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

Башкатова Анна Денисовна

ОТЧЕТ

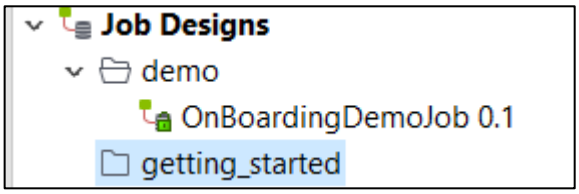
по дисциплине «Инструменты для хранения и обработки больших данных»
Тема: «02 Работа в ETL-системе Talend»

Направление подготовки (специальность) 38.03.05 – бизнес-информатика
Направленность (профиль) образовательной программы «Аналитика данных и эффективное управление»

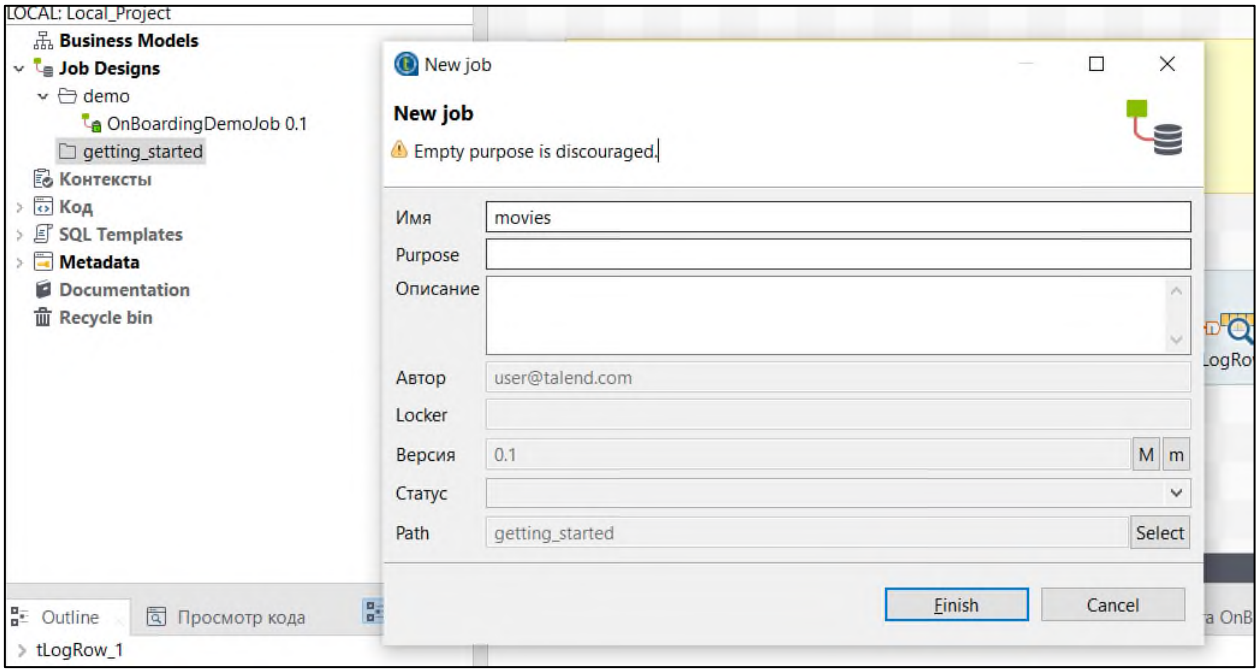
Курс обучения: 3
Форма обучения: очная
Руководитель: Босенко Т. М.

Москва
2023

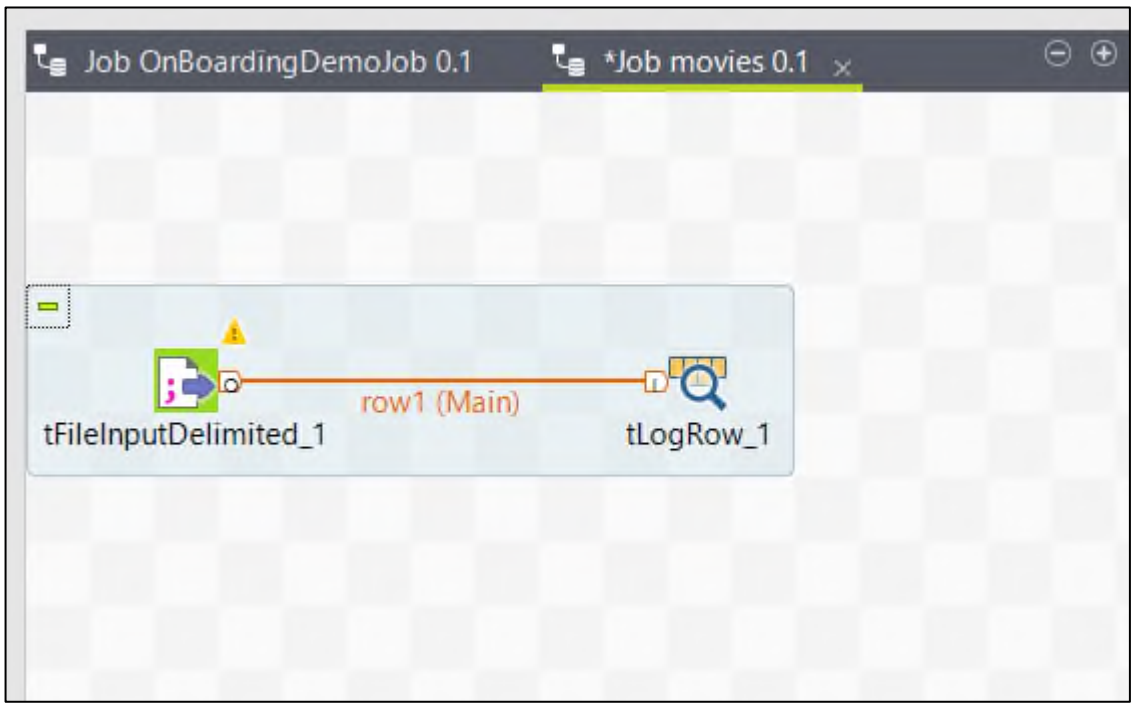
1. Создание проекта.
2. Создали новый проект; создали папку:



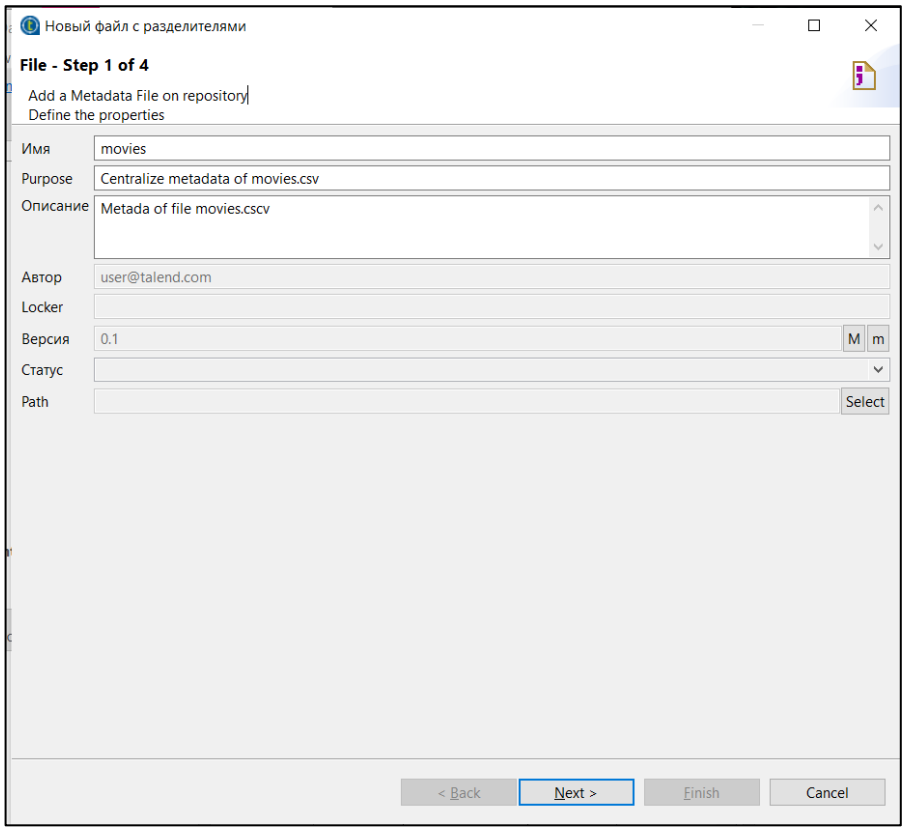
3. Создали работу:

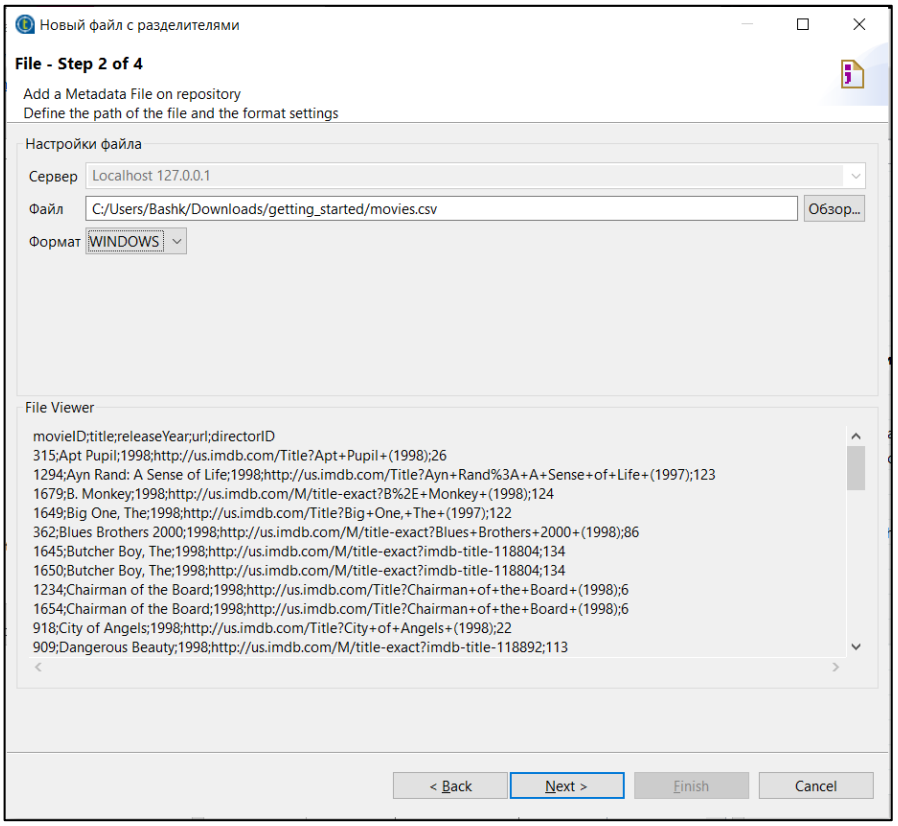


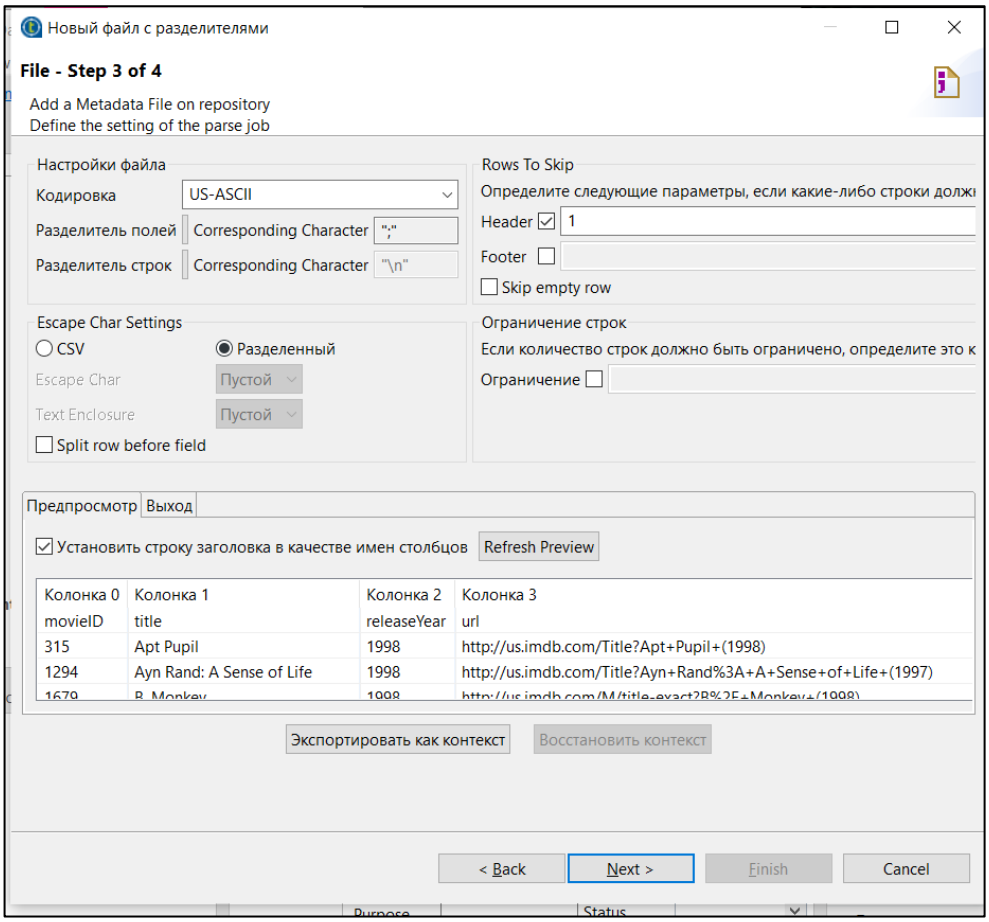
4. Добавили элементы и соединили их:

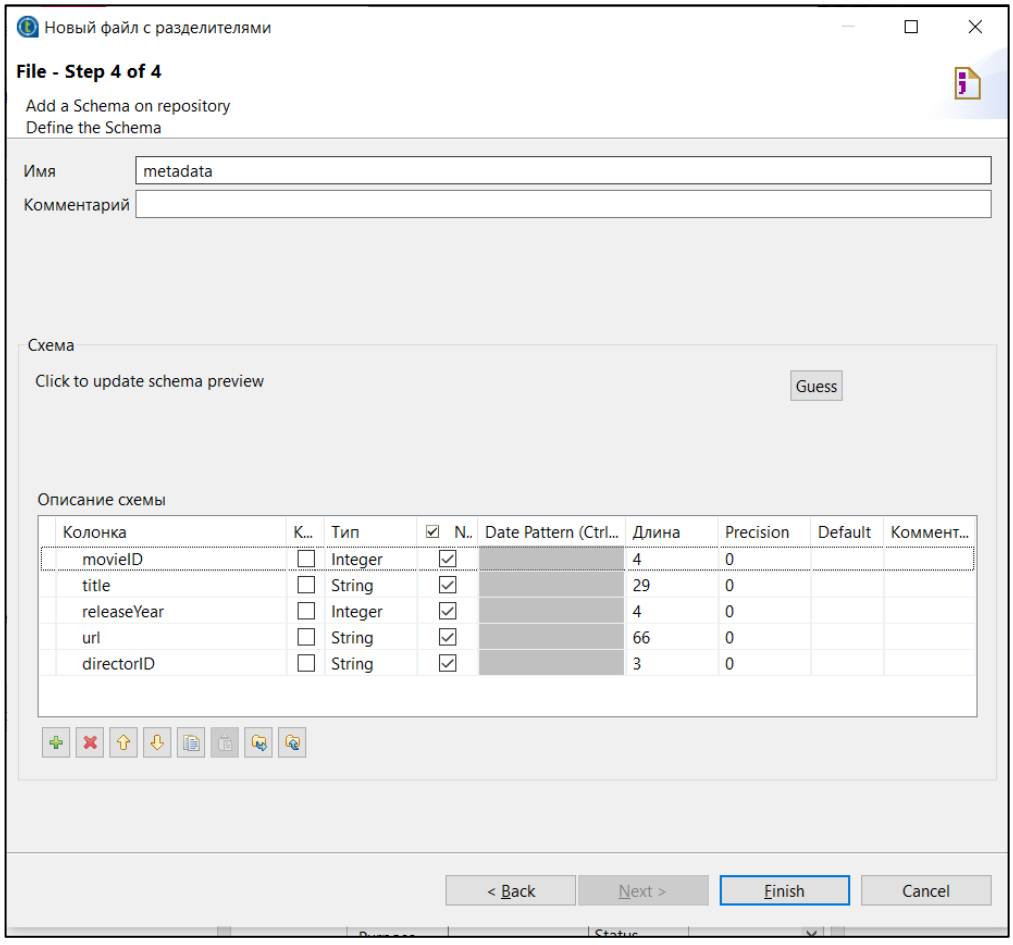


5. Создали файл с разделителями:

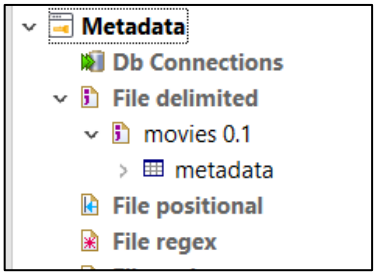

Новый файл с разделителями
File - Step 1 of 4
Add a Metadata File on repository
Define the properties
Имя
movies
Purpose
Centralize metadata of movies.csv
Описание
Metada of file movies.cscv
Автор
user@talend.com
Locker
Версия
0.1
M
m
Статус
Path
Select
< Back
Next >
Finish
Cancel


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/Bashk/Downloads/getting_started/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

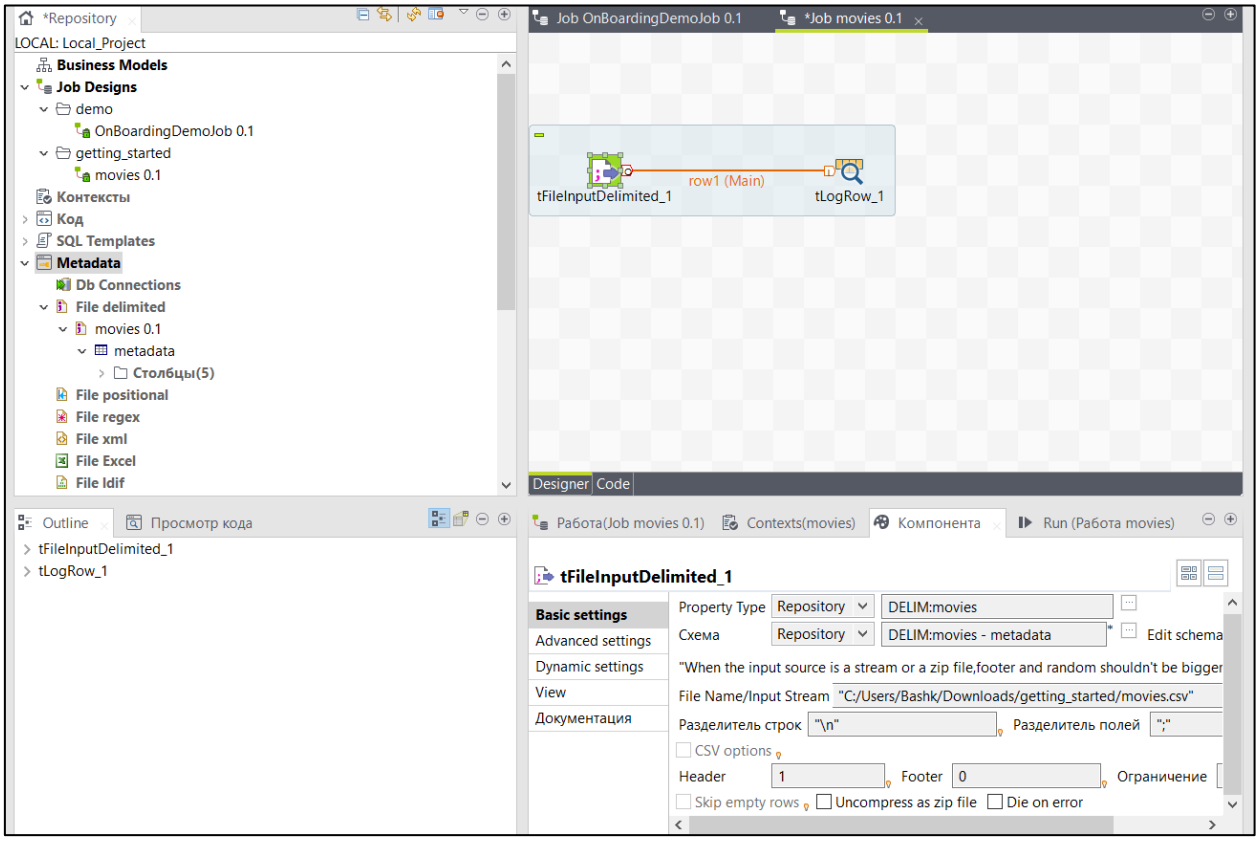

Новый файл с разделителями
File - Step 3 of 4
Add a Metadata File on repository
Define the setting of the parse job
Настройки файла
Кодировка
US-ASCII
Разделитель полей
Corresponding Character
";"
Разделитель строк
Corresponding Character
"\n"
Rows To Skip
Определите следующие параметры, если какие-либо строки должн
Header
1
Footer
Skip empty row
Escape Char Settings
CSV
Разделенный
Escape Char
Пустой
Text Enclosure
Пустой
Split row before field
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
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)
Экспортировать как контекст
Восстановить контекст
< Back
Next >
Finish
Cancel

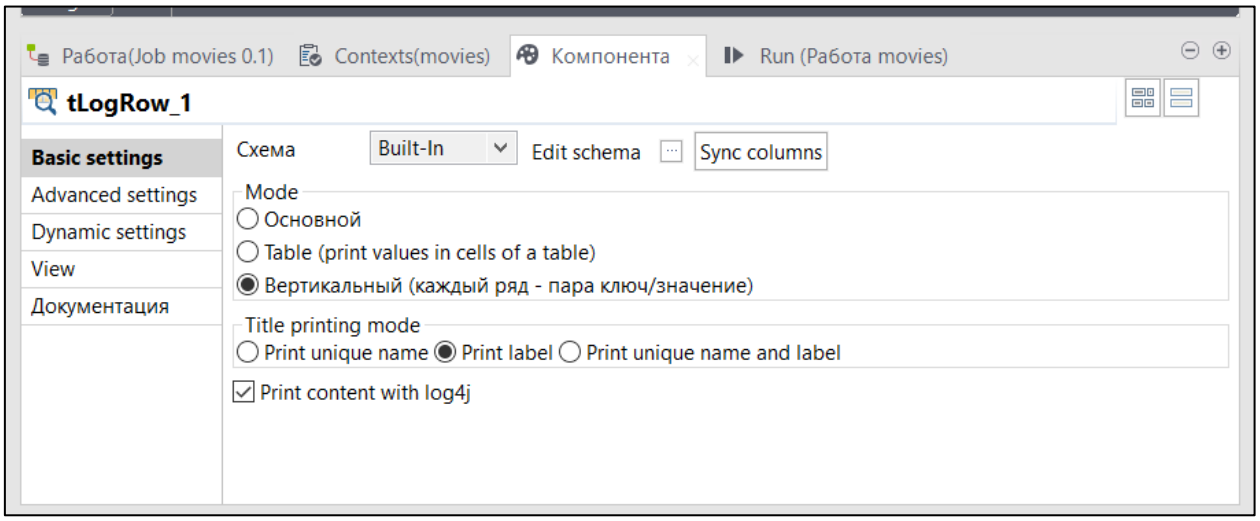

Новый файл с разделителями
File - Step 4 of 4
Add a Schema on repository
Define the Schema
Имя
metadata
Комментарий
Схема
Click to update schema preview
Guess
Описание схемы
Колонка
K...
Тип
N..
Date Pattern (Ctrl...
Длина
Precision
Default
Коммент...
movieID
Integer
4
0
title
String
29
0
releaseYear
Integer
4
0
url
String
66
0
directorID
String
3
0
< Back
Next >
Finish
Cancel

6. Схема с загруженными метаданными:

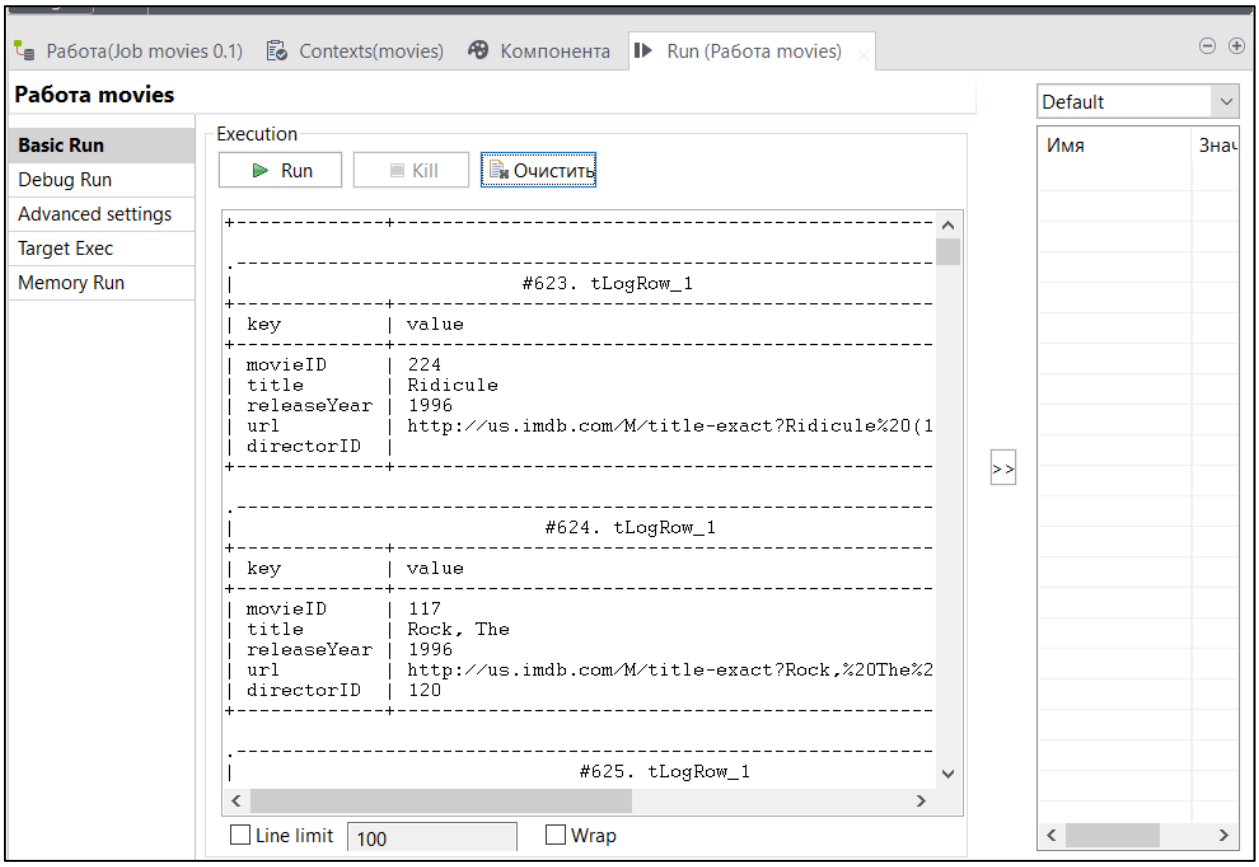


7. Настройка и выполнение работы:

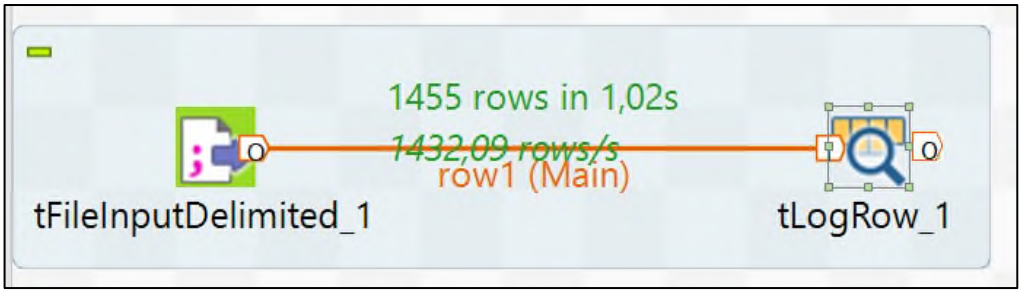




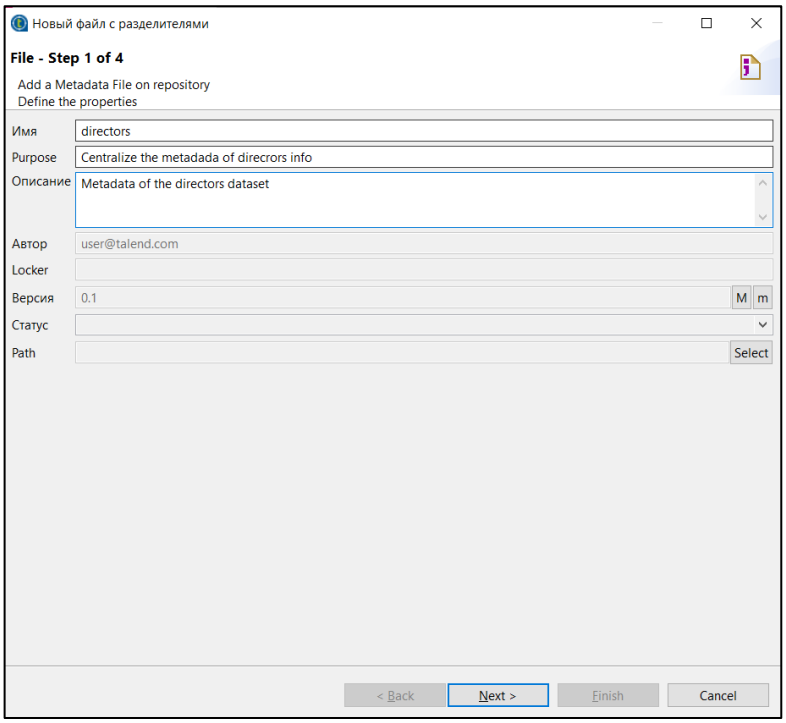
8. Консоль Запуска отображает информацию о фильмах, считанную из исходного файла:

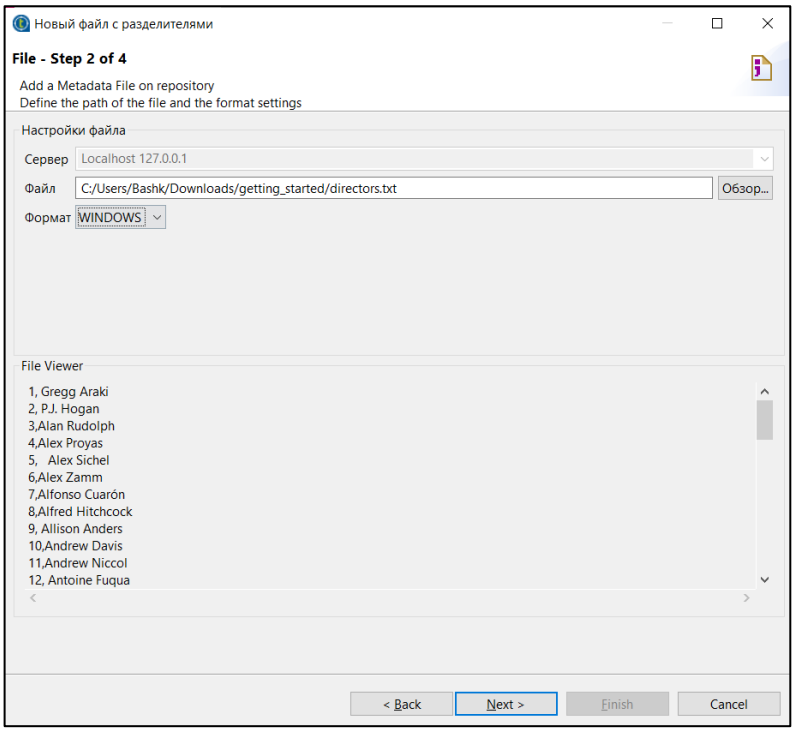


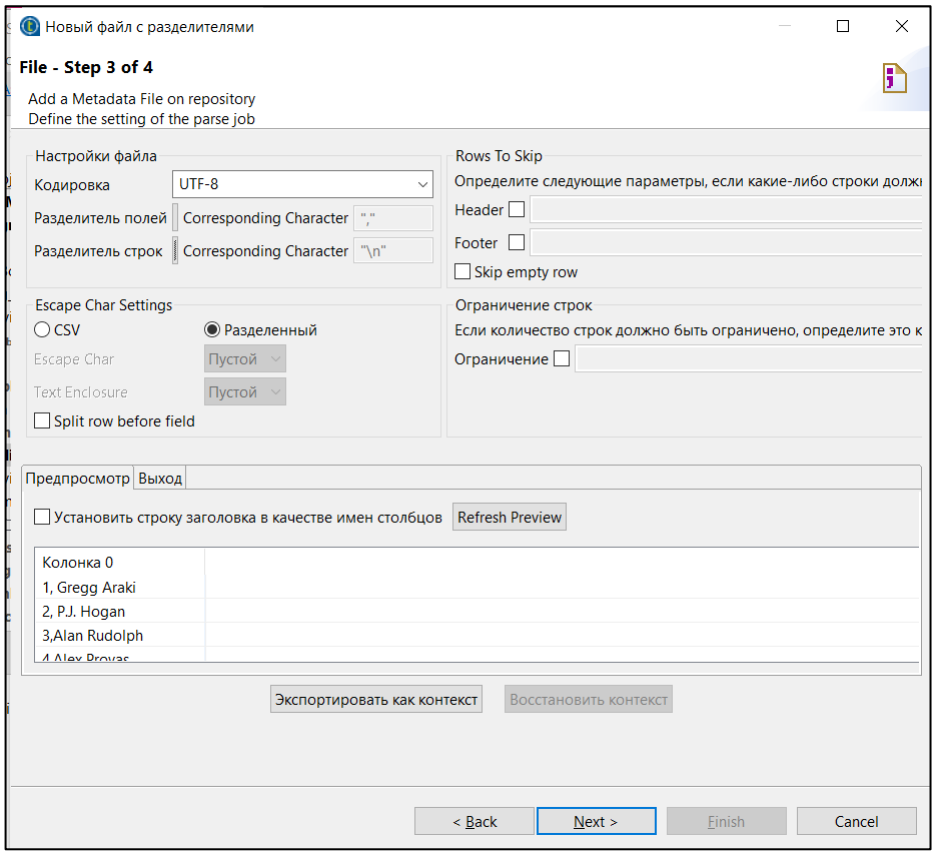
9. Фильтрация информации о фильмах:

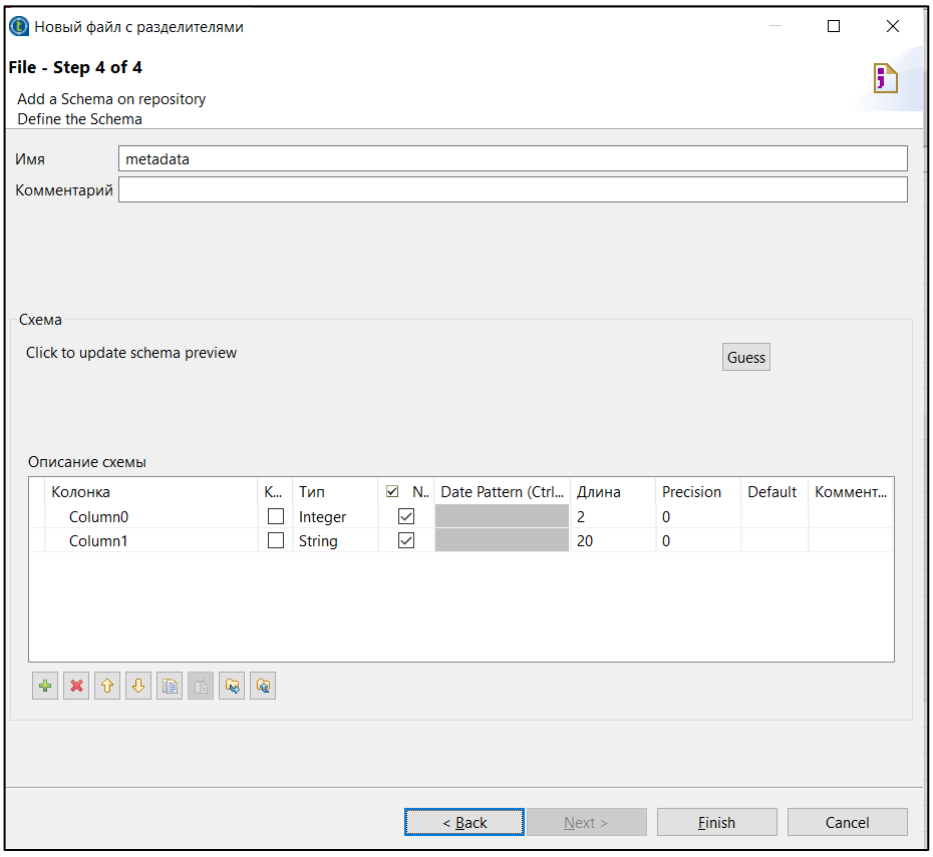


10.Подготовка метаданных:

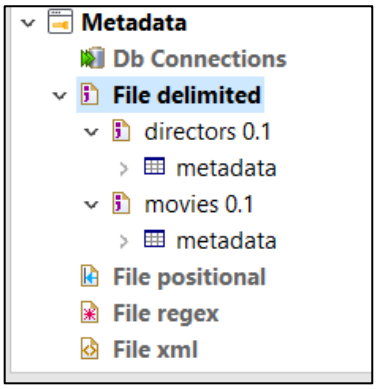

Новый файл с разделителями
File - Step 1 of 4
Add a Metadata File on repository
Define the properties
Имя
directors
Purpose
Centralize the metadada of direcrors info
Описание
Metadata of the directors dataset
Автор
user@talend.com
Locker
Версия
0.1
M
m
Статус
Path
Select
< Back
Next >
Finish
Cancel


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/Bashk/Downloads/getting_started/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
10,Andrew Davis
11,Andrew Niccol
12, Antoine Fuqua
< Back
Next >
Finish
Cancel

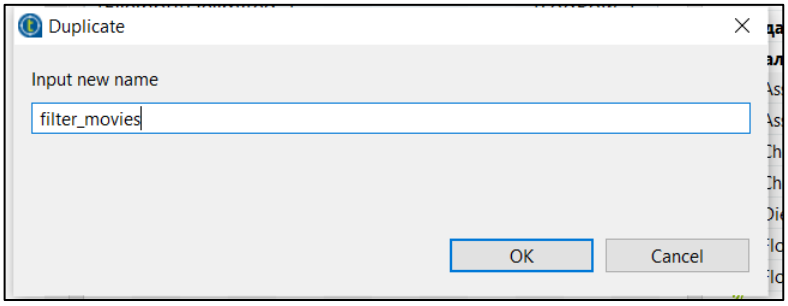




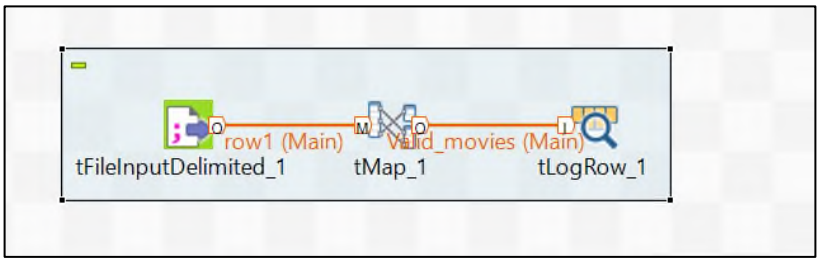
11.Представления репозитория:



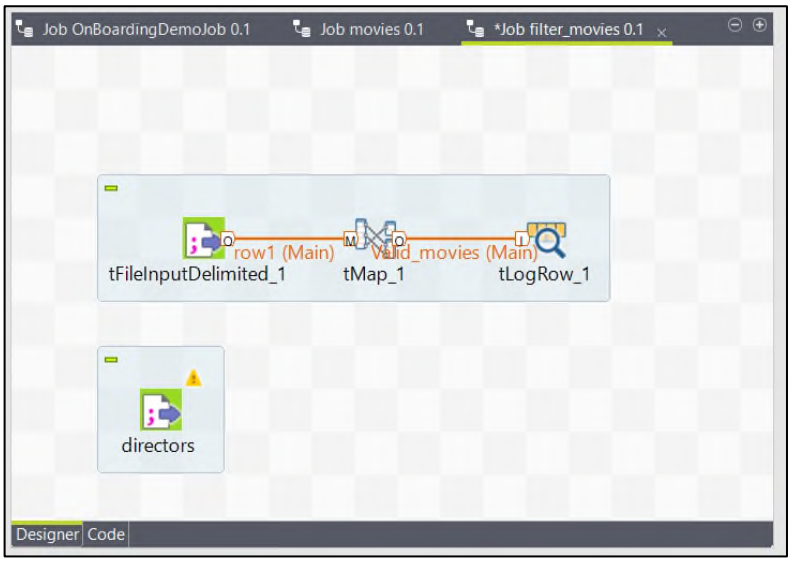
12.Дублирование задания:

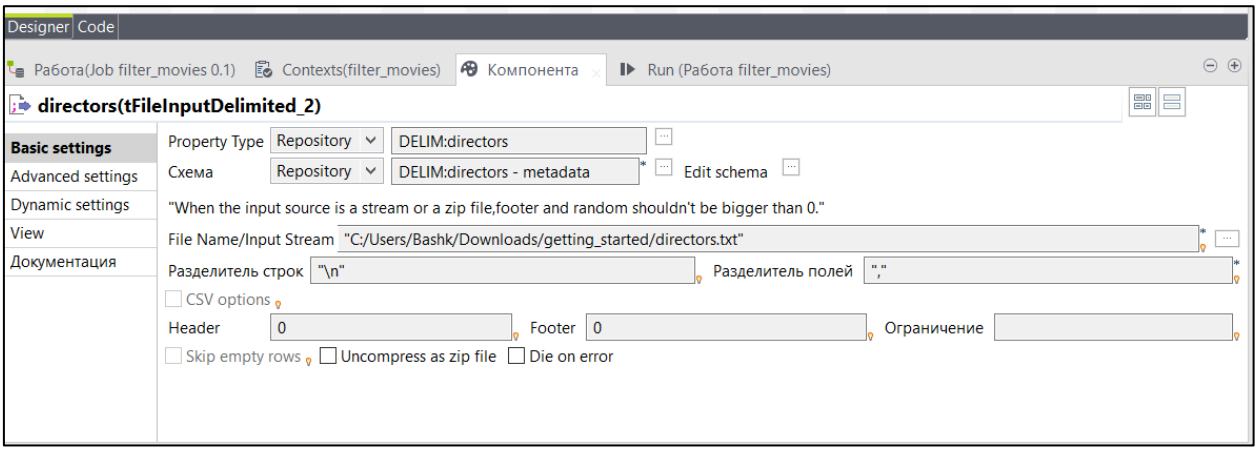


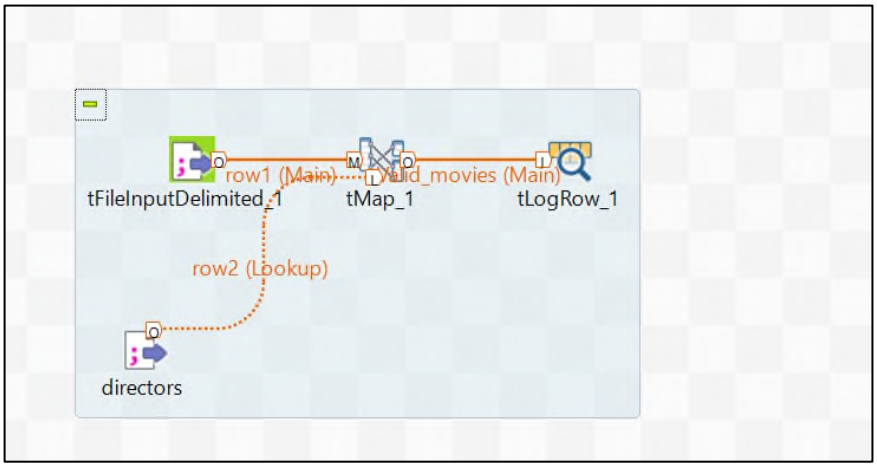
13.Добавление компонента сопоставления:

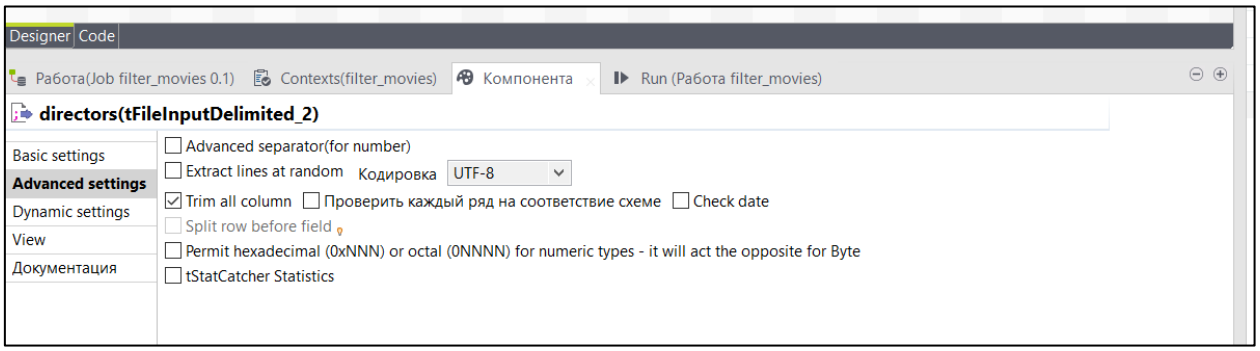


14.Добавление компонента поиска:

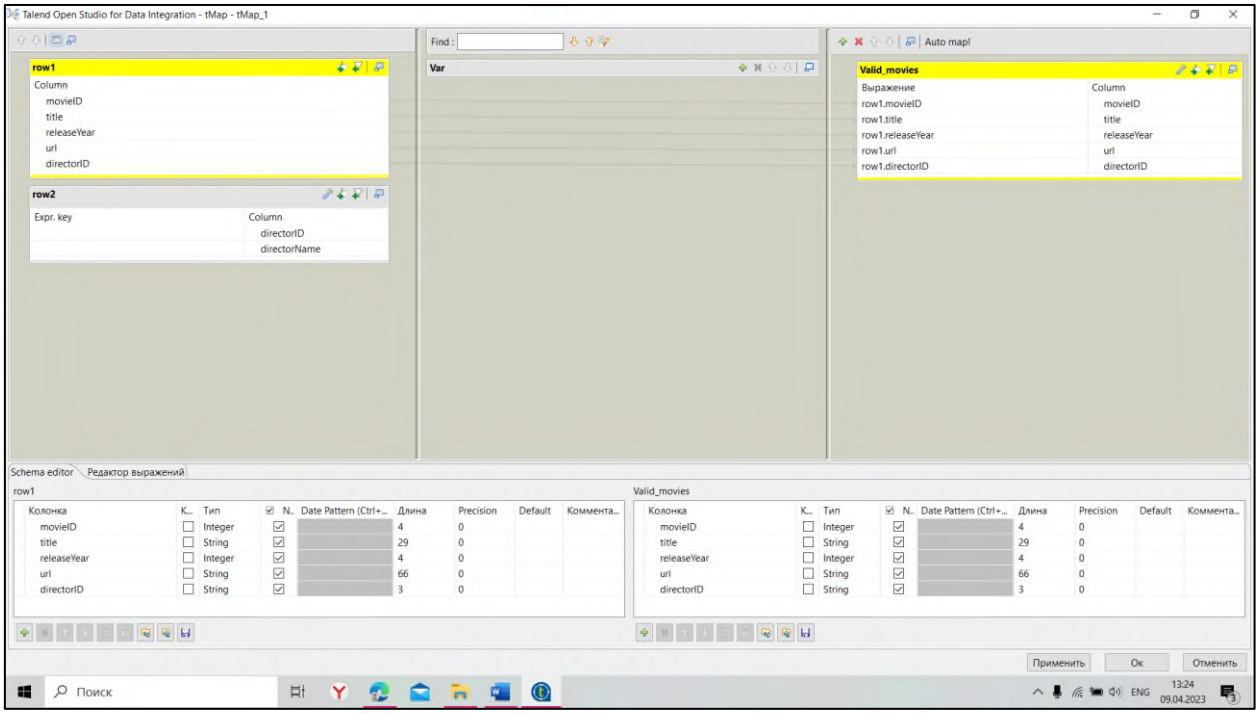

Job OnBoardingDemoJob 0.1
Job movies 0.1
*Job filter_movies 0.1
row1 (Main)
valid_movies (Main)
tFileInputDelimited_1
tMap_1
tLogRow_1
directors
Designer
Code

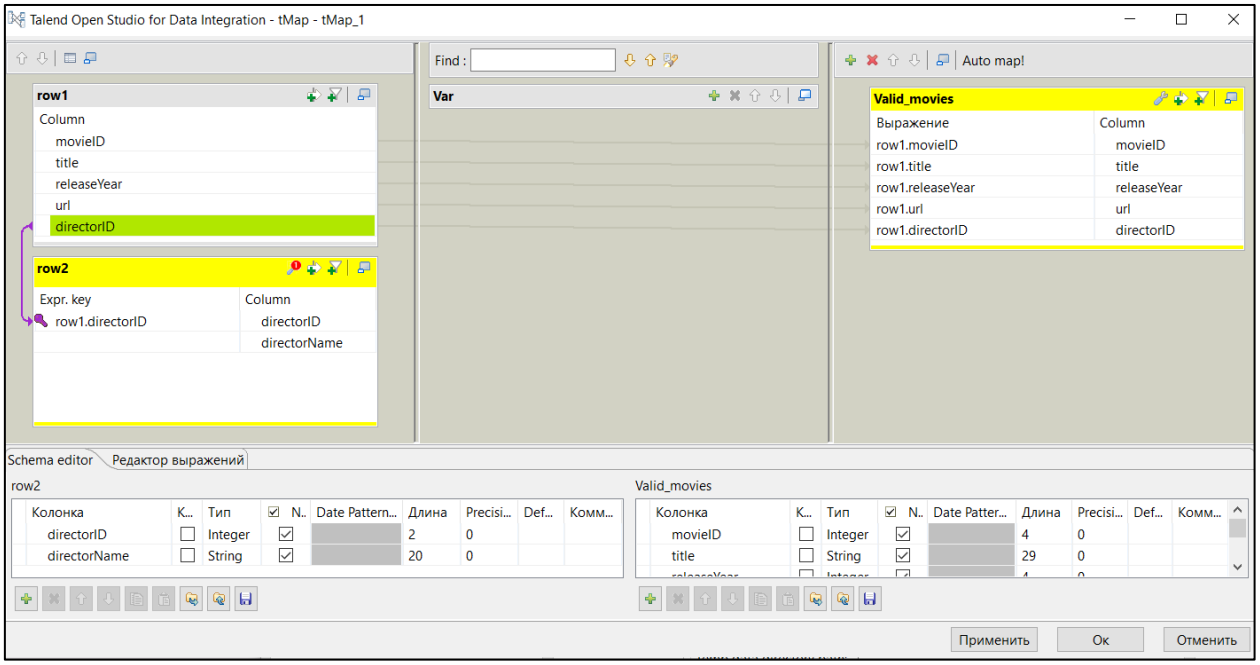

Designer
Code
Работа(Job filter_movies 0.1)
Contexts(filter_movies)
Компонента
Run (Работа filter_movies)
directors(tFileInputDelimited_2)
Basic settings
Advanced settings
Dynamic settings
View
Документация
Property Type
Repository
DELIM:directors
Схема
Repository
DELIM:directors - metadata
Edit schema
"When the input source is a stream or a zip file,footer and random shouldn't be bigger than 0."
File Name/Input Stream
"C:/Users/Bashk/Downloads/getting_started/directors.txt"
Разделитель строк
"\n"
Разделитель полей
","
CSV options
Header
0
Footer
0
Ограничение
Skip empty rows
Uncompress as zip file
Die on error

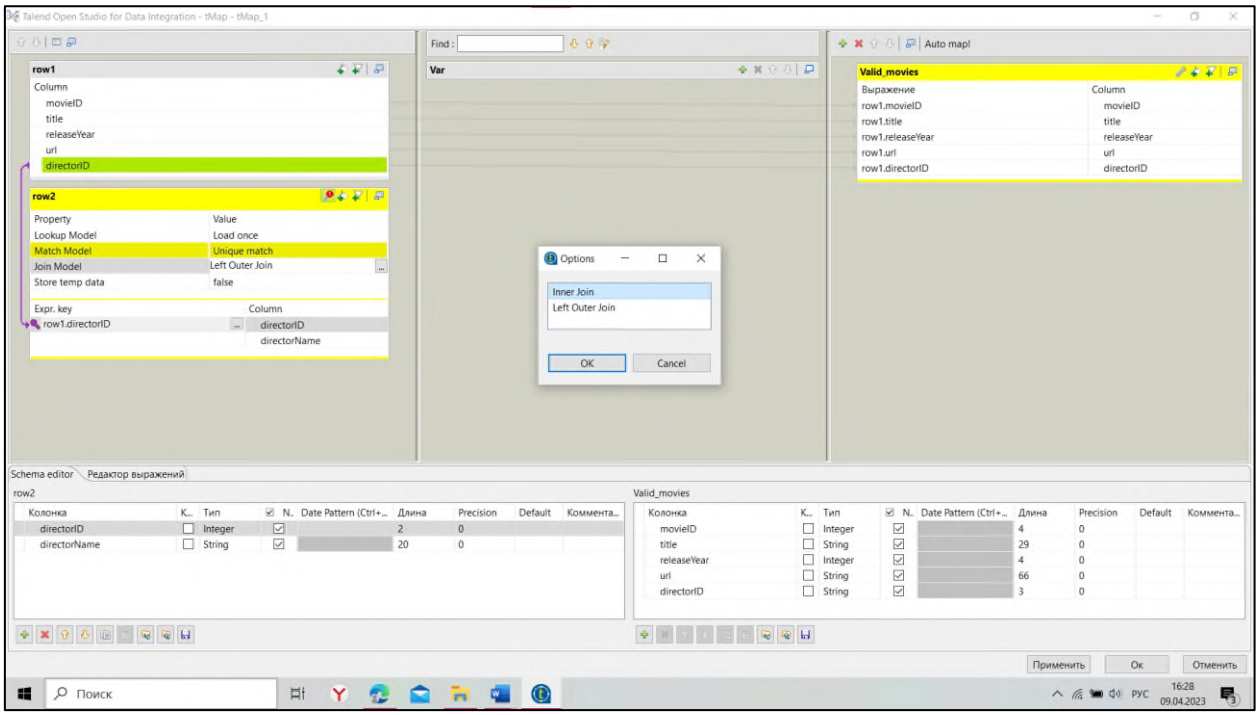

row1 (Main)
valid_movies (Main)
tFileInputDelimited_1
tMap_1
tLogRow_1
row2 (Lookup)
directors


Designer
Code
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

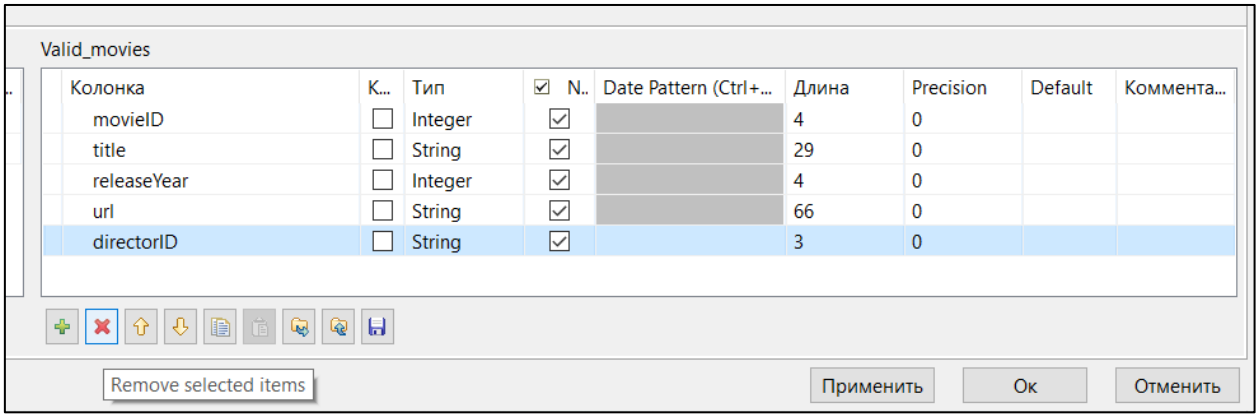
15.Настройка сопоставлений и выполнение:

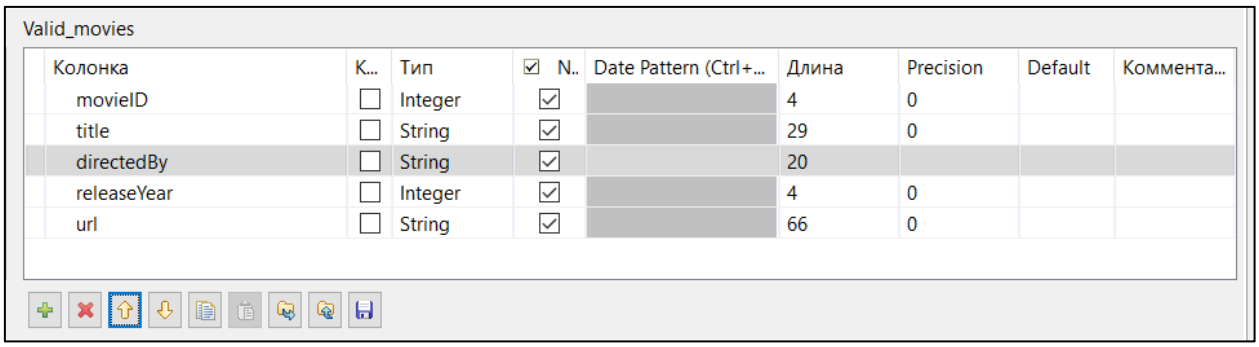


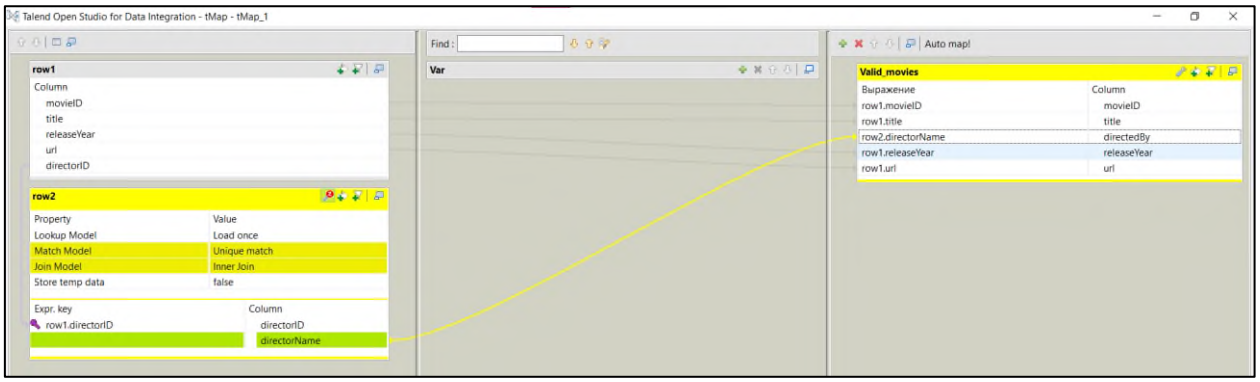




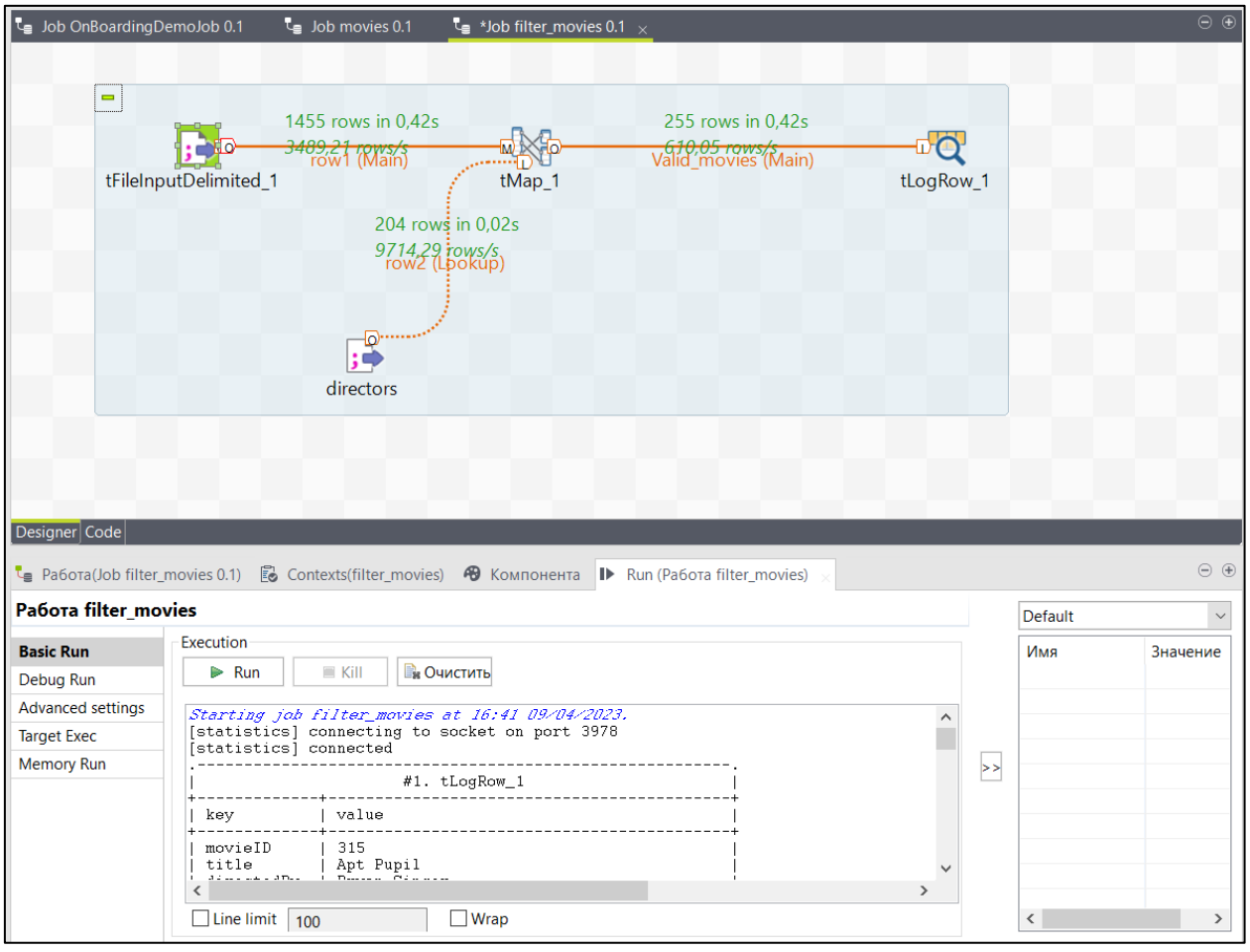
16. Удаление и добавление строк



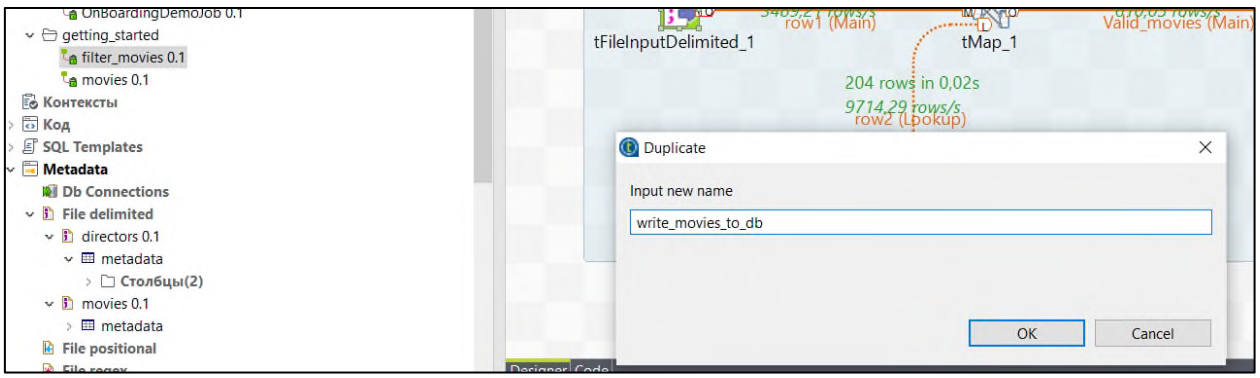


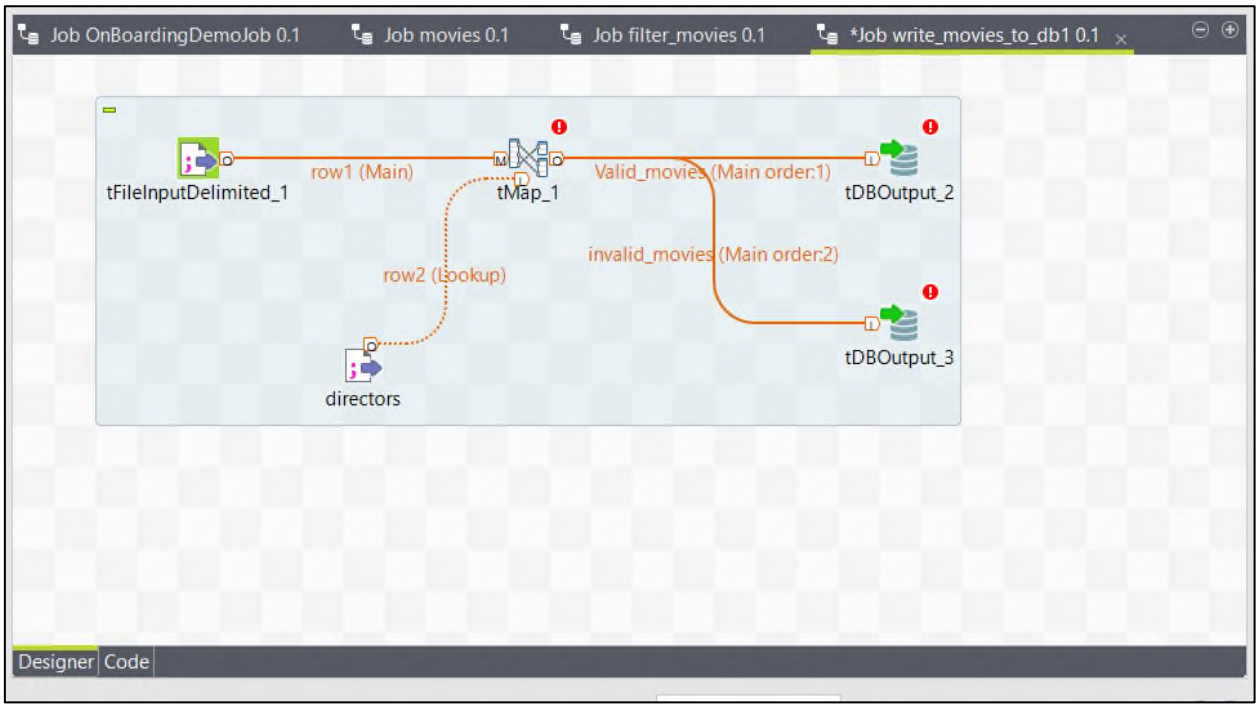


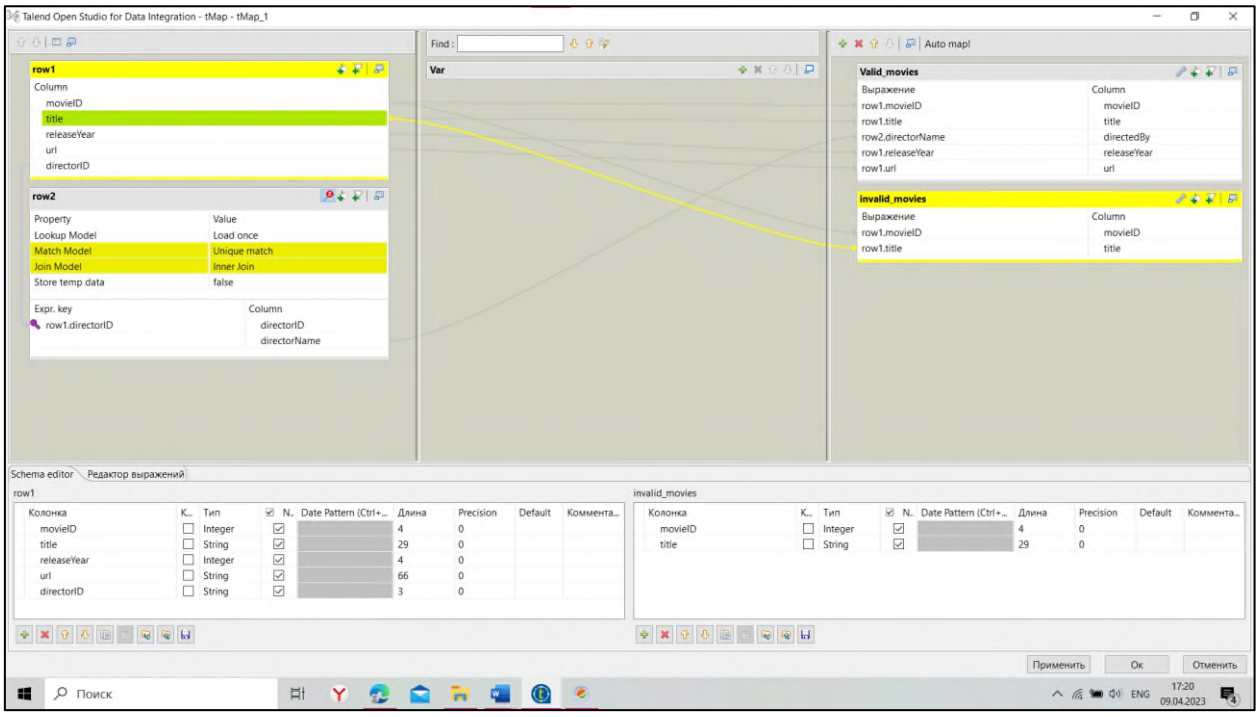
17.Выполнение:

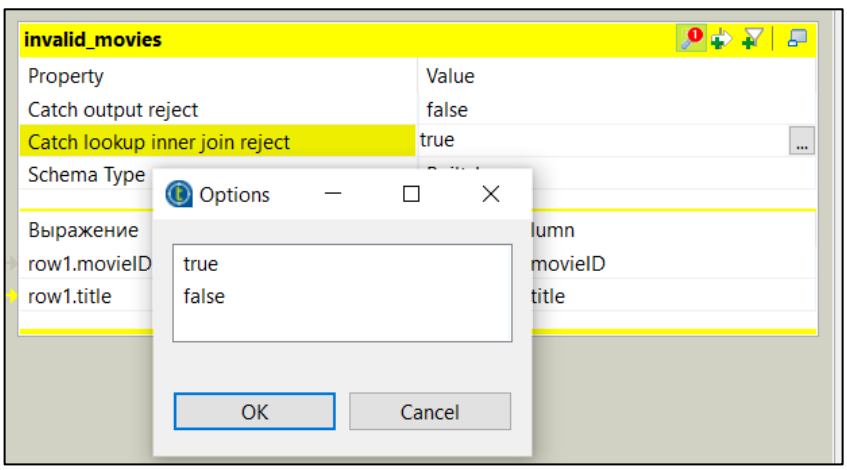


18.Сбор информации об отклоненных фильмах и сохранение результатов обработки в базе данных

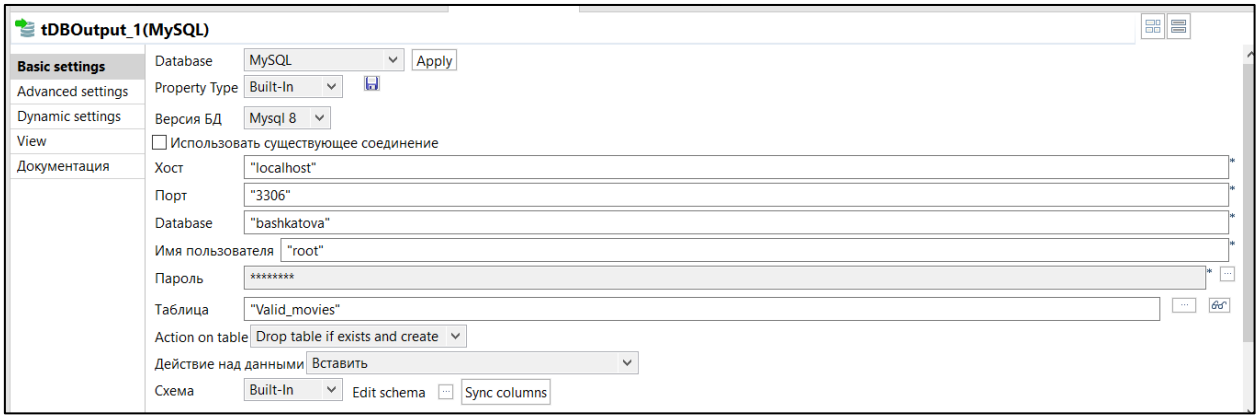


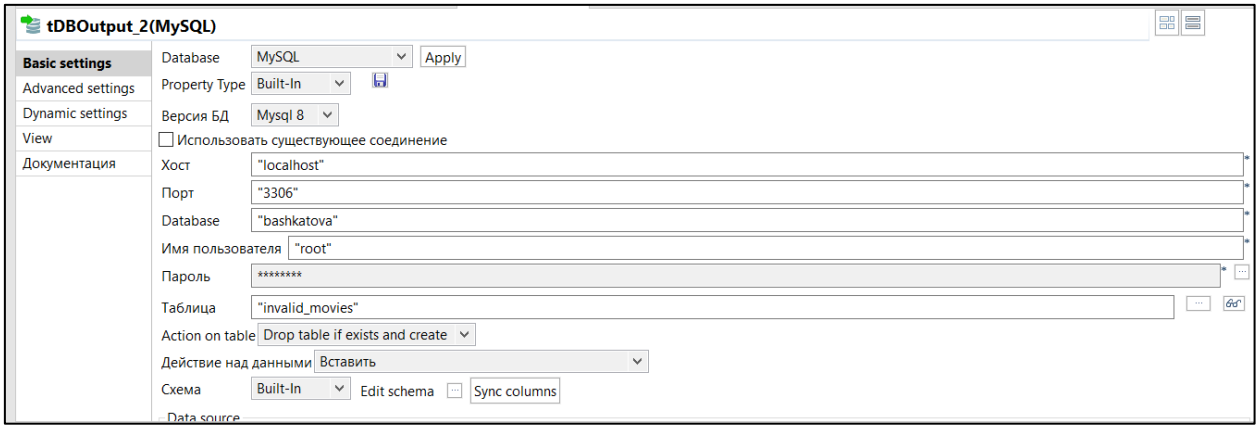

Job OnBoardingDemoJob 0.1
Job movies 0.1
Job filter_movies 0.1
*Job write_movies_to_db1 0.1
tFileInputDelimited_1
row1 (Main)
tMap_1
Valid_movies (Main order:1)
tDBOutput_2
invalid_movies (Main order:2)
row2 (Lookup)
tDBOutput_3
directors
Designer
Code

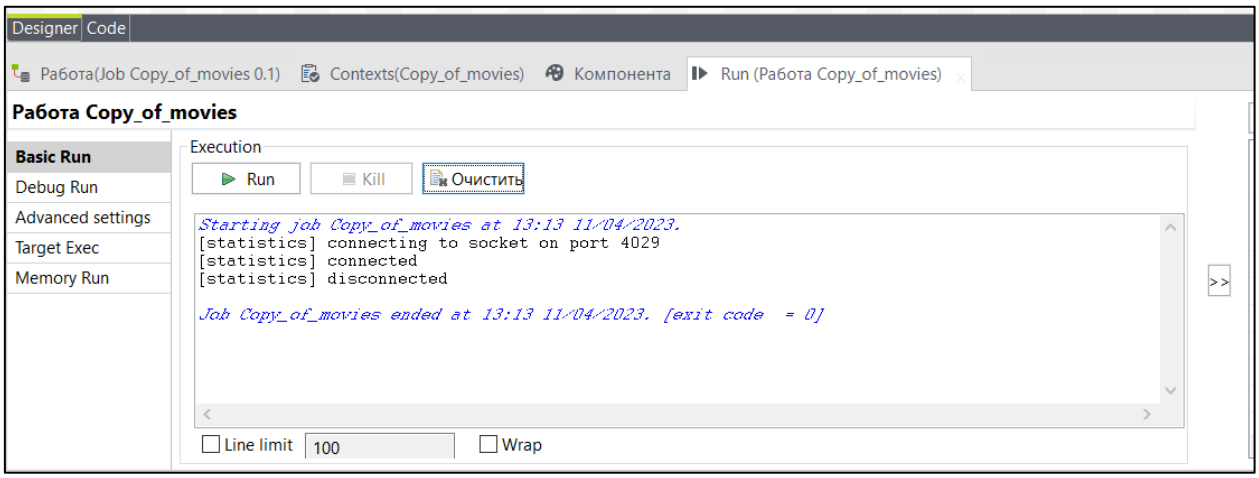

Talend Open Studio for Data Integration - tMap - tMap_1
row1
row2
Var
Valid_movies
invalid_movies
Schema editor
Редактор выражений

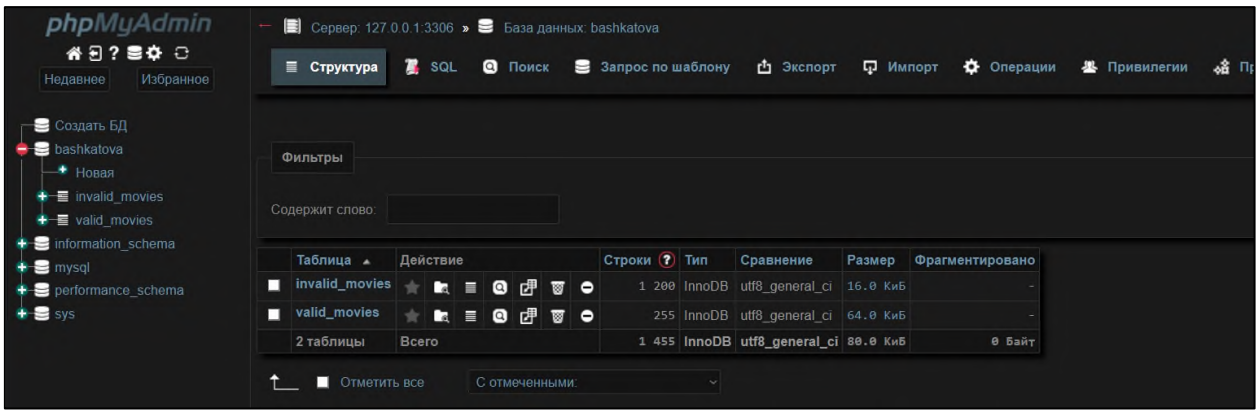

invalid_movies
Property
Value
Catch output reject
false
Catch lookup inner join reject
true
Schema Type
Options
true
false
OK
Cancel

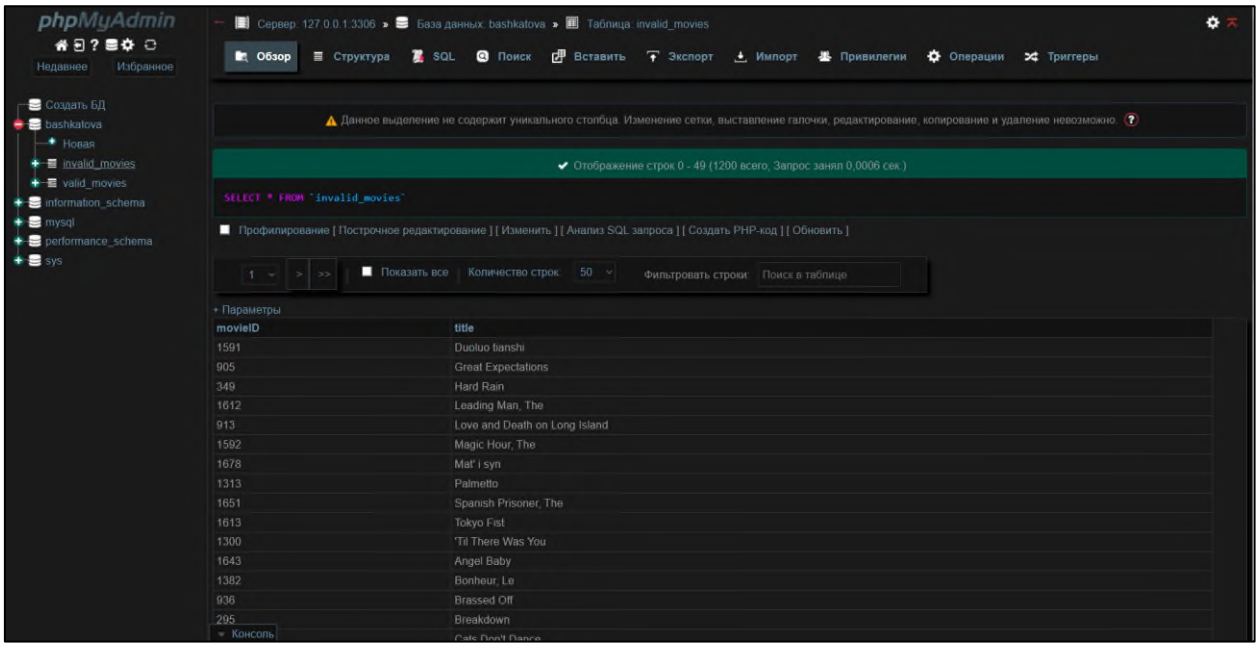
19.Настройка выходных данных базы данных MySQL










phpMyAdmin
Недавнее
Избранное
Создать БД
bashkatova
Новая
invalid_movies
valid_movies
information_schema
mysql
performance_schema
sys
Сервер: 127.0.0.1:3306 » База данных: bashkatova » Таблица: invalid_movies
Обзор
Структура
SQL
Поиск
Вставить
Экспорт
Импорт
Привилегии
Операции
Триггеры
Данное выделение не содержит уникального столбца. Изменение сетки, выставление галочки, редактирование, копирование и удаление невозможно.
Отображение строк 0 - 49 (1200 всего, Запрос занял 0,0006 сек.)
SELECT * FROM `invalid_movies`
Профилирование [Построчное редактирование] [Изменить] [Анализ SQL запроса] [Создать PHP-код] [Обновить]
Показать все
Количество строк:
50
Фильтровать строки:
Поиск в таблице
+ Параметры
movieID
title
1591 Duoluo tianshi
905 Great Expectations
349 Hard Rain
1612 Leading Man, The
913 Love and Death on Long Island
1592 Magic Hour, The
1678 Mat' i syn
1313 Palmetto
1651 Spanish Prisoner, The
1613 Tokyo Fist
1300 'Til There Was You
1643 Angel Baby
1382 Bonheur, Le
936 Brassed Off
295 Breakdown
Консоль