

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

Макарова Екатерина Павловна

ЛАБОРАТОРНАЯ РАБОТА 2

Динамические соединения с базами данных

Проектный практикум по разработке ETL-решений

Направление подготовки

38.03.05 Бизнес-информатика

Профиль подготовки

Аналитика данных и эффективное управление

Курс обучения: 4

Форма обучения: очная

Преподаватель: кандидат технических наук,
доцент Босенко Тимур Муртазович

Москва

2025

Цель работы: получить практические навыки создания ETL-процесса для интеграции данных из различных источников с использованием динамических соединений в Pentaho Data Integration, включая обработку повторяющихся данных.

Задачи:

- Создать динамические подключения к различным источникам данных.
- Разработать процесс выявления и обработки дублирующихся записей.
- Реализовать механизм объединения данных в единое хранилище.
- Настроить обработку ошибок при выполнении трансформации.

Задачи по индивидуальному заданию вариант 8:

- Фильтр по доставке: только Standard Class;
- Статистика продаж;
- Анализ по городам;

Ход работы:

1. Подготовка базы данных:

1) Создание таблицы «orders» (Рис. 1).

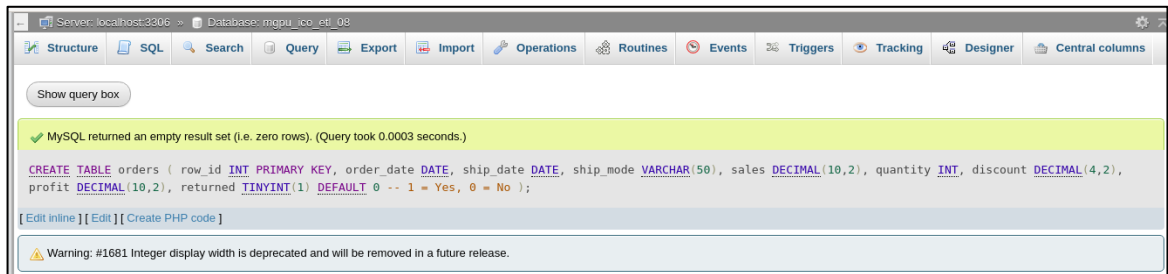


Рис. 1

2) Создание таблицы «customers» (Рис. 2).

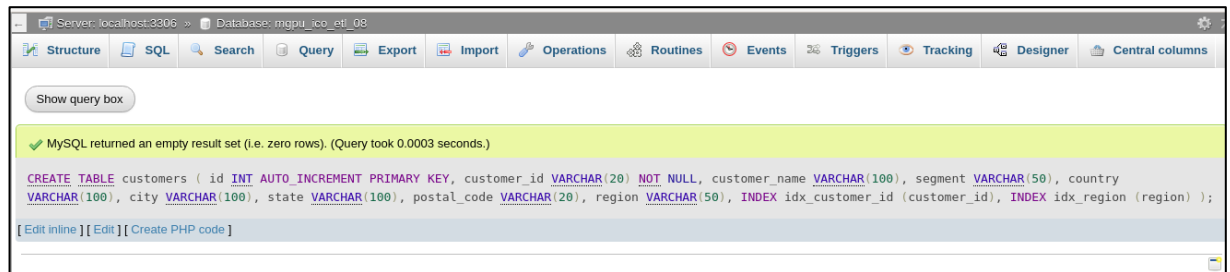


Рис. 2

3) Создание таблицы «products» (Рис. 3).

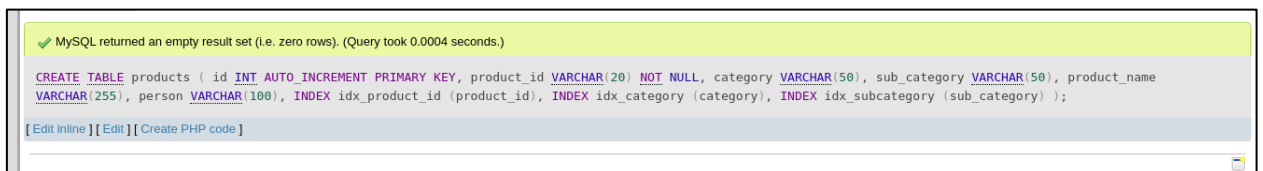


Рис. 3

4) Создание индексов для оптимизации запросов (Рис. 4).



Рис. 4

2. Проверка работоспособности трансформаций

1) Выполнение трансформации на обработку и загрузку данных по заказам в базу данных (Рис. 5).

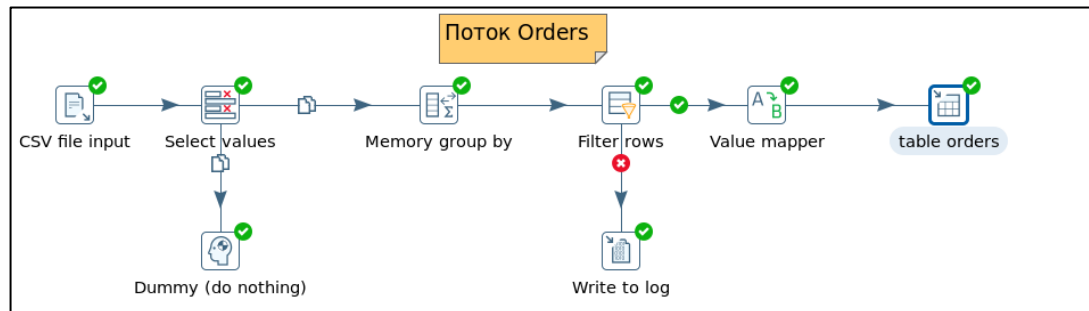


Рис. 5

Запрос к базе данных на проверку импорта данных в таблицу «orders» (Рис. 6).

Showing rows 0 - 24 (9994 total, Query took 0.0003 seconds.)

SELECT * FROM `orders`

☐ Profiling [[Edit inline](#)] [[Edit](#)] [[Explain SQL](#)] [[Create PHP code](#)] [[Refresh](#)]

1 > >> | Number of rows: 25 | Filter rows: Search this table | Sort by key: None

Extra options

				row_id	order_date	ship_date	ship_mode	sales	quantity	discount	profit	returned
<input type="checkbox"/>	Edit	Copy	Delete	1	2018-11-08	2018-11-11	Second Class	261.96	2	0.00	41.91	NULL
<input type="checkbox"/>	Edit	Copy	Delete	2	2018-11-08	2018-11-11	Second Class	731.94	3	0.00	219.58	NULL
<input type="checkbox"/>	Edit	Copy	Delete	3	2018-06-12	2018-06-16	Second Class	14.62	2	0.00	6.87	NULL
<input type="checkbox"/>	Edit	Copy	Delete	4	2017-10-11	2017-10-18	Standard Class	957.58	5	0.40	-383.03	NULL
<input type="checkbox"/>	Edit	Copy	Delete	5	2017-10-11	2017-10-18	Standard Class	22.37	2	0.20	2.52	NULL
<input type="checkbox"/>	Edit	Copy	Delete	6	2016-06-09	2016-06-14	Standard Class	48.86	7	0.00	14.17	NULL
<input type="checkbox"/>	Edit	Copy	Delete	7	2016-06-09	2016-06-14	Standard Class	7.28	4	0.00	1.97	NULL
<input type="checkbox"/>	Edit	Copy	Delete	8	2016-06-09	2016-06-14	Standard Class	907.15	6	0.20	90.72	NULL
<input type="checkbox"/>	Edit	Copy	Delete	9	2016-06-09	2016-06-14	Standard Class	18.50	3	0.20	5.78	NULL
<input type="checkbox"/>	Edit	Copy	Delete	10	2016-06-09	2016-06-14	Standard Class	114.90	5	0.00	34.47	NULL
<input type="checkbox"/>	Edit	Copy	Delete	11	2016-06-09	2016-06-14	Standard Class	1706.18	9	0.20	85.31	NULL
<input type="checkbox"/>	Edit	Copy	Delete	12	2016-06-09	2016-06-14	Standard Class	911.42	4	0.20	68.36	NULL
<input type="checkbox"/>	Edit	Copy	Delete	13	2019-04-15	2019-04-20	Standard Class	15.55	3	0.20	5.44	NULL

Рис. 6

2) Выполнение трансформации на обработку и загрузку данных по заказам в базу данных (Рис. 7).

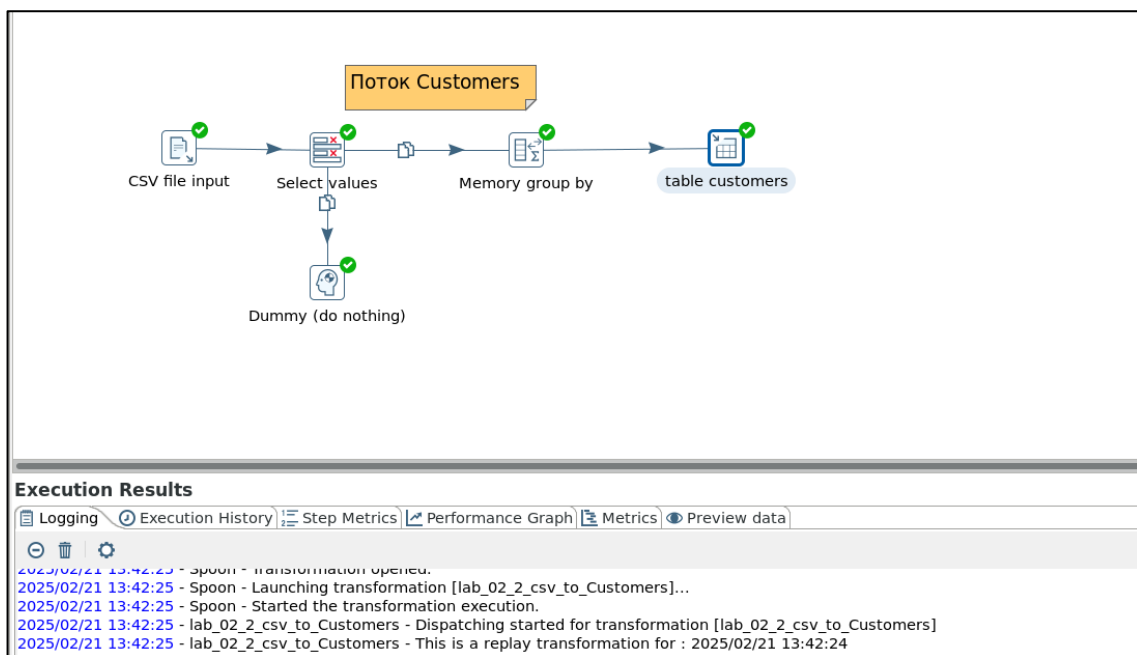


Рис. 7

Выполнение запроса к базе данных на импорт данных о клиентах (Рис. 8).

Showing rows 0 - 24 (4910 total, Query took 0.0002 seconds.)

SELECT * FROM `customers`

Profiling

Edit inline

Edit

Explain SQL

Create PHP code

Refresh

1

>

>>

Number of rows:

25

Filter rows:

Search this table

Sort by key:

None

Extra options

Рис. 8

3) Выполнение трансформации на обработку и загрузку данных по продуктам в базу данных (Рис. 9).

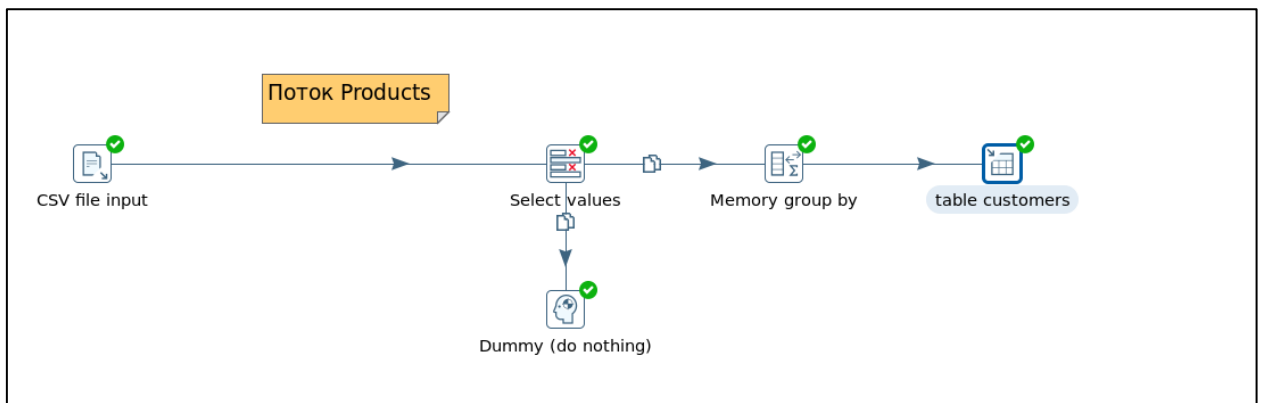


Рис. 9

Выполнение запроса на проверку импорта данных в таблицу «products» (Рис. 10).

Showing rows 0 - 24 (5371 total, Query took 0.0002 seconds.)

`SELECT * FROM `products``

☐ Profiling [\[Edit inline \]](#) [\[Edit \]](#) [\[Explain SQL \]](#) [\[Create PHP code \]](#) [\[Refresh \]](#)

1 > >> Number of rows: 25 Filter rows: Search this table Sort by key: None

Extra options

					id	product_id	category	sub_category	product_name	person
<input type="checkbox"/>	Edit	Copy	Delete		1	OFF-AP-10002578	Office Supplies	Appliances	Fellowes Premier Superior Surge Suppressor, 10-Out...	Chuck Magee
<input type="checkbox"/>	Edit	Copy	Delete		2	OFF-PA-10000575	Office Supplies	Paper	Wirebound Message Books, Four 2 3/4 x 5 White Form...	Chuck Magee
<input type="checkbox"/>	Edit	Copy	Delete		3	TEC-MA-10002790	Technology	Machines	NeatDesk Desktop Scanner & Digital Filing System	Kelly Williams
<input type="checkbox"/>	Edit	Copy	Delete		4	OFF-AR-10000255	Office Supplies	Art	Newell 328	Kelly Williams
<input type="checkbox"/>	Edit	Copy	Delete		5	TEC-PH-10001061	Technology	Phones	Apple iPhone 5C	Cassandra Brandow
<input type="checkbox"/>	Edit	Copy	Delete		6	OFF-AR-10003179	Office Supplies	Art	Dixon Ticonderoga Core-Lock Colored Pencils	Anna Andreadi
<input type="checkbox"/>	Edit	Copy	Delete		7	OFF-AP-10003040	Office Supplies	Appliances	Fellowes 8 Outlet Superior Workstation Surge Prote...	Anna Andreadi
<input type="checkbox"/>	Edit	Copy	Delete		8	OFF-BI-10004654	Office Supplies	Binders	VariCap6 Expandable Binder	Cassandra Brandow
<input type="checkbox"/>	Edit	Copy	Delete		9	FUR-CH-10001802	Furniture	Chairs	Hon Every-Day Chair Series Swivel Task Chairs	Anna Andreadi
<input type="checkbox"/>	Edit	Copy	Delete		10	OFF-PA-10000675	Office Supplies	Paper	Xerox 1919	Chuck Magee
<input type="checkbox"/>	Edit	Copy	Delete		11	FUR-CH-10004698	Furniture	Chairs	Padded Folding Chairs, Black, 4/Cartron	Cassandra Brandow
<input type="checkbox"/>	Edit	Copy	Delete		12	OFF-PA-10000241	Office Supplies	Paper	IBM Multi-Purpose Copy Paper, 8 1/2 x 11", Case	Anna Andreadi
<input type="checkbox"/>	Edit	Copy	Delete		13	FUR-FU-10003268	Furniture	Furnishings	Eldon Radial Chair Mat for Low to Medium Pile Carp...	Cassandra Brandow
<input type="checkbox"/>	Edit	Copy	Delete		14	OFF-AR-10003560	Office Supplies	Art	Zebra Zazzle Fluorescent Highlighters	Anna Andreadi

Рис. 10

После этого таблицы были очищены с помощью запроса TRUNCATE TABLE “название таблицы” (orders, customers, products).

4) Проверка работоспособности Job на импорт данных из трансформаций (Рис. 11).

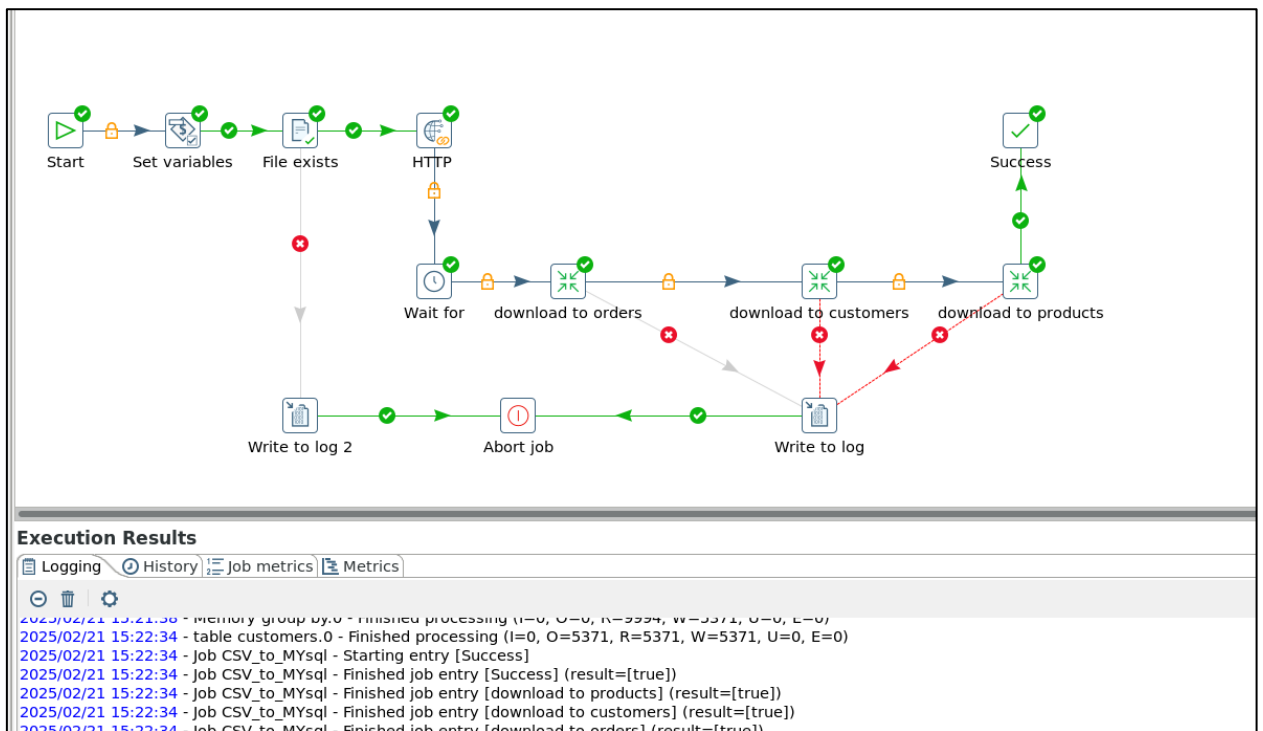


Рис. 11

Запрос к таблице «customers» для проверки успешности импорта данных (Рис. 12).

Showing rows 0 - 24 (2533 total, Query took 0.0002 seconds.)

SELECT * FROM `customers`

Profiling [Edit inline] [Edit] [Explain SQL] [Create PHP code] [Refresh]

1 > >> Number of rows: 25 Filter rows: Search this table Sort by key: None

Extra options

	id	customer_id	customer_name	segment	country	city	state	postal_code	region
<input type="checkbox"/> Edit <input type="checkbox"/> Copy <input type="checkbox"/> Delete	1	CC-12670	Craig Carreira	Consumer	United States	Chicago	Illinois	60610	Central
<input type="checkbox"/> Edit <input type="checkbox"/> Copy <input type="checkbox"/> Delete	2	SO-20335	Sean O'Donnell	Consumer	United States	Fort Lauderdale	Florida	33311	South
<input type="checkbox"/> Edit <input type="checkbox"/> Copy <input type="checkbox"/> Delete	3	RF-19840	Roy Franz-sisch	Consumer	United States	Chesapeake	Virginia	23320	South
<input type="checkbox"/> Edit <input type="checkbox"/> Copy <input type="checkbox"/> Delete	4	JE-15745	Joel Eaton	Consumer	United States	Newark	Ohio	43055	East
<input type="checkbox"/> Edit <input type="checkbox"/> Copy <input type="checkbox"/> Delete	5	SJ-20215	Sarah Jordan	Consumer	United States	Columbia	Tennessee	38401	South
<input type="checkbox"/> Edit <input type="checkbox"/> Copy <input type="checkbox"/> Delete	6	MM-18055	Michelle Moray	Consumer	United States	Aurora	Colorado	80013	West
<input type="checkbox"/> Edit <input type="checkbox"/> Copy <input type="checkbox"/> Delete	7	AC-10450	Amy Cox	Consumer	United States	Seattle	Washington	98105	West
<input type="checkbox"/> Edit <input type="checkbox"/> Copy <input type="checkbox"/> Delete	8	KH-16360	Katherine Hughes	Consumer	United States	Chicago	Illinois	60623	Central
<input type="checkbox"/> Edit <input type="checkbox"/> Copy <input type="checkbox"/> Delete	9	PG-18895	Paul Gonzalez	Consumer	United States	Richmond	Virginia	23223	South
<input type="checkbox"/> Edit <input type="checkbox"/> Copy <input type="checkbox"/> Delete	10	BD-11605	Brian Dahlen	Consumer	United States	Springfield	Virginia	22153	South
<input type="checkbox"/> Edit <input type="checkbox"/> Copy <input type="checkbox"/> Delete	11	FO-14305	Frank Olsen	Consumer	United States	New York City	New York	10035	East
<input type="checkbox"/> Edit <input type="checkbox"/> Copy <input type="checkbox"/> Delete	12	CS-11950	Carlos Soltero	Consumer	United States	New York City	New York	10024	East

Рис. 12

Запрос к таблице «orders» для проверки успешности импорта данных (Рис. 13).

Showing rows 0 - 24 (9994 total, Query took 0.0005 seconds.)

SELECT * FROM `orders`

Profiling [Edit inline] [Edit] [Explain SQL] [Create PHP code] [Refresh]

1 > >> Number of rows: 25 Filter rows: Search this table Sort by key: None

Extra options

		row_id	order_date	ship_date	ship_mode	sales	quantity	discount	profit	returned
<input type="checkbox"/>	Edit Copy Delete	1	2018-11-08	2018-11-11	Second Class	261.96	2	0.00	41.91	NULL
<input type="checkbox"/>	Edit Copy Delete	2	2018-11-08	2018-11-11	Second Class	731.94	3	0.00	219.58	NULL
<input type="checkbox"/>	Edit Copy Delete	3	2018-06-12	2018-06-16	Second Class	14.62	2	0.00	6.87	NULL
<input type="checkbox"/>	Edit Copy Delete	4	2017-10-11	2017-10-18	Standard Class	957.58	5	0.40	-383.03	NULL
<input type="checkbox"/>	Edit Copy Delete	5	2017-10-11	2017-10-18	Standard Class	22.37	2	0.20	2.52	NULL
<input type="checkbox"/>	Edit Copy Delete	6	2016-06-09	2016-06-14	Standard Class	48.86	7	0.00	14.17	NULL
<input type="checkbox"/>	Edit Copy Delete	7	2016-06-09	2016-06-14	Standard Class	7.28	4	0.00	1.97	NULL
<input type="checkbox"/>	Edit Copy Delete	8	2016-06-09	2016-06-14	Standard Class	907.15	6	0.20	90.72	NULL
<input type="checkbox"/>	Edit Copy Delete	9	2016-06-09	2016-06-14	Standard Class	18.50	3	0.20	5.78	NULL
<input type="checkbox"/>	Edit Copy Delete	10	2016-06-09	2016-06-14	Standard Class	114.90	5	0.00	34.47	NULL
<input type="checkbox"/>	Edit Copy Delete	11	2016-06-09	2016-06-14	Standard Class	1706.18	9	0.20	85.31	NULL
<input type="checkbox"/>	Edit Copy Delete	12	2016-06-09	2016-06-14	Standard Class	911.42	4	0.20	68.36	NULL
<input type="checkbox"/>	Edit Copy Delete	13	2019-04-15	2019-04-20	Standard Class	15.55	3	0.20	5.44	NULL

Рис. 13

Запрос к таблице «products» для проверки успешности импорта данных (Рис. 14).

Showing rows 0 - 24 (5371 total, Query took 0.0002 seconds.)

SELECT * FROM `products`

Profiling [Edit inline] [Edit] [Explain SQL] [Create PHP code] [Refresh]

1 > >> Number of rows: 25 Filter rows: Search this table Sort by key: None

Extra options

		id	product_id	category	sub_category	product_name	person
<input type="checkbox"/>	Edit Copy Delete	1	OFF-AP-10002578	Office Supplies	Appliances	Fellowes Premier Superior Surge Suppressor, 10-Out...	Chuck Magee
<input type="checkbox"/>	Edit Copy Delete	2	OFF-PA-10000575	Office Supplies	Paper	Wirebound Message Books, Four 2 3/4 x 5 White Form...	Chuck Magee
<input type="checkbox"/>	Edit Copy Delete	3	TEC-MA-10002790	Technology	Machines	NeatDesk Desktop Scanner & Digital Filing System	Kelly Williams
<input type="checkbox"/>	Edit Copy Delete	4	OFF-AR-10000255	Office Supplies	Art	Newell 328	Kelly Williams
<input type="checkbox"/>	Edit Copy Delete	5	TEC-PH-10001061	Technology	Phones	Apple iPhone 5C	Cassandra Brandow
<input type="checkbox"/>	Edit Copy Delete	6	OFF-AR-10003179	Office Supplies	Art	Dixon Ticonderoga Core-Lock Colored Pencils	Anna Andreadi
<input type="checkbox"/>	Edit Copy Delete	7	OFF-AP-10003040	Office Supplies	Appliances	Fellowes 8 Outlet Superior Workstation Surge Prote...	Anna Andreadi
<input type="checkbox"/>	Edit Copy Delete	8	OFF-BI-10004654	Office Supplies	Binders	VariCap6 Expandable Binder	Cassandra Brandow
<input type="checkbox"/>	Edit Copy Delete	9	FUR-CH-10001802	Furniture	Chairs	Hon Every-Day Chair Series Swivel Task Chairs	Anna Andreadi
<input type="checkbox"/>	Edit Copy Delete	10	OFF-PA-10000675	Office Supplies	Paper	Xerox 1919	Chuck Magee
<input type="checkbox"/>	Edit Copy Delete	11	FUR-CH-10004698	Furniture	Chairs	Padded Folding Chairs, Black, 4/Cartron	Cassandra Brandow
<input type="checkbox"/>	Edit Copy Delete	12	OFF-PA-10000241	Office Supplies	Paper	IBM Multi-Purpose Copy Paper, 8 1/2 x 11", Case	Anna Andreadi
<input type="checkbox"/>	Edit Copy Delete	13	FUR-FU-10003268	Furniture	Furnishings	Eldon Radial Chair Mat for Low to Medium Pile Carp...	Cassandra Brandow
<input type="checkbox"/>	Edit Copy Delete	14	OFF-AR-10003560	Office Supplies	Art	Zebra Zazzle Fluorescent Highlighters	Anna Andreadi
<input type="checkbox"/>	Edit Copy Delete	15	OFF-EN-10002986	Office Supplies	Envelopes	#10-4 1/8" x 9 1/2" Premium Diagonal Seam Envelope	Chuck Magee

Рис. 14

Выполнение индивидуального задания:

1. Фильтр по доставке: только Standard Class.

Атрибут «ship_mode» относится к таблице orders, поэтому был добавлен фильтр к трансформации «orders». Настройка компонента «filter rows» (Рис. 15).

Filter rows

Step name: Filter ship_mode

Send 'true' data to step: Value mapper

Send 'false' data to step: Write to log 2

The condition:

ship_mode = Standard Class (String)

Help OK Cancel

Рис. 15

Добавлен компонент для записи логов, если компонент с фильтрацией будет выполнен с ошибкой (Рис. 16).

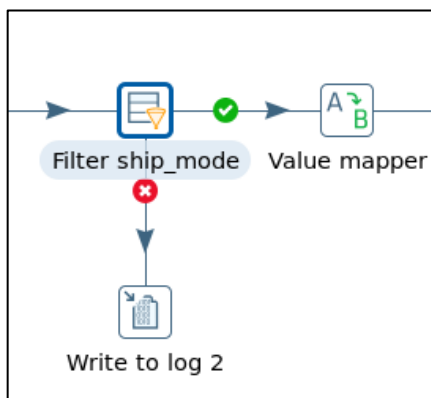


Рис. 16

Общий вид трансформации «orders» (Рис. 17).

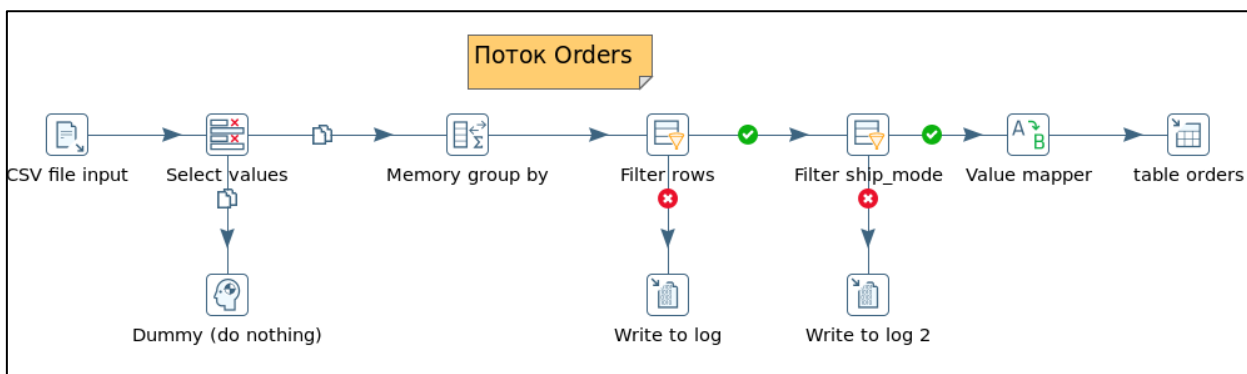


Рис. 17

Была запущена Job, после её выполнения был выполнен SQL запрос на просмотр уникальных значений в таблице «orders» по атрибуту «ship_mode» (Рис. 18).



Рис. 18

Таким образом, данные были успешно отфильтрованы по ship_mode.

2. Статистика продаж:

1) Добавления компонента «CSV file input» для импорта CSV файла (Рис. 19).

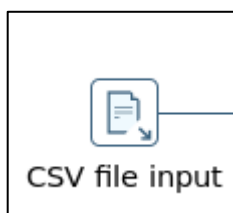


Рис. 19

Настройка компонента (Рис. 20).

Name	Type	Format	Length	Precision	Currency	Decimal	Group	Trim type
1 Row ID	Integer	#	15	0	\$	-	-	none
2 Order ID	String		14		\$	-	-	none
3 Order Date	Date	dd/MM/yyyy			\$	-	-	none
4 Ship Date	Date	dd/MM/yyyy			\$	-	-	none
5 Ship Mode	String		14		\$	-	-	none
6 Customer ID	String		8		\$	-	-	none
7 Customer Name	String		19		\$	-	-	none
8 Segment	String		11		\$	-	-	none
9 Country	String		13		\$	-	-	none
10 City	String		13		\$	-	-	none
11 State	String		12		\$	-	-	none
12 Postal Code	Integer		15	0	\$	-	-	none
13 Region	String		7		\$	-	-	none
14 Product ID	String		15		\$	-	-	none
15 Category	String		15		\$	-	-	none
16 Sub-Category	String		11		\$	-	-	none

Рис. 20

2) Добавление компонента «Replace in string» (Рис. 21).

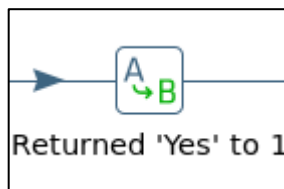


Рис. 21

Настройка компонента: для того, чтобы в дальнейшем проанализировать количество возвратов, необходимо преобразовать значения атрибута «Returned» в числовой тип данных, для этого были заменены значения «Yes» в 1 (Рис. 22).

In stream field	Out stream field	use RegEx	Search	Replace with	Set empty string?	Replace with field	Whole Word	Case sensitive	Is Unicode
1 Returned		N	Yes	1	N		N	N	N

Рис. 22

3) Добавление компонента «Select values» (Рис. 23).

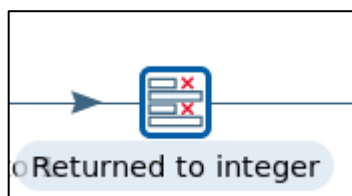


Рис. 23

Настройка компонента: преобразование атрибута «Returned» в тип данных integer (Рис. 24).

Select values

Step nameReturned to integer

Select & AlterRemoveMeta-data

Fields to alter the meta-data for :

Fieldname	Rename to	Type	Length	Precision	Binary to Normal?	Format	Date Format Lenient?	Date Locale
1Returned		Integer	1		N		N	

Рис. 24

4) Добавления компонента «In field value is null» (Рис. 25).

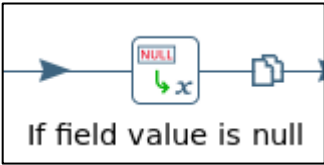


Рис. 25

Настройка компонента: если значение в атрибуте «Returned» пропущено, то заменять его на 0 (Рис. 26).

If field value is null

Step nameIf field value is null

Replace Null for all fields

Replace by value

Set empty string?

Mask (Date)

Select fields

Select value type

Value types

Type	Replace by value	Conversion mask (Date)	Set empty string?
------	------------------	------------------------	-------------------

Fields

Field	Replace by value	Conversion mask (Date)	Set empty string?
1Returned	0		N

Рис. 26

5) Добавление компонента «Calculator» и запись логов при ошибке (Рис. 27).

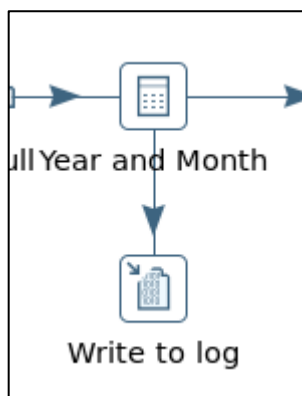


Рис. 27

Настройка компонента: из атрибута «Order Date» получаем два новых атрибута «Year» и «Month» типа данных integer, которые будут использованы при анализе по годам и месяцам (Рис. 28).

Calculator									
Step name									
Year and Month									
<input checked="" type="checkbox"/> Throw an error on non existing files									
Fields:									
	New field	Calculation	Field A	Field B	Field C	Value type	Length	Precision	Ren
1	Year	Year of date A	Order Date			Integer			N
2	Month	Month of date A	Order Date			Integer			N

Рис. 28

6) Добавление компонента «Group by» (Рис. 29).

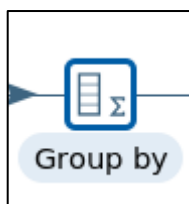


Рис. 29

Настройка компонента: группировка данных по атрибутам регион, категория, год, месяц, возврат (Рис. 30). Агрегирование данных:

- сумма по продажам;
- сумма по выручке;
- количество заказов;
- средняя стоимости скидки;

- сумма по количеству возвратов.

Group by

Step name:

Include all rows? ☐

Temporary files directory:

TMP-file prefix:

Add line number, restart in each group ☐

Line number field name:

Always give back a result row ☐

The fields that make up the group:

Group field

- 1 Region
- 2 Category
- 3 Year
- 4 Month
- 5 Returned

Get Fields

Aggregates :

	Name	Subject	Type
1	total_sales	Sales	Sum
2	total_profit	Profit	Sum
3	orders_count	Order ID	Number of Distinct Values (N)
4	avg_discount	Discount	Average (Mean)
5	total_quantity	Quantity	Sum
6	returns_count	Returned	Sum

Get lookup fields

Рис. 30

7) Добавление компонента «Калькулятор» для расчёта новых метрик и логирование ошибок (Рис. 31).

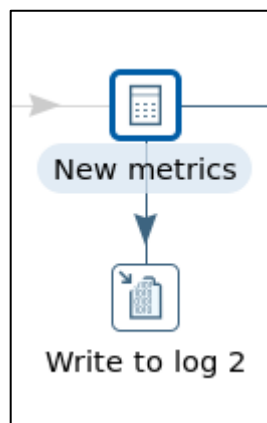


Рис. 31

Настройка компонента: расчёт новых метрик с помощью формул встроенных в калькулятор (Рис. 32):

- средний чек;
- процент возвратов;
- рентабельность.

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	Co
1	avg_order_value	A / B	total_sales	orders_count		Number			N	
2	return_rate	A / B	returns_count	orders_count		Number			N	
3	profitability	100 * A / B	total_profit	total_sales		Number			N	

Рис. 32

8) Добавление компонента «Table output» (Рис. 33).

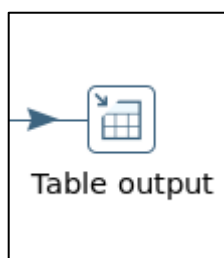


Рис. 33

Настройка компонента: подключение к БД для записи данных в таблицу «sales_statistics» (Рис. 34).

Table output

Step name

Connection

Target schema

Target table

Commit size

Truncate table ☐

Ignore insert errors ☐

Specify database fields ☐

Рис. 34

Общий вид трансформации (Рис. 35).

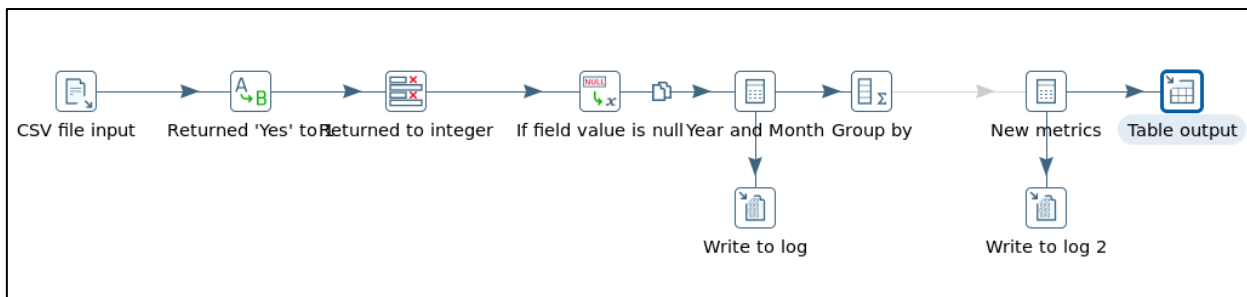


Рис. 35

9) Добавление компонента трансформации для статистики продаж в Job (Рис. 36).

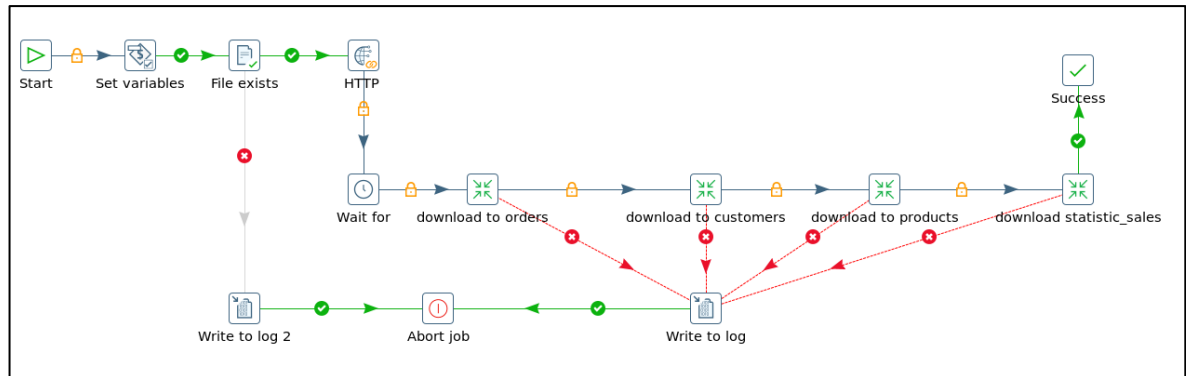


Рис. 36

10) Создание таблицы «sales_statistics» для записи данных в БД (Рис. 37).

```

1 CREATE TABLE sales_statistics (
2     id INT AUTO_INCREMENT PRIMARY KEY,
3     Region VARCHAR(50),
4     Category VARCHAR(50),
5     `Year` INT,
6     Month INT,
7     Returned DECIMAL(5,2),
8     total_sales DECIMAL(12,2),
9     total_profit DECIMAL(12,2),
10    orders_count INT,
11    avg_discount DECIMAL(5,2),
12    return_rate DECIMAL(5,2),
13    profitability DECIMAL(5,2)
14 );
    
```

Рис. 37

Создание индексов (Рис. 38).

```

CREATE INDEX idx_region ON sales_statistics (`Region`);
CREATE INDEX idx_category ON sales_statistics (`Category`);
CREATE INDEX idx_year_month ON sales_statistics (`Year`, `Month`);
    
```

Рис. 38

11) Запросы к базе данных для проверки импорта данных:

- Топ 3 региона по сумме продаж (Рис. 39).

Showing rows 0 - 2 (3 total, Query took 0.0027 seconds.) [total_sales: 23459.78... - 13999.96...]

```
SELECT region, total_sales FROM sales_statistics ORDER BY total_sales DESC LIMIT 3;
```

Profiling [Edit inline] [Edit] [Explain SQL] [Create PHP code] [Refresh]

Extra options

	region	total_sales
<input type="checkbox"/> Edit Copy Delete	South	23459.78
<input type="checkbox"/> Edit Copy Delete	Central	17499.95
<input type="checkbox"/> Edit Copy Delete	West	13999.96

Рис. 39

- Сумма продаж по годам и месяцам (Рис. 40).

Showing rows 0 - 24 (48 total, Query took 0.0043 seconds.) [year: 2016... - 2018...] [month: 1... - 1...]

```
SELECT year, month, SUM(total_sales) AS total_sales FROM sales_statistics GROUP BY year, month ORDER BY year, month;
```

Profiling [Edit inline] [Edit] [Explain SQL] [Create PHP code] [Refresh]

1 > >> ☐ Show all Number of rows: 25 Filter rows: Search this table

Extra options

	year	month	total_sales
<input type="checkbox"/> Edit Copy Delete	2016	1	14236.90
<input type="checkbox"/> Edit Copy Delete	2016	2	4519.90
<input type="checkbox"/> Edit Copy Delete	2016	3	55691.02
<input type="checkbox"/> Edit Copy Delete	2016	4	28295.35
<input type="checkbox"/> Edit Copy Delete	2016	5	23648.29
<input type="checkbox"/> Edit Copy Delete	2016	6	34595.13
<input type="checkbox"/> Edit Copy Delete	2016	7	33946.36
<input type="checkbox"/> Edit Copy Delete	2016	8	27909.46
<input type="checkbox"/> Edit Copy Delete	2016	9	81777.34
<input type="checkbox"/> Edit Copy Delete	2016	10	31453.38
<input type="checkbox"/> Edit Copy Delete	2016	11	78628.75
<input type="checkbox"/> Edit Copy Delete	2016	12	69545.63
<input type="checkbox"/> Edit Copy Delete	2017	1	18174.09
<input type="checkbox"/> Edit Copy Delete	2017	2	11951.39
<input type="checkbox"/> Edit Copy Delete	2017	3	38726.25
<input type="checkbox"/> Edit Copy Delete	2017	4	34195.25

Рис. 40

- Вывод категории с наибольшим процентом возвратов (Рис. 41).

Showing rows 0 - 0 (1 total, Query took 0.0023 seconds.) [return_rate: 6.00... - 6.00...]

```
SELECT category, return_rate FROM sales_statistics WHERE return_rate > 5 ORDER BY return_rate DESC;
```

Profiling [Edit inline] [Edit] [Explain SQL] [Create PHP code] [Refresh]

☐ Show all Number of rows: 25 Filter rows: Search this table

Extra options

	category	return_rate
<input type="checkbox"/> Edit Copy Delete	Office Supplies	6.00

Рис. 41

3. Анализ по городам:

1) Добавление компонента для импорта данных из CSV (Рис. 42).

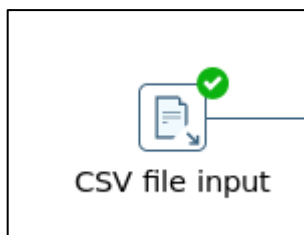


Рис. 42

Настройка компонента для импорта данных (Рис. 43).

The screenshot shows the 'CSV file input' configuration window. The 'Step name' is 'CSV file input'. The 'Filename' is '\${CSV_FILE_PATH}'. The 'Delimiter' is ';'. The 'Enclosure' is '"'. The 'NIO buffer size' is '50000'. The 'Lazy conversion?' checkbox is checked. The 'Header row present?' checkbox is checked. The 'Add filename to result' checkbox is unchecked. The 'The row number field name (optional)' field is empty. The 'Running in parallel?' checkbox is unchecked. The 'New line possible in fields?' checkbox is unchecked. The 'Format' is 'mixed'. The 'File encoding' is empty. Below these settings is a table with 16 rows and 9 columns: Name, Type, Format, Length, Precision, Currency, Decimal, Group, and Trim type.

	Name	Type	Format	Length	Precision	Currency	Decimal	Group	Trim type
1	Row ID	Integer	#	15	0	\$	-	,	none
2	Order ID	String		14		\$	-	,	none
3	Order Date	Date	dd/MM/yyyy			\$	-	,	none
4	Ship Date	Date	dd/MM/yyyy			\$	-	,	none
5	Ship Mode	String		14		\$	-	,	none
6	Customer ID	String		8		\$	-	,	none
7	Customer Name	String		19		\$	-	,	none
8	Segment	String		11		\$	-	,	none
9	Country	String		13		\$	-	,	none
10	City	String		13		\$	-	,	none
11	State	String		12		\$	-	,	none
12	Postal Code	Integer	#	15	0	\$	-	,	none
13	Region	String		7		\$	-	,	none
14	Product ID	String		15		\$	-	,	none
15	Category	String		15		\$	-	,	none
16	Gift_Catename	String		11		\$	-	,	none

Рис. 43

2) Добавление компонента для валидации данных и запись логов при ошибке (Рис. 44).

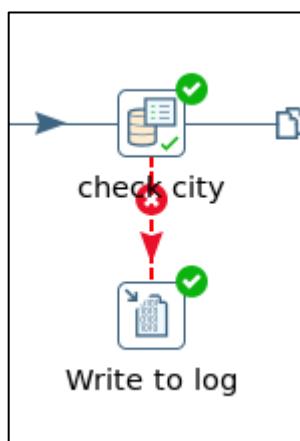


Рис. 44

Настройка компонента: значения в атрибуте «City» не могут быть пустыми, а также длина значения должна быть больше 2 (Рис. 45).

The screenshot shows the 'Data validator' configuration window. The 'Stepname' is 'check city'. The 'Validation description' is 'check city'. The 'Name of field to validate' is 'City'. The 'Error code' is '400' and the 'Error description' is 'error city'. Under the 'Type' section, 'Verify data type?' is unchecked, 'Data type' is 'None', and 'Conversion mask', 'Decimal Symbol', and 'Grouping Symbol' are empty. Under the 'Data' section, 'Null allowed?', 'Only null values allowed?', and 'Only numeric data expected' are all unchecked. 'Max string length' is empty, 'Min string length' is '2', and 'Maximum value', 'Minimum value', 'Expected start string', 'Expected end string', and 'Not allowed start string' are all empty.

Рис. 45

3) Добавления компонента замены пустых значений (Рис. 46).

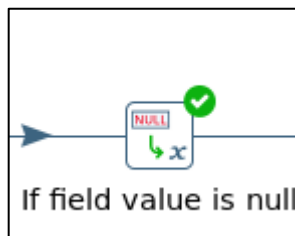


Рис. 46

Настройка компонента: заменить пустое значение в атрибуте «City» на «Unknown» (Рис. 47).

If field value is null [X]

Step name:

Replace Null for all fields

Replace by value:

Set empty string? ☐

Mask (Date):

Select fields ☒

Select value type ☐

Value types

Type	Replace by value	Conversion mask (Date)	Set empty string?

Fields

Field	Replace by value	Conversion mask (Date)	Set empty string?
1 City	Unknown		N

[?] Help [OK] [Get Fields] [Cancel]

Рис. 47

- 4) Добавления компонента «Калькулятор» (Рис. 48).

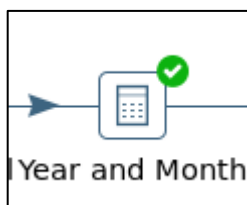


Рис. 48

Настройка компонента: из атрибута «Order Date» берутся значения «Year», «Month» и записываются в новые атрибуты (Рис. 49). Эти атрибуты потребуются для дальнейшего анализа.

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	Conversion mask	Decimal symbol	Group
1	Year	Year of date A	Order Date			Integer			N			
2	Month	Month of date A	Order Date			Integer			N			

Рис. 49

- 5) Добавление компонента «Group by» (Рис. 50).



Рис. 50

Настройка компонента: группировка данных по атрибутам город, месяц, год. Агрегирование данных (Рис. 51):

- Сумма продаж;
- Сумма по выручке;
- Количество заказов;
- Средняя стоимость скидки.

Group by

Step name:

Include all rows? ☐

Temporary files directory:

TMP-file prefix:

Add line number, restart in each group ☐

Line number field name:

Always give back a result row ☐

The fields that make up the group:

▲ Group field

1 City

2 Month

3 Year

Aggregates :

	Name	Subject	Type	Value
1	total_sales	Sales	Sum	
2	total_profit	Profit	Sum	
3	orders_count	Order ID	Number of Distinct Values (N)	
4	avg_discount	Discount	Average (Mean)	

Рис. 51

6) Добавление компонента «Калькулятор» для расчёта новой метрики (Рис. 52).

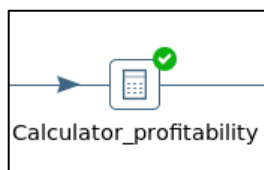


Рис. 52

Настройка компонента: расчёт новой метрики – рентабельность (Рис. 53).

Calculator

Step name
Calculator_profitability

☒ Throw an error on non existing files

Fields:

	New field	Calculation	Field A	Field B	Field C	Value type	Length	Precision	Remove
1	profitability	100 * A / B	total_profit	total_sales		Number			N

Рис. 53

7) Добавление компонента для экспорта данных (Рис. 54).

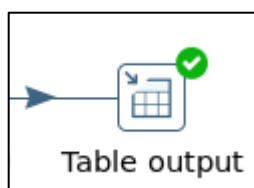


Рис. 54

Настройка компонента: подключение к БД для записи данных в таблицу «city_stats» (Рис. 55).

Table output

Step name
Table output

Connection
city_stats

Target schema
mgpu_ico_etl_08

Target table
city_stats

Commit size
1000

Truncate table ☐

Ignore insert errors ☐

Specify database fields ☐

Main options Database fields

Partition data over tables ☐

Partitioning field

Partition data per month ☒

Partition data per day ☐

Use batch update for inserts ☒

Is the name of the table defined in a field? ☐

Field that contains name of table:

Store the tablename field ☒

Return auto-generated key ☐

Name of auto-generated key field

Help OK Cancel SQL

Рис. 55

Общий вид трансформации (Рис. 56).

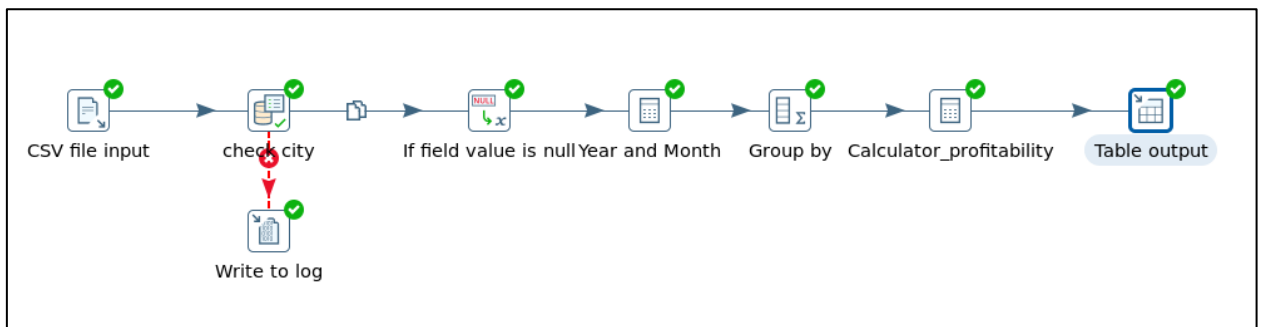


Рис. 56

8) Добавление компонента трансформации в Job (Рис. 57).

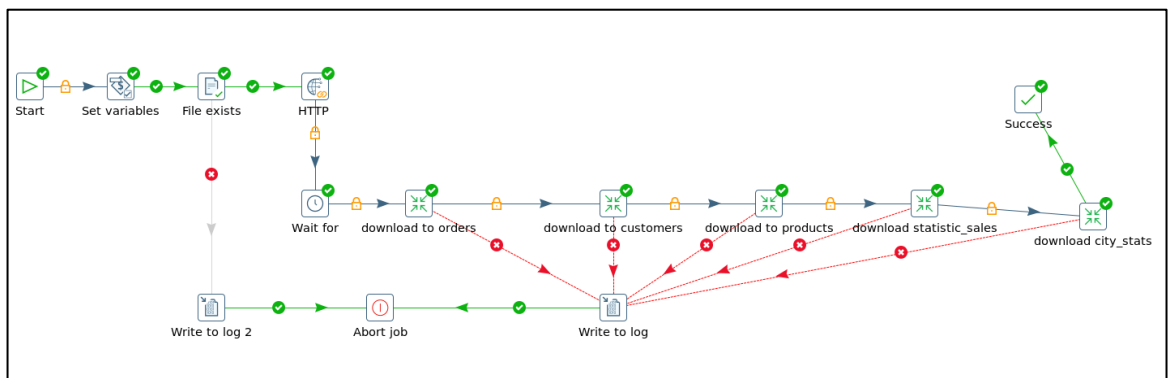


Рис. 57

Настройка компонента трансформации (Рис. 58).

Transformation

Entry Name:

Transformation:

Options | Logging | Arguments | Parameters

Run configuration:

Execution

- ☐ Execute every input row
- ☐ Clear results rows before execution
- ☐ Clear results files before execution
- ☒ Wait for remote transformation to complete
- ☐ Follow local abort to remote transformation
- ☐ Suppress result data from remote transformation

Рис. 58

9) Создание таблицы «city_stats» для экспорта данных в БД (Рис. 59).

```
1 CREATE TABLE city_stats (  
2   city VARCHAR(100) PRIMARY KEY,  
3   `Year` INT,  
4   `Month` int,  
5   total_sales DECIMAL(12,2),  
6   total_profit DECIMAL(12,2),  
7   orders_count INT,  
8   avg_discount DECIMAL(5,2),  
9   return_rate DECIMAL(5,2),  
10  profitability DECIMAL(5,2)  
11 );
```

Рис. 59

Создание индексов (Рис. 60).

```
1 CREATE INDEX idx_city ON city_stats (city);  
2 CREATE INDEX idx_profitability ON city_stats (profitability);
```

Рис. 60

10) Запрос к базе данных для проверки экспорта данных Рис. 61).

The screenshot shows a database query interface. At the top, a SQL query is entered: `SELECT * FROM 'city_stats'`. Below the query bar, there are options for Profiling, Edit inline, Edit, Explain SQL, Create PHP code, and Refresh. A toolbar shows the first page selected, navigation arrows, and settings for Number of rows (25), Filter rows (Search this table), and Sort by key (None). An 'Extra options' button is also present. The main area displays a table with 8 columns: city, Year, Month, total_sales, total_profit, orders_count, avg_discount, and profitability. The table contains 15 rows of data for various cities in 2016.

city	Year	Month	total_sales	total_profit	orders_count	avg_discount	profitability
Houston	2016	4	697.07	61.79	1	0.20	8.86
Jackson	2016	11	508.62	219.08	1	0.00	43.07
Chicago	2016	12	2.39	-6.34	1	0.80	-265.00
El Paso	2016	12	803.96	-30.96	1	0.20	-3.85
Houston	2016	10	5.31	-1.59	1	0.60	-30.00
Indianapolis	2016	12	1107.66	498.52	1	0.00	45.01
Chicago	2016	9	331.54	-82.88	1	0.20	-25.00
Huntsville	2016	10	1488.77	117.40	1	0.50	7.89
Chicago	2016	9	10.90	3.41	1	0.20	31.25
El Paso	2016	2	11.36	-4.52	1	0.50	-39.80
Houston	2016	1	16.45	5.55	1	0.20	33.75
Fort Worth	2016	5	225.02	-70.73	1	0.30	-31.43
Laredo	2016	12	609.61	63.25	1	0.30	10.38
Houston	2016	12	60.42	6.04	1	0.20	10.00

Рис. 61

11) Запросы к таблице для аналитики по городам:

- Анализ суммы продаж и суммы выручки по городам (Рис. 62).

Showing rows 0 - 24 (531 total. Query took 0.0058 seconds.)

```
SELECT city, SUM(total_sales) AS total_sales, SUM(total_profit) AS total_profit FROM city_stats GROUP BY city ORDER BY total_sales DESC;
```

☐ Profiling [\[Edit inline \]](#) [\[Edit \]](#) [\[Explain SQL \]](#) [\[Create PHP code \]](#) [\[Refresh \]](#)

1 > >> | Number of rows: 25 Filter rows:

Extra options

city	total_sales	total_profit
New York City	256368.12	62036.94
Los Angeles	175851.36	30440.85
Seattle	119540.73	29156.03
San Francisco	112669.13	17507.38
Philadelphia	109077.12	-13837.81
Houston	64504.77	-10153.51
Chicago	48539.62	-6654.58
San Diego	47521.03	6377.20
Jacksonville	44713.19	-2323.81
Springfield	43054.31	6200.71
Detroit	42446.94	13181.83
Columbus	38706.25	5897.11
Newark	28576.14	5793.78
Columbia	25283.31	5606.13

Рис. 62

- Анализ количества заказов по города и размер скидки (Рис. 63).

```
SELECT city, COUNT(orders_count) AS total_orders, AVG(avg_discount) AS average_discount FROM city_stats GROUP BY city ORDER BY total_orders DESC;
```

☐ Profiling [\[Edit inline \]](#) [\[Edit \]](#) [\[Explain SQL \]](#) [\[Create PHP code \]](#) [\[Refresh \]](#)

1 > >> | Number of rows: 25 Filter rows:

Extra options

city	total_orders	average_discount
New York City	434	0.054608
Los Angeles	379	0.072559
San Francisco	263	0.066160
Philadelphia	260	0.318462
Seattle	210	0.066190
Houston	186	0.377957
Chicago	167	0.389222
Columbus	109	0.158716
San Diego	87	0.077011
Dallas	79	0.354430

Рис. 63

- Анализ рентабельности по городам (Рис. 64).

```
SELECT city, SUM(`profitability`) as profitability FROM city_stats GROUP BY city ORDER BY profitability DESC;
```

☐ Profiling [\[Edit inline \]](#) [\[Edit \]](#) [\[Explain SQL \]](#) [\[Create PHP code \]](#) [\[Refresh \]](#)

1 > >> | Number of rows: 25 Filter rows:

Extra options

city	profitability
New York City	12014.80
Los Angeles	9389.40
San Francisco	7077.08
Seattle	5248.16
San Diego	1976.28
Detroit	1624.45
Columbus	1461.21
Richmond	1438.63
Jackson	1069.32

Рис. 64

Выводы по работе:

В ходе выполнения лабораторной работы были успешно реализованы ключевые этапы проектирования ETL-процесса в Pentaho Data Integration. Основной целью работы являлось освоение навыков интеграции данных из различных источников с использованием динамических соединений. Для этого были созданы подключения к CSV-файлам и базе данных MySQL, а также настроены трансформации для обработки данных, включая фильтрацию, замену значений и агрегацию. Результаты работы подтвердили корректность настройки компонентов: данные были загружены в таблицы orders, customers, products, а затем преобразованы и сохранены в аналитических таблицах sales_statistics и city_stats.

Проверка результатов осуществлялась через SQL-запросы, которые подтвердили отсутствие ошибок в данных и корректность агрегации. Например, запросы на определение топ-3 регионов по продажам и анализ рентабельности по городам продемонстрировали работоспособность реализованных процессов. Для оптимизации производительности были созданы индексы в таблицах БД.