа в ETL-системе Talend.

Выполнила: Соколова М. С., группа: АДЭУ-201

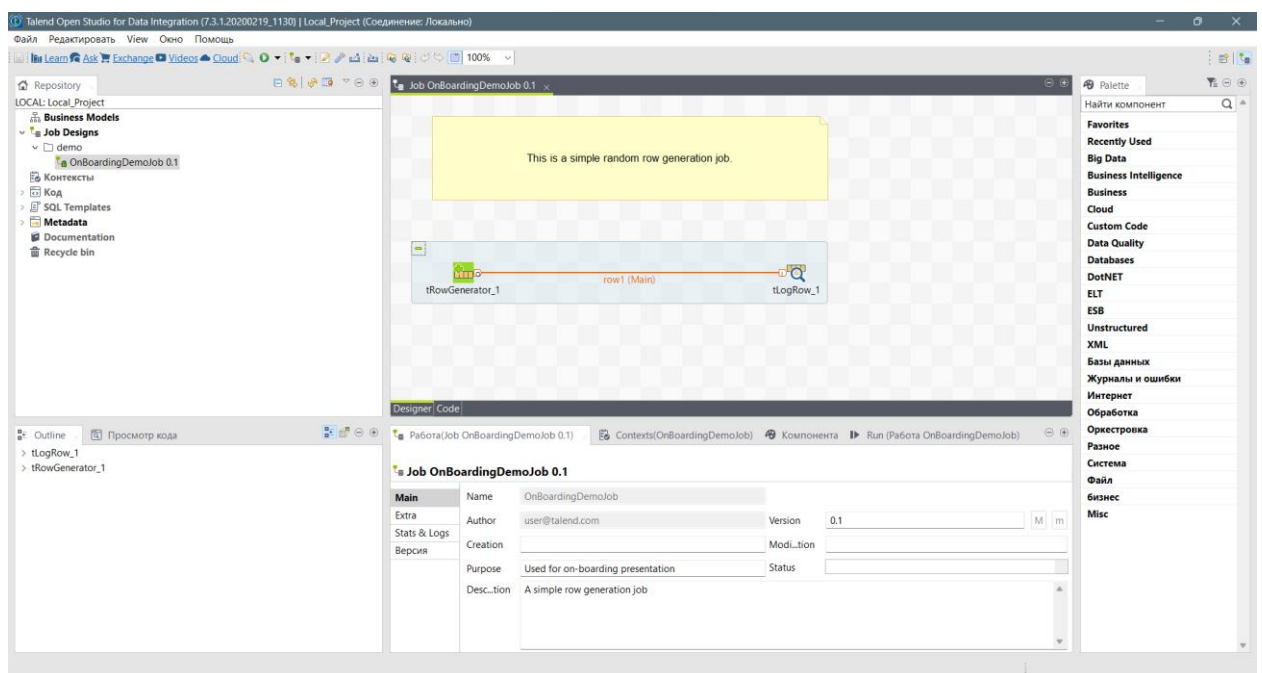
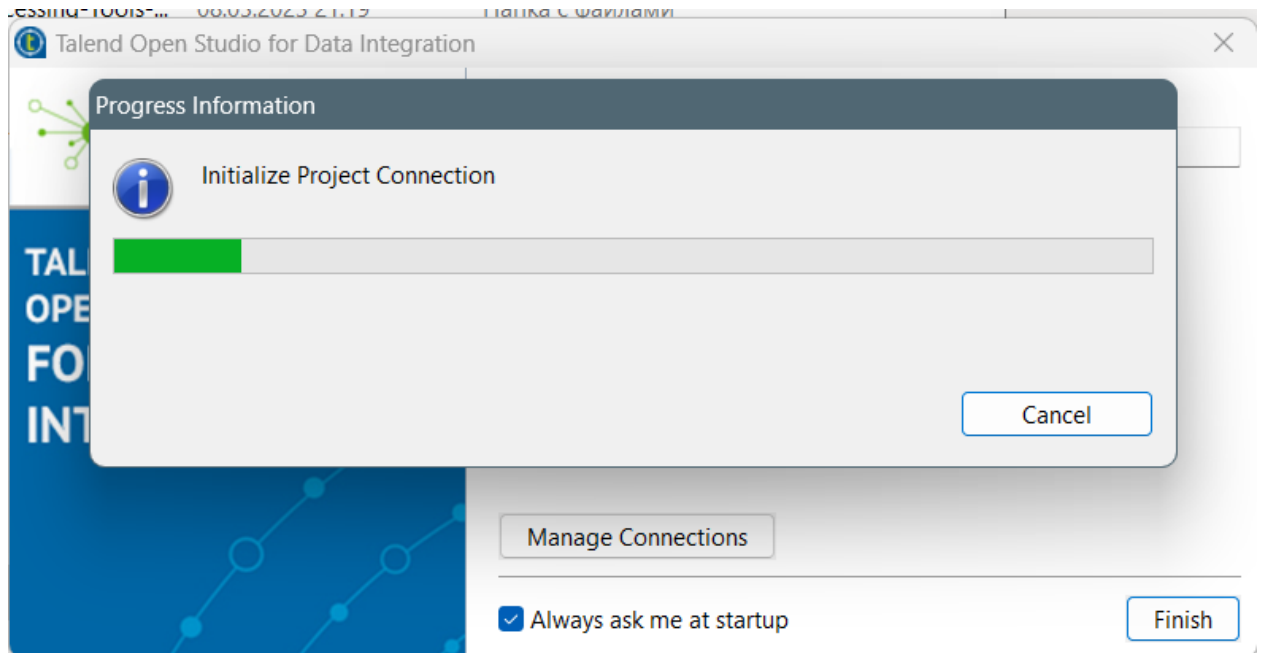
Преподаватель: Босенко Т. М.

Москва

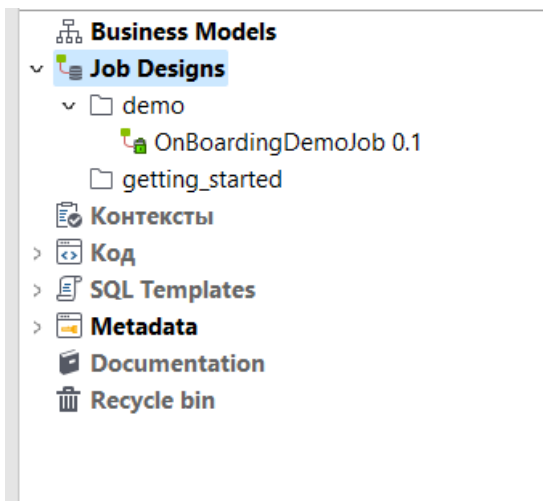
2023

Была скачена Java 8 [Download Java for Windows](#)

Создан новый проект.



В узле Job Designs создаем папку getting_started. Затем создаем работу movies.



В правой панели Palette выбираем Файл-Вход-tFileInputDelimited, переносим на нашу рабочую зону. Далее выбираем Журналы и ошибки – tLogRow. Переносим и соединяем нажатием на центр первого элемента и тянем на центр второго элемента.



В этом разделе создаем файл с разделителем.



Новый файл с разделителями

File - Step 1 of 4

Add a Metadata File on repository
Define the properties

Имя: movies

Purpose: Centralize metadata of movies.csv

Описание: Metadata of file movies.csv

Автор: user@talend.com

Locker:

Версия: 0.1 M m

Статус:

Path: Select

File - Step 2 of 4

Add a Metadata File on repository
Define the path of the file and the format settings

Настройки файла

Сервер: Localhost 127.0.0.1

Файл: C:/Users/marin/Downloads/movies.csv Обзор...

Формат: WINDOWS

File Viewer

movieID;title;releaseYear;url;directorID
315;Apt Pupil;1998;http://us.imdb.com/Title?Apt+Pupil+(1998);26
1294;Ayn Rand: A Sense of Life;1998;http://us.imdb.com/Title?Ayn+Rand%3A+A+Sense+of+Life+(1997);123
1679;B. Monkey;1998;http://us.imdb.com/M/title-exact?B%2E+Monkey+(1998);124
1649;Big One, The;1998;http://us.imdb.com/Title?Big+One,+The+(1997);122
362;Blues Brothers 2000;1998;http://us.imdb.com/M/title-exact?Blues+Brothers+2000+(1998);86
1645;Butcher Boy, The;1998;http://us.imdb.com/M/title-exact?imdb-title-118804;134
1650;Butcher Boy, The;1998;http://us.imdb.com/M/title-exact?imdb-title-118804;134
1234;Chairman of the Board;1998;http://us.imdb.com/Title?Chairman+of+the+Board+(1998);6
1654;Chairman of the Board;1998;http://us.imdb.com/Title?Chairman+of+the+Board+(1998);6
918;City of Angels;1998;http://us.imdb.com/Title?City+of+Angels+(1998);22
909;Dangerous Beauty;1998;http://us.imdb.com/M/title-exact?imdb-title-118892;113

< Back Next > Finish Cancel

File - Step 3 of 4

Add a Metadata File on repository

Define the setting of the parse job

Настройки файла

Кодировка

US-ASCII

Разделитель полей

Corresponding Character

","

Разделитель строк

Corresponding Character

"\n"

Escape Char Settings

CSV

Разделенный

Escape Char

Пустой

Text Enclosure

Пустой

Split row before field

Rows To Skip

Определите следующие параметры, если какие-либо строки должны быть пропущены

Header

☒

2

Footer

☐

Skip empty row

Ограничение строк

Если количество строк должно быть ограничено, определите это количество

Ограничение

☐

Предпросмотр

Выход

☒ Установить строку заголовка в качестве имен столбцов

Refresh Preview

Колонка 0	Колонка 1	Колонка 2	Колонка 3
movieID	title	releaseYear	url
315	Apt Pupil	1998	http://us.imdb.com/Title?Apt+Pupil+(1998)
1294	Ayn Rand: A Sense of Life	1998	http://us.imdb.com/Title?Ayn+Rand%3A+A+Sense+of+Life+(1997)
1679	B. Monkey	1998	http://us.imdb.com/M/title-exact?B%2E+Monkey+(1998)

Экспортировать как контекст

Восстановить контекст

< Back

Next >

Finish

Cancel

File - Step 4 of 4

Add a Schema on repository

Define the Schema

Имя

metadata

Комментарий

Схема

Click to update schema preview

Guess

Описание схемы

Колонка	К...	Тип	<input checked="" type="checkbox"/> N.	Date Pattern (Ctrl...	Длина	Precision	Default	Коммент...
movieID	<input type="checkbox"/>	Integer	<input checked="" type="checkbox"/>		4	0		
title	<input type="checkbox"/>	String	<input checked="" type="checkbox"/>		29	0		
releaseYear	<input type="checkbox"/>	Integer	<input checked="" type="checkbox"/>		4	0		
url	<input type="checkbox"/>	String	<input checked="" type="checkbox"/>		66	0		
directorID	<input type="checkbox"/>	String	<input checked="" type="checkbox"/>		3	0		

+

-

↕

↕

📄

📄

🔍

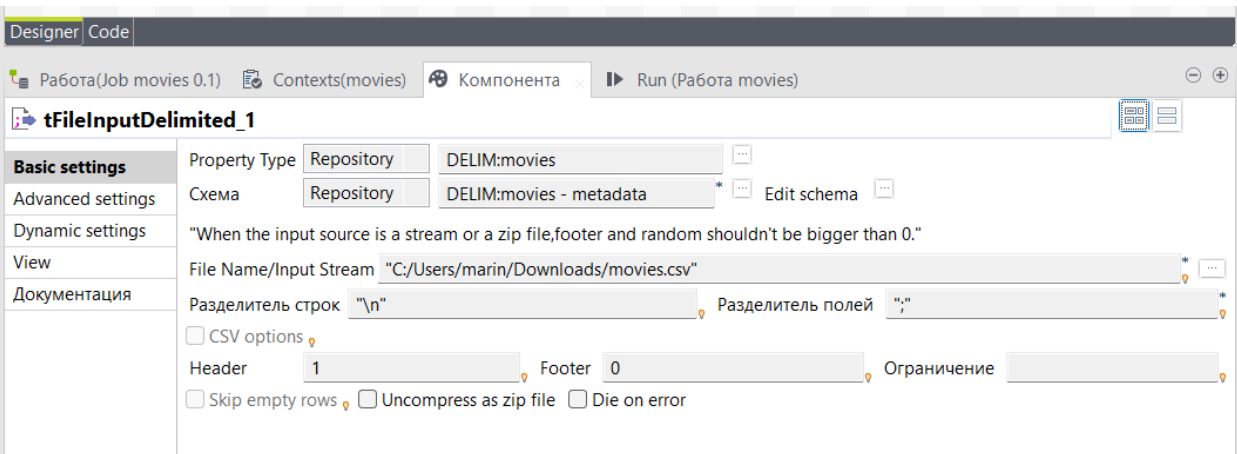
🔍

< Back

Next >

Finish

Cancel



The screenshot displays the JetBrains DataGrip IDE interface. At the top, a data flow diagram shows a connection from 'tFileInputDelimited_1' to 'tLogRow_1'. The flow is labeled with '1455 rows in 0,55s' and '2626,35 rows/s row1 (Main)'. Below the diagram, the 'Designer' tab is active, showing a project named 'Работа (Job movies 0.1)'. The 'Run (Работа movies)' button is highlighted. On the right, the 'Default' tab is selected, showing a table with columns 'Имя' and 'Значение'. The table is empty. Below the table, the 'Execution' section is visible, showing a console window with the output of the 'tLogRow_1' component. The console output includes a header row with 'key' and 'value', followed by a row with 'movieID' and '224', and another row with 'title' and 'Ridicule'. The 'releaseYear' is listed as '1996'. The console window also shows a 'Line limit' of 100 and a 'Wrap' option.

File - Step 1 of 4



Add a Metadata File on repository
Define the properties

Имя	<input type="text" value="directors"/>		
Purpose	<input type="text" value="Centralize metadata of directors info"/>		
Описание	<input type="text" value="Metadata of directors dataset"/>		
Автор	<input type="text" value="user@talend.com"/>		
Locker	<input type="text"/>		
Версия	<input type="text" value="0.1"/>	<input type="button" value="M"/>	<input type="button" value="m"/>
Статус	<input type="text"/>		
Path	<input type="text"/>		<input type="button" value="Select"/>

File - Step 2 of 4



Add a Metadata File on repository
Define the path of the file and the format settings

Настройки файла

Сервер	<input type="text" value="localhost 127.0.0.1"/>	<input type="button" value="v"/>
Файл	<input type="text" value="C:/Users/marin/Downloads/directors.txt"/>	<input type="button" value="Обзор..."/>
Формат	<input type="text" value="WINDOWS"/> <input type="button" value="v"/>	

File Viewer


1, Gregg Araki
2, P.J. Hogan
3, Alan Rudolph
4, Alex Proyas
5, Alex Sichel
6, Alex Zamm
7, Alfonso Cuarón
8, Alfred Hitchcock
9, Allison Anders
10, Andrew Davis
11, Andrew Niccol
12, Antoine Fuqua

< Back

Next >

Finish

Cancel

File - Step 3 of 4


Add a Metadata File on repository
Define the setting of the parse job

Настройки файла

Кодировка

UTF-8

Разделитель полей

Corresponding Character " "

Разделитель строк

Corresponding Character "\n"

Escape Char Settings

☐ CSV
☒ Разделенный

Escape Char

Пустой

Text Enclosure

Пустой

☐ Split row before field

Rows To Skip

Определите следующие параметры, если какие-либо строки должны быть пропущены

Header
☐

Footer
☐

☐ Skip empty row

Ограничение строк

Если количество строк должно быть ограничено, определите это количество

Ограничение
☐

Предпросмотр

Выход

☐ Установить строку заголовка в качестве имен столбцов

Refresh Preview

Колонка 0	
1, Gregg Araki	
2, P.J. Hogan	
3, Alan Rudolph	
4, Alex Proyas	

Экспортировать как контекст


Восстановить контекст

< Back

Next >

Finish

Cancel

File - Step 4 of 4


Add a Schema on repository
Define the Schema

Имя

metadata

Комментарий

Схема

Click to update schema preview

Guess

Описание схемы

Колонка	К...	Тип	<input checked="" type="checkbox"/> N..	Date Pattern (Ctrl...	Длина	Precision	Default	Коммент...
Column0	<input type="checkbox"/>	Integer	<input checked="" type="checkbox"/>		2	0		
Column1	<input type="checkbox"/>	String	<input checked="" type="checkbox"/>		20	0		

+

×

↑

↓

📄

🔍

🔄

🗑️

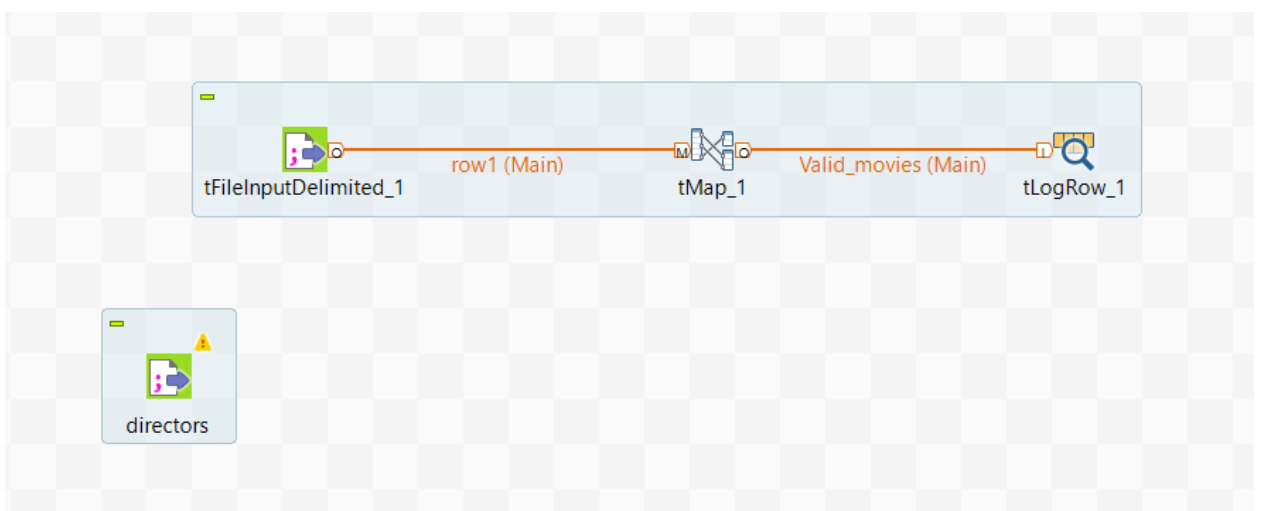
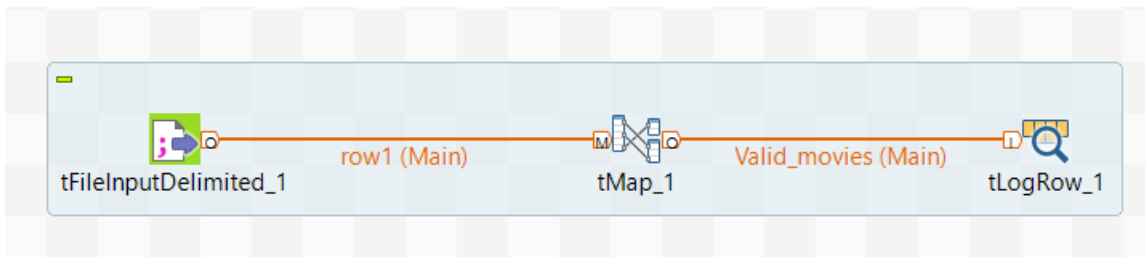
< Back

Next >

Finish

Cancel

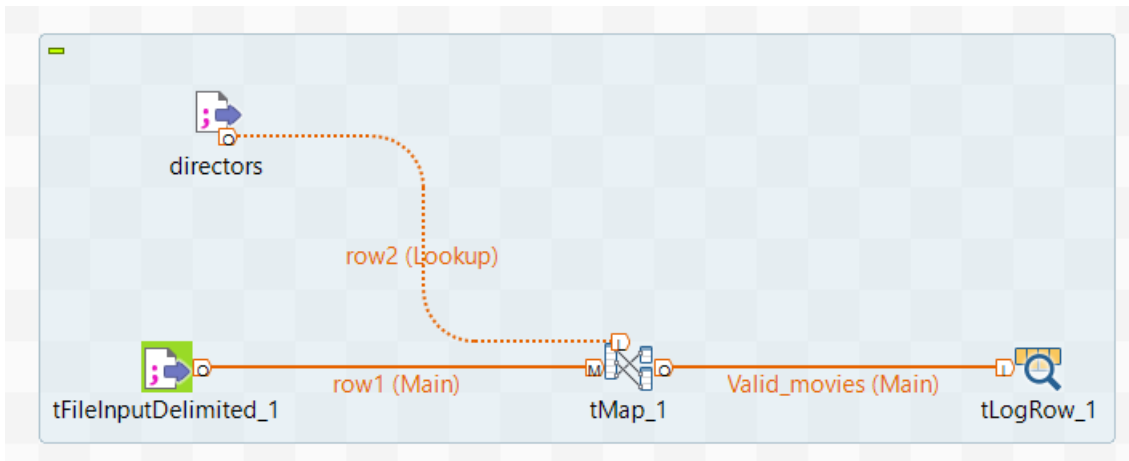
- File delimited
 - directors 0.1
 - metadata
 - Столбцы(2)
 - directorID
 - directorName



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

directors(tFileInputDelimited_2)

Basic settings	Property Type	Repository	DELIM:directors
Advanced settings	Схема	Repository	DELIM:directors - metadata Edit schema
Dynamic settings	"When the input source is a stream or a zip file,footer and random shouldn't be bigger than 0."		
View	File Name/Input Stream "C:/Users/marin/Downloads/directors.txt"		
Документация	Разделитель строк "\n" Разделитель полей ";"		
	<input type="checkbox"/> CSV options Header 0 Footer 0 Ограничение <input type="checkbox"/> Skip empty rows <input type="checkbox"/> Uncompress as zip file <input type="checkbox"/> Die on error		



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

directors(tFileInputDelimited_2)

Basic settings

Advanced settings

Dynamic settings

View

Документация

☐ Advanced separator(for number)

☐ Extract lines at random Кодировка UTF-8

☒ Trim all column ☐ Проверить каждый ряд на соответствие схеме ☐ Check date

☐ Split row before field

☐ Permit hexadecimal (0xNNN) or octal (0NNNN) for numeric types - it will act the opposite for Byte

☐ tStatCatcher Statistics

Talend Open Studio for Data Integration - tMap - tMap_1

Find: Var

Auto map!

row1

Column	movieID	title	releaseYear	url	directorID
Expr. key					

row2

Expr. key	Column
	directorID
	directorName

Valid_movies

Выражение	Column
row1.movieID	movieID
row1.title	title
row1.releaseYear	releaseYear
row1.url	url
row1.directorID	directorID

Schema editor Редактор выражений

row1

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"/>		29	0		
releaseYear		Integer	<input checked="" type="checkbox"/>		4	0		
url		String	<input checked="" type="checkbox"/>		66	0		
directorID		String	<input checked="" type="checkbox"/>		3	0		

Valid_movies

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"/>		29	0		
releaseYear		Integer	<input checked="" type="checkbox"/>		4	0		
url		String	<input checked="" type="checkbox"/>		66	0		
directorID		String	<input checked="" type="checkbox"/>		3	0		

row1

Column	movieID	title	releaseYear	url	directorID
Expr. key					

row2

Expr. key	Column
row1.directorID	directorID
	directorName

Valid_movies

Выражение	Column
row1.movieID	movieID
row1.title	title
row1.releaseYear	releaseYear
row1.url	url
row1.directorID	directorID

row1

Column
movieID
title
releaseYear
url
directorID

row2

Property	Value
Lookup Model	Load once
Match Model	Unique match
Join Model	Left Outer Join
Store temp data	false

Expr. key	Column
row1.directorID	directorID
	directorName

Var

Options

Inner Join
Left Outer Join

OKCancel

row1

Column
movieID
title
releaseYear
url
directorID

row2

Property	Value
Lookup Model	Load once
Match Model	Unique match
Join Model	Inner join
Store temp data	false

Expr. key	Column
row1.directorID	directorID
	directorName

Var

Valid_movies

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

directors

204 rows in 0,03s
6375 rows/s
row2 (lookup)

tFileInputDelimited_1

1455 rows in 0,52s
2825,24 rows/s
row1 (Main)

tMap_1

255 rows in 0,52s
494,19 rows/s
valid_movies (Main)

tLogRow_1

DesignerCode

Pa60ra(Job write_movies_to_db 0.1)КомпонентаRun (Pa60ra write_movies_to_db)

Pa60ra write_movies_to_db

Basic Run
Debug Run
Advanced settings
Target Exec
Memory Run

Execution

RunKillОчистить

title	Flubber
directedBy	Les Mayfield
releaseYear	1997
url	http://us.imdb.com/M/title-exact?imdb-title-119137

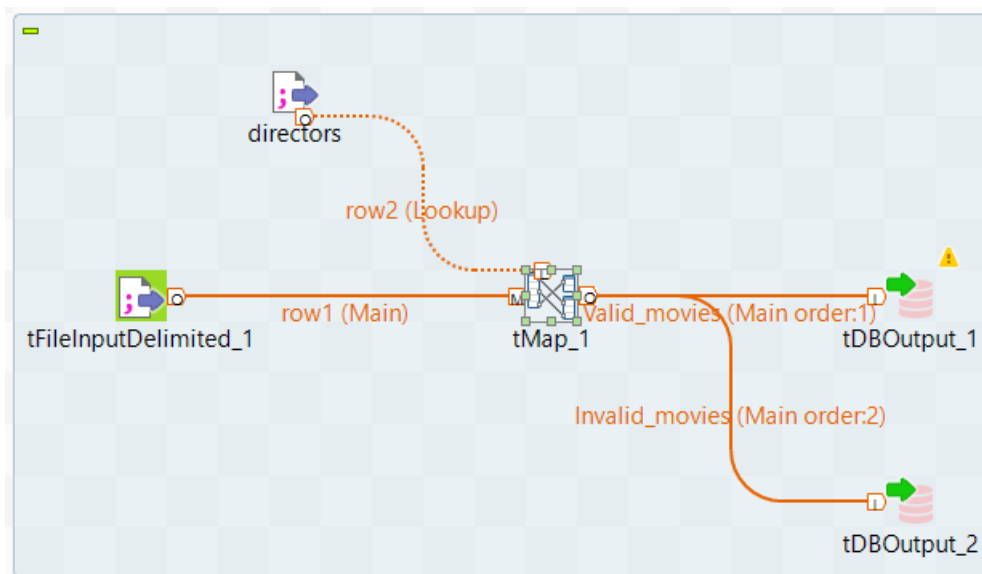
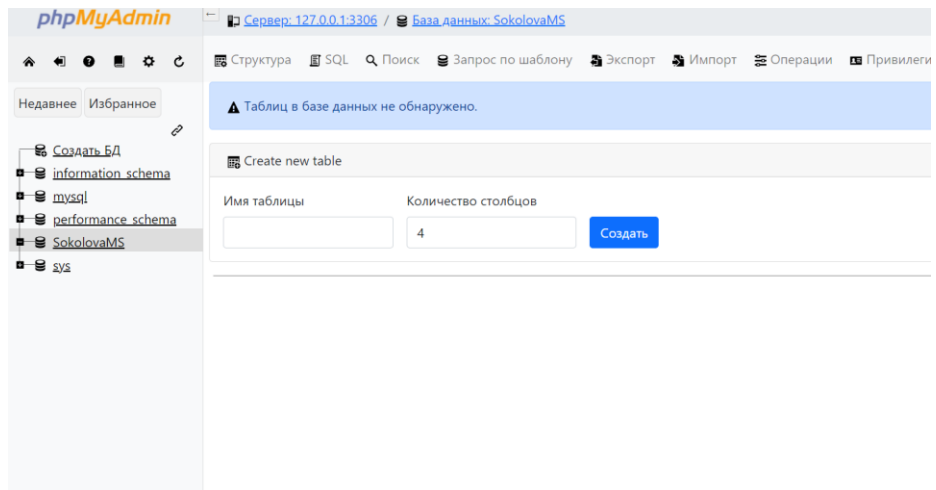
#125. tLogRow_1

key	value
movieID	269
title	Full Monty, The

☐ Line limit 100☐ Wrap

Default

Имя	Значение
-----	----------



row1

Column	Value
movieID	
title	
releaseYear	
url	
directorID	

row2

Property	Value
Lookup Model	Load once
Match Model	Unique match
Join Model	Inner join
Store temp data	false

Expr. key: row1.directorID

Column: directorID, directorName

Valid_movies

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

Invalid_movies

Property	Value
Catch output reject	false
Catch lookup inner join reject	true
Schema Type	Built-In

Выражение: row1.movieID, row1.title

Column: movieID, title

tDBOutput_1(MySQL)

Basic settings

Database: MySQL Apply

Property Type: Built-In

Версия БД: Mysql 8

☐ Использовать существующее соединение

Хост: "localhost" Порт: "3306"

Database: "SokolovaMS"

Имя пользователя: "root" Пароль: "*****"

Таблица: "valid_movies"

Action on table: Drop and create table Действие над данными: Вставить

Схема: Built-In Edit schema Sync columns

Data source

This option only applies when deploying and running in the Talend Runtime

Создать БД

information_schema

mysql

performance_schema

SokolovaMS

Новая

invalid_movies

valid_movies

sys

Филтры

Содержит слово:

Таблица	Действие	Строки	Тип	Сравнение	Размер	Фрагментировано
<input type="checkbox"/> invalid_movies	☆	1 200	InnoDB	utf8mb3_general_ci	96.0 КиБ	-
<input type="checkbox"/> valid_movies	☆	255	InnoDB	utf8mb3_general_ci	16.0 КиБ	-
2 таблицы	Всего	1 455	InnoDB	utf8mb3_general_ci	112.0 КиБ	0 Байт

↑

☐ Отметить все

С отмеченными:

Печать

Словарь данных

Create new table

```

graph LR
    A[tFileInputDelimited_1] -- "1455 rows in 0,1s  
14405,94 rows/s  
row1 (Main)" --> B[tMap_1]
    B -- "204 rows in 0,04s  
4533,33 rows/s  
row2 (lookup)" --> C[tDBOutput_1]
    B -- "255 rows in 1,14s  
224,47 rows/s  
Valid_movies (Main order:1)" --> C
    B -- "1200 rows in 0,15s  
7792,21 rows/s  
Invalid_movies (Main order:2)" --> D[tDBOutput_2]
  
```

designer Code

Работа(Job write_movies_to_db 0.1)

Компонента

Run (Работа write_movies_to_db)

Работа write_movies_to_db

Basic Run

Debug Run

Advanced settings

Target Exec

Memory Run

Execution

Run

Kill

Очистить

Starting job write_movies_to_db at 17:59 13/04/2023.

[statistics] connecting to socket on port 3493

[statistics] connected

[statistics] disconnected

Job write_movies_to_db ended at 17:59 13/04/2023. [exit code = 0]