

Building Sales Data mart Using **Pentaho** **Part 2 (continued)**

By Naheed Anjum Arafat

Task 2.1

Pos_details file → stage_sales

Objective



POS header

TRX_ID - Unique Transaction id

JRSTORE - Store ID, variable length, **take first five digits only**

JRREGNUM - Terminal # within the store, variable length

JRTRX - Transaction# at the terminal

JRDATE JRTIME - Start of transaction

JRCASHIER - the Shift Manager on duty

JRCUST - customer_id if the customer is a member

Source Table: POS_header

TRX_ID	JRSTORE	JRREGNUM	JRDATE	JRTIME	JRCASHIER	JRCUST
35802164	65010	001	2015-01-01 ...	12:08	06745	0
35802166	65010	001	2015-01-01 ...	14:48	06745	1116535
35802171	65010	001	2015-01-01 ...	19:00	06745	699665
35802173	65010	001	2015-01-01 ...	19:17	06745	174126
35802175	65010	001	2015-01-01 ...	20:17	06745	1997482
35802177	650101	002	2015-01-01 ...	12:48	06745	1998069
35802179	65010	001	2015-01-01 ...	12:00	06745	1397804
35802180	65010	001	2015-01-01 ...	14:10	06745	994042
35802184	65010	001	2015-01-01 ...	15:10	06745	1997480
35802186	65010	001	2015-01-01 ...	16:36	06745	0
35802194	650101	002	2015-01-01 ...	12:47	06745	0

POS detail

JOURNAL_ID - Row ID unique

TRX_ID - transaction ID

JRSTORE - first five digits is the Store ID

JRREGNUM - Terminal # within the store, variable length

JRTRX - Transaction# at the terminal

JRQTY - Quantity sold

JRPRICE - Retail price to be sold at

JREXTEN - Qty X Retail price amount

JRDISC - Amount of discount

JRITEM - Barcode, unique product code.

JRDATE - Date of the transaction


JRTIME - time of the transaction


JRCOST - unit cost (the price at which bought from the suppliers)

Source Table: POS_Details

DataSet: SELECT extract_pos_details_20150101_20150110.csv.JOURNAL_ID, extract_pos_details_20150101_20150110.csv.TRX_ID, extract_pos_details_20150101_20150110...

JOURNAL_ID	TRX_ID	JRSTORE	JRREGNUM	JRTRX	JRQTY	JRPRICE	JREXTEN	JRDISC	JRITEM	JRDATE	JRTIME	JRCOST
261865863	35806461	65035	001	152	1	19.9	19.9	0	101010007	2015-01-01...	18:03	15.92
261825335	35832815	65036	001	141	1	19.9	19.9	0	101010007	2015-01-01...	20:30	15.92
261803895	35807490	650431	002	18	1	19.9	19.9	0	101010007	2015-01-01...	11:51	15.92
261805864	35808539	650601	002	19	1	18.9	18.9	0	130180235	2015-01-01...	13:52	15.12
261867026	35832830	65037	001	52	1	23.9	23.9	0	130180243	2015-01-01...	15:48	19.12
261784662	35806840	650381	002	75	1	23.9	23.9	0	130180243	2015-01-01...	20:12	19.12
261867384	35806581	65037	001	111	1	19.9	19.9	0	130180251	2015-01-01...	19:47	15.92
261791697	35806469	650351	002	25	1	8.9	8.9	0	112040027	2015-01-01...	14:15	7.12
261791737	35806476	650351	002	32	1	6.9	6.9	0	130180350	2015-01-01...	15:16	5.52
261870859	35808814	650631	002	65	1	13.9	13.9	0	101080125	2015-01-01...	13:34	11.12
261865907	35806107	65035	001	162	1	16.9	16.9	3.38	134010917	2015-01-01...	18:51	10.816
261871642	35808943	650631	002	202	1	29.9	29.9	0	134010636	2015-01-01...	19:29	23.92
261865779	35832767	65035	001	134	1	29.9	29.9	0	134010644	2015-01-01...	17:08	23.92
261792367	35808440	65058	001	58	1	29.9	29.9	0	134010644	2015-01-01...	14:49	23.92

 Previous page

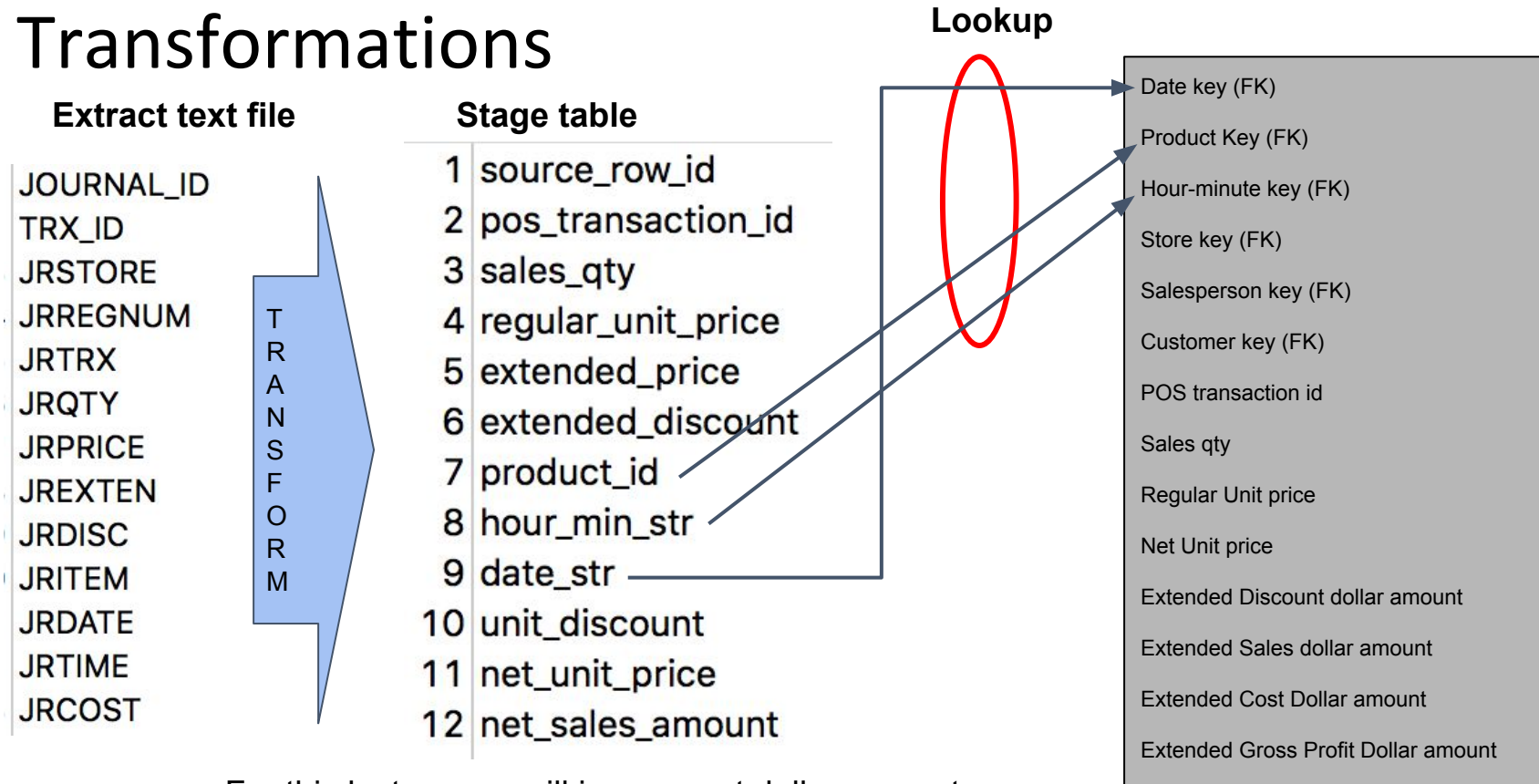
Next page 

Task 2 - Sales Fact Table



Design Transformation

Transformations



For this lecture, we will ignore cost dollar amount

Lookups in Dimension Tables

dim_product: barcode("123456789") → dim_product_key

dim_date: date_str("yyyyMMdd") → dim_date_key

dim_hour_min: hour_min_str("HH:mm") → dim_hour_min_key

JRDATE → date_str → dim_date_key

JRDATE is a STRING ,e.g. "2011-01-02 00:00:00.000000000"

Transformation:

JRDATE to date_str with format "yyyyMMdd"

Using **Java script step**

- Convert JRDATE to date using `str2date()`
- Convert date to date_str using `date2str()`

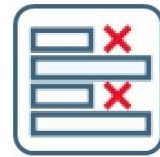
JRITEM \rightarrow product_id \rightarrow dim_product_key

JRITEM is a STRING, e.g. "180111339"

This is the barcode to lookup dim_product_key

Transformation:

Rename JRITEM \rightarrow product_id



Select values

JRTIME \rightarrow hour_min \rightarrow dim_hour_min_key

JRTIME is a STRING of length 5 format "15:46"

This is the hour_min to lookup dim_hour_min_key

Transformation:

Rename JRTIME \rightarrow hour_min



Select values

TRX_ID → pos_transaction_id

TRX_ID is a BIGINT (big integer)

Transform

Rename TRX_ID → pos_transaction_id



Select values

JOURNAL_ID → source_row_id

JOURNAL_ID is a BIGINT (big integer) unique row id from source

Transformation:

Rename JOURNAL_ID → source_row_id



Select values

Measures

Rename:

JRQTY → sales_qty

JRPRICE → regular_unit_price

JREXTEN → extended_price

JRDISC → extended_discount

Formula:

$\text{unit_discount} = \text{extended_discount} / \text{sales_qty}$

$\text{net_unit_price} = \text{regular_unit_price} - \text{unit_discount}$

$\text{net_sales_amount} = \text{net_unit_price} * \text{sales_qty}$



Calculator

Unused source fields

JRREGNUM - Terminal number unique in store

JRTRX - transaction # at the terminal

JRSTORE - store id

Task 2.1

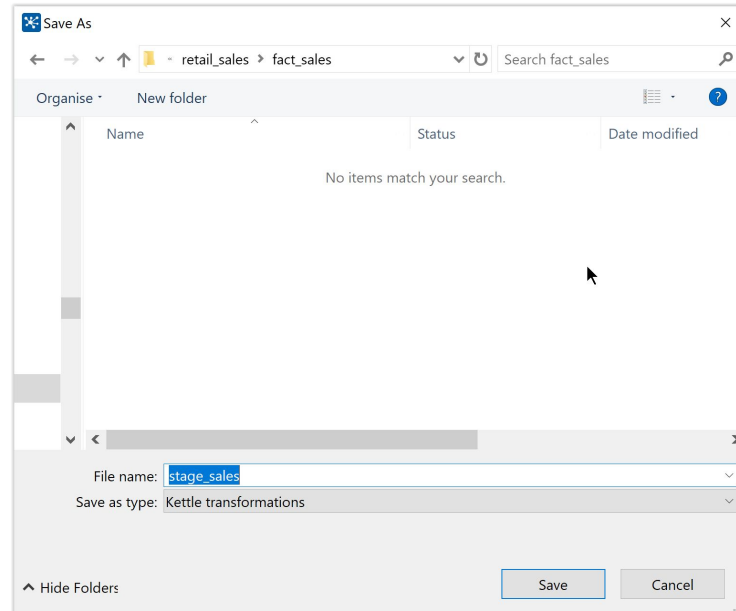
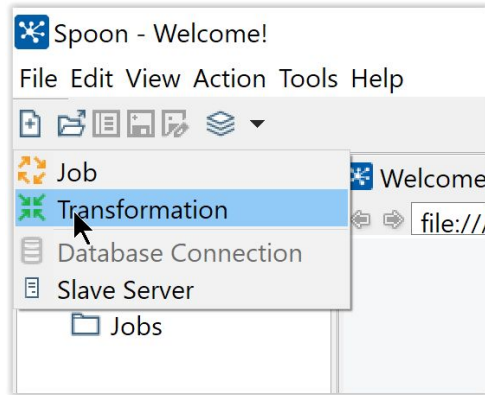
extract file → stage_sales

Objective



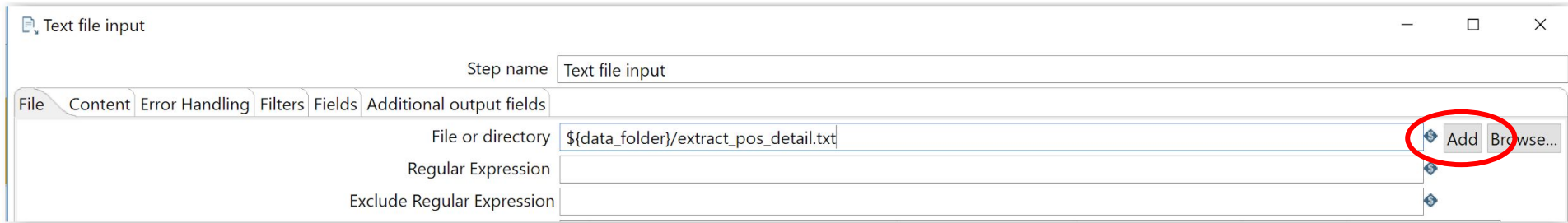
Task 2.1 - Create Transformation

- New transformation
- Save as stage_sales.ktr in folder fact_sales



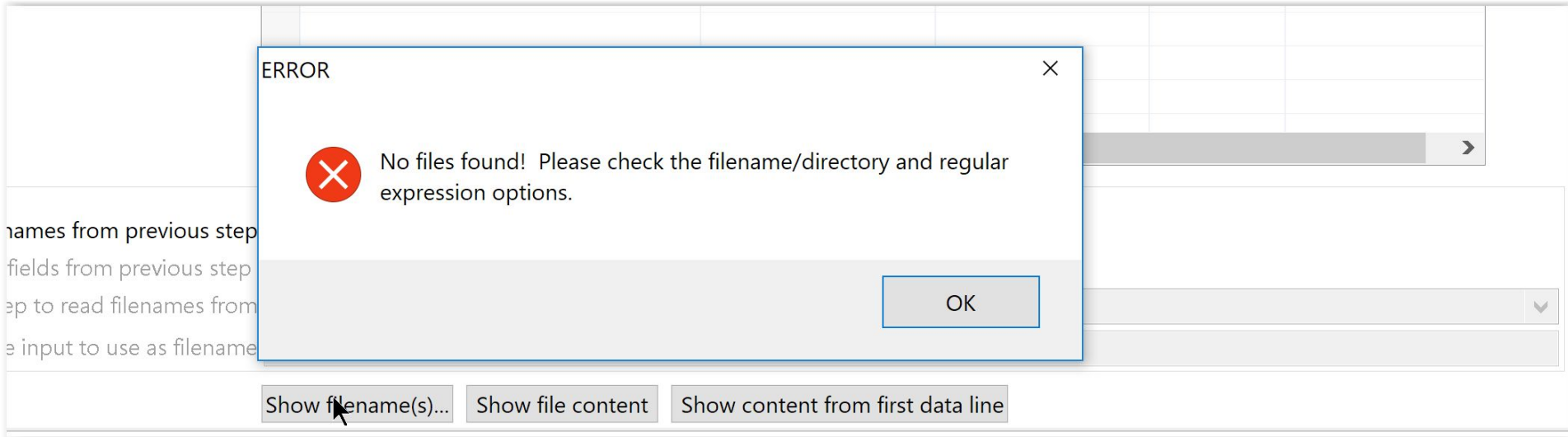
Text file input

- Drag text file input to canvas
- Rename step name `extract_pos_detail`
- Set filename to path of extract file
 - `${data_folder}/extract_pos_details.txt`
- Click Add



The screenshot shows a configuration window titled "Text file input". At the top, there is a "Step name" field containing "Text file input". Below this is a tabbed interface with tabs for "File", "Content", "Error Handling", "Filters", "Fields", and "Additional output fields". The "File" tab is currently selected. Inside the "File" tab, there are three input fields: "File or directory" containing the path `$(data_folder)/extract_pos_detail.txt`, "Regular Expression", and "Exclude Regular Expression". To the right of the "File or directory" field, there are two buttons: "Add" and "Browse...". The "Add" button is circled in red.

Check filename for error



Click on Show filename
If you see error, you have mistyped the filename

Show file content

<

previous step ☐

previous step ☐

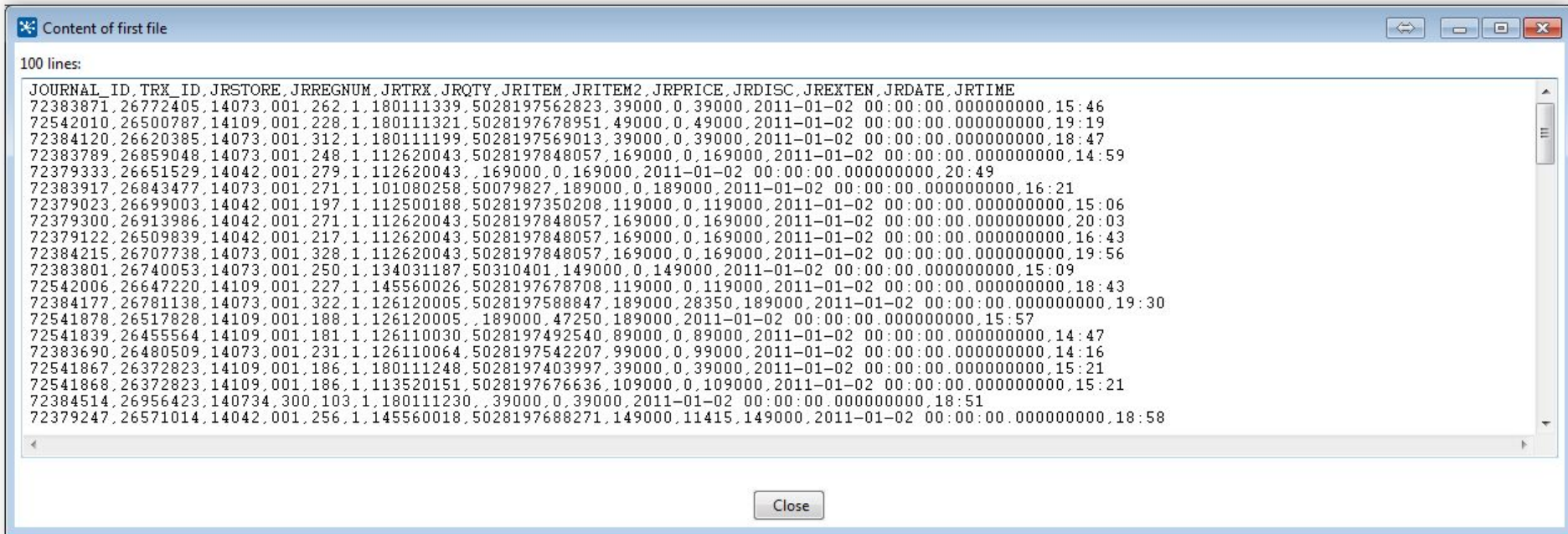
names from

as filename

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

OK Preview rows Cancel

Show file content



The screenshot shows a window titled "Content of first file" with standard Windows window controls. Below the title bar, it says "100 lines:". The main area contains 100 lines of text, each representing a data record. The records are separated by commas, indicating a CSV format. The data includes various numerical values and dates. A "Close" button is located at the bottom center of the window.

```
JOURNAL_ID,TRX_ID,JRSTORE,JRREGNUM,JRTRX,JRQTY,JRITEM,JRITEM2,JRPRICE,JRDISC,JREXTEN,JRDATE,JRTIME
72383871,26772405,14073,001,262,1,180111339,5028197562823,39000,0,39000,2011-01-02 00:00:00.000000000,15:46
72542010,26500787,14109,001,228,1,180111321,5028197678951,49000,0,49000,2011-01-02 00:00:00.000000000,19:19
72384120,26620385,14073,001,312,1,180111199,5028197569013,39000,0,39000,2011-01-02 00:00:00.000000000,18:47
72383789,26859048,14073,001,248,1,112620043,5028197848057,169000,0,169000,2011-01-02 00:00:00.000000000,14:59
72379333,26651529,14042,001,279,1,112620043,169000,0,169000,2011-01-02 00:00:00.000000000,20:49
72383917,26843477,14073,001,271,1,101080258,50079827,189000,0,189000,2011-01-02 00:00:00.000000000,16:21
72379023,26699003,14042,001,197,1,112500188,5028197350208,119000,0,119000,2011-01-02 00:00:00.000000000,15:06
72379300,26913986,14042,001,271,1,112620043,5028197848057,169000,0,169000,2011-01-02 00:00:00.000000000,20:03
72379122,26509839,14042,001,217,1,112620043,5028197848057,169000,0,169000,2011-01-02 00:00:00.000000000,16:43
72384215,26707738,14073,001,328,1,112620043,5028197848057,169000,0,169000,2011-01-02 00:00:00.000000000,19:56
72383801,26740053,14073,001,250,1,134031187,50310401,149000,0,149000,2011-01-02 00:00:00.000000000,15:09
72542006,26647220,14109,001,227,1,145560026,5028197678708,119000,0,119000,2011-01-02 00:00:00.000000000,18:43
72384177,26781138,14073,001,322,1,126120005,5028197588847,189000,28350,189000,2011-01-02 00:00:00.000000000,19:30
72541878,26517828,14109,001,188,1,126120005,189000,47250,189000,2011-01-02 00:00:00.000000000,15:57
72541839,26455564,14109,001,181,1,126110030,5028197492540,89000,0,89000,2011-01-02 00:00:00.000000000,14:47
72383690,26480509,14073,001,231,1,126110064,5028197542207,99000,0,99000,2011-01-02 00:00:00.000000000,14:16
72541867,26372823,14109,001,186,1,180111248,5028197403997,39000,0,39000,2011-01-02 00:00:00.000000000,15:21
72541868,26372823,14109,001,186,1,113520151,5028197676636,109000,0,109000,2011-01-02 00:00:00.000000000,15:21
72384514,26956423,14073,300,103,1,180111230,39000,0,39000,2011-01-02 00:00:00.000000000,18:51
72379247,26571014,14042,001,256,1,145560018,5028197688271,149000,11415,149000,2011-01-02 00:00:00.000000000,18:58
```

Notice the field separator is “,” (comma)

Text file input

Click on Content tab

Text file input

Step name: Text file input

File **Content** Error Handling Filters Fields Additional output fields

File or directory: Add Browse...

Regular Expression:

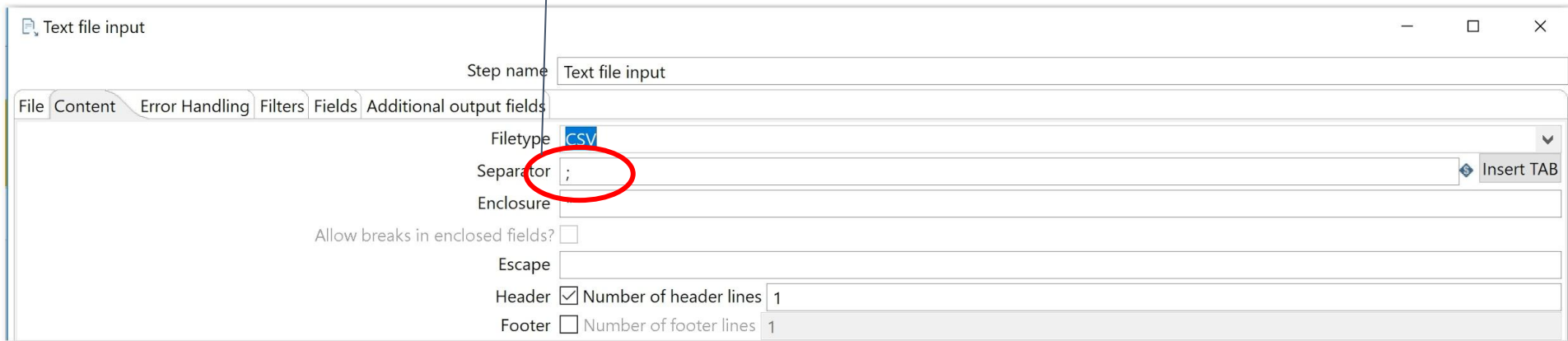
Exclude Regular Expression:

Selected files:

#	File/Directory	Wildcard (RegExp)	Exclude wildcard	Required	Include subfolders	
1	\${data_folder}/extract_pos_detail.txt			N	N	Delete
						Edit

Text file input

Change Separator to “,”



The screenshot shows a configuration window titled "Text file input". It has a tabbed interface with "File", "Content", "Error Handling", "Filters", "Fields", and "Additional output fields". The "File" tab is active. The "Step name" is "Text file input". The "Filetype" is set to "CSV". The "Separator" field, which currently contains a semicolon ";", is circled in red. A blue line points from the text "Change Separator to “,”" to this field. To the right of the separator field is a button labeled "Insert TAB". The "Enclosure" field is empty. The "Allow breaks in enclosed fields?" checkbox is unchecked. The "Escape" field is empty. The "Header" section has a checked checkbox for "Number of header lines" with a value of "1". The "Footer" section has an unchecked checkbox for "Number of footer lines" with a value of "1".

Text file input

Step name Text file input

File Content Error Handling Filters Fields Additional output fields

Filetype CSV

Separator ; Insert TAB

Enclosure

Allow breaks in enclosed fields? ☐

Escape

Header ☒ Number of header lines 1

Footer ☐ Number of footer lines 1

Get fields

The screenshot shows a 'Text file input' dialog box. The 'Fields' tab is selected and circled in red. A red box contains the text 'Click on Fields tab' and 'Click on Get Fields', with a line pointing to the 'Get Fields' button, which is also circled in red. The dialog includes a 'Step name' field, a tabbed interface, a table with columns for field details, and buttons for 'OK', 'Preview rows', and 'Cancel'.

Text file input

Step name Text file input

File Content Error Handling **Fields** Additional output fields

#	Name	Type	Format	Position	Length	Precision	Repeat
1							

Click on Fields tab
Click on Get Fields

Get Fields Minimal width

OK Preview rows Cancel

Help

Get fields

Set the Data Types
to these values

#	Name	Type	Format	Position	Length	Decimal	Group	Null if	Default	Trim type	Repeat
1	JOURNAL_ID	Integer								none	N
2	TRX_ID	Integer								none	N
3	JRSTORE	String								none	N
4	JRREGNUM	String								none	N
5	JRTRX	Integer								none	N
6	JRQTY	Number			15	2				none	N
7	JRPRICE	Number			15	2			\$	none	N
8	JREXTEN	Number			15	2			\$	none	N
9	JRDISC	Number			15	2			\$	none	N
10	JRITEM	Integer								none	N
11	JRDATE	String								none	N
12	JRTIME	String								none	N
13	JRCOST	Number			15	2			\$	none	N



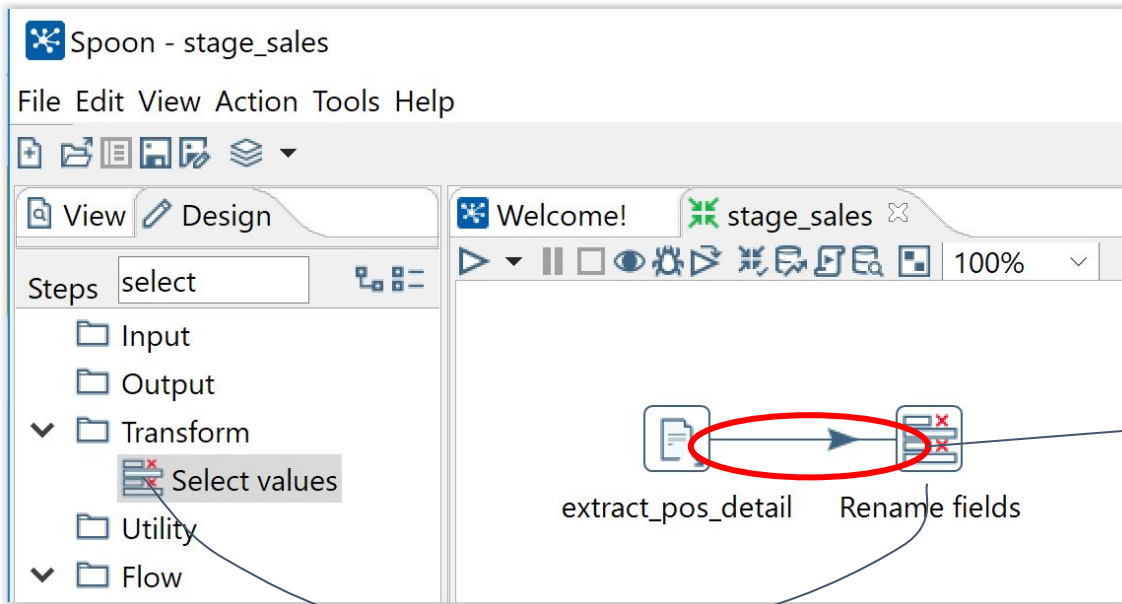
extract_pos_detail

Step: Select Values

Select values step to rename fields

- a. Rename JRITEM → product_id
- b. Rename JRTIME → hour_min_str
- c. Rename TRX_ID → pos_transaction_id
- d. Rename JOURNAL_ID → source_row_id
- e. JRQTY → sales_qty
JRPRICE → regular_unit_price
JREXTEN → extended_gross_sales_amount
JRDISC → extended_discount_amount

Select Values Step



2. Link the 2 steps

1. Drag and drop into
canvas

Select values

Select & Alter			Remove	Meta-data
Fields :				
#	Fieldname	Rename to		
1	JOURNAL_ID	source_row_id		
2	TRX_ID	pos_transaction_id		
3	JRSTORE			
4	JRREGNUM			
5	JRTRX			
6	JRQTY	sales_qty		
7	JRPRICE	regular_unit_price		
8	JREXTEN	extended_price		
9	JRDISC	extended_discount		
10	JRITEM	product_id		
11	JRDATE			
12	JRTIME	hour_min_str		
13	JRCOST			

Include

The diagram shows a mapping interface. On the left, there is a table with three columns: #, Fieldname, and Rename to. The table contains 13 rows of data. A red circle is drawn around the table. To the right of the table, there are two buttons: 'Get fields to select' and 'Edit Mapping'. The 'Get fields to select' button is also circled in red. An arrow points from the text 'Click Get fields to select' to the 'Get fields to select' button. Another arrow points from the text 'Set the new field names' to the table.

Click Get fields to select

Set the new field names



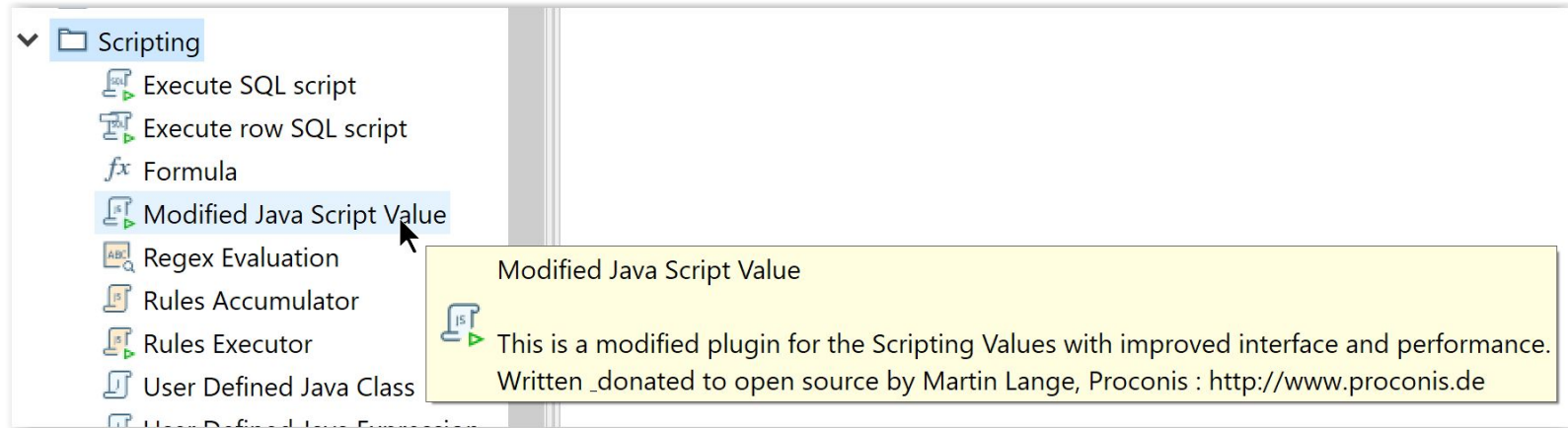
extract_pos_detail



Rename fields

Modified Java Script Value

Navigate to Scripting/Modified Java Script Value step



Modified Java Script Value step

Script Values / Mod

Step name: **Extract date string**

Java script functions :

- Transform Scripts
- Transform Constants
- Transform Functions
- Input fields
 - source_row_id
 - pos_transaction_id
 - JRSTORE
 - JRREGNUM
 - JRTRX
 - sales_qty
 - regular_unit_price
 - extended_price
 - extended_discount
 - product_id
 - JRDATE
 - hour_min_str
- Output fields
 - Please use the 'Replace value'

Java script :

```
Script 1
//Convert JRDATE to yyyyMMdd
// Sample JRDATE "2011-01-02 00:00:00.000000000"

var date_str = date2str(str2date(JRDATE, "yyyy-MM-dd HH:mm:ss.SSSSSS"), "yyyyMMdd");
```

Linernr: 0
Compatibility mode? ☐ Optimization level 9

#	Fieldname	Rename to	Type	Length	Precision	Replace value 'Fieldname' or 'Rename to'
1	date_str		String			N

Buttons: ? Help, OK, Cancel, Get variables, Test script

Change step name to "date_str"

Enter the script from next slide into this box

Set Fieldname to date_str
Set Type to String

Enter the following script into the box

```
//Convert JRDATE to yyyyMMdd  
// Sample JRDATE "2011-01-02 00:00:00.000000000"  
  
var date_str = date2str(str2date(JRDATE,"yyyy-MM-dd HH:mm:ss.SSSSSS"),"yyyyMMdd");
```



extract_pos_detail

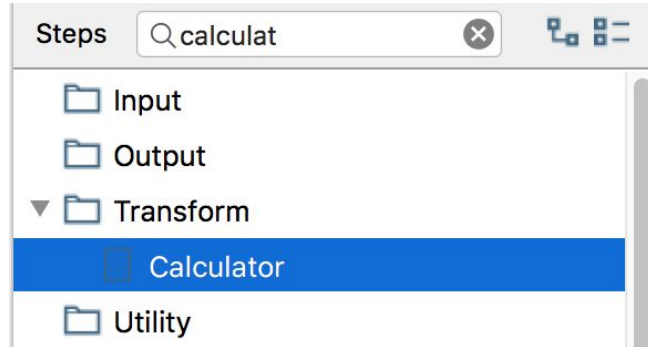


Rename fields

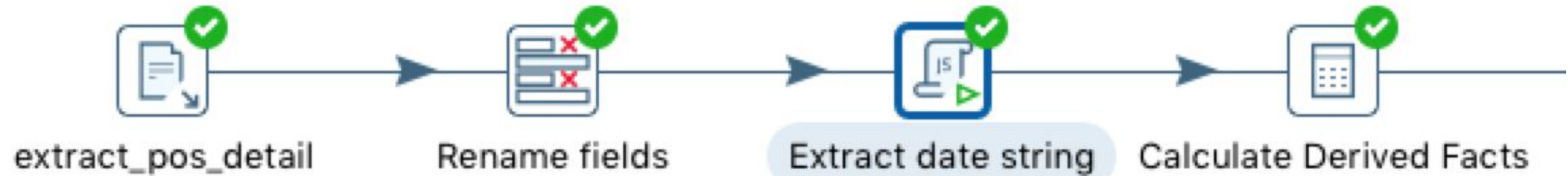


Extract date string

Calculator



Find the Calculator Step.
Add it.
Rename it.



Recall that

Rename:

JRQTY → sales_qty

JRPRICE → regular_unit_price

JREXTEN → extended_price (=regular_unit_price * sales_qty)

JRDISC → extended_discount (=unit_discount * sales_qty)

Formula:

unit_discount = extended_discount/sales_qty

net_unit_price = regular_unit_price - unit_discount

net_sales_amount = net_unit_price * sales_qty

Calculator

Click on Calculation column
Dialog box with all the formulas will open up

Calculator

Step 1 Filter

Select the calculation type to perform

Fields:

#	New field	Calculation	Field A	Field B
1	unit_discount	A / B	extended_discount	sales

-
- Set field to constant value A
- Create a copy of field A
- A + B
- A - B
- A * B
- A / B

Double-Click on "A / B"

Calculator

Field A

Step name

Fields:

#	New field	Calculation	Field A
1	unit_discount	A / B	extended_discount
			JRDATE
			JRREGNUM
			JRSTORE
			JRTRX
			date_str
			extended_discount
			extended_price
			hour min str

Field B

Step name Calculate Deri

Fields:

#	New field	Calculation	Field A	Field B	Field C
1	unit_discount	A / B	extended_discount	sales_qty	
				JRDATE	
				JRREGNUM	
				JRSTORE	
				JRTRX	
				date_str	
				extended_discount	
				extended_price	
				hour_min_str	
				pos_transaction_id	
				product_id	
				regular_unit_price	
				sales_qty	
				source_row_id	

Help

Calculator

Step name

Calculate Derived Facts

Fields:

#	New field	Calculation	Field A	Field B	Field C	Value type	Length
1	unit_discount	A / B	extended_discount	sales_qty		<div>Number</div> <div> <div>Number</div> <div>String</div> <div>Date</div> </div>	

#	New field	Calculation	Field A	Field B	Field C	Value type	Length
1	unit_discount	A / B	extended_discount	sales_qty		Number	
						<div> Number String Date </div>	

Calculator - define 3 fields

unit_discount
net_unit_price
net_sales_amount

Fields:

#	New field	Calculation	Field A	Field B	Field C	Value type	Length	Precision	Remove	C
1	unit_discount	A / B	extended_discount	sales_qty		Number			N	
2	net_unit_price	A - B	regular_unit_price	unit_discount		Number			N	
3	net_sales_amount	A * B	net_unit_price	sales_qty		Number			N	

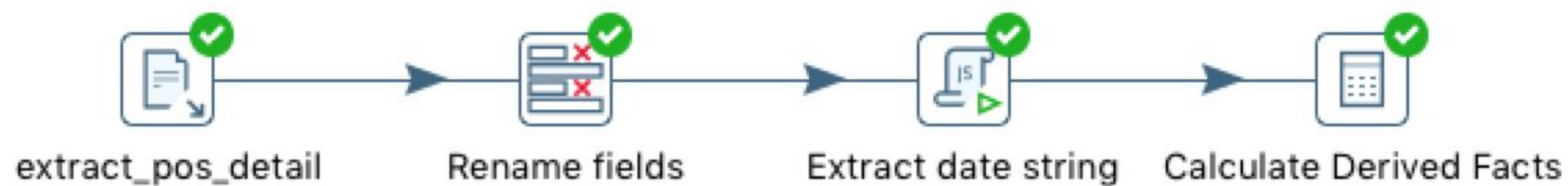


Table output

- Microsoft Excel Output
- Pentaho Reporting Output
- Properties Output
- RSS Output
- S3 File Output
- SQL File Output
- Table output**
- Text file output
- XML Output

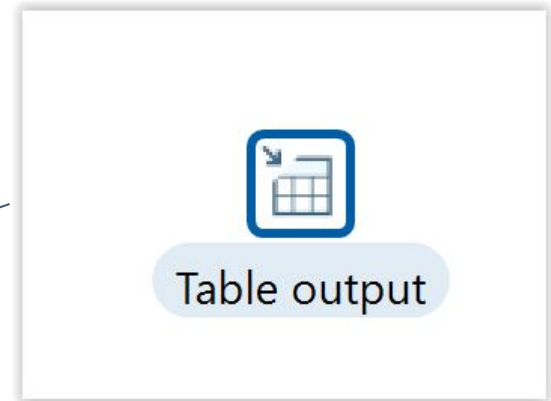
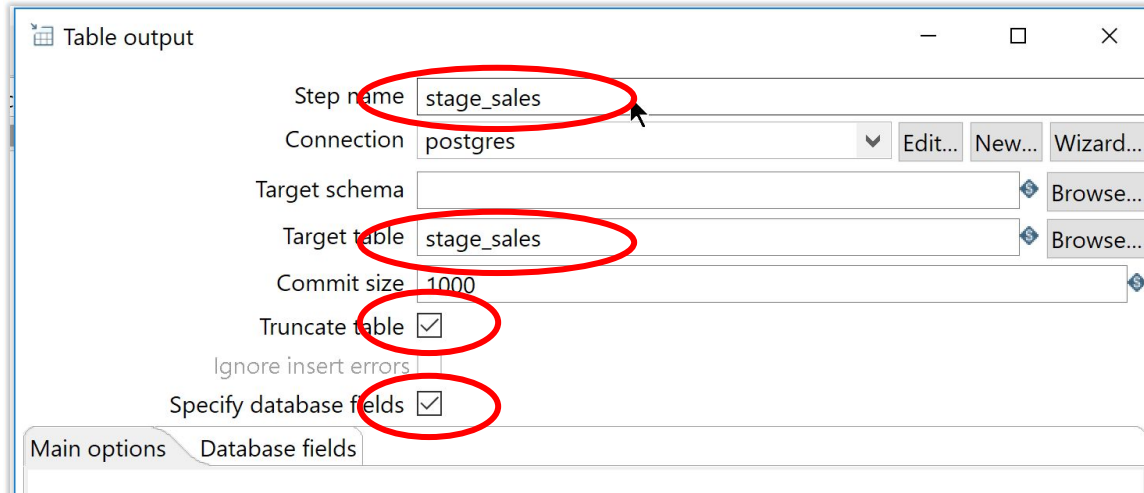


Table output



Table output



The screenshot shows the 'Table output' configuration window. The following settings are highlighted with red circles:

- Step name: stage_sales
- Target table: stage_sales
- Truncate table: ☒
- Specify database fields: ☒

Other visible settings include:

- Connection: postgres
- Target schema: (empty)
- Commit size: 1000
- Ignore insert errors: (unchecked)

The window has two tabs at the bottom: 'Main options' (selected) and 'Database fields'.

1. Set Step name to stage_sales
2. Set table name to stage_sales
3. Set Truncate table to YES
4. Set Specify database fields to YES

Table output

Main options

Database fields

Fields to insert:


#	Table field	Stream field	
1	source_row_id	source_row_id	
2	pos_transaction_id	pos_transaction_id	
3	JRSTORE	JRSTORE	
4	JRREGNUM	JRREGNUM	
5	JRTRX	JRTRX	
6	sales_qty	sales_qty	
7	regular_unit_price	regular_unit_price	
8	extended_price	extended_price	
9	extended_discount	extended_discount	
10	product_id	product_id	
11	JRDATE	JRDATE	
12	hour_min_str	hour_min_str	
13	date_str	date_str	
14	unit_discount	unit_discount	
15	net_unit_price	net_unit_price	
16	net_sales_amount	net_sales_amount	

Get fields

Enter field mapping

Click on get fields

Delete unwanted fields

Specify database fields 

Main options Database fields

Fields to insert:

#	Table field	Stream field	
1	source_row_id	source_row_id	
2	pos_transaction_id	pos_transaction_id	
3	JRSTORE	JRSTORE	
4	JRREGNUM	JRREGNUM	
5	JRTRX	JRTRX	
6	sales_qty	sales_qty	
7	regular_unit_price	regular_unit_price	
8	extended_price	extended_price	
9	extended_discount	extended_discount	
10	product_id	product_id	
11	JRDATE	JRDATE	
12	hour_min_str	hour_min_str	
13	JRCOST	JRCOST	
14	date_str	date_str	
15	unit_discount	unit_discount	
16	net_unit_price	net_unit_price	
17	net_sales_amount	net_sales_amount	

Insert before this row
Insert after this row

Move up ⌘↑
Move down ⌘↓
Optimal Column size incl. header F3
Optimal Column size excl. header F4


Clear all

Select all rows ⌘A
Clear selection ⌘
Filtered selection ⌘F

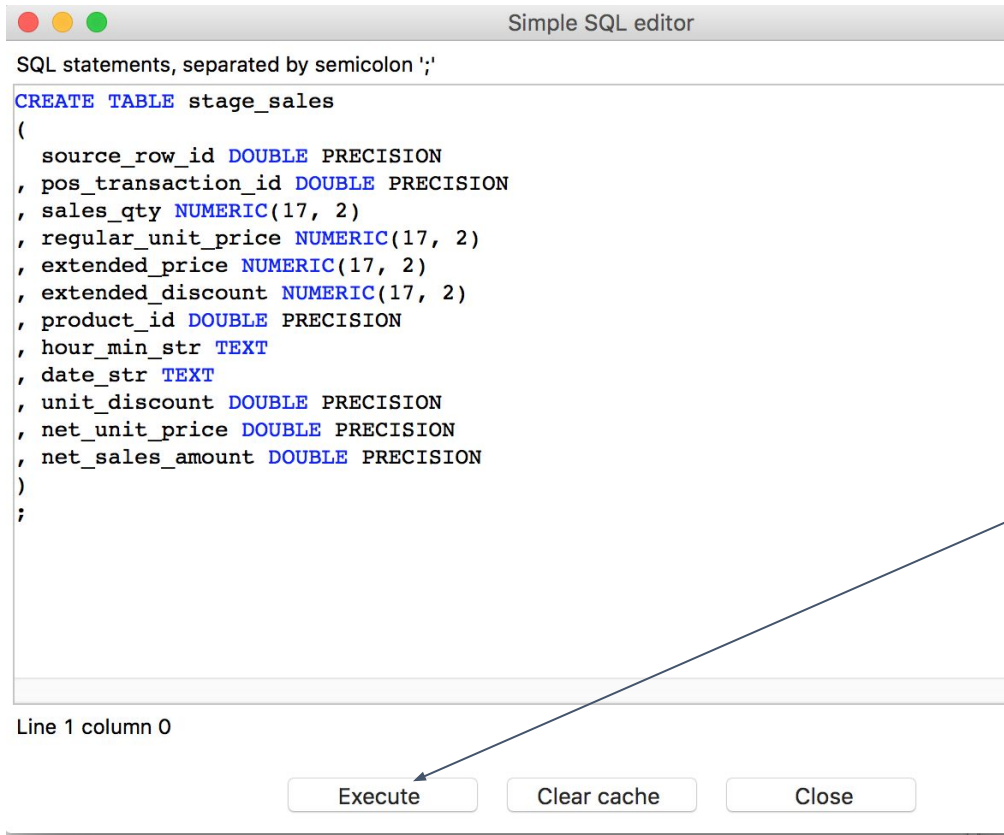
Copy selected lines to clipboard ⌘C
Paste clipboard to table ⌘V
Cut selected lines ⌘X
Delete selected lines ⌘X
Keep only selected lines ⌘K

Copy field value to all rows

Undo : not available ⌘Z
Redo : not available ⌘Y

 Help

Create table: Click SQL and Execute



SQL statements, separated by semicolon ';'

```
CREATE TABLE stage_sales
(
  source_row_id DOUBLE PRECISION
, pos_transaction_id DOUBLE PRECISION
, sales_qty NUMERIC(17, 2)
, regular_unit_price NUMERIC(17, 2)
, extended_price NUMERIC(17, 2)
, extended_discount NUMERIC(17, 2)
, product_id DOUBLE PRECISION
, hour_min_str TEXT
, date_str TEXT
, unit_discount DOUBLE PRECISION
, net_unit_price DOUBLE PRECISION
, net_sales_amount DOUBLE PRECISION
)
;
```

Line 1 column 0

Execute Clear cache Close

Execute

Finally



Run Transformation

Run Transformation



Execution Results

[Execution History](#) [Logging](#) [Step Metrics](#) [Performance Graph](#) [Metrics](#) [Preview data](#)

#	Stepname	Copynr	Read	Written	Input	Output	Updated	Rejected	Errors	Active	Time	Speed (r/s)	input/output
1	extract_pos_detail	0	0	73168	73169	0	1	0	0	Finished	2.3s	31,647	-
2	Rename fields	0	73168	73168	0	0	0	0	0	Finished	3.1s	23,602	-
3	Extract date string	0	73168	73168	0	0	0	0	0	Finished	3.7s	19,991	-
4	Calculate Derived Facts	0	73168	73168	0	0	0	0	0	Finished	4.4s	16,476	-
5	stage_sales	0	73168	73168	0	73168	0	0	0	Finished	5.1s	14,299	-

End of Task 2.1