

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: «Домашняя работа 1 “ETL Компоненты и начало работы с ETL на
примере Pentaho Data Integration”»**

Направление подготовки (специальность) 38.03.05 – бизнес-информатика

Направленность (профиль) образовательной программы «Аналитика данных
и эффективное управление»

Курс обучения: 3

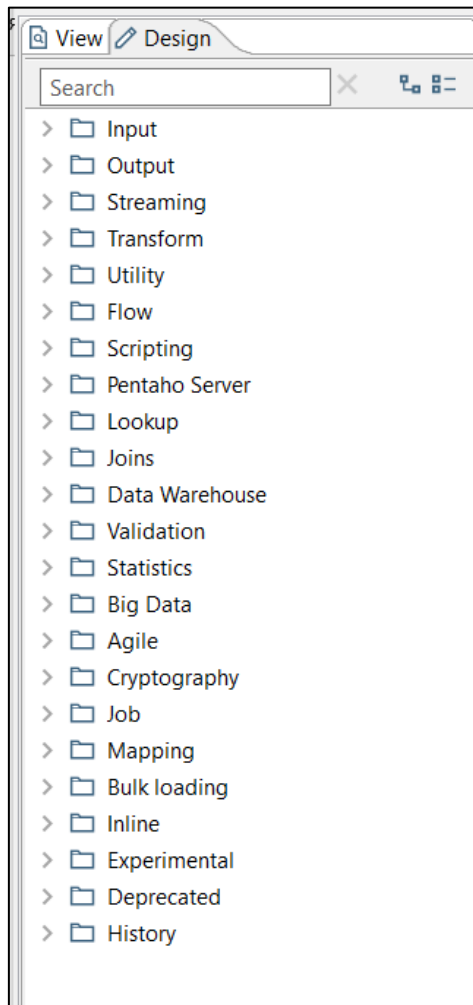
Форма обучения: очная

Руководитель: Босенко Т. М.

Москва
2023

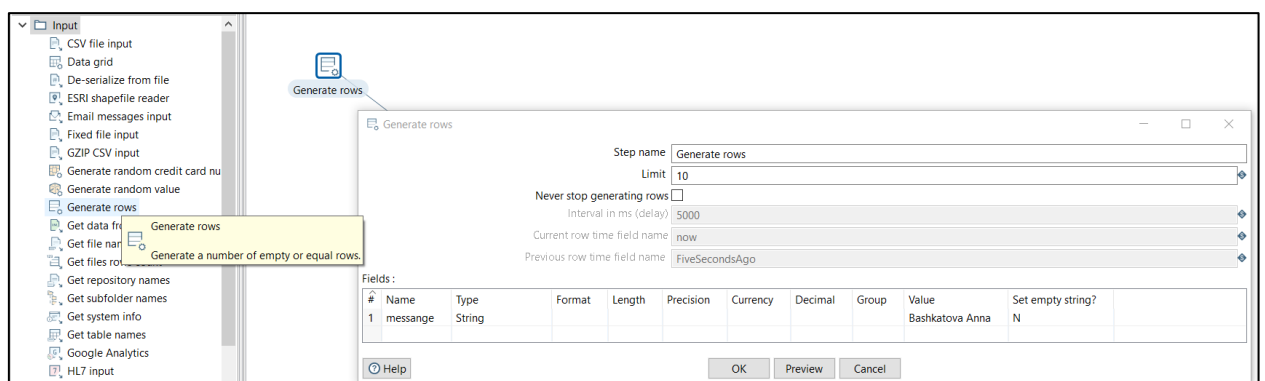
Работа по видео: <https://www.youtube.com/watch?v=-oCBttnefMQ>

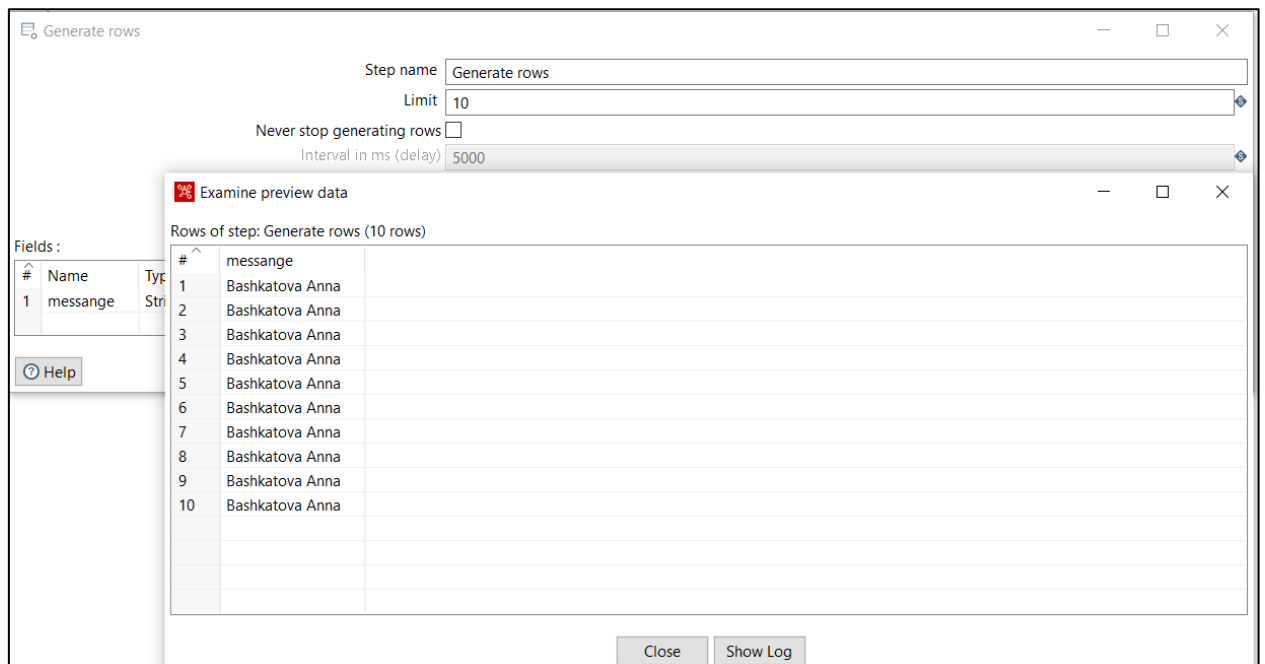
Компоненты:



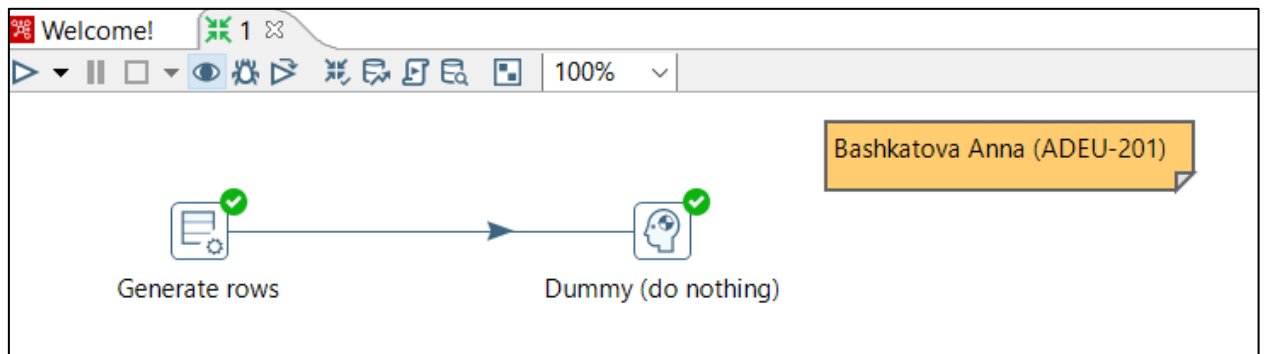
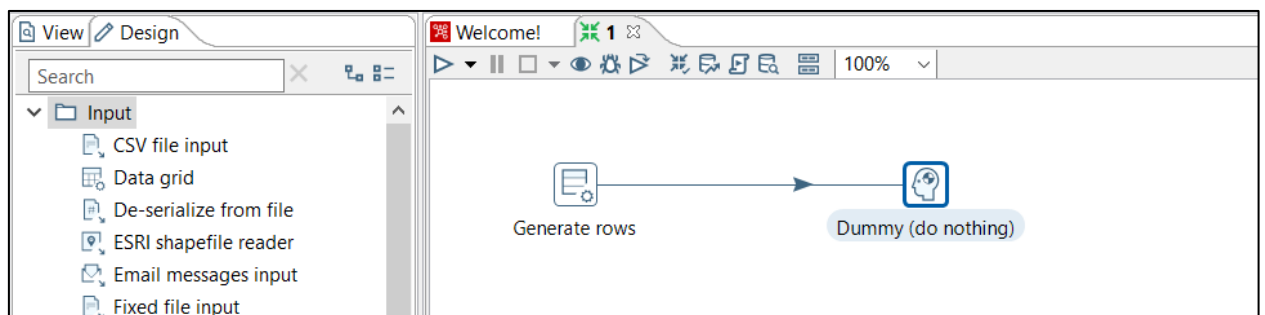
Создание job1.

Добавляем первый компонент – generate rows.

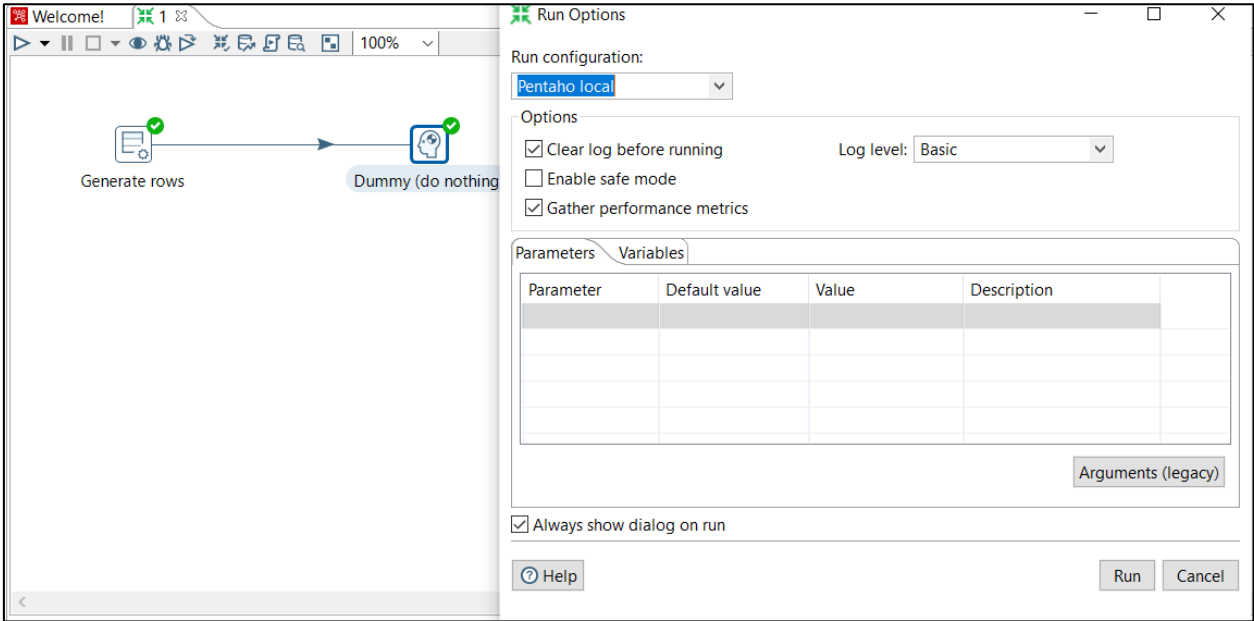


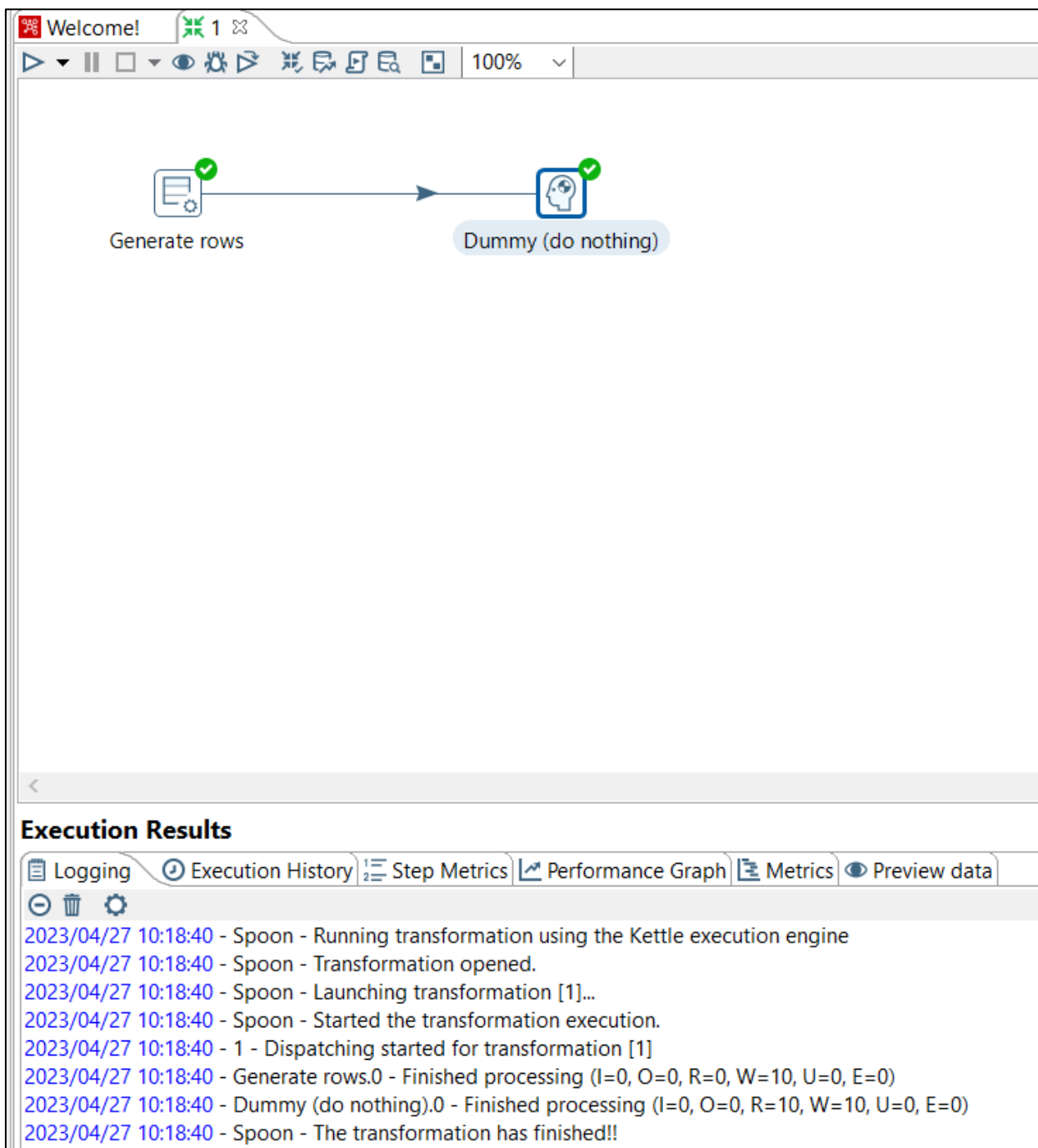


Далее добавляем dummy



Запуск

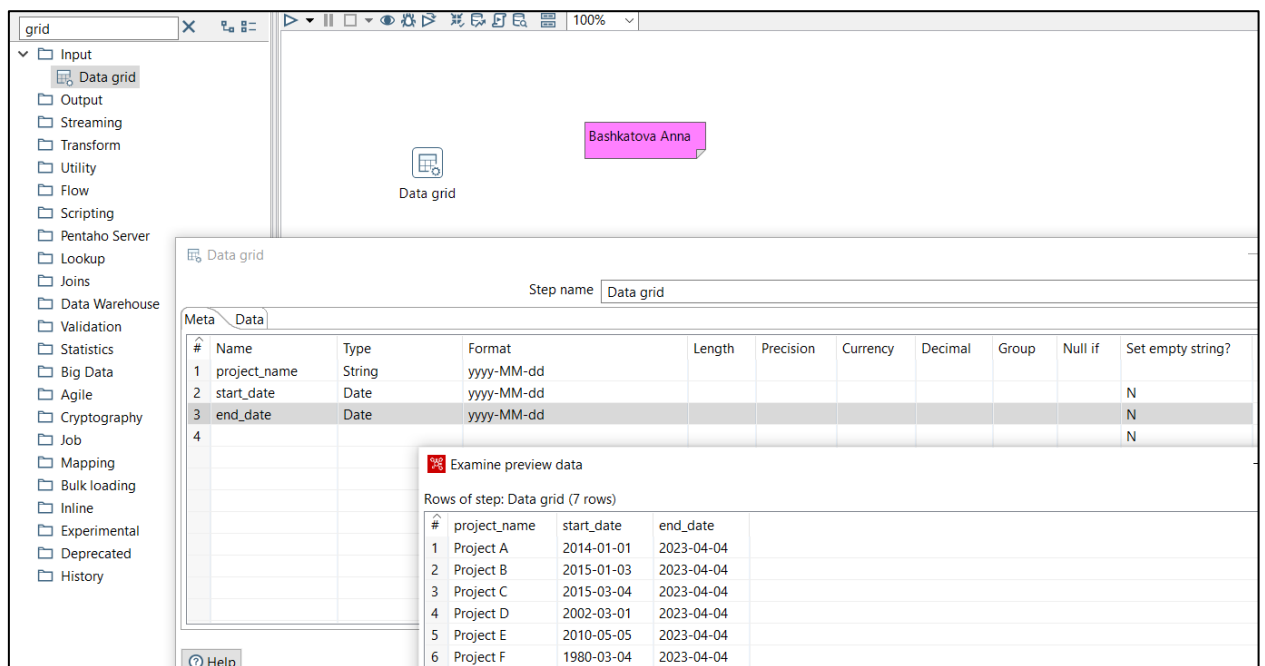
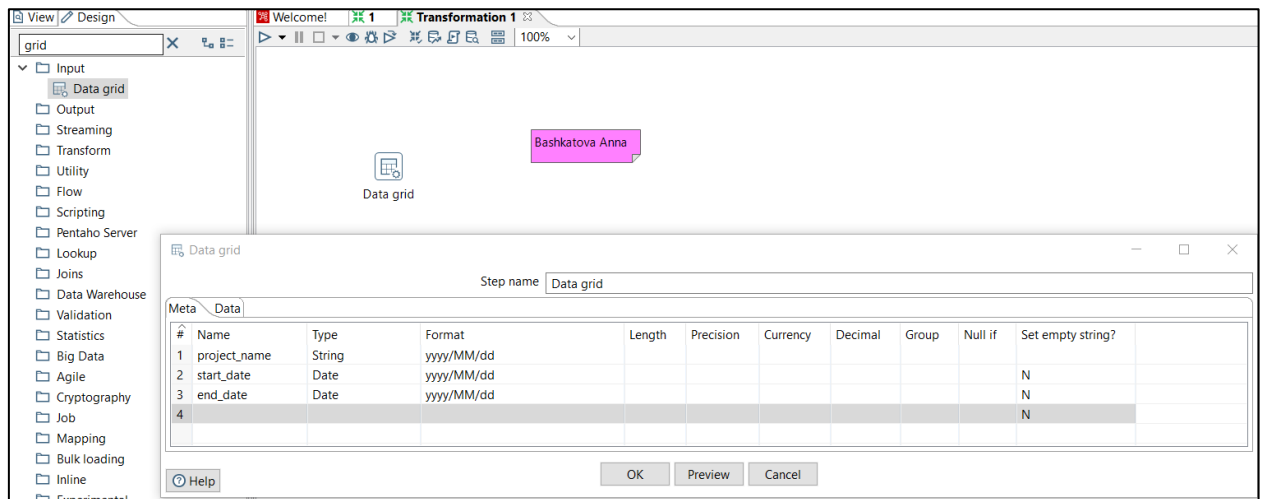




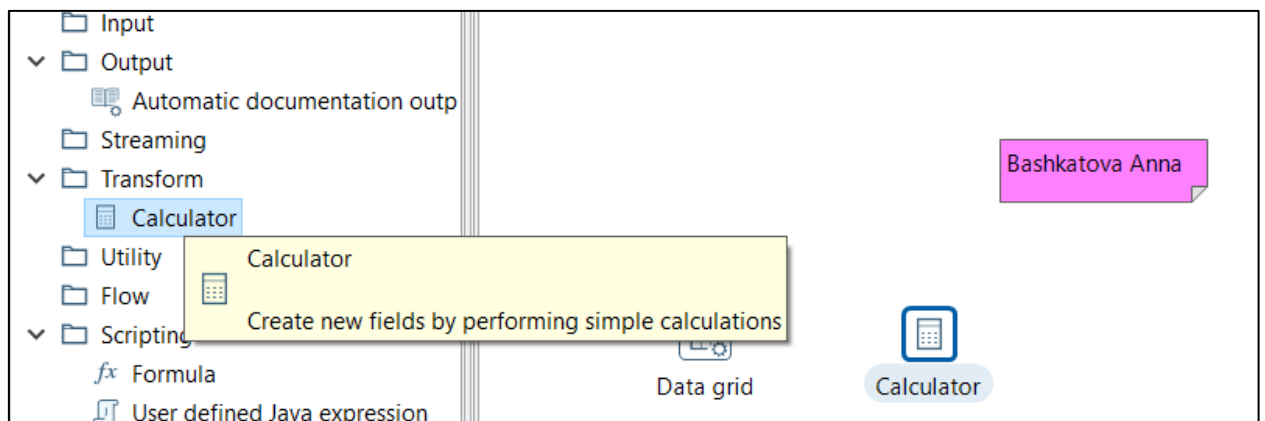
Execution Results													
<div> <div>Logging</div> <div>Execution History</div> <div>Step Metrics</div> <div>Performance Graph</div> <div>Metrics</div> <div>Preview data</div> </div>													
#	Stepname	Copynr	Read	Written	Input	Output	Updated	Rejected	Errors	Active	Time	Speed (r/s)	input/output
1	Generate rows	0	0	10	0	0	0	0	0	Finished	0.0s	3 333	-
2	Dummy (do nothing)	0	10	10	0	0	0	0	0	Finished	0.0s	435	-

Создание job2.

Добавление компонента data grid (создает статические значения)



Добавляем Calculator



Automatic documentation outp

Streaming

Transform

Calculator

Utility

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	Conversion mask	Decimal symbol	Grouping symbol	Currency symbol
1	diff-dates	Date A - Date B (in days)	end_date	start_date		Integer							

Help

OK

Cancel

Добавляем number range (создает последовательность чисел).

Number range

Step name

Number range

Input field

diff-dates

Output field

range

Default value(if no range)

unknown

Ranges (min <= x< max):

#	Lower Bound	Upper Bound	Value ^
1		30.0	excellent
2	30.0	180.0	very good
3	180.0	360.0	good
4	360.0		poor
5			

Help

OK

Cancel

Добавляем User defined Java expression

User defined Java expression

Step name

User defined Java expression

Fields:

#	New field	Java expression	Value type	Length	Precision	Replace value
1	duration	(diff_dates == null)?"unknown":diff_dates + " days"	String			
2	message	"The performance was " + Number range	String			

Help

OK

Cancel

Bashkatova Anna

Data grid

Calculator

1..3...
2...5

performance

message

Examine preview data

Rows of step: message (6 rows)

#	project_name	start_date	end_date	diff_dates	range
1	Project A	2014-01-01	2023-04-04	3380	poor
2	Project B	2022-01-03	2023-04-04	456	poor
3	Project C	2023-03-03	2023-04-04	32	very good
4	Project D	2002-03-01	2023-04-04	7704	poor
5	Project E	2023-03-05	2023-04-04	30	very good
6	Project F	1980-03-04	2023-04-04	15736	poor

Execution Results

Logging Execution History Step

2023/04/27 10:57:04 - job2 - Dispatching sta

2023/04/27 10:57:04 - Data grid.0 - Finished

2023/04/27 10:57:04 - Calculator.0 - Finished

2023/04/27 10:57:04 - performance.0 - Finish

2023/04/27 10:57:05 - message.0 - Finished p

2023/04/27 10:57:05 - Spoon - The transform

Close

Создаем job3.

Добавляем data grid

Bashkatova Anna

Data grid

Data grid

Step name projects

Meta Data

#	Name	Type	Format	Length	Precision	Currency	Decimal	Group	Null if	Set empty string?
1	project_name	String								N
2	start_date	Date	yyyy-MM-dd							N
3	end_date	Date	yyyy-MM-dd							N
4	estimated	String								N

Help OK Preview Cancel

Bashkatova Anna

projects

Data grid

Step name projects

Meta Data

#	project_name	start_date	end_date	estimated
1	Project A	2023-01-01	2023-04-04	30
2	Project B	2023-03-23	2023-04-04	180
3	Project C	2023-04-01	2023-04-04	180
4	Project D	2020-12-30	2023-04-04	700
5	Project E	2022-12-12	2023-04-04	700
6	Project F	2023-03-01	2023-04-04	---

Examine preview data

Rows of step: projects (6 rows)

#	project_name	start_date	end_date	estimated
1	Project A	2023-01-01	2023-04-04	30
2	Project B	2023-03-23	2023-04-04	180
3	Project C	2023-04-01	2023-04-04	180
4	Project D	2020-12-30	2023-04-04	700
5	Project E	2022-12-12	2023-04-04	700
6	Project F	2023-03-01	2023-04-04	---

Help

Добавляем Select values (изменяем тип)

Bashkatova Anna

projects → Select values

Select values

Step name Select values

Select & Alter Remove Meta-data

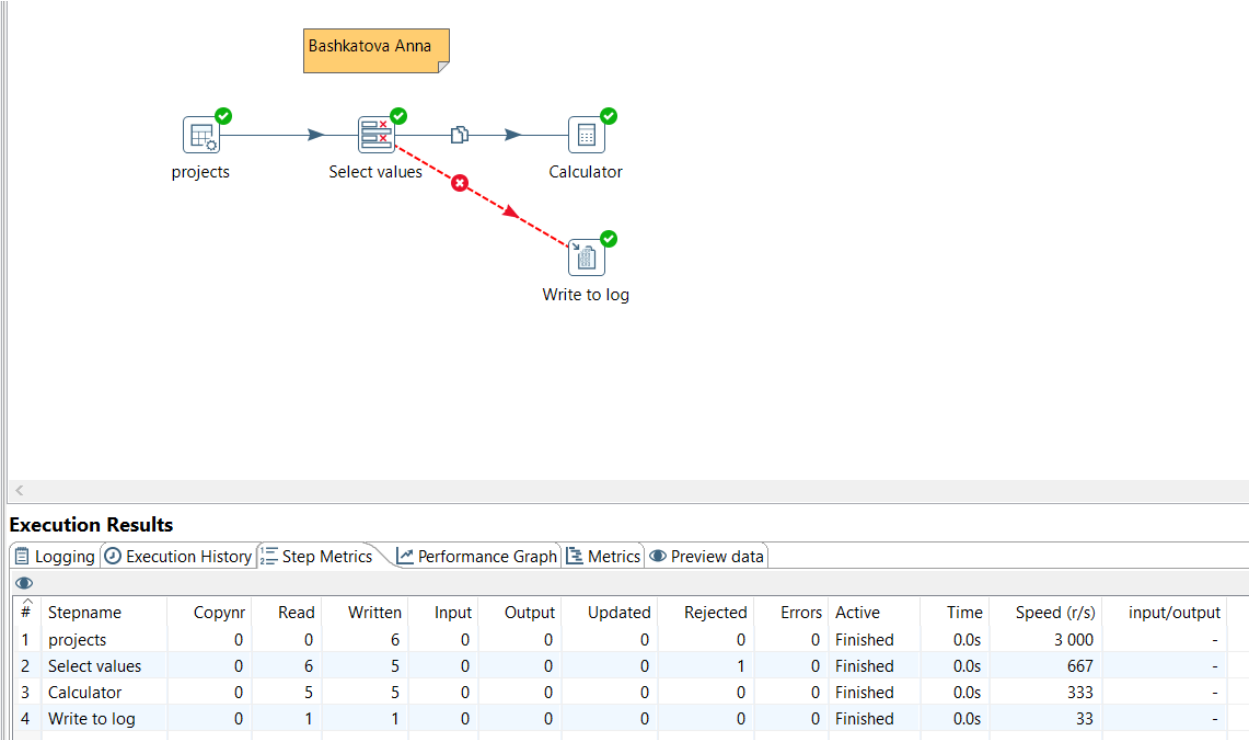
Fields to alter the meta-data for :

#	Fieldname	Rename to	Type	Length	Prec
1	estimated		Integer		

Get fields to change

Help OK Cancel

Добавляем calculator и write to log



Создаем job4.

Добавляем Microsoft Excel input

Microsoft Excel input

Step name: Microsoft Excel input

Add Field(s)

Files | Sheets | Content | Error Handling | Fields | Additional output fields

Spread sheet type (engine): Excel 97-2003 XLS (XLS)

File or directory:

Regular Expression:

Exclude Regular Expression:

Password:

Selected files:

#	File/Directory	Wildcard (RegExp)	Exclude wildcard	Required	Inc
1	D:\Ubuntu\sources\temp\sample-superstore.xls				

Accept filenames from previous steps

Accept filenames from previous step: ☐

Step to read filenames from: Step 1: Read from file

Field in the input to use as filename: File/Directory

Show filename(s)...

Help OK Preview rows Cancel

[illegible]

Microsoft Excel input

Step name

Microsoft Excel input

Files

Sheets

Content

Error Handling

Fields

Additional output fields

#	Name	Type	Length	Precision	Trim type	Repeat	Format	Currency	Decimal	Grouping
1	Row ID	Number								
2	Order ID	String								
3	Order Date	Date								
4	Ship Date	Date								
5	Ship Mode	String								
6	Customer ID	String								
7	Customer Name	String								
8	Segment	String								
9	Country	String								
1..	City	String								
1..	State	String								
1..	Postal Code	Number								
1..	Region	String								
1..	Product ID	String								
1..	Category	String								
1..	Sub-Category	String								
1..	Product Name	String								
1..	Sales	Number								

Examine preview data

Rows of step: Microsoft Excel input (1000 rows)

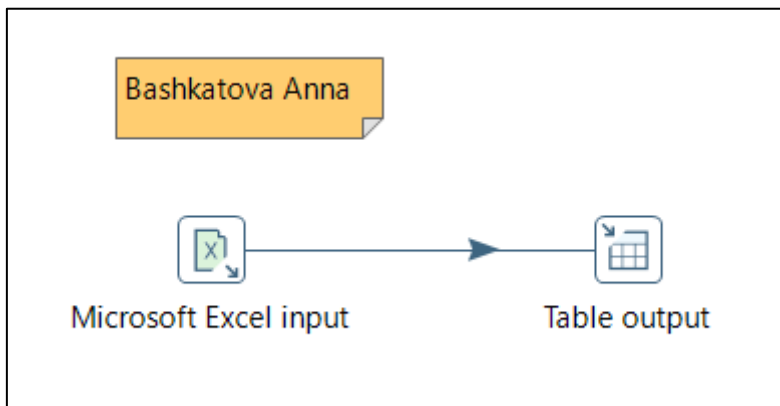
#	Row ID	Order ID	Order Date	Ship Date	Ship Mode	Customer ID	Customer Name
1	1,0	CA-2018-152156	2018/11/08 00:00:00.000	2018/11/11 00:00:00.000	Second Class	CG-12520	Claire Gute
2	2,0	CA-2018-152156	2018/11/08 00:00:00.000	2018/11/11 00:00:00.000	Second Class	CG-12520	Claire Gute
3	3,0	CA-2018-138688	2018/06/12 00:00:00.000	2018/06/16 00:00:00.000	Second Class	DV-13045	Darrin Van Huf
4	4,0	US-2017-108966	2017/10/11 00:00:00.000	2017/10/18 00:00:00.000	Standard Class	SO-20335	Sean O'Donnel
5	5,0	US-2017-108966	2017/10/11 00:00:00.000	2017/10/18 00:00:00.000	Standard Class	SO-20335	Sean O'Donnel
6	6,0	CA-2016-115812	2016/06/09 00:00:00.000	2016/06/14 00:00:00.000	Standard Class	BH-11710	Brosina Hoffm.
7	7,0	CA-2016-115812	2016/06/09 00:00:00.000	2016/06/14 00:00:00.000	Standard Class	BH-11710	Brosina Hoffm.
8	8,0	CA-2016-115812	2016/06/09 00:00:00.000	2016/06/14 00:00:00.000	Standard Class	BH-11710	Brosina Hoffm.
9	9,0	CA-2016-115812	2016/06/09 00:00:00.000	2016/06/14 00:00:00.000	Standard Class	BH-11710	Brosina Hoffm.
1..	10,0	CA-2016-115812	2016/06/09 00:00:00.000	2016/06/14 00:00:00.000	Standard Class	BH-11710	Brosina Hoffm.
1..	11,0	CA-2016-115812	2016/06/09 00:00:00.000	2016/06/14 00:00:00.000	Standard Class	BH-11710	Brosina Hoffm.
1..	12,0	CA-2016-115812	2016/06/09 00:00:00.000	2016/06/14 00:00:00.000	Standard Class	BH-11710	Brosina Hoffm.
1..	13,0	CA-2019-114412	2019/04/15 00:00:00.000	2019/04/20 00:00:00.000	Standard Class	AA-10480	Andrew Allen
1..	14,0	CA-2018-161389	2018/12/05 00:00:00.000	2018/12/10 00:00:00.000	Standard Class	IM-15070	Irene Maddox
1..	15,0	US-2017-118983	2017/11/22 00:00:00.000	2017/11/26 00:00:00.000	Standard Class	HP-14815	Harold Pawlan
1..	16,0	US-2017-118983	2017/11/22 00:00:00.000	2017/11/26 00:00:00.000	Standard Class	HP-14815	Harold Pawlan
1..	17,0	CA-2016-105893	2016/11/11 00:00:00.000	2016/11/18 00:00:00.000	Standard Class	PK-19075	Pete Kriz
1..	18,0	CA-2016-167164	2016/05/13 00:00:00.000	2016/05/15 00:00:00.000	Second Class	AG-10270	Alejandro Grov
1..	19,0	CA-2016-143336	2016/08/27 00:00:00.000	2016/09/01 00:00:00.000	Second Class	ZD-21925	Zuschuss Dona
2	20,0	CA-2016-143336	2016/08/27 00:00:00.000	2016/09/01 00:00:00.000	Second Class	ZD-21925	Zuschuss Dona

Help

Close


Show Log

Подключение к БД



Создаем job

Создаем временные переменные к рабочим папкам.



Set Environment Variables

Please enter the values of the variables or create new ones

#	Name	Value
1	HOME	D:\Ubuntu\sources\sources\temp
2	WORKFOLDER	D:\Ubuntu\sources\sources\introduction_pentaho
3		

OK Cancel

Башкатова Анна

HTTP

Start

HTTP

Name of job entry: HTTP

General

Headers

URL:

Run for every result row? ☐

Input field which contains URL:

Input field which contains upload file name:

Input field which contains destination file name:

Authentication

Username:

Password:

Proxy server for upload:

Proxy port:

Ignore proxy for hosts:

Upload file

Upload file:

Browse...

Websaver reply

Target file:

Browse...

Append to specified target file? ☐

Add date and time to file name? ☐

Target file extension:

Help

OK

Cancel

The screenshot displays the Databricks Jobs interface. At the top, a job plan shows a sequence of steps: 'Start' (green play button icon), followed by a lock icon, and then 'HTTP' (globe icon with a green checkmark). An orange callout box labeled 'Башкатова Анна' points to the 'HTTP' step. Below the job plan, the 'Execution Results' section is visible, with tabs for 'Logging', 'History', 'Job metrics', and 'Metrics'. The 'History' tab is selected, showing a list of log entries for the job execution on 2023/04/27 at 14:56:30. The log entries include: 'Spoon - Save file as...', 'Spoon - Starting job...', 'sample-super - Start of job execution', 'sample-super - Starting entry [HTTP]', 'HTTP - Start of HTTP job entry', 'HTTP - Connecting to URL: https://github.com/Data-Learn/data-engineering/blob/5fcf30ba0f67f4b7739519ef382cb51d8313b8ab/DE-101%20Modules/Module01/DE%20-%20...', 'HTTP - Resource type: Content-Type: text/html; charset=utf-8, last modified on: Thu, 27 Apr 2023 11:56:30 GMT.', 'HTTP - Finished writing 146198 bytes to result file [D:\Ubuntu\sources\sources\temp\sample-super.xls]', 'sample-super - Finished job entry [HTTP] (result=[true])', 'sample-super - Job execution finished', and 'Spoon - Job has ended.'

```
graph LR; Start[Start] --> HTTP[HTTP]; Start --> Shell[Shell];
```

The diagram illustrates a simple workflow with three nodes: 'Start', 'HTTP', and 'Shell'. The 'Start' node is at the top left, followed by two parallel paths leading to 'HTTP' and 'Shell' nodes. All nodes are marked as successful with green checkmarks. A yellow callout box labeled 'Башкатова Анна' is positioned near the 'HTTP' node.

Добавляем эксель файл

Step name

Output fields

Sheets to read	#	Sheet name	Start row	Start column
	1	Orders		

Microsoft Excel input

Step nameMicrosoft Excel input


FilesSheetsContentError HandlingFieldsAdditional output fields

#	Name	Type	Length	Precision	Trim type	Repeat	Format	Currency	Decimal	Grouping	
1	Row ID	Integer			none	N	#				
2	Order ID	String			none	N					
3	Order Date	Date			none	N	dd/MM/yyyy				
4	Ship Date	Date			none	N	dd/MM/yyyy				
5	Ship Mode	String			none	N					
6	Customer ID	String			none	N					
7	Customer Name	String			none	N					
8	Segment	String			none	N					
9	Country	String			none	N					
1..	City	String			none	N					
1..	State	String			none	N					
1..	Postal Code	String			none	N	#				
1..	Region	String			none	N					
1..	Product ID	String			none	N					
1..	Category	String			none	N					
1..	Sub-Category	String			none	N					
1..	Product Name	String			none	N					
1..	Sales	Number			none	N					
1..	Quantity	Number			none	N					
2..	Discount	Number			none	N					
2..	Profit	Number			none	N					


Get fields from header row...

Help


OKPreview rowsCancel



Orders Excel input



Returns Excel input 2



People Excel input 2 2

Sort rows

- □ ×

Step name

Sort returns

Sort directory

%java.io.tmpdir%

Browse...

TMP-file prefix

out

Sort size (rows in memory)

1000000

Free memory threshold (in %)

Compress TMP Files?

☐

Only pass unique rows? (verifies keys only)

☒

Fields:

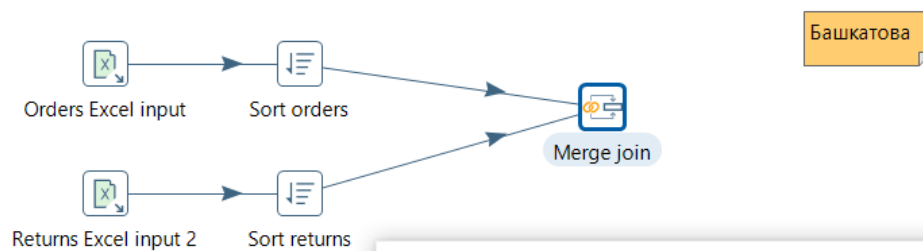
Fieldname	Ascending	Case sensitive compare?	Sort based on current locale?	Collator Strength	Presorted?
Order ID	Y				

? Help

OK

Cancel

Get Fields



Step name: Merge join

First Step: Sort orders

Second Step: Sort returns

Join Type: LEFT OUTER

Keys for 1st step:

#	Key field
1	Order ID

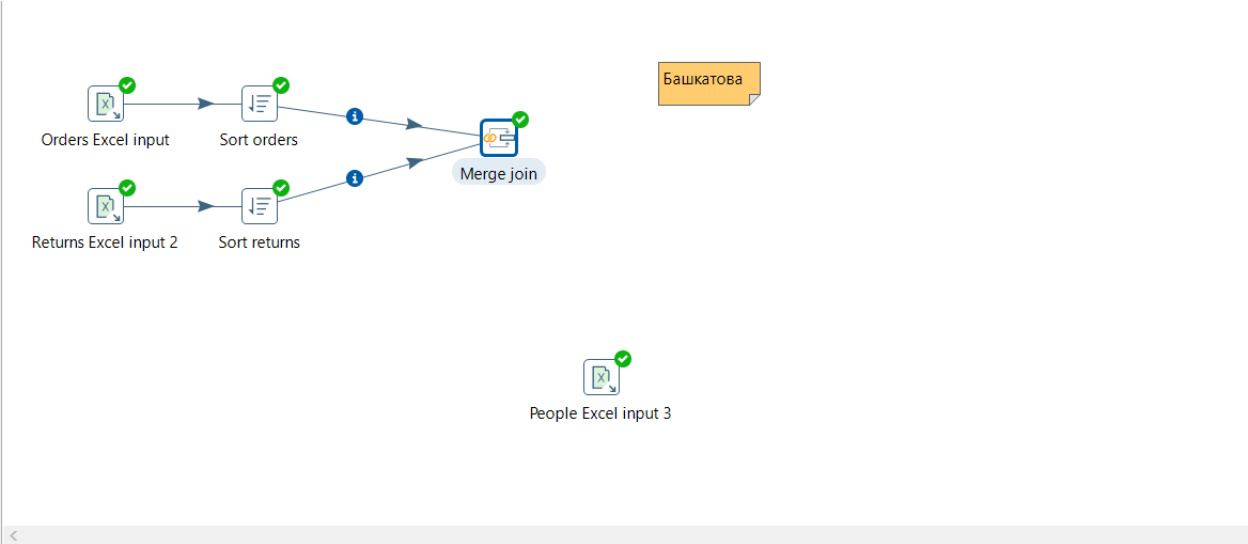
Keys for 2nd step:

#	Key field
1	Order ID

Get key fields (left)

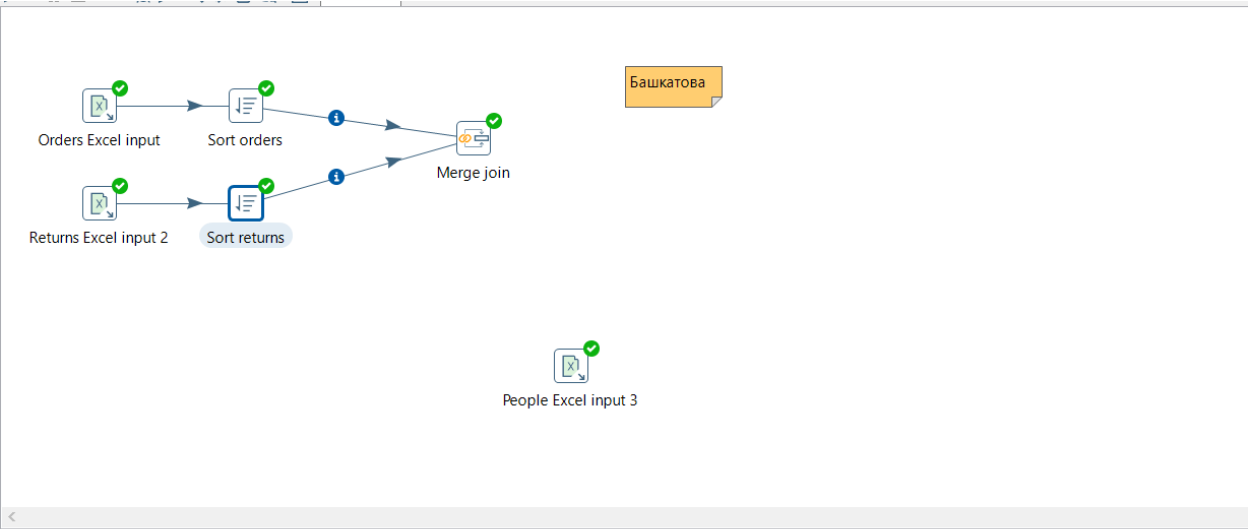
Get key fields (right)

Help OK Cancel



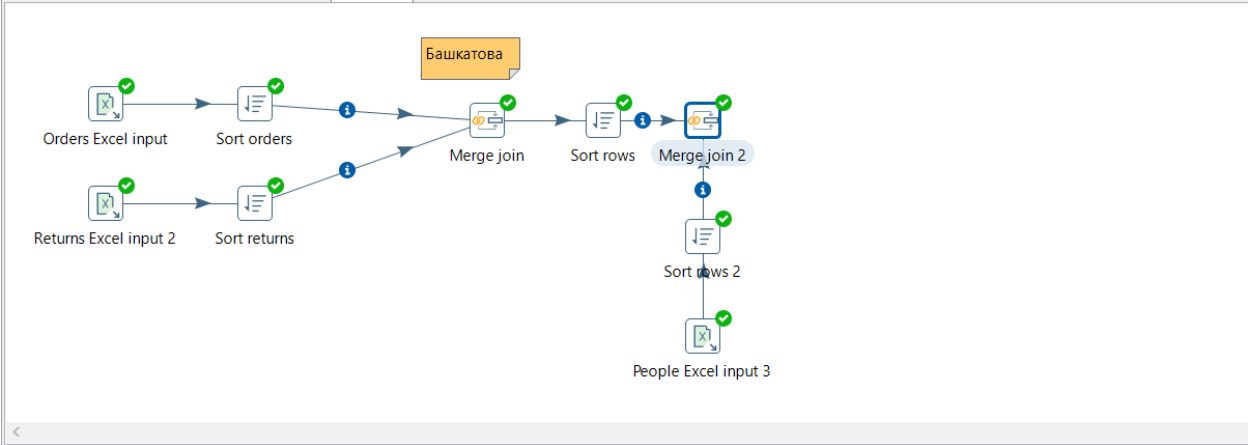
Execution Results

Logging Execution History Step Metrics Performance Graph Metrics Preview data										
First rows Last rows Off										
#	Row ID	Order ID	Order Date	Ship Date	Ship Mode	Customer ID	Customer Name	Segment	Country	City
1	2718	CA-2016-100006	07/09/2016	13/09/2016	Standard Class	DK-13375	Dennis Kane	Consumer	United States	New York
2	6288	CA-2016-100090	08/07/2016	12/07/2016	Standard Class	EB-13705	Ed Braxton	Corporate	United States	San Francisco
3	6289	CA-2016-100090	08/07/2016	12/07/2016	Standard Class	EB-13705	Ed Braxton	Corporate	United States	San Francisco
4	9515	CA-2016-100293	14/03/2016	18/03/2016	Standard Class	NF-18475	Neil Franzusich	Home Office	United States	Jacksonville
5	3084	CA-2016-100328	28/01/2016	03/02/2016	Standard Class	JC-15340	Jasper Cacioppo	Consumer	United States	New York
6	3836	CA-2016-100363	08/04/2016	15/04/2016	Standard Class	JM-15655	Jim Mitchum	Corporate	United States	Glendale
7	3837	CA-2016-100363	08/04/2016	15/04/2016	Standard Class	JM-15655	Jim Mitchum	Corporate	United States	Glendale
8	9441	CA-2016-100391	25/05/2016	29/05/2016	Standard Class	BW-11065	Barry Weirich	Consumer	United States	New York
9	6560	CA-2016-100670	10/01/2016	22/01/2016	Standard Class	KA-16730	Karen Miller	Consumer	United States	Mountain View



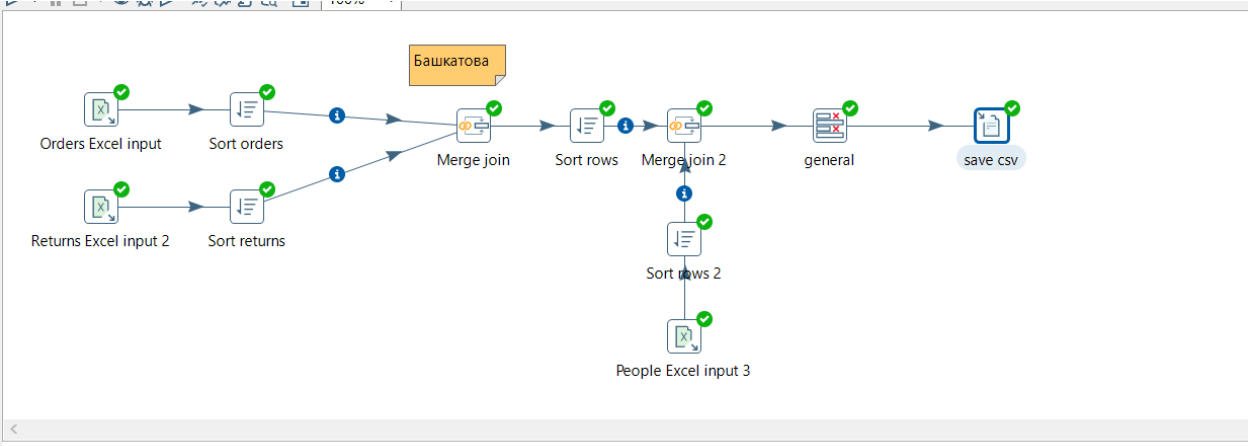
Execution Results

Logging Execution History Step Metrics Performance Graph Metrics Preview data													
#	Stepname	Copynr	Read	Written	Input	Output	Updated	Rejected	Errors	Active	Time	Speed (r/s)	input/output
1	Orders Excel input	0	0	9994	9994	0	0	0	0	Finished	1.8s	5 656	-
2	People Excel input 3	0	0	4	4	0	0	0	0	Finished	1.2s	3	-
3	Returns Excel input 2	0	0	800	800	0	0	0	0	Finished	1.2s	654	-
4	Sort returns	0	800	296	0	0	0	0	0	Finished	1.4s	566	-
5	Sort orders	0	9994	9994	0	0	0	0	0	Finished	1.8s	5 467	-
6	Merge join	0	10290	9994	0	0	0	0	0	Finished	2.2s	4 650	-



Execution Results

Logging Execution History Step Metrics Performance Graph Metrics Preview data													
#	Stepname	Copynr	Read	Written	Input	Output	Updated	Rejected	Errors	Active	Time	Speed (r/s)	input/output
1	People Excel input 3	0	0	4	4	0	0	0	0	Finished	0.8s	5	-
2	Orders Excel input	0	0	9994	9994	0	0	0	0	Finished	1.6s	6 223	-
3	Returns Excel input 2	0	0	800	800	0	0	0	0	Finished	1.2s	691	-
4	Sort orders	0	9994	9994	0	0	0	0	0	Finished	1.6s	6 068	-
5	Sort rows 2	0	4	4	0	0	0	0	0	Finished	1.2s	3	-
6	Sort returns	0	800	296	0	0	0	0	0	Finished	1.3s	606	-
7	Merge join	0	10290	9994	0	0	0	0	0	Finished	2.0s	5 247	-
8	Sort rows	0	9994	9994	0	0	0	0	0	Finished	2.0s	5 002	-
9	Merge join 2	0	9998	9994	0	0	0	0	0	Finished	2.4s	4 197	-



Execution Results

Logging Execution History Step Metrics Performance Graph Metrics Preview data													
#	Stepname	Copynr	Read	Written	Input	Output	Updated	Rejected	Errors	Active	Time	Speed (r/s)	input/output
1	Returns Excel input 2	0	0	800	800	0	0	0	0	Finished	1.1s	700	-
2	People Excel input 3	0	0	4	4	0	0	0	0	Finished	1.1s	3	-
3	Orders Excel input	0	0	9994	9994	0	0	0	0	Finished	1.6s	6 196	-
4	Sort returns	0	800	296	0	0	0	0	0	Finished	1.3s	603	-
5	Sort orders	0	9994	9994	0	0	0	0	0	Finished	1.7s	6 050	-
6	Sort rows 2	0	4	4	0	0	0	0	0	Finished	1.3s	3	-
7	Merge join	0	10290	9994	0	0	0	0	0	Finished	2.0s	5 245	-
8	Sort rows	0	9994	9994	0	0	0	0	0	Finished	2.0s	4 972	-
9	Merge join 2	0	9998	9994	0	0	0	0	0	Finished	2.4s	4 231	-
1..	general	0	9994	9994	0	0	0	0	0	Finished	2.4s	4 192	-
1..	save csv	0	9994	9994	0	9994	0	0	0	Finished	2.4s	4 135	-