

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 обработки и хранения больших данных**

Отчет

Тема:

Работа в ETL-системе Talend

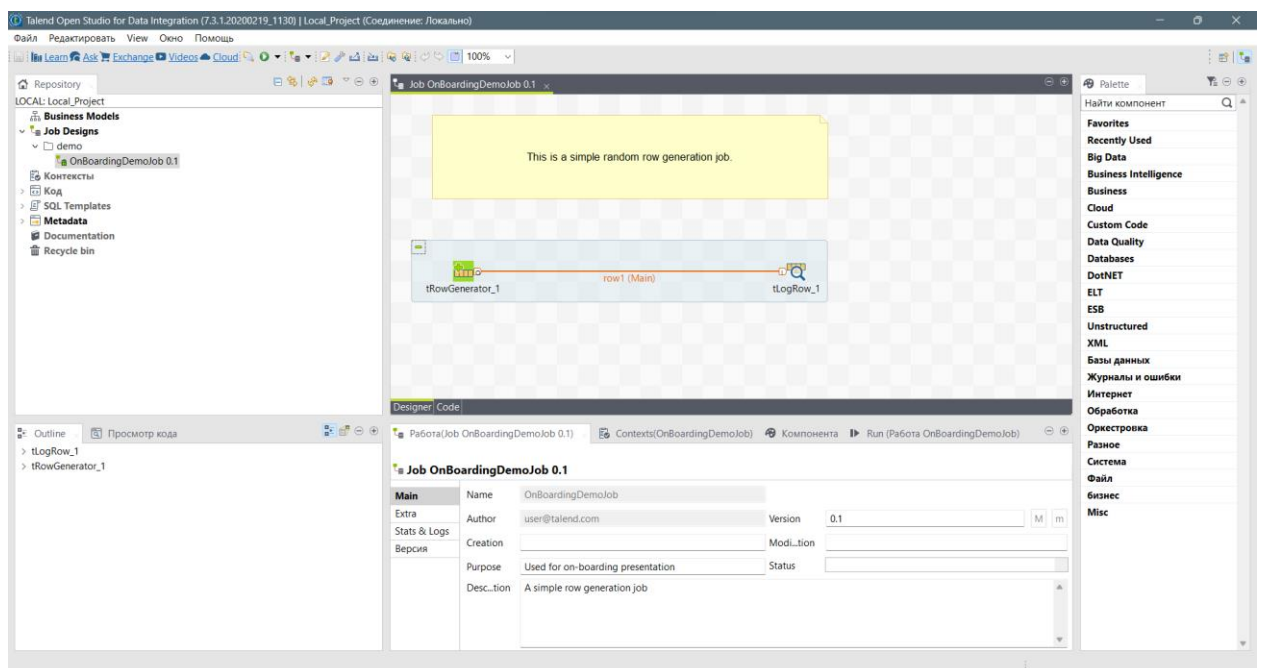
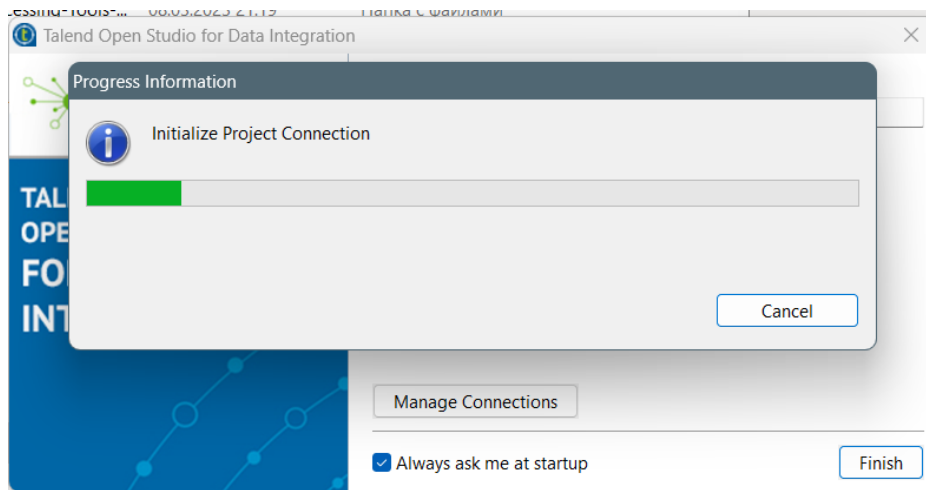
Выполнила: Олифир А. А., группа: АДЭУ-201

Преподаватель: Босенко Т. М.

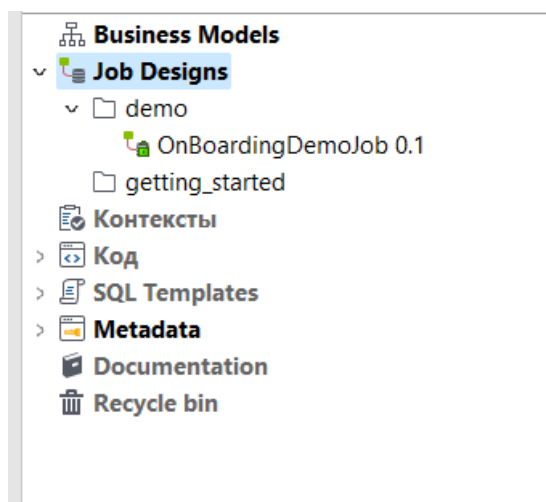
Москва

2023

Скачивание Java 8 [Download Java for Windows](#) и создание нового проекта.



В узле Job Designs создаем папку getting\_started. Затем создаем работу movies.



**New job**

⚠ Empty purpose is discouraged.

Имя:

Purpose:

Описание:

Автор:

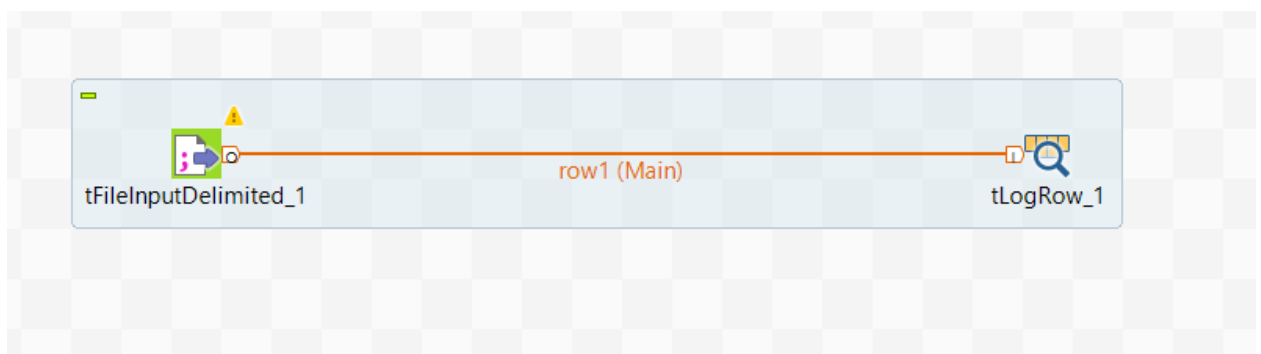
Locker:

Версия:

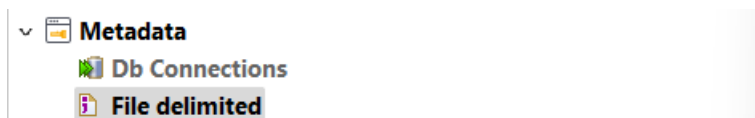
Статус:

Path:

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:

Purpose:

Описание:

Автор:

Locker:

Версия:

Статус:

Path:

## 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/aalin/Downloads/data\_for\_ex\_02/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);  
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

&lt; Back

Next &gt;

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 ☒ 2Footer ☐☐ 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

&lt; Back

Next &gt;

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	K...	Тип	✓ 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

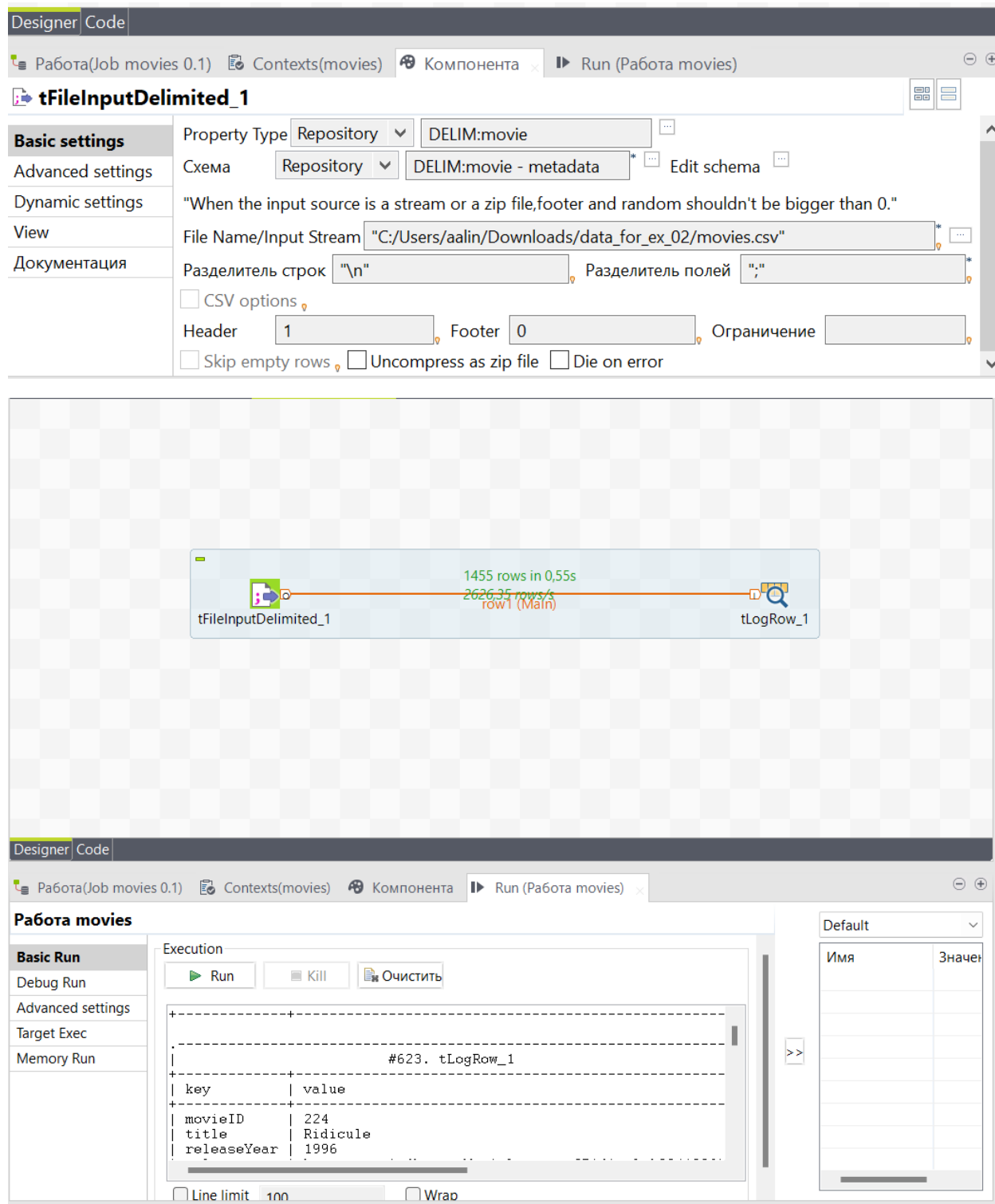
Next >

Finish


Cancel

- Metadata
  - Db Connections
    - File delimited
      - movies 0.1
        - metadata
          - Столбцы(5)
            - directorID
            - movieID
            - releaseYear
            - title
            - url

### Настройка и выполнение работы:





## Подготовка файла метаданных directors

**File - Step 1 of 4**

Add a Metadata File on repository  
Define the properties

Имя	directors		
Purpose	Centralize metadata of directors info		
Описание	Metadata of directors dataset		
Автор	user@talend.com		
Locker			
Версия	0.1	M	m
Статус			
Path			Select

 Новый файл с разделителями

**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/aalin/Downloads/data_for_ex_02/directors.txt	Обзор...	
Формат	WINDOWS ▼		

**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

< 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 долж

Header

Footer

☐ Skip empty row

Ограничение строк

Если количество строк должно быть ограничено, определите это к

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

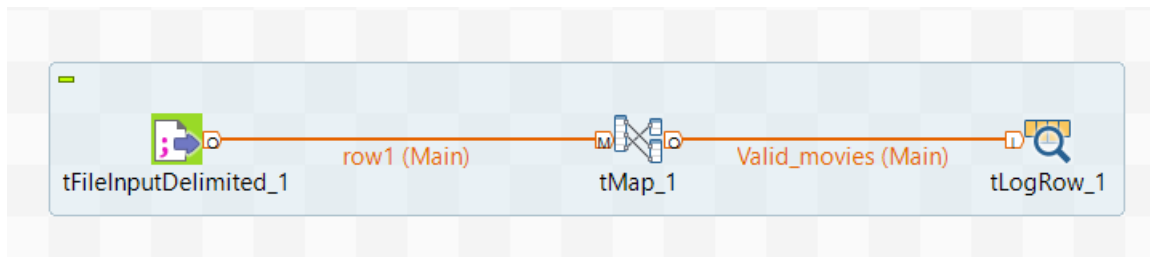
Next >

Finish

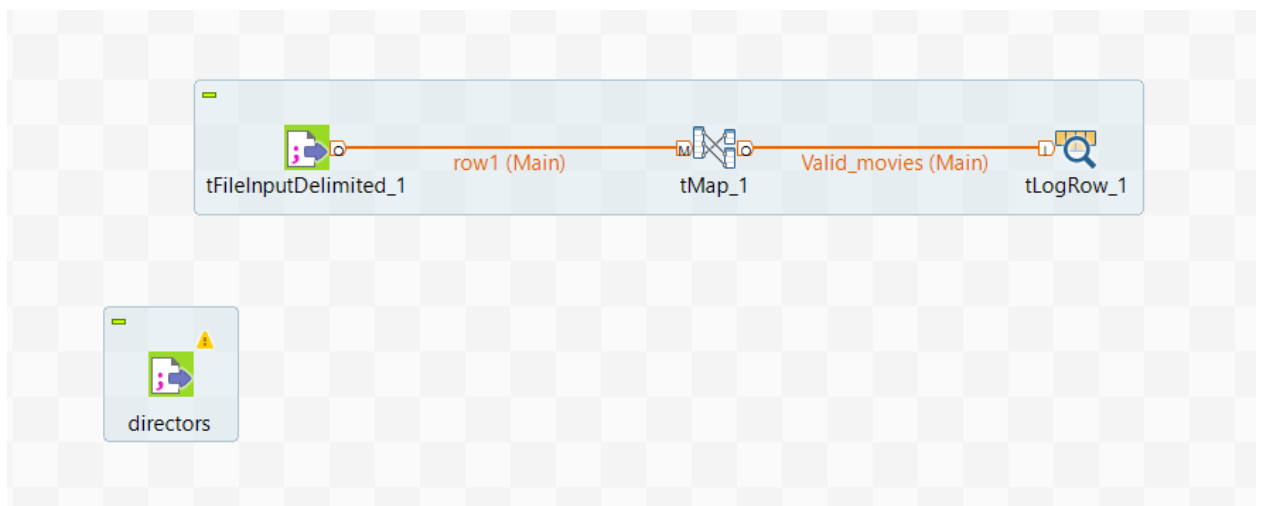
Cancel



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



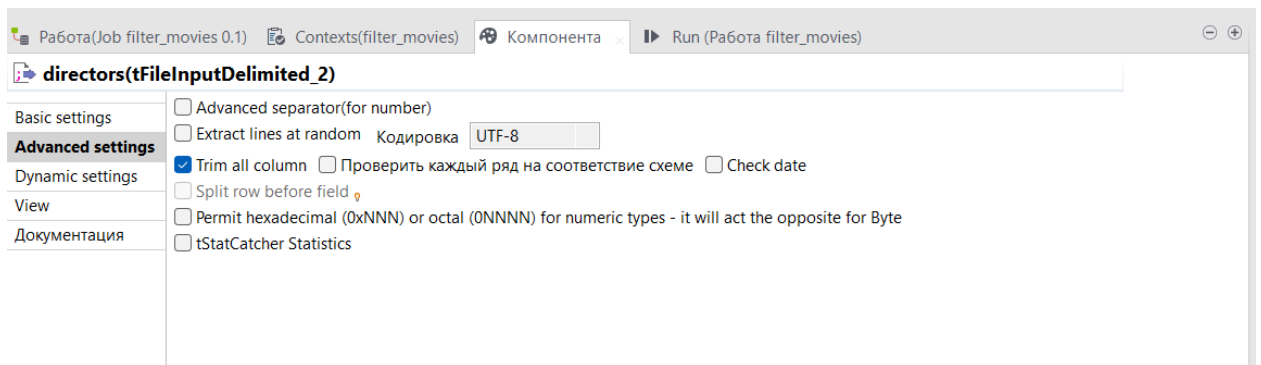
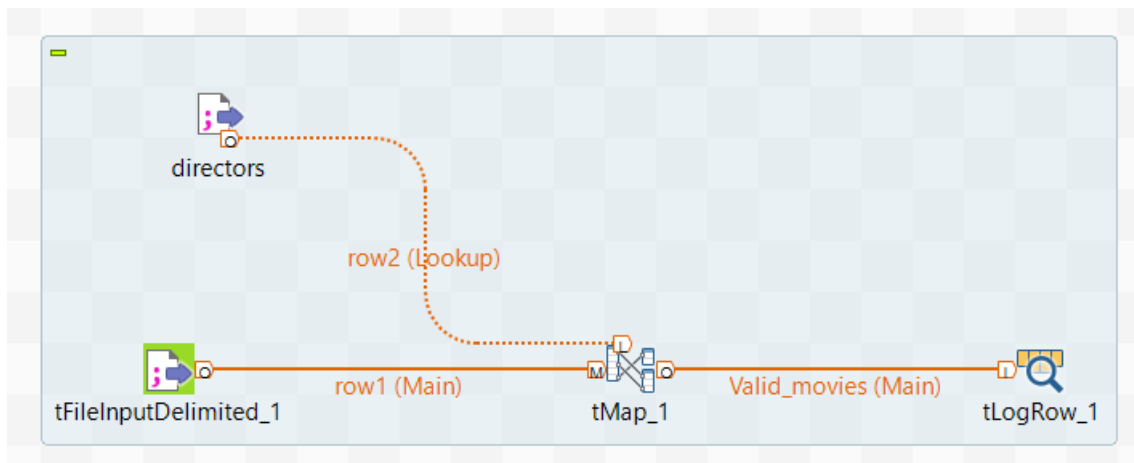
Добавление компонента



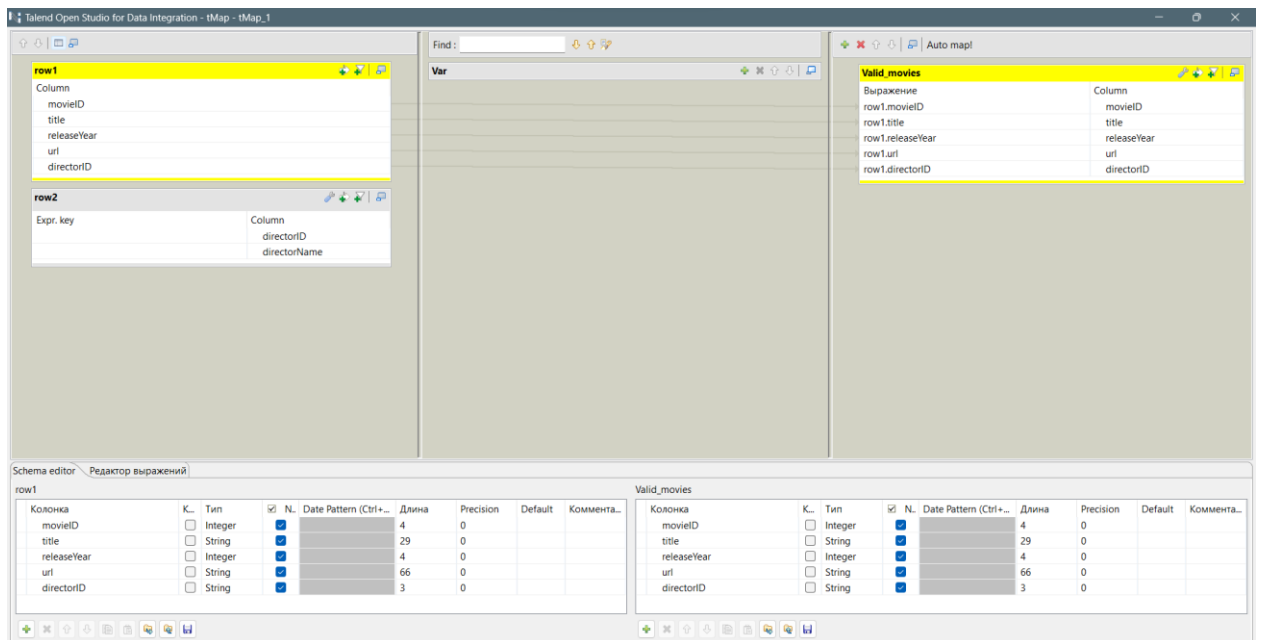
Работа(Job filter\_movies 0.1) Contexts(filter\_movies) Компонента Run (Работа filter\_movies)

**directors(tFileInputDelimited\_2)**

Basic settings	Property Type	Repository	DELIM:directors
Advanced settings	Схема	Repository	DELIM:directors - metadata
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
	Разделитель полей	","	
	Header	0	Footer
	Footer	0	Ограничение
	<input type="checkbox"/> CSV options <input type="checkbox"/> Skip empty rows <input type="checkbox"/> Uncompress as zip file <input type="checkbox"/> Die on error		



## Конфигурация компонентов и создание проекта



row1

Column	
movieID	
title	
releaseYear	
url	
directorID	

row2

Expr. key	Column
row1.directorID	directorID
	directorName

Var

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

Schema editor

Редактор выражений

row1

Колонка	K...	Тип	✓ N	Date Pattern (C...	Длина	Precision	Default	Коммен...
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		

valid\_movies

Колонка	K...	Тип	✓ N	Date Pattern (C...	Длина	Precision	Default	Коммен...
movieID	<input type="checkbox"/>	Integer	<input checked="" type="checkbox"/>		4	0		
title	<input type="checkbox"/>	String	<input checked="" type="checkbox"/>		29	0		
directedBy	<input type="checkbox"/>	String	<input checked="" type="checkbox"/>		20			
releaseYear	<input type="checkbox"/>	Integer	<input checked="" type="checkbox"/>		4	0		

Job movies 0.1

Designers Code

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

### Работа movies

**Basic Run**

Debug Run

Advanced settings

Target Exec

Memory Run

Execution

Run Kill Очистить

```

172|Empire Strikes Back, The|Irvin Kershner|19:
192|Raging Bull|Martin Scorsese|1980|http://us
200|Shining, The|Stanley Kubrick|1980|http://u:
[statistics] disconnected

Job movies ended at 15:15 09/04/2023. [exit co

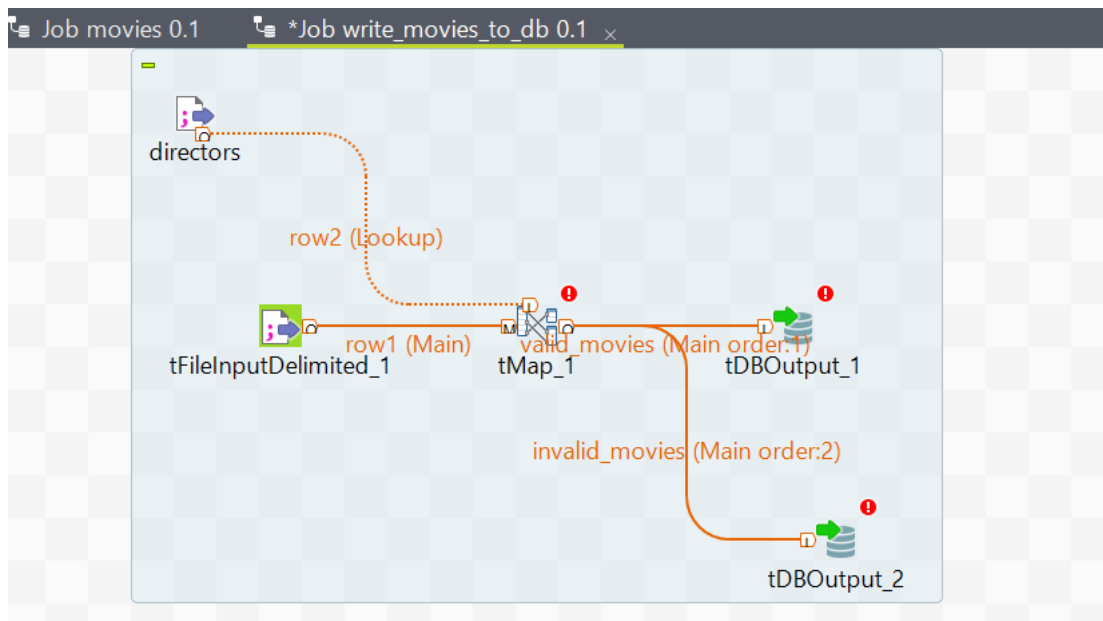
```

Default	
Имя	Значение

Добавление компонентов в проект

Job movies 0.1 \*Job write\_movies\_to\_db 0.1

tFileInputDelimited\_1 row1 (Main) tDBOutput\_1

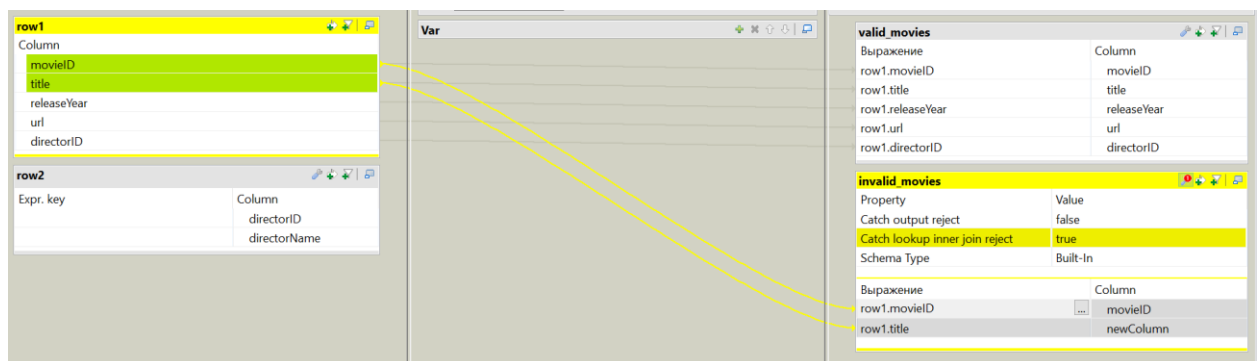


Конфигурация сопоставлений для отклоненных данных

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

invalid_movies	
Выражение	Column



Конфигурация базы данных

Database MySQL ▼ Apply

Property Type Built-In ▼ 📁

Версия БД Mysql 5 ▼

☐ Использовать существующее соединение

Хост "localhost" \* Порт "3306" \*

Database "gettingstarted" \*

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

## Результаты

movieID	title	directedBy	releaseYear	url
315	Apt Pupil	Bryan Singer	1998	<a href="http://us.imdb.com/Title?Apt+Pupil+(1998)">http://us.imdb.com/Title?Apt+Pupil+(1998)</a>
1294	Ayn Rand: A Sense of Life	Michael Paxton	1998	<a href="http://us.imdb.com/Title?Ayn+Rand%3A+A+Sense+of+Li...">http://us.imdb.com/Title?Ayn+Rand%3A+A+Sense+of+Li...</a>
1679	B. Monkey	Michael Radford	1998	<a href="http://us.imdb.com/M/title-exact?B%2E+Monkey+(1998...">http://us.imdb.com/M/title-exact?B%2E+Monkey+(1998...</a>
1649	Big One, The	Michael Moore	1998	<a href="http://us.imdb.com/Title?Big+One,+The+(1997)">http://us.imdb.com/Title?Big+One,+The+(1997)</a>
362	Blues Brothers 2000	John Landis	1998	<a href="http://us.imdb.com/M/title-exact?Blues+Brothers+20...">http://us.imdb.com/M/title-exact?Blues+Brothers+20...</a>
1645	Butcher Boy, The	Neil Jordan	1998	<a href="http://us.imdb.com/M/title-exact?imdb-title-118804">http://us.imdb.com/M/title-exact?imdb-title-118804</a>
1650	Butcher Boy, The	Neil Jordan	1998	<a href="http://us.imdb.com/M/title-exact?imdb-title-118804">http://us.imdb.com/M/title-exact?imdb-title-118804</a>
1234	Chairman of the Board	Alex Zamm	1998	<a href="http://us.imdb.com/Title?Chairman+of+the+Board+(19...">http://us.imdb.com/Title?Chairman+of+the+Board+(19...</a>
1654	Chairman of the Board	Alex Zamm	1998	<a href="http://us.imdb.com/Title?Chairman+of+the+Board+(19...">http://us.imdb.com/Title?Chairman+of+the+Board+(19...</a>
918	City of Angels	Brad Silberling	1998	<a href="http://us.imdb.com/Title?City+of+Angels+(1998)">http://us.imdb.com/Title?City+of+Angels+(1998)</a>
909	Dangerous Beauty	Marshall Herskovitz	1998	<a href="http://us.imdb.com/M/title-exact?imdb-title-118892">http://us.imdb.com/M/title-exact?imdb-title-118892</a>