

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

Домашнее задание 1

Тема:

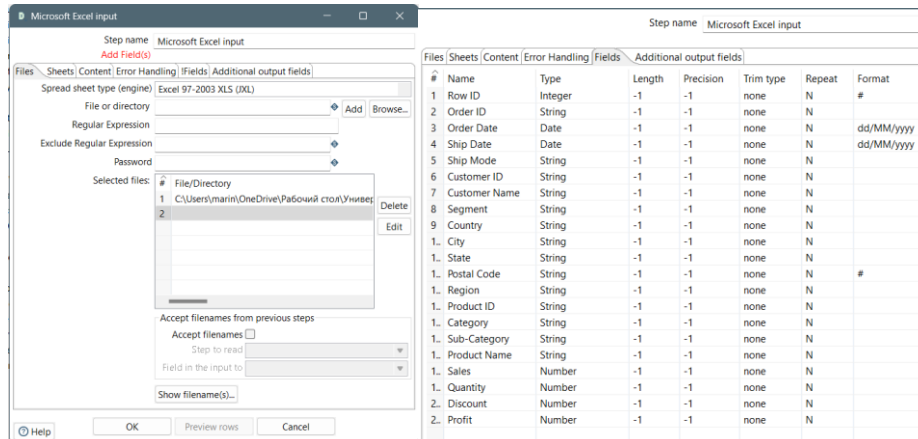
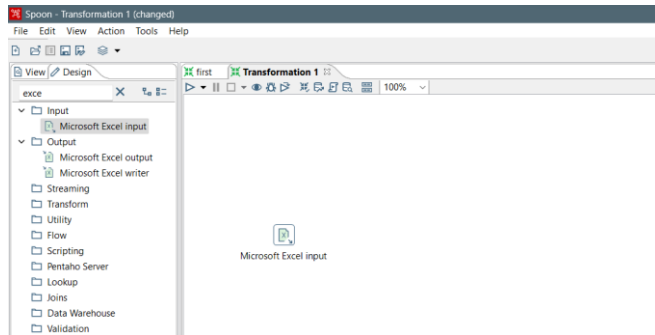
Pentaho.

Выполнила: Соколова М. С., группа: АДЭУ-201

Преподаватель: Босенко Т. М.

Москва

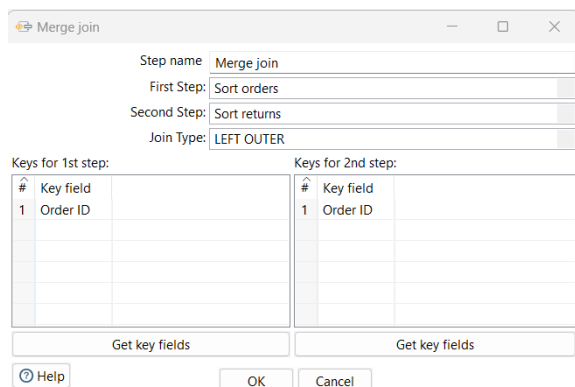
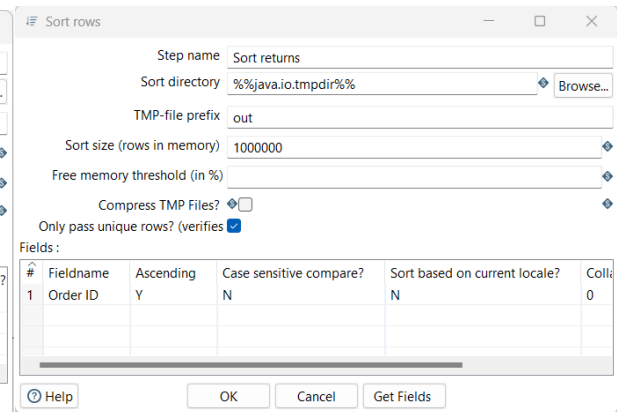
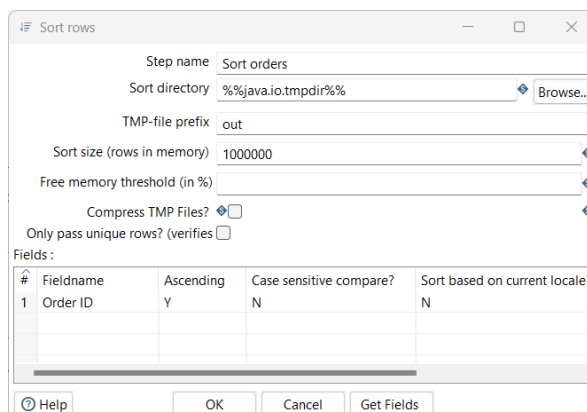
2023



Orders Excel input

People Excel input 2

Returns Excel input 3



Sort rows

Step name: Sort rows

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?
1	Region	Y	N	N	0	N

Help OK Cancel Get Fields

Sort rows

Step name: Sort people

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?
1	Region	Y	N	N	0	N

Help OK Cancel Get Fields

Merge join

Step name: Merge join 2

First Step: Sort rows

Second Step: Sort people

Join Type: LEFT OUTER

Keys for 1st step:

#	Key field
1	Region

Get key fields

Keys for 2nd step:

#	Key field
1	Region

Get key fields

Help OK Cancel

Select values

Step name: general

Select & Alter Remove Meta-data

Fields to remove:

#	Fieldname
1	OrderID_1
2	Region_1

Get fields to remove

Help OK Cancel

Text file output

Step name: Save samplestore general csv

File Content Fields

Filename: C:\Users\marini\OneDrive\Рабочий стол\Универ\3 курс\3 семестр\Устройства Browse...

Pass output to servlet? ☐

Create Parent folder? ☒

Do not create file at start? ☐

Accept file name from field? ☐

File name field:

Extension: csv

Include stepnr in filename? ☐

Include partition nr in filename? ☐

Include date in filename? ☐

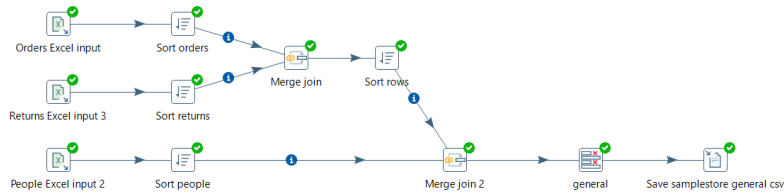
Include time in filename? ☐

Specify Date time format:

Show filename(s):

Add filenames to result? ☒

Help OK Cancel



Text file input

Step name: General text file

File (Content \ Error Handling \ Filters \ Fields \ Additional output fields)

File or directory

Regular Expression

Exclude Regular Expression

Selected files:

- 1. C:\Users\marin\OneDrive\Рабочий стол\Универ\3 курс\3 курс 2 семестр\Инструменты для
- 2.

Accept filenames from previous steps

Accept filenames from previous step ☐

Pass through fields from previous step ☐

Step to read filenames from

Field in the input to use as filename

Show filename(s) Show file content Show content from first data line

Help OK Preview rows Cancel

Text file input

Step name: General text file

File (Content \ Error Handling \ Filters \ Fields \ Additional output fields)

Filetype: CSV

Separator: ;

Enclosure: "

Escape: \

Header: ☒ Number of header lines: 1

Footer: ☐ Number of footer lines: 1

Wrapped lines: ☐ Number of times wrapped: 1

Paged layout (printout): ☐ Number of lines per page: 80

Document header lines: 0

Compression: None

No empty rows: ☒

Include filename in output: ☐ Filename fieldname

Rownum in output: ☐ Rownum fieldname

Rownum by file? ☐

Format: mixed

Encoding: mixed

Length: Characters

Limit: 0

Be lenient when parsing dates: ☒

The date format locale: ru\_RU

Base file filename

Help OK Preview rows Cancel

Text file input

Step name: General text file

File (Content \ Error Handling \ Filters \ Fields \ Additional output fields)

#	Name	Type	Format	Position	Length	Precision	Currency	Decimal	Group	Null if	Default
1	Row_ID	Integer		15	15	0	py6.	-	-	-	-
2	Order_ID	String		14	14		py6.	-	-	-	-
3	Order_Date	Date	dd/MM/yyyy				py6.	-	-	-	-
4	Ship_Date	Date	dd/MM/yyyy				py6.	-	-	-	-
5	Ship_Mode	String		14	14		py6.	-	-	-	-
6	Customer_ID	String		8	8		py6.	-	-	-	-
7	Customer_Name	String		19	19		py6.	-	-	-	-
8	Segment	String		11	11		py6.	-	-	-	-
9	Country	String		13	13		py6.	-	-	-	-
1.	City	String		13	13		py6.	-	-	-	-
1.	State	String		12	12		py6.	-	-	-	-
1.	Postal_Code	Integer	#	15	15	0	py6.	-	-	-	-
1.	Region	String		7	7		py6.	-	-	-	-
1.	Product_ID	String		15	15		py6.	-	-	-	-
1.	Category	String		15	15		py6.	-	-	-	-
1.	Sub_Category	String		11	11		py6.	-	-	-	-
1.	Product_Name	String		85	85		py6.	-	-	-	-
1.	Sales	Number	##	15	15	0	py6.	-	-	-	-
1.	Quantity	Number	##	15	15	0	py6.	-	-	-	-
2.	Discount	Number	##	15	15	0	py6.	-	-	-	-

Get Fields Minimal width

Help OK Preview rows Cancel

Examine preview data

Rows of step: General text file (100 rows)

#	Row_ID	Order_ID	Order_Date	Ship_Date	Ship_Mode	Customer_ID	Customer_Name	Seg
1	6569	CA-2016-100678	18/04/2016	22/04/2016	Standard Class	KM-16720	Kunst Miller	Con
2	6570	CA-2016-100678	18/04/2016	22/04/2016	Standard Class	KM-16720	Kunst Miller	Con
3	6571	CA-2016-100678	18/04/2016	22/04/2016	Standard Class	KM-16720	Kunst Miller	Con
4	6572	CA-2016-100678	18/04/2016	22/04/2016	Standard Class	KM-16720	Kunst Miller	Con
5	6315	CA-2016-100762	24/11/2016	29/11/2016	Standard Class	NG-18355	Nat Gilpin	Con
6	6316	CA-2016-100762	24/11/2016	29/11/2016	Standard Class	NG-18355	Nat Gilpin	Con
7	6317	CA-2016-100762	24/11/2016	29/11/2016	Standard Class	NG-18355	Nat Gilpin	Con
8	6318	CA-2016-100762	24/11/2016	29/11/2016	Standard Class	NG-18355	Nat Gilpin	Con
9	6252	CA-2016-101147	02/12/2016	04/12/2016	First Class	MC-17575	Matt Collins	Con
1.	1575	CA-2016-101602	15/12/2016	18/12/2016	First Class	MC-18100	Mick Crebaggia	Con
1.	1576	CA-2016-101602	15/12/2016	18/12/2016	First Class	MC-18100	Mick Crebaggia	Con
1.	9558	CA-2016-103086	17/10/2016	19/10/2016	Second Class	EB-14170	Evan Bailliet	Con
1.	4283	CA-2016-103100	20/12/2016	23/12/2016	First Class	AB-10105	Adrian Barton	Con
1.	4284	CA-2016-103100	20/12/2016	23/12/2016	First Class	AB-10105	Adrian Barton	Con
1.	5725	CA-2016-103191	22/09/2016	27/09/2016	Standard Class	VG-21805	Vivek Grady	Con
1.	7546	CA-2016-103492	10/10/2016	15/10/2016	Standard Class	CM-12715	Craig Molinari	Con
1.	7547	CA-2016-103492	10/10/2016	15/10/2016	Standard Class	CM-12715	Craig Molinari	Con

Close Show Log

Sort rows

Step name: Sort rows

Sort directory: %java.io.tmpdir%\%

TMP-file prefix: out

Sort size (rows in memory): 1000000

Free memory threshold (in %):

Compress TMP files? ☐

Only pass unique rows? verifies ☒

Fields:

#	Fieldname	Ascending	Case sensitive compare?	Sort based on current locale?
1	Product_ID	Y	N	N

Help OK Cancel Get Fields

JSON output

Step name: JSON output

General \ Fields

Operation: Write to file

Settings

Json bloc name: data

Nr rows in a bloc: 1000000

Output Value: outputValue

Compatibility mode: ☐

Output File

Filename: C:\Users\marin\OneDrive\Рабочий стол\Ун\ Browse...

Append: ☐

Create Parent folder: ☒

Do not open create at start: ☐

Extension: js

Encoding: UTF-8

Help OK Cancel



products\_0.js

Файл Изменить Просмотр

```

{
  "data": [
    {
      "Sub_Category": "Bookcases",
      "Category": "Furniture",
      "Product_ID": "FUR-B0-10000112",
      "Product_Name": "Bush Birmingham Collection Bookcase, Dark Cherry"
    },
    {
      "Sub_Category": "Bookcases",
      "Category": "Furniture",
      "Product_ID": "FUR-B0-10000330",
      "Product_Name": "Sauder Camden County Barrister Bookcase, Planked Cherry Finish"
    },
    {
      "Sub_Category": "Bookcases",
      "Category": "Furniture",
      "Product_ID": "FUR-B0-10000362",
      "Product_Name": "Sauder Inglewood Library Bookcases"
    },
    {
      "Sub_Category": "Bookcases",
      "Category": "Furniture",
      "Product_ID": "FUR-B0-10000468",
      "Product_Name": "O'Sullivan 2-Shelf Heavy-Duty Bookcases"
    },
    {
      "Sub_Category": "Bookcases",
      "Category": "Furniture",
      "Product_ID": "FUR-B0-10000711",
      "Product_Name": "Hon Metal Bookcases, Gray"
    },
    {
      "Sub_Category": "Bookcases",
      "Category": "Furniture",
      "Product_ID": "FUR-B0-10000780",
      "Product_Name": "O'Sullivan Plantations 2-Door Library in Landvery Oak"
    },
    {
      "Sub_Category": "Bookcases",
      "Category": "Furniture",
      "Product_ID": "FUR-B0-10001337",
      "Product_Name": "O'Sullivan Living Dimensions 2-Shelf Bookcases"
    },
    {
      "Sub_Category": "Bookcases",
      "Category": "Furniture",
      "Product_ID": "FUR-B0-10001519",
      "Product_Name": "O'Sullivan 3-Shelf Heavy-Duty Bookcases"
    },
    {
      "Sub_Category": "Bookcases",
      "Category": "Furniture",
      "Product_ID": "FUR-B0-10001567",
      "Product_Name": "Bush Westfield Collection Bookcases, Dark Cherry Finish, Fully Assembled"
    },
    {
      "Sub_Category": "Bookcases",
      "Category": "Furniture",
      "Product_ID": "FUR-B0-10001601",
      "Product_Name": "Sauder Mission Library with Doors, Fruitwood Finish"
    },
    {
      "Sub_Category": "Bookcases",
      "Category": "Furniture",
      "Product_ID": "FUR-B0-10001608",
      "Product_Name": "Hon Metal Bookcases, Black"
    },
    {
      "Sub_Category": "Bookcases",
      "Category": "Furniture",
      "Product_ID": "FUR-B0-10001619",
      "Product_Name": "O'Sullivan Cherrywood Estates Traditional Bookcase"
    },
    {
      "Sub_Category": "Bookcases",
      "Category": "Furniture",
      "Product_ID": "FUR-B0-10001798",
      "Product_Name": "Bush Somerset Collection Bookcase"
    },
    {
      "Sub_Category": "Bookcases",
      "Category": "Furniture",
      "Product_ID": "FUR-B0-10001811",
      "Product_Name": "Atlantic Metals Mobile 5-Shelf Bookcases, Custom Colors"
    },
    {
      "Sub_Category": "Bookcases",
      "Category": "Furniture",
      "Product_ID": "FUR-B0-10001918",
      "Product_Name": "Sauder Forest Hills Library with Doors, Woodland Oak Finish"
    },
    {
      "Sub_Category": "Bookcases",
      "Category": "Furniture",
      "Product_ID": "FUR-B0-10001972",
      "Product_Name": "O'Sullivan 4-Shelf Bookcase in Odessa"
    }
  ]
}
  
```

Sort rows

Step name: **Sort rows 2**

Sort directory: %%java.io.tmpdir%%

TMP-file prefix: out

Sort size (rows in memory): 1000000

Free memory threshold (in %):

Compress TMP Files? ☐

Only pass unique rows? (verifies keys) ☒

Fields:

#	Fieldname	Ascending	Case sensitive compare?	Sort based on current locale?	Collator
1	Order_ID	Y	N	N	0

Filter rows

Step name: **Filter rows**

Send 'true' data to step: XML output

Send 'false' data to step:

The condition:

XML output

Step name: **XML output**

File Content Fields

Filename: C:\Users\marin\OneDrive\Рабочий стол\Универ\3 курс\3

Do not create file at start ☐

Pass output to servlet ☐

Extension: xml

Include stepnr in filename? ☐

Include date in filename? ☐

Include time in filename? ☐

Specify Date time format:

Date time format:

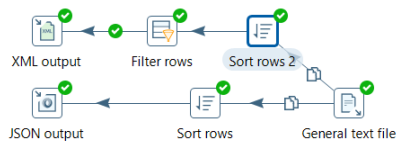
Add filenames to result ☐

XML output

Step name: **XML output**

File Content Fields

#	Fieldname	Element name	Content type	Type	Format
1	Order_ID		Element	String	
2	Returned		Element	String	



```

<?xml version='1.0' encoding='UTF-8'?>
<Rows>
<Row><Order_ID>CA-2016-100762</Order_ID> <Returned>Yes</Returned></Row>
<Row><Order_ID>CA-2016-100867</Order_ID> <Returned>Yes</Returned></Row>
<Row><Order_ID>CA-2016-102652</Order_ID> <Returned>Yes</Returned></Row>
<Row><Order_ID>CA-2016-103373</Order_ID> <Returned>Yes</Returned></Row>
<Row><Order_ID>CA-2016-103744</Order_ID> <Returned>Yes</Returned></Row>
<Row><Order_ID>CA-2016-103940</Order_ID> <Returned>Yes</Returned></Row>
<Row><Order_ID>CA-2016-104829</Order_ID> <Returned>Yes</Returned></Row>
<Row><Order_ID>CA-2016-105270</Order_ID> <Returned>Yes</Returned></Row>
<Row><Order_ID>CA-2016-108609</Order_ID> <Returned>Yes</Returned></Row>
<Row><Order_ID>CA-2016-108861</Order_ID> <Returned>Yes</Returned></Row>
<Row><Order_ID>CA-2016-109918</Order_ID> <Returned>Yes</Returned></Row>
<Row><Order_ID>CA-2016-110786</Order_ID> <Returned>Yes</Returned></Row>
<Row><Order_ID>CA-2016-111871</Order_ID> <Returned>Yes</Returned></Row>
<Row><Order_ID>CA-2016-116785</Order_ID> <Returned>Yes</Returned></Row>
<Row><Order_ID>CA-2016-123225</Order_ID> <Returned>Yes</Returned></Row>
<Row><Order_ID>CA-2016-123253</Order_ID> <Returned>Yes</Returned></Row>
<Row><Order_ID>CA-2016-123498</Order_ID> <Returned>Yes</Returned></Row>
<Row><Order_ID>CA-2016-124688</Order_ID> <Returned>Yes</Returned></Row>
<Row><Order_ID>CA-2016-126361</Order_ID> <Returned>Yes</Returned></Row>
<Row><Order_ID>CA-2016-126403</Order_ID> <Returned>Yes</Returned></Row>
<Row><Order_ID>CA-2016-126522</Order_ID> <Returned>Yes</Returned></Row>
<Row><Order_ID>CA-2016-127012</Order_ID> <Returned>Yes</Returned></Row>
<Row><Order_ID>CA-2016-127131</Order_ID> <Returned>Yes</Returned></Row>
<Row><Order_ID>CA-2016-133690</Order_ID> <Returned>Yes</Returned></Row>
<Row><Order_ID>CA-2016-134726</Order_ID> <Returned>Yes</Returned></Row>
<Row><Order_ID>CA-2016-135657</Order_ID> <Returned>Yes</Returned></Row>
<Row><Order_ID>CA-2016-135699</Order_ID> <Returned>Yes</Returned></Row>
<Row><Order_ID>CA-2016-140816</Order_ID> <Returned>Yes</Returned></Row>
<Row><Order_ID>CA-2016-141726</Order_ID> <Returned>Yes</Returned></Row>
<Row><Order_ID>CA-2016-142769</Order_ID> <Returned>Yes</Returned></Row>
<Row><Order_ID>CA-2016-143336</Order_ID> <Returned>Yes</Returned></Row>
<Row><Order_ID>CA-2016-143840</Order_ID> <Returned>Yes</Returned></Row>
<Row><Order_ID>CA-2016-148614</Order_ID> <Returned>Yes</Returned></Row>
<Row><Order_ID>CA-2016-148950</Order_ID> <Returned>Yes</Returned></Row>
<Row><Order_ID>CA-2016-151162</Order_ID> <Returned>Yes</Returned></Row>
<Row><Order_ID>CA-2016-152345</Order_ID> <Returned>Yes</Returned></Row>
<Row><Order_ID>CA-2016-153150</Order_ID> <Returned>Yes</Returned></Row>
  
```

Filter rows

Step name: Filter rows 2

Send 'true' data to step: Central Excel output

Send 'false' data to step:

The condition:

Region = Central (String)

Help OK Cancel

Microsoft Excel output

Step name: Central Excel output

File Content Custom Fields

Filename: C:\Users\marin\OneDrive\Рабочий стол\ Browse...

Create Parent folder ☒

Do not create file at start ☐

Extension: xls

Include stepnr in filename? ☐

Include date in filename? ☐

Include time in filename? ☐

Specify Date time format ☐

Date time format:

Show filename(s)...

Add filenames to result ☒

Help OK Cancel

Microsoft Excel output

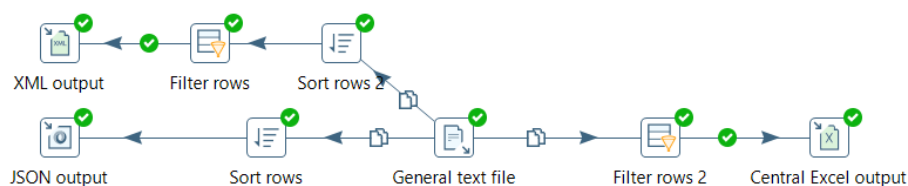
Step name: Central Excel output

File Content Custom Fields

#	Name	Type	Format
1	Order_ID	String	
2	Order_Date	Date	dd/MM/yyyy
3	Ship_Date	Date	dd/MM/yyyy
4	Ship_Mode	String	
5	Customer_ID	String	
6	Customer_Name	String	
7	Segment	String	
8	Country	String	
9	City	String	
10	State	String	
11	Postal_Code	Integer	

Get Fields Minimal width

Help OK Cancel



Filter rows

Step name: Filter rows 3

Send 'true' data to step: South file output

Send 'false' data to step:

The condition:

Region = South (String)

Help OK Cancel

Text file output

Step name: South file output

File Content Fields

#	Name	Type	Format	Length	Precis
1	Order_ID	String			
2	Order_Date	Date	dd/MM/yyyy		
3	Ship_Date	Date	dd/MM/yyyy		
4	Ship_Mode	String			
5	Customer_ID	String			
6	Customer_Name	String			
7	Segment	String			
8	Country	String			
9	City	String			
10	State	String			
11	Postal_Code	Integer			
12	Region	String			
13	Product_ID	String			

Get Fields Minimal width

Help OK Cancel

Text file output

Step name: South file output

File Content Fields

Filename: C:\Users\marin\OneDrive\Рабочий стол\Универ3 курс\ Browse...

Pass output to servlet ☐

Create Parent folder ☒

Do not create file at start ☐

Accept file name from field? ☐

File name field:

Extension: csv

Include stepnr in filename? ☐

Include partition nr in filename? ☐

Include date in filename? ☐

Include time in filename? ☐

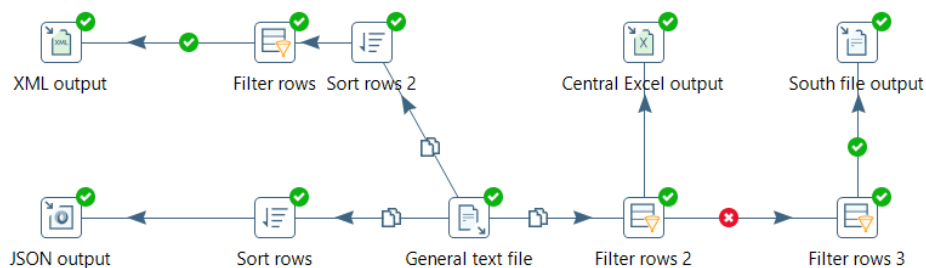
Specify Date time format ☐

Date time format:

Show filename(s)...

Add filenames to result ☒

Help OK Cancel



**Filter rows**

Step name: Filter rows 3 2

Send 'true' data to step: East file output

Send 'false' data to step:

The condition:

Region = East (String)

Help OK Cancel

**Text file output**

Step name: East file output

File Content Fields

Filename: C:\Users\marin\OneDrive\Рабочий стол\Универ\3 к Browse...

Pass output to servlet: ☐

Create Parent folder: ☒

Do not create file at start: ☐

Accept file name from field: ☐

File name field:

Extension: dat

Include stepnr in filename: ☐

Include partition nr in filename: ☐

Include date in filename: ☐

Include time in filename: ☐

Specify Date time format:

Date time format:

Show filename(s):

Add filenames to result: ☒

Help OK Cancel

Workflow diagram showing data processing steps: XML output, Filter rows, Sort rows 2, Central Excel output, South file output, East file output, West Arizona file output, West Washington file output, West California file output, JSON output, Sort rows, General text file, Filter rows 2, Filter rows 3, Filter rows 3 2, Filter rows 3 2 2, Filter rows 3 2 2 2, and Filter rows 3 2 2 2 2. Arrows indicate the flow between these steps.

**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	General text file	0	0	29982	9995	0	1	0	0	Finished	0.9s	32 273	-
2	Filter rows 2	0	9994	9994	0	0	0	0	0	Finished	0.9s	10 677	-
3	Filter rows 3	0	7671	7671	0	0	0	0	0	Finished	0.9s	8 152	-
4	Sort rows 2	0	9994	5009	0	0	0	0	0	Finished	1.0s	9 974	-
5	Filter rows 3 2	0	6051	6051	0	0	0	0	0	Finished	0.9s	6 376	-
6	Filter rows	0	5009	296	0	0	0	0	0	Finished	1.0s	4 863	-
7	Filter rows 3 2 2	0	3203	3203	0	0	0	0	0	Finished	1.0s	3 357	-
8	Filter rows 3 2 2 2	0	2979	2473	0	0	0	0	0	Finished	1.0s	3 077	-
9	South file output	0	1620	1620	0	1621	0	0	0	Finished	0.9s	1 712	-
10	Sort rows	0	9994	1862	0	0	0	0	0	Finished	1.0s	9 703	-
11	XML output	0	296	296	0	296	0	0	0	Finished	1.1s	279	-
12	West Washington file output	0	0	0	0	0	0	0	0	Finished	1.0s	0	-
13	Central Excel output	0	2323	2323	0	2323	0	0	0	Finished	2.4s	974	-
14	West Arizona file output	0	224	224	0	225	0	0	0	Finished	1.0s	232	-