

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:

Проектный практикум по разработке ETL-решений

Лабораторная работа №2-1

Тема:

«Динамические соединения с базами данных»

Выполнил(а): Морозова Валерия АДЭУ-211

Преподаватель:

Москва

2025

Цель работы: получить практические навыки создания ETL-процесса для загрузки данных из CSV-файла в базу данных MySQL с использованием Pentaho Data Integration.

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.

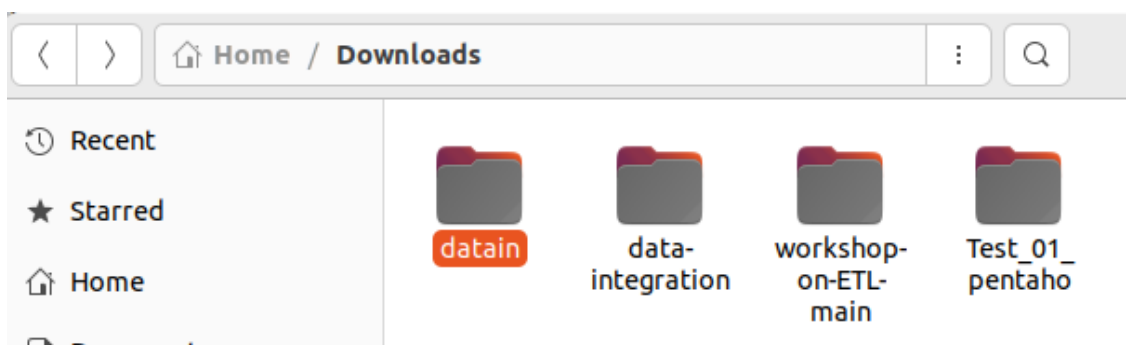


Рисунок 1. Создание директории datain

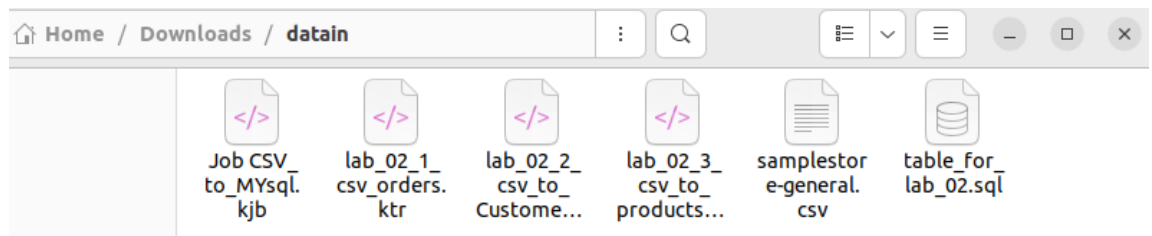


Рисунок 2. Пополнение директория необходимыми файлами

Теперь необходимо открыть их и запустить в Pentaho предварительно изменив пути

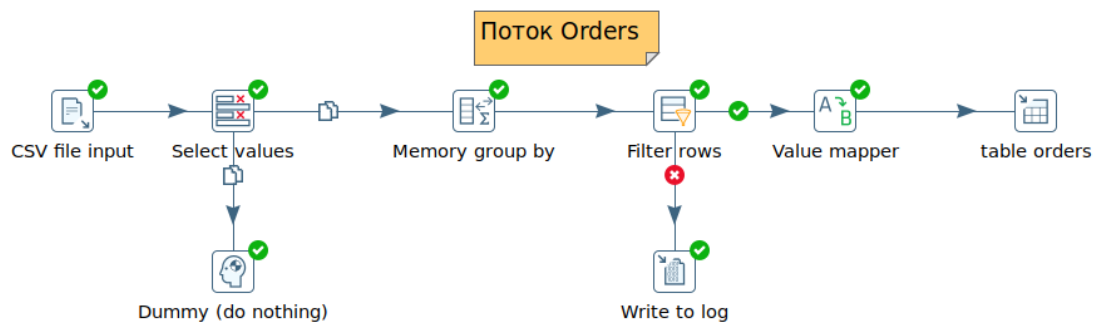


Рисунок 3. Трансформация lab_02_1_csv_orders.ktr

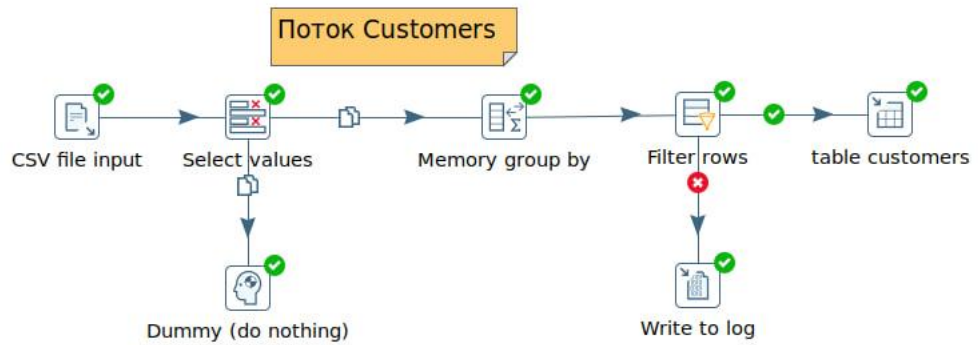


Рисунок 4. Трансформация lab_02_2_csv_customers.ktr

Трансформация с добавленным фильтром по стране: только United States (Вариант 10).

Рисунок 5. Фильтр только United States

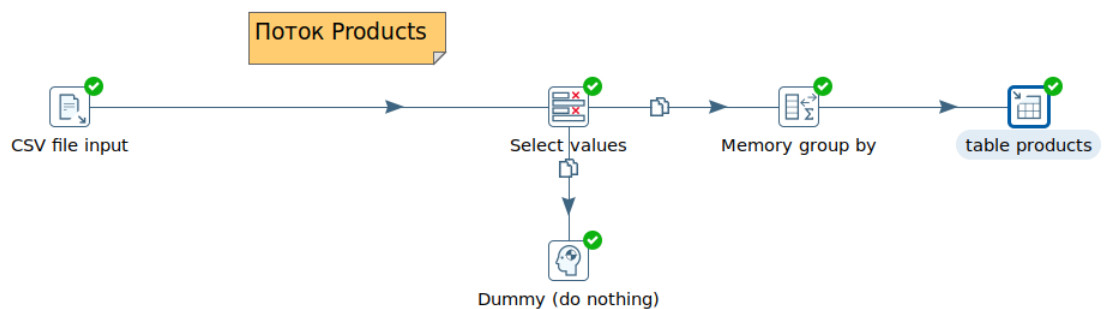


Рисунок 6. Трансформация lab_02_3_csv_products.ktr

Server: localhost:3306 » Database: mgpu_ico_etl_10 » Table: orders

Showing rows 0 - 24 (9994 total, Query took 0.0002 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

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

Рисунок 7. В MySQL была создана и заполнена данными таблица orders

Server: localhost:3306 » Database: mgpu_ico_etl_10 » Table: customers

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

	id	customer_id	customer_name	segment	country	city	state	postal_code	region
<input type="checkbox"/>	1	CC-12670	Craig Carreira	Consumer	United States	Chicago	Illinois	60610	Central
<input type="checkbox"/>	2	SO-20335	Sean O'Donnell	Consumer	United States	Fort Lauderdale	Florida	33311	South
<input type="checkbox"/>	3	BS-11590	Brendan Sweed	Corporate	United States	Columbus	Indiana	47201	Central
<input type="checkbox"/>	4	RF-19840	Roy Franz	Consumer	United States	Chesapeake	Virginia	23320	South
<input type="checkbox"/>	5	DR-12880	Dan Reichenbach	Corporate	United States	Inglewood	California	90301	West
<input type="checkbox"/>	6	JE-15745	Joel Eaton	Consumer	United States	Newark	Ohio	43055	East
<input type="checkbox"/>	7	SJ-20215	Sarah Jordan	Consumer	United States	Columbia	Tennessee	38401	South

Рисунок 8. В MySQL была создана и заполнена данными таблица customers

Server: localhost:3306 » Database: mgpu_ico_etl_10 » Table: products

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

SELECT * FROM `products`

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

	id	product_id	category	sub_category	product_name	person
<input type="checkbox"/>	1	OFF-AP-10002578	Office Supplies	Appliances	Fellowes Premier Superior Surge Suppressor, 10-Out...	Chuck Magee
<input type="checkbox"/>	2	OFF-PA-10000575	Office Supplies	Paper	Wirebound Message Books, Four 2 3/4 x 5 White Form...	Chuck Magee
<input type="checkbox"/>	3	TEC-MA-10002790	Technology	Machines	NeatDesk Desktop Scanner & Digital Filing System	Kelly Williams
<input type="checkbox"/>	4	OFF-AR-10000255	Office Supplies	Art	Newell 328	Kelly Williams
<input type="checkbox"/>	5	TEC-PH-10001061	Technology	Phones	Apple iPhone 5C	Cassandra Brandow
<input type="checkbox"/>	6	OFF-AR-10003179	Office Supplies	Art	Dixon Ticonderoga Core-Lock Colored Pencils	Anna Andreadi
<input type="checkbox"/>	7	OFF-AP-	Office	Appliances	Fellowes 8 Outlet Superior Workstation	Anna Andreadi

Рисунок 9. В MySQL была создана и заполнена данными таблица products

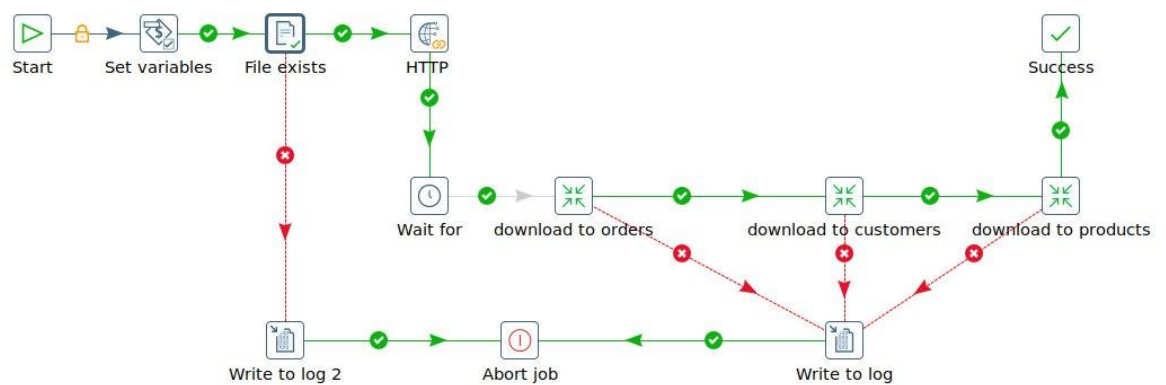


Рисунок 10. Job CSV_to_Mysql отработал после загрузки файла по протоколу HTTP

Индивидуальные задания

Анализ скидок

Server: localhost:3306 » Database: mgpu_ico_etl_10 » Table: analys_discounts

MySQL returned an empty result set (i.e. zero rows). (Query took 0.0003 seconds.)

SELECT * FROM `analys_discounts`

row_id	order_date	customer_id	customer_name	segment	product_name	discount	max_discount
--------	------------	-------------	---------------	---------	--------------	----------	--------------

Рисунок 11. Создала таблицу с нужными столбцами для анализа скидок

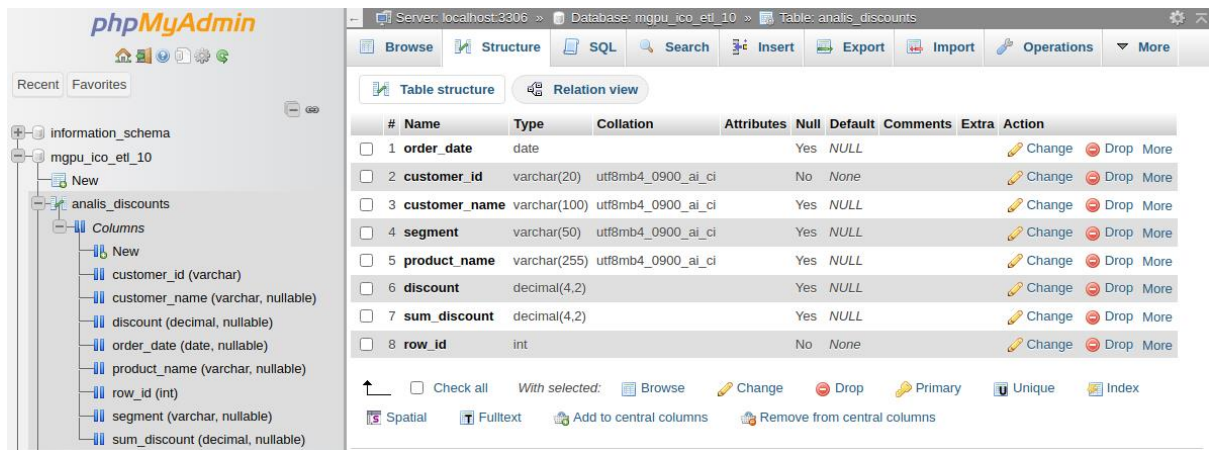


Рисунок 12. Скорректировала типы данных

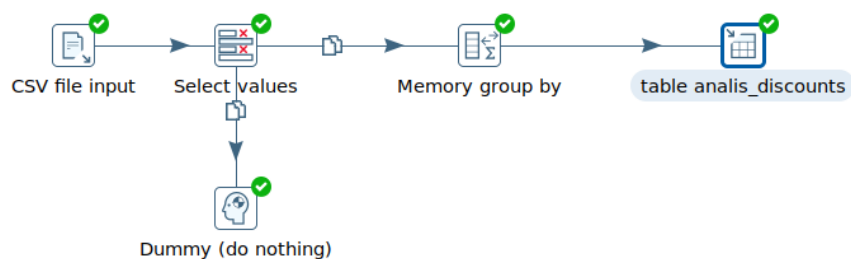


Рисунок 13. Трансформация lab_02_4

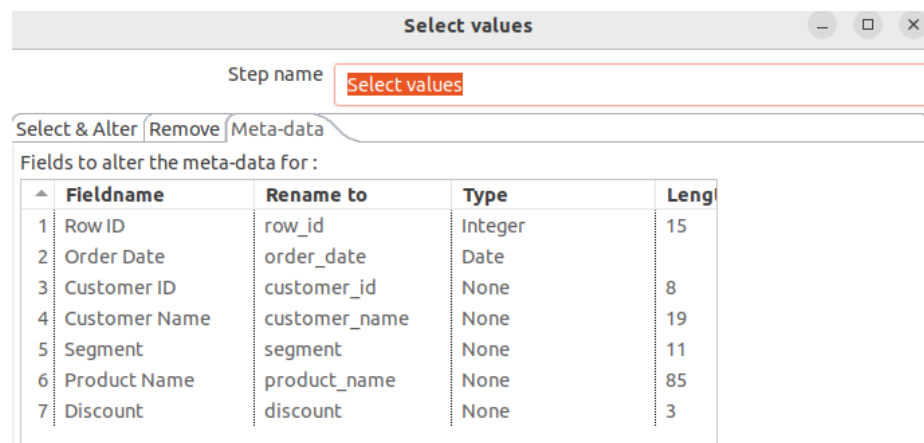


Рисунок 14. Выбор нужных столбцов

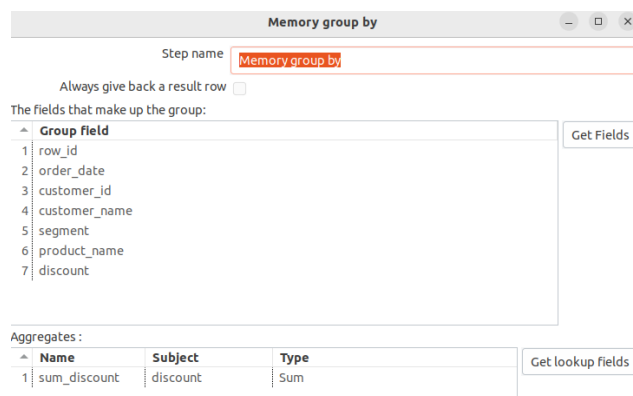


Рисунок 15. Группировка по сумме скидок

Server: localhost:3306 Database: mgpu_ico_etl_10 Table: analys_discounts

Current selection does not contain a unique column. Grid edit, checkbox, Edit, Copy and Delete features are not available.

Showing rows 0 - 24 (9994 total, Query took 0.0069 seconds.) [discount: 0.80... - 0.80...]

SELECT * FROM `analys_discounts` ORDER BY `discount` DESC

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

Number of rows: 25 Filter rows: Search this table

Extra options

order_date	customer_id	customer_name	segment	product_name	discount	sum_discount	row_id
2018-10-10	WB-21850	William Brown	Consumer	Wilson Jones Ledger-Size, Piano-Hinge Binder, 2", ...	0.80	0.80	9836
2019-10-02	SC-20695	Steve Chapman	Corporate	Fellowes Superior 10 Outlet Split Surge Protector	0.80	0.80	9420
2019-06-19	ZC-21910	Zuschuss Carroll	Consumer	Hoover Replacement Belt for Commercial Guardsman H...	0.80	0.80	4102
2019-12-23	MM-17920	Michael Moore	Consumer	Acco 6 Outlet Guardian Premium Plus Surge Suppress...	0.80	0.80	3546
2019-04-30	TC-21475	Tony Chapman	Home Office	GBC Linen Binding Covers	0.80	0.80	4758

Рисунок 16. Данные загружены в таблицу analys_discount

SELECT segment, COUNT(sum_discount) AS discount_count FROM analys_discounts GROUP BY segment ORDER BY discount_count DESC;

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

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

Extra options

segment	discount_count
Consumer	5191
Corporate	3020
Home Office	1783

Рисунок 17. Количество скидок для разных сегментов потребителей

Как видно, на первом месте по количеству получаемых скидок и акций на товары у обычных потребителей, на втором месте корпорации и на последнем домашний офис.

Showing rows 0 - 4 (5 total, Query took 0.0092 seconds.)

SELECT product_name, COUNT(sum_discount) AS discount_count FROM analys_discounts GROUP BY product_name ORDER BY discount_count DESC LIMIT 5;

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

Extra options

product_name	discount_count
Staple envelope	48
Easy-staple paper	46
Staples	46
Avery Non-Stick Binders	20
Staples in misc. colors	19

Рисунок 18. На какие товары чаще всего устанавливают скидки

Отчет по регионам

Step name:

Select & Alter Remove Meta-data

Fields to alter the meta-data for:

	Fieldname	Rename to	Type	Length
1	Row ID	row_id	Integer	15
2	Region	region	None	7
3	Sales	sales	None	7
4	Profit	profit	None	9
5	State	state	None	

Рисунок 19. Выбор нужных столбцов для анализа

Step name:

Always give back a result row ☐

The fields that make up the group:

	Group field
1	row_id
2	region
3	sales
4	profit
5	state

Get Fields

Aggregates:

	Name	Subject	Type	Value
1	sum_sales	sales	Sum	

Get lookup fields

Рисунок 20. Группировка значений по сумме продаж

Step name:

Connection: Edit... New... Wizard...

Target schema: Browse...

Target table: Browse...

Commit size:

Truncate table ☐

Ignore insert errors ☒

Specify database fields ☒

Main options Database fields

Fields to insert:

	Table field	Stream field
1	row_id	row_id
2	region	region
3	sales	sales
4	profit	profit
5	state	state
6	sum_sales	sum_sales

Get fields

Enter field mapping

Рисунок 21. Выгрузка данных в таблицу regions_report

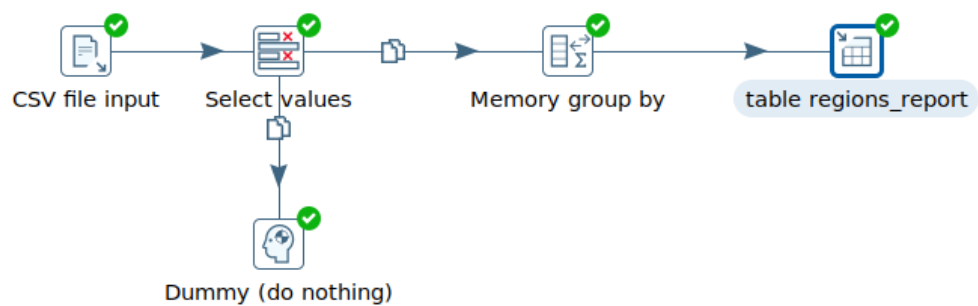


Рисунок 22. Трансформация lab_02_5

phpMyAdmin interface showing the structure and data of the `regions_report` table in the `mgpu_ico_etl_10` database.

Table Structure:

- `profit` (decimal, nullable)
- `region` (varchar, nullable)
- `row_id` (PRI, int)
- `sales` (decimal, nullable)
- `state` (varchar, nullable)
- `sum_sales` (decimal)

Table Data (Rows 0 - 24):

row_id	state	region	sales	profit	sum_sales
1	Kentucky	South	262	42	262
2	Kentucky	South	732	220	732
3	California	West	15	7	15
4	Florida	South	958	-383	958
5	Florida	South	22	3	22
6	California	West	49	14	49
7	California	West	7	2	7
8	California	West	907	91	907
9	California	West	19	6	19

Рисунок 23. Данные загружены в таблицу regions_report

SQL query results showing profit by region:

```

SELECT region, SUM(profit) AS sum_profit FROM regions_report GROUP BY region ORDER BY `sum_profit` DESC
    
```

Table Data (Rows 0 - 3):

region	sum_profit
West	108386
East	91521
South	46721
Central	39719

Рисунок 24. Прибыль по регионам

Согласно sql запросу видно, что больше всего выручки делает западный регион, а меньше всего центральный.

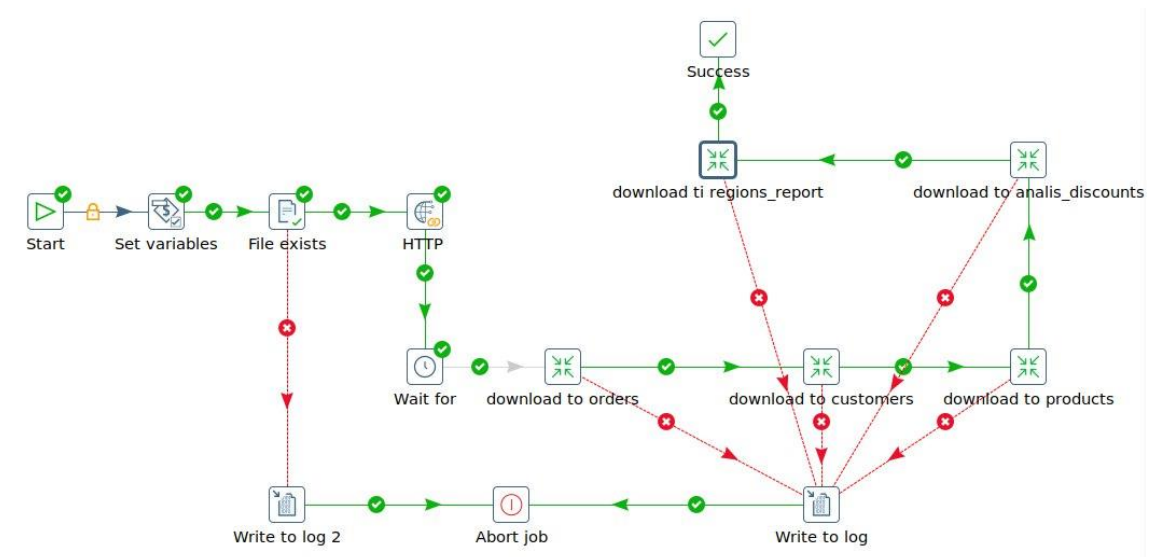


Рисунок 25. Итоговый Job CSV_to_Mysql с двумя новыми трансформациями

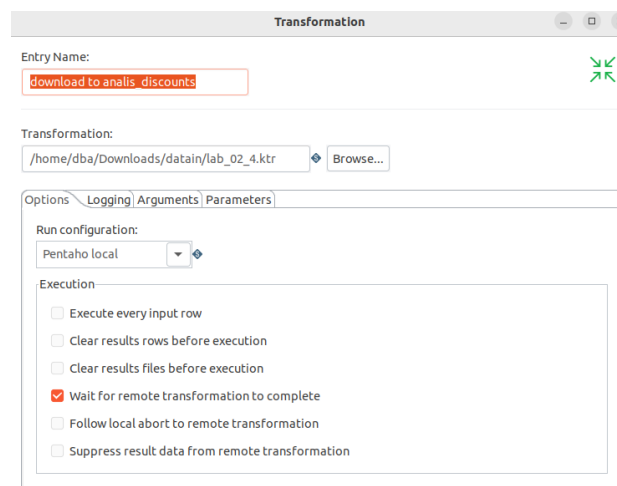


Рисунок 26. Настройка степа трансформации по анализу скидок

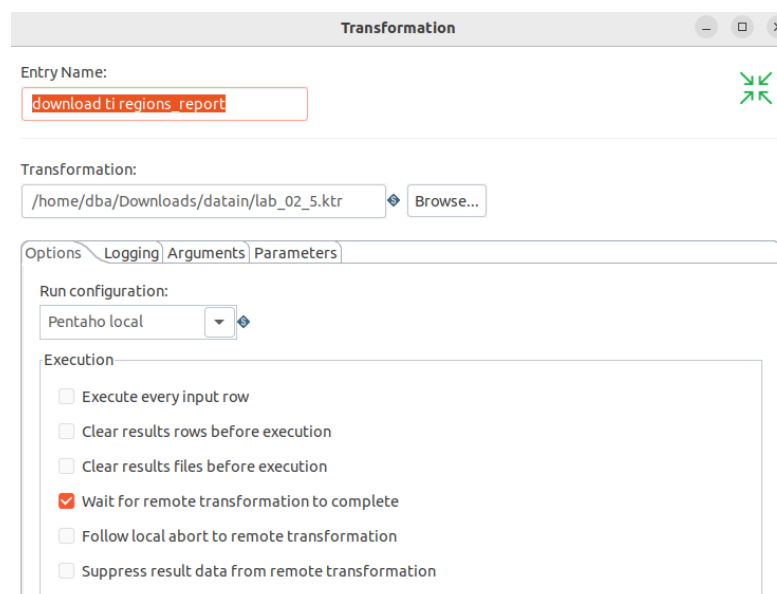


Рисунок 26. Настройка степа трансформации по отчету по регионам

Выводы

Выполнены все трансформации

Успешно выполнена и дополнена двумя трансформациями, согласно индивидуальному заданию, структура job:

Start

- |— Set Variables (FILE_PATH)
- |— Check File Exists
- |— HTTP Download
- |— Transform: Load Orders (lab_02_1)
- |— Transform: Load Customers (lab_02_2)
- |— Transform: Load Products (lab_02_3)
- |— Transform: Analytics 1 (lab_02_4)
- |— Transform: Analytics 2 (lab_02_5)

Проведен анализ скидок, согласно которому, видно, что более всего потребителями и корпорациями востребованы срочные конверты, легко скрепляемая бумага, скобы и антипригарные папки, соответственно больше всего скидок именно на эти товары.

Сделан отчет по регионам, где видно, что больше всего выручки делает западный регион, а меньше всего центральный.

Контрольные вопросы

1. Что такое динамические соединения в Pentaho Data Integration?

Динамические соединения в Pentaho Data Integration позволяют создавать подключения к базам данных или другим источникам данных на лету, основываясь на значениях, полученных во время выполнения трансформации. Это особенно полезно, когда нужно подключаться к различным источникам данных в зависимости от условий, таких как параметры, заданные пользователем, или результаты предыдущих шагов трансформации. Используя динамические соединения, можно избежать необходимости жестко задавать параметры подключения в трансформации.

2. Как организовать обработку ошибок в трансформации?

Обработка ошибок в PDI может быть организована с помощью нескольких методов:

"Error Handling": Многие компоненты (например, "Table Output" или "Text File Output") имеют встроенные опции для обработки ошибок, такие как запись ошибок в отдельный файл или таблицу.

"Error Rows": Вы можете настроить поток данных так, чтобы ошибки направлялись в отдельный поток, например, через "Filter Rows" или "Switch/Case", где можно обрабатывать их отдельно.

"Try-Catch": В PDI также доступны "Try-Catch" блоки, которые позволяют обрабатывать исключения и выполнять альтернативные действия в случае ошибок.

3. Какие методы выявления дублей существуют?

Существует несколько методов выявления дублей в PDI:

"Unique Rows": Этот шаг позволяет удалить дубликаты на основе заданных полей.

"Group By": С помощью этого шага можно агрегировать данные и выявить дубликаты, используя группировку по определённым полям.

"Row Normalizer": Этот шаг позволяет нормализовать строки и выявить дублирующиеся записи.

Сравнение данных: Можно использовать "Merge Join" или "Join" для сравнения данных из разных источников и выявления дублей.

4. Как настроить параметризацию подключений?

Параметризацию подключений в PDI можно настроить следующим образом:

- ❖ **Создание параметров**: В диалоговом окне "Transformation" или "Job" можно создать переменные, которые будут использоваться в качестве параметров.

- ❖ **Использование переменных**: В настройках подключения к базе данных указать параметры подключения (например, имя пользователя,

пароль, URL) с использованием переменных, например, `${db_user}` или `${db_password}`.

❖ Передача значений: При запуске трансформации или задания можно передавать значения для этих переменных, что позволит динамически изменять параметры подключения.

5. Какие компоненты Pentaho Data Integration используются для объединения данных?

Для объединения данных в PDI используются следующие компоненты:

"Merge Join": Позволяет объединять данные из двух потоков на основе общего поля.

"Join Rows": Этот шаг также позволяет объединять строки из разных источников, используя заданные ключи.

"Union": Объединяет строки из нескольких источников в один поток.

"Group By": Хотя в первую очередь используется для агрегации, этот компонент также может помочь в объединении данных, группируя их по определённым полям.