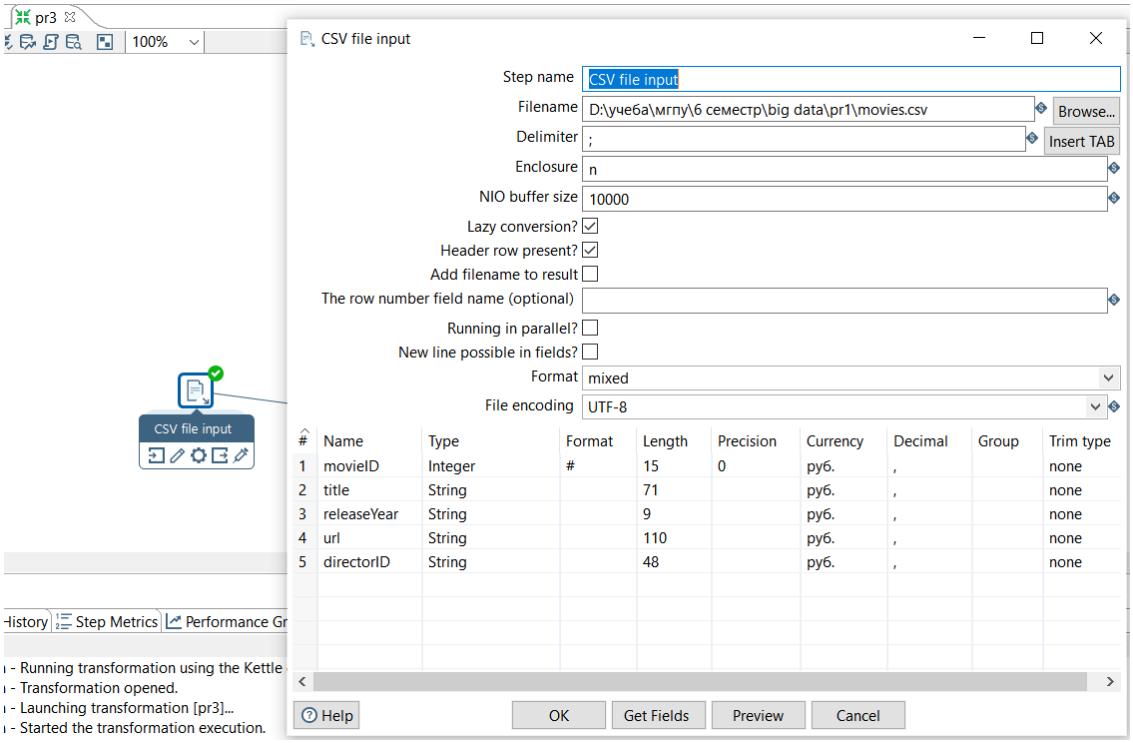
Департамент образования и науки города Москвы
Государственное автономное образовательное учреждение
высшего образования города Москвы
«Московский городской педагогический университет»
Институт цифрового образования
Департамент информатики, управления и технологий

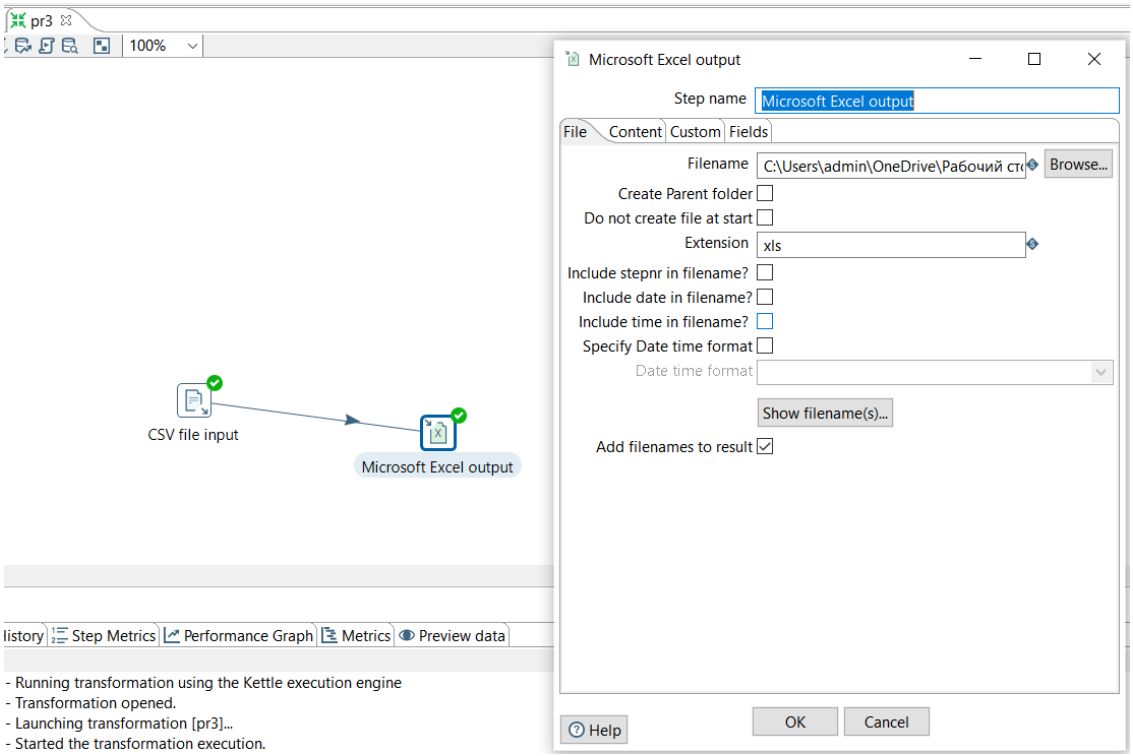
ДИСЦИПЛИНА:
«Инструменты для хранения и обработки больших данных»
пр_03
Тема:
«Архитектура хранилищ данных: традиционная и облачная. Практическая часть»

Выполнила: Николаева С.Г., АДЭУ-201
Преподаватель: Босенко Тимур Муртазович

Москва
2023

Задача: загрузить данные из файла csv, оставить 3 столбца и сохранить результат в другой файл.





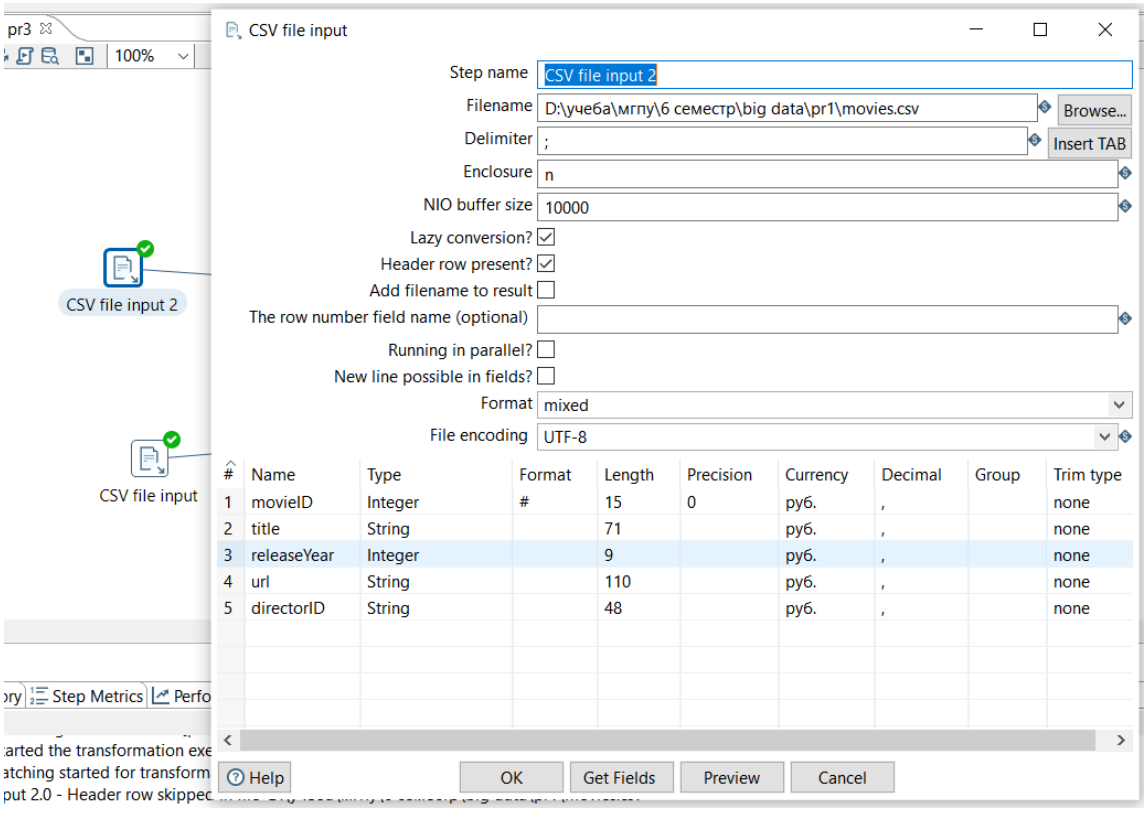
До

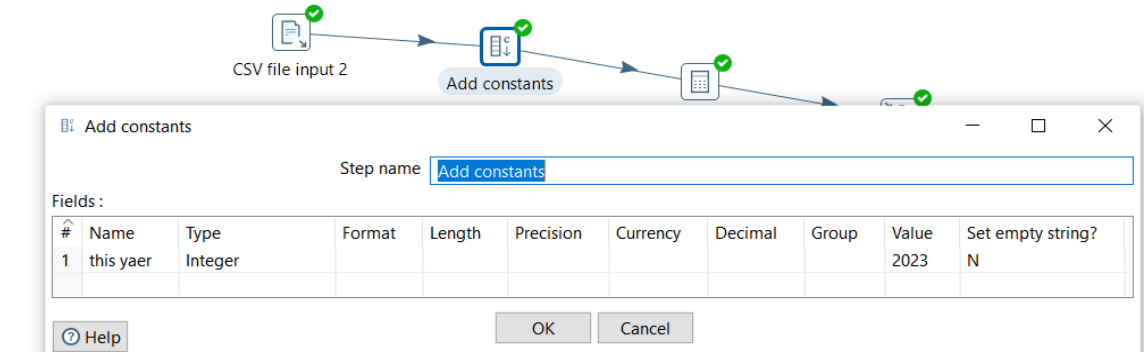
	A	B	C	D	E	F	G	H
1	movieID	title	releaseYea	url	directorID			
2	315	Apt Pupil	1998	http://us.i	26			
3	1294	Ayn Rand:	1998	http://us.i	123			
4	1679	B. Monkey	1998	http://us.i	124			
5	1649	Big One, Tl	1998	http://us.i	122			
6	362	Blues Brot	1998	http://us.i	86			
7	1645	Butcher Bc	1998	http://us.i	134			
8	1650	Butcher Bc	1998	http://us.i	134			
9	1234	Chairman	1998	http://us.i	6			
10	1654	Chairman	1998	http://us.i	6			
11	918	City of Ang	1998	http://us.i	22			
12	909	Dangerous	1998	http://us.i	113			
13	691	Dark City	1998	http://us.i	4			
14	353	Deep Risin	1998	http://us.i	186			
15	329	Desperate	1998	http://us.i	14			
16	348	Desperate	1998	http://us.i	13			
17	1591	Duoluo tia	1998	http://us.imdb.com/M/title-exact?imdb-title-112913				
18	1594	Everest	1998	http://us.i	39			
19	350	Fallen	1998	http://us.i	64			
20	1105	Firestorm	1998	http://us.i	44			

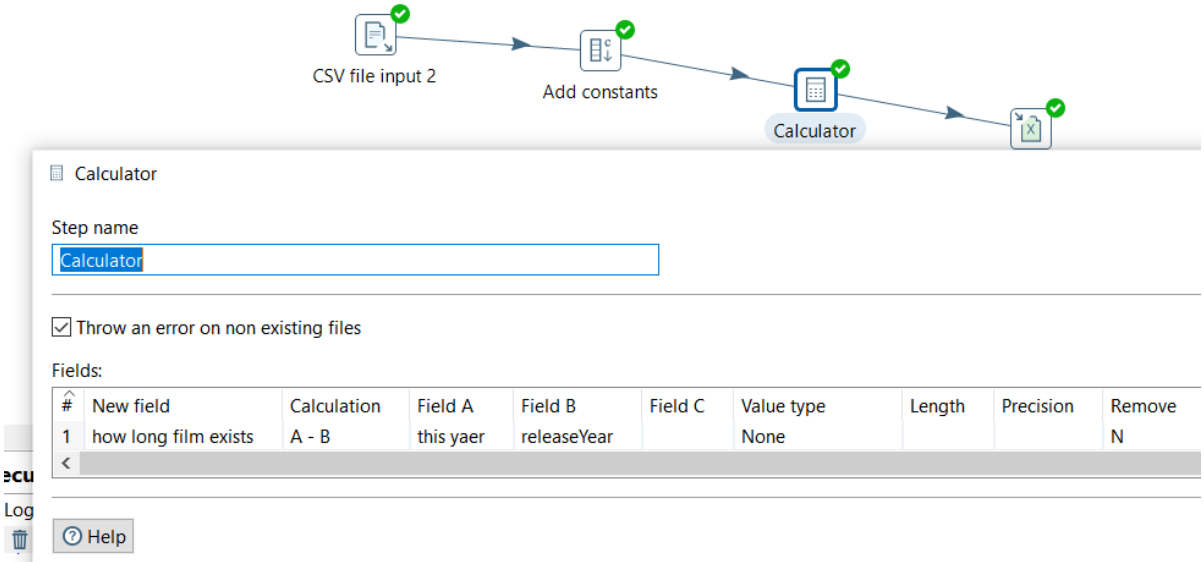
После

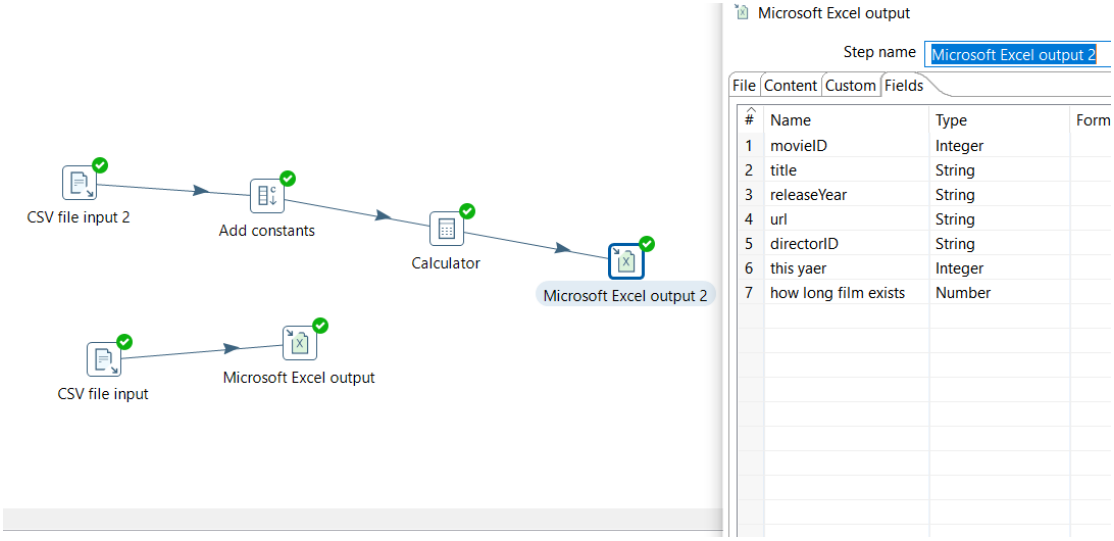
	A	B	C	D	E
1	movieID	title	releaseYear		
2	315,00	Apt Pupil	1998		
3	1 294,00	Ayn Rand:	1998		
4	1 679,00	B. Monkey	1998		
5	1 649,00	Big One, T	1998		
6	362,00	Blues Brot	1998		
7	1 645,00	Butcher Bc	1998		
8	1 650,00	Butcher Bc	1998		
9	1 234,00	Chairman	1998		
10	1 654,00	Chairman	1998		
11	918,00	City of Ang	1998		
12	909,00	Dangerous	1998		
13	691,00	Dark City	1998		
14	353,00	Deep Risin	1998		
15	329,00	Desperate	1998		
16	348,00	Desperate	1998		
17	1 591,00	Duoluo tia	1998		
18	1 594,00	Everest	1998		
19	350,00	Fallen	1998		
20	1 105,00	Firestorm	1998		
21	1 062,00	Four Days	1998		

Эксперимент: хочу узнать, сколько лет существует фильм.


pr3
100%
CSV file input 2
CSV file input
Step Metrics
CSV file input
Step name
CSV file input 2
Filename
D:\учеба\мгпу\6 семестр\big data\pr1\movies.csv
Browse...
Delimiter
;
Insert TAB
Enclosure
n
NIO buffer size
10000
Lazy conversion?
Header row present?
Add filename to result
The row number field name (optional)
Running in parallel?
New line possible in fields?
Format
mixed
File encoding
UTF-8
Name Type Format Length Precision Currency Decimal Group Trim type
1 movieID Integer # 15 0 руб. , none
2 title String 71 руб. , none
3 releaseYear Integer 9 руб. , none
4 url String 110 руб. , none
5 directorID String 48 руб. , none
Help
OK
Get Fields
Preview
Cancel


CSV file input 2
Add constants
Add constants
Step name
Add constants
Fields :
Name Type Format Length Precision Currency Decimal Group Value Set empty string?
1 this yaer Integer 2023 N
Help
OK
Cancel


CSV file input 2
Add constants
Calculator
Calculator
Step name
Calculator
Throw an error on non existing files
Fields:
New field Calculation Field A Field B Field C Value type Length Precision Remove
1 how long film exists A - B this yaer releaseYear None N
Help



Результат

	A	B	C	D	E	F	G	H
1	movieID	title	releaseYe	url	directorID	this yaer	how long film exists	
2	315,00	Apt Pupil	1 998,00	http://us.in	26	2 023,00	25,00	
3	1 294,00	Ayn Rand:	1 998,00	http://us.in	123	2 023,00	25,00	
4	1 679,00	B. Monkey	1 998,00	http://us.in	124	2 023,00	25,00	
5	1 649,00	Big One, T	1 998,00	http://us.in	122	2 023,00	25,00	
6	362,00	Blues Brot	1 998,00	http://us.in	86	2 023,00	25,00	
7	1 645,00	Butcher Bo	1 998,00	http://us.in	134	2 023,00	25,00	
8	1 650,00	Butcher Bo	1 998,00	http://us.in	134	2 023,00	25,00	
9	1 234,00	Chairman	1 998,00	http://us.in	6	2 023,00	25,00	
10	1 654,00	Chairman	1 998,00	http://us.in	6	2 023,00	25,00	
11	918,00	City of Ang	1 998,00	http://us.in	22	2 023,00	25,00	
12	909,00	Dangerous	1 998,00	http://us.in	113	2 023,00	25,00	
1425	89,00	Blade Run	1 982,00	http://us.in	155	2 023,00	41,00	
1426	674,00	Cat People	1 982,00	http://us.in		2 023,00	41,00	
1427	423,00	E.T. the Ex	1 982,00	http://us.in	187	2 023,00	41,00	
1428	527,00	Gandhi	1 982,00	http://us.in		2 023,00	41,00	
1429	1 037,00	Grease 2	1 982,00	http://us.in		2 023,00	41,00	
1430	414,00	My Favorit	1 982,00	http://us.in		2 023,00	41,00	
1431	214,00	Pink Floyd	1 982,00	http://us.in		2 023,00	41,00	
1432	638,00	Return of M	1 982,00	http://us.in		2 023,00	41,00	
1433	632,00	Sophie's C	1 982,00	http://us.in		2 023,00	41,00	
1434	228,00	Star Trek:	1 982,00	http://us.in		2 023,00	41,00	