

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»

дз_01

Тема:

**«ETL Компоненты и начало работы с ETL на примере Pentaho Data
Integration»**

Выполнила: Николаева С.Г., АДЭУ-201

Преподаватель: Босенко Тимур Муртазович

Москва

2023



Упражнения на Pentaho DI

- Запустим Pentaho DI (если еще не сделали)
- Посмотрим примеры трансформация (копию будет на git)
- Подключимся к PostgreSQL (localhost)

#	message	time	id
1	Nikolaeva Sofiya	2019-01-01	1
2	Nikolaeva Sofiya	2019-01-01	2
3	Nikolaeva Sofiya	2019-01-01	3
4	Nikolaeva Sofiya	2019-01-01	4
5	Nikolaeva Sofiya	2019-01-01	5
6	Nikolaeva Sofiya	2019-01-01	6
7	Nikolaeva Sofiya	2019-01-01	7

Задание 1



Generate rows

Generate rows

Step name: Generate rows

Limit: 10

Never stop generating rows: ☐

Interval in ms (delay): 5000

Current row time field name: now

Previous row time field name: FiveSecondsAgo

Fields:

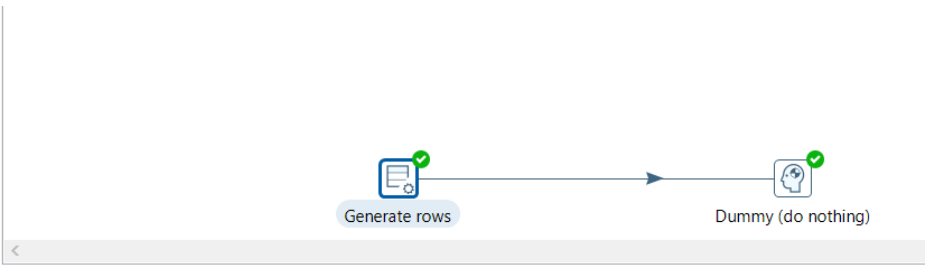
#	Name	Type	Format	Length	Precision	Currency	Decimal	Group	Value	Set empty
1	message	String							Nikolaeva Sofiya	

Help OK Preview Cancel

Examine preview data

Rows of step: Generate rows (10 rows)

#	message
1	Nikolaeva Sofiya
2	Nikolaeva Sofiya
3	Nikolaeva Sofiya
4	Nikolaeva Sofiya
5	Nikolaeva Sofiya
6	Nikolaeva Sofiya
7	Nikolaeva Sofiya
8	Nikolaeva Sofiya
9	Nikolaeva Sofiya
10	Nikolaeva Sofiya



Execution Results

Logging	Execution History	Step Metrics	Performance Graph	Metrics	Preview data
<input checked="" type="radio"/> First rows	<input type="radio"/> Last rows	<input type="radio"/> Off			
#	message				
1	Nikolaeva Sofiya				
2	Nikolaeva Sofiya				
3	Nikolaeva Sofiya				
4	Nikolaeva Sofiya				
5	Nikolaeva Sofiya				
6	Nikolaeva Sofiya				
7	Nikolaeva Sofiya				
8	Nikolaeva Sofiya				
9	Nikolaeva Sofiya				
1..	Nikolaeva Sofiya				

ЗАДАНИЕ 2




Data grid

Data grid

Step name

Meta Data

#	project_name	start_date	end_date
1	nikol_1	2000-01-03	2023-04-04
2	nikol_2	2020-01-03	2023-04-04
3	nikol_3	2005-01-03	2023-04-04
4	nikol_4	2009-01-03	2023-04-04
5	nikol_5	2021-01-03	2023-04-04
6			

 Examine preview data

Rows of step: Data grid (6 rows)

#	project_name	start_date	end_date
1	nikol_1	2000-01-03	2023-04-04
2	nikol_2	2020-01-03	2023-04-04
3	nikol_3	2005-01-03	2023-04-04
4	nikol_4	2009-01-03	2023-04-04
5	nikol_5	2021-01-03	2023-04-04



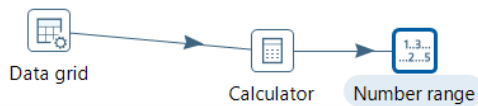
Calculator

Step name

☒ Throw an error on non existing files

Fields:

#	New field	Calculation	Field A	Field B	Field C	Value type	Length	Precision	Remove
1	diff_dates	Date A - Date B (in days)	end_date	start_date		Integer			N



Number range

Step name:

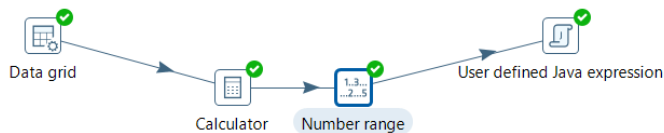
Input field:

Output field:

Default value(if no range):

Ranges (min <= x < max):

#	Lower Bound	Upper Bound	Value
1		1200.0	excellent
2	1200.0	3000.0	very good
3	3000.0	5000.0	good
4	5000.0		poor



Execution Results

Logging Execution History Step Metrics Performance Graph Metrics Preview data

☒ First rows ☐ Last rows ☐ Off

#	project_name	start_date	end_date	diff_dates	range
1	nikol_1	2000-01-03	2023-04-04	8492	poor
2	nikol_2	2020-01-03	2023-04-04	1187	excellent
3	nikol_3	2005-01-03	2023-04-04	6665	poor
4	nikol_4	2009-01-03	2023-04-04	5204	poor
5	nikol_5	2021-01-03	2023-04-04	821	excellent
6	nikol_6	2001-04-01	2023-01-01	7945	poor



Data grid

Data grid

Step name Data grid

Meta Data

#	Name	Type	Format
1	project_name	String	
2	start_date	Date	yyyy-MM-dd
3	end_date	Date	yyyy-MM-dd
4	estimated	String	



Data grid

Data grid

Step name Data grid

Meta Data

#	project_name	start_date	end_date	estimated
1	nik_1	2022-04-04	2023-04-04	366
2	nik_2	2021-04-04	2023-04-04	731
3	nik_3	2020-04-04	2023-04-04	1096



Data grid



Select values

Select values

Step name Select values

Select & Alter Remove Meta-data

Fields to alter the meta-data for :

#	Fieldname	Rename to	Type	Ler
1	estimated		Integer	▼
<				

Help

OK

Diagram showing a flow from **Data grid** to **Select values**.

Data grid configuration:

Step name: Data grid

#	project_name	start_date	end_date	estimated
1	nik_1	2022-04-04	2023-04-04	366
2	nik_2	2021-04-04	2023-04-04	731
3	nik_3	2020-04-04	2023-04-04	1096
4	nik_3	2020-04-04	2023-04-04	hello

Diagram showing a flow from **Data grid** to **Select values** to **Calculator**.

Calculator configuration:

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	diff	Date A - Date B (in days)	end_date	start_date		None			N

Diagram showing a flow from **Data grid** to **Select values** to **Calculator** to **Write to log**.

Write to log configuration:

Step name: Write to log

Log level: Basic

Print header: ☒

Limit rows?: ☐

Nr of rows to print: 0

Write to log:

#	Field
1	project_name
2	start_date
3	end_date
4	estimated

Buttons: Help, OK, Get Fields, Cancel

Log output:

```

j.0 - -----> Linenr 1-----
j.0 - project_name = nik_3
j.0 - start_date = 2020-04-04
j.0 - end_date = 2023-04-04
  
```

РАЗДЕЛЕНИЕ ПОТОКА: ОШИБКА – В ЛОГ, КОГДА ВСЕ ХОРОШО СРАБОТАЛО - ПРОХОДИТ

2023/04/04 16:09:37 - Write to log.0 - estimated = hello

Execution Results

Logging	Execution History	Step Metrics	Performance Graph	Metrics	Preview data
<input checked="" type="radio"/> First rows <input type="radio"/> Last rows <input type="radio"/> Off					
#	project_name	start_date	end_date	estimated	diff
1	nik_1	2022-04-04	2023-04-04	366	365
2	nik_2	2021-04-04	2023-04-04	731	730
3	nik_3	2020-04-04	2023-04-04	1096	1095

Задание 4

Практика

- Скачаем [sample-superstore.xls](#) из 1 модуля. (1 job)
- Объединим данные из 3 таблиц в одну. (1 transformation)
- Разобьем данные на разные форматы (2 transformation)
 - Информацию по продуктам сохраним в JSON формате
 - Информацию о возвратах сохраним в формате XML
 - Информацию о заказах разобьем по регионам:
 - CENTRAL – Одним файлом в формате Excel (xls)
 - WEST – Несколько файлов разбитых по штатам в csv
 - SOUTH – Один файл формата csv в zip архиве
 - EAST – текстовый файл с расширением .dat
- Добавляем "ошибки" для большего реализма :D (3 transformation)
 - WEST – разные названия страны (US, United States, USA), лишние символы в поле City
 - EAST – добавляем опечатки в названиях городов (сложно прогнозируемые для ручного исправления)
 - SOUTH – добавляем дубли заказов



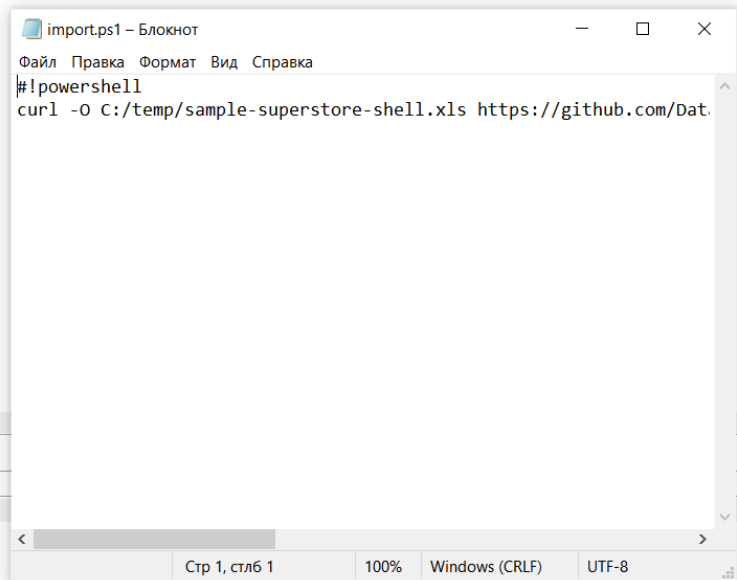
Set Environment Variables

Please enter the values of the variables or create new ones

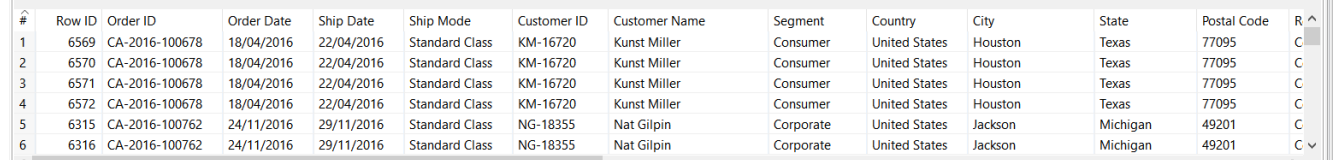
#	Name	Value
1	HOME	D:\учеба\мгпу\6 семестр\big data\homework1\sources\temp
2	WORKFOLDER	D:\учеба\мгпу\6 семестр\big data\homework1\sources\introduction_pentaho

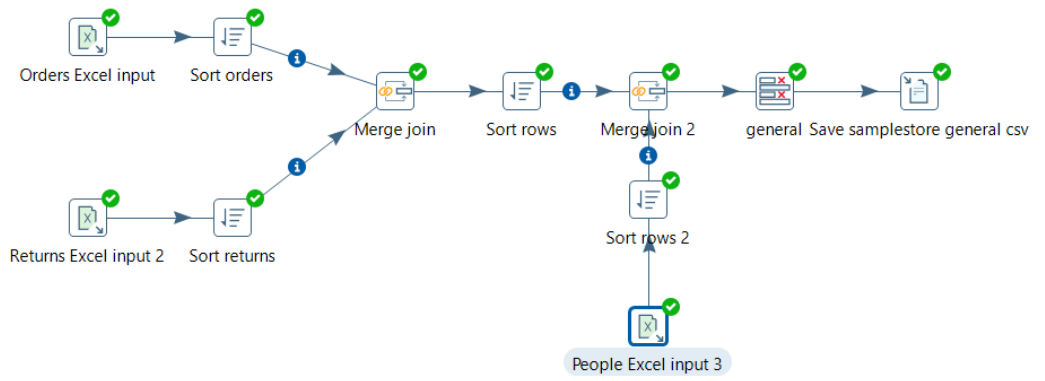
OK

Cancel



```
023/04/04 17:15:43 - Spoon - Starting job...
023/04/04 17:15:43 - job_download_samplestore - Start of job execution
023/04/04 17:15:43 - job_download_samplestore - Starting entry [HTTP]
023/04/04 17:15:43 - HTTP - Start of HTTP job entry.
023/04/04 17:15:43 - HTTP - Connecting to URL: https://github.com/Data-Learn/data-engineering/raw/master/DE-101%20Modules/Module01/DE%20-%20101%20Lab%201.1/Sample%20-%20
023/04/04 17:15:45 - HTTP - Resource type: Content-Type: application/octet-stream, last modified on: Tue, 04 Apr 2023 14:15:44 GMT.
023/04/04 17:15:46 - HTTP - Finished writing 3384320 bytes to result file [D:\ye6a6\mryny6 cemeacr\big data\homework1\sources\temp\sample-superstore.xls]
023/04/04 17:15:46 - job_download_samplestore - Finished job entry [HTTP] (result=[true])
023/04/04 17:15:46 - job_download_samplestore - Starting entry [Shell]
023/04/04 17:15:46 - Shell - Running on platform: Windows 10
023/04/04 17:15:46 - Shell - Executing command : cmd.exe /C ""D:\ye6a6\mryny6 cemeacr\big data\homework1\sources\introduction_pentaho\import.ps1""
```





Execution Results

Logging				Execution History				Step Metrics				Performance Graph				Metrics				Preview data			
<input checked="" type="radio"/> First rows				<input type="radio"/> Last rows				<input type="radio"/> Off															
#	Person	Region																					
1	Anna Andreadi	West																					
2	Chuck Magee	East																					
3	Kelly Williams	Central																					
4	Cassandra Brandow	South																					