

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 №1-1

Тема:

«Установка и настройка ETL-инструмента. Создание конвейеров
данных»

Выполнил(а): Морозова Валерия АДЭУ-211

Преподаватель:

Москва

2025

Цель работы: изучение основных принципов работы с ETL-инструментами на примере Pentaho Data Integration (PDI), настройка конвейера обработки данных, фильтрация и замена значений в Excel файле, а также выгрузка обработанных данных в базу данных MySQL/PostgreSQL.

Задачи:

– Настроить среду для работы с Pentaho Data Integration (PDI):

Запуск виртуальной машины с Ubuntu 22.04 в VirtualBox.

Развертывание Pentaho Data Integration.

– Создать ETL-конвейер:

Загрузить данные из CSV-файла.

Очистить, преобразовать и отфильтровать данные.

Выполнить замену значений.

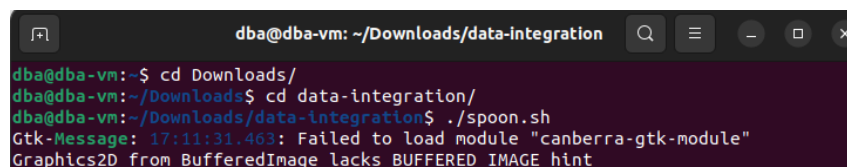
Выгрузить обработанные данные в MySQL или PostgreSQL.

– Проверить корректность обработки:

Выполнить SQL-запросы для проверки результата.

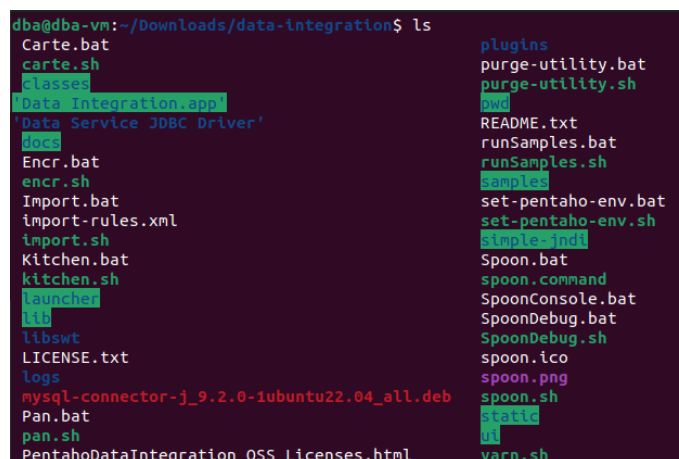
Подготовить отчет с описанием проделанных шагов.

Вариант 10. Анализ банковских транзакций: выявление паттернов, сегментация



```
dba@dba-vm: ~/Downloads/data-integration
dba@dba-vm:~$ cd Downloads/
dba@dba-vm:~/Downloads$ cd data-integration/
dba@dba-vm:~/Downloads/data-integration$ ./spoon.sh
Gtk-Message: 17:11:31.463: Failed to load module "canberra-gtk-module"
Graphics2D from BufferedImage lacks BUFFERED_IMAGE hint
```

Рисунок 1. Переход в нужную папку и запуск Pentaho Spoon



```
dba@dba-vm:~/Downloads/data-integration$ ls
Carte.bat          plugins
Carte.sh           purge-utility.bat
classes            purge-utility.sh
Data Integration.app  runSamples.bat
Data Service JDBC Driver'  runSamples.sh
Docs               samples
Encr.bat           set-pentaho-env.bat
encr.sh            set-pentaho-env.sh
Import.bat         simple-ndf
import-rules.xml   Spoon.bat
import.sh          spoon.command
Kitchen.bat        SpoonConsole.bat
kitchen.sh         SpoonDebug.bat
launcher           SpoonDebug.sh
lib               spoon.ico
libswt             spoon.png
LICENSE.txt        spoon.sh
logs               static
mysql-connector-j_9.2.0-1ubuntu22.04_all.deb  ui
Pan.bat            yarn.sh
pan.sh
PentahoDataIntegration_OSS_Licenses.html
```

Рисунок 2. Проверка установки коннектора mysql

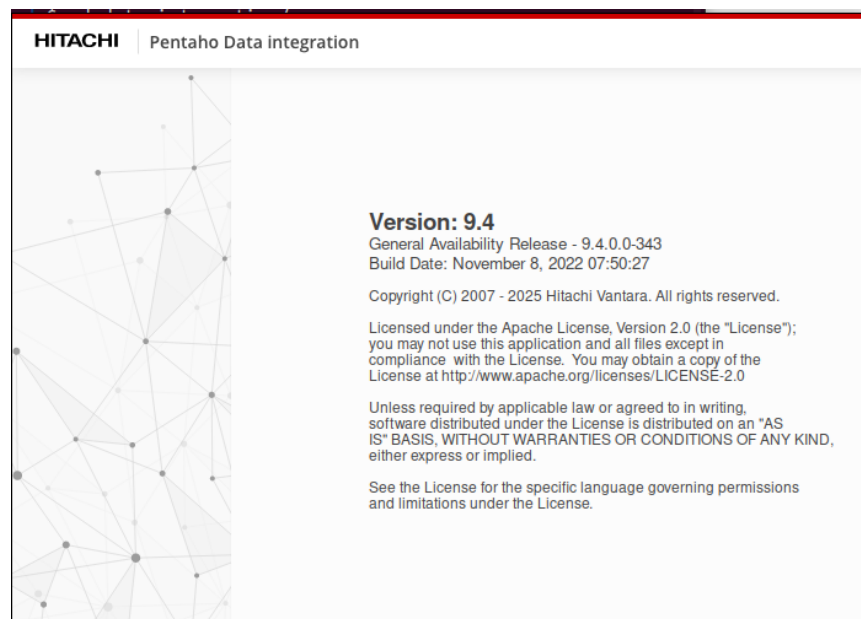


Рисунок 3. Pentaho успешно запущен

CSV file input

Step name:

Filename:

Delimiter:

Enclosure:

NIO buffer size:

Lazy conversion? ☒

Header row present? ☒

Add filename to result ☐

The row number field name (optional)

Running in parallel? ☐

New line possible in fields? ☐

Format:

File encoding:

	Name	Type	Format	Length	Precision	Currency	Decimal	Group	Trim type
1	Account_No	String		13		\$.	,	none
2	DATE	Date	yyyy-MM-dd			\$.	,	none
3	TRANSACTION DETAILS	String		32		\$.	,	none
4	CHQ.NO.	Boolean				\$.	,	none
5	VALUE DATE	Date	yyyy-MM-dd			\$.	,	none
6	DEPOSIT_AMT	Integer	#	15	0	\$.	,	none
7	BALANCE_AMT	Integer	#	15	0	\$.	,	none

Рисунок 4. Загрузка csv файла

Examine preview data							
Rows of step: CSV file input (1000 rows)							
	Account_No	DATE	TRANSACTION DETAILS	CHQ.NO.	VALUE DATE	DEPOSIT_AMT	BALANCE_AMT
1	409000611074'	2017-06-29	TRF FROM Indiaforensic SERVICES	<null>	2017-06-29	1000000	1000000
2	409000611074'	2017-07-05	TRF FROM Indiaforensic SERVICES	<null>	2017-07-05	1000000	2000000
3	409000611074'	2017-07-18	FDRL/INTERNAL FUND TRANSFE	<null>	2017-07-18	500000	2500000
4	409000611074'	2017-08-01	TRF FRM Indiaforensic SERVICES	<null>	2017-08-01	3000000	5500000
5	409000611074'	2017-08-16	FDRL/INTERNAL FUND TRANSFE	<null>	2017-08-16	500000	6000000
6	409000611074'	2017-08-16	FDRL/INTERNAL FUND TRANSFE	<null>	2017-08-16	500000	6500000
7	409000611074'	2017-08-16	FDRL/INTERNAL FUND TRANSFE	<null>	2017-08-16	500000	7000000
8	409000611074'	2017-08-16	FDRL/INTERNAL FUND TRANSFE	<null>	2017-08-16	500000	7500000
9	409000611074'	2017-08-16	FDRL/INTERNAL FUND TRANSFE	<null>	2017-08-16	500000	8000000
10	409000611074'	2017-08-16	FDRL/INTERNAL FUND TRANSFE	<null>	2017-08-16	500000	8500000
11	409000611074'	2017-08-16	INDO GIBL Indiaforensic STL01071	<null>	2017-08-16	<null>	8366100
12	409000611074'	2017-08-16	INDO GIBL Indiaforensic STL02071	<null>	2017-08-16	<null>	8348100
13	409000611074'	2017-08-16	INDO GIBL Indiaforensic STL03071	<null>	2017-08-16	<null>	8343100
14	409000611074'	2017-08-16	INDO GIBL Indiaforensic STL04071	<null>	2017-08-16	<null>	8147300
15	409000611074'	2017-08-16	INDO GIBL Indiaforensic STL05071	<null>	2017-08-16	<null>	8065700
16	409000611074'	2017-08-16	INDO GIBL Indiaforensic STL06071	<null>	2017-08-16	<null>	8023900
17	409000611074'	2017-08-16	INDO GIBL Indiaforensic STL07071	<null>	2017-08-16	<null>	7925400
18	409000611074'	2017-08-16	INDO GIBL Indiaforensic STL10071	<null>	2017-08-16	<null>	7781600
19	409000611074'	2017-08-16	INDO GIBL Indiaforensic STL11071	<null>	2017-08-16	<null>	7449950
20	409000611074'	2017-08-16	INDO GIBL Indiaforensic STL12071	<null>	2017-08-16	<null>	7320950

Рисунок 5. Предпросмотр загруженных данных

Select values

Step name

Select & AlterRemoveMeta-data

Fields:

	Fieldname	Rename to	Length	Precision
1	Account_No			
2	DATE			
3	DEPOSIT_AMT			
4	BALANCE_AMT			

Get fields to select

Edit Mapping

Рисунок 6. Выбор столбцов для дальнейшего анализа

Select values

Step name

Select & AlterRemoveMeta-data

Fields to remove:

	Fieldname
1	TRANSACTION DETAILS
2	CHQ.NO.
3	VALUE DATE

Рисунок 7. Столбцы, выбранные для удаления

Value mapper

Step name:

Fieldname to use:

Target field name (empty=overwrite):

Default upon non-matching:

Field values:

	Source value	Target value
1	'	

Рисунок 8. Замена лишнего символа в номере аккаунта

Filter rows

Step name:

Send 'true' data to step:

Send 'false' data to step:

The condition:

DEPOSIT_AMT IS NOT NULL

AND

BALANCE_AMT IS NOT NULL

Рисунок 9. Фильтрация нулевых значений в двух столбцах

Memory group by

Step name:

Always give back a result row: ☐

The fields that make up the group:

Group field

- Account_No
- DATE
- DEPOSIT_AMT
- BALANCE_AMT

Aggregates :

Name	Subject	Type
1 Balance_Sum	BALANCE_AMT	Sum

Рисунок 10. Группировка данных по сумме баланса

phpMyAdmin

Server: localhost:3306 > Database: mgpu_ico_etl_10 > Table: bank_segment

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

SELECT * FROM `bank_segment`

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

id	Account_No	DATE	DEPOSIT_AMT	BALANCE_AMT	Balance_Sum
----	------------	------	-------------	-------------	-------------

Query results operations

Рисунок 11. Создание таблицы в MySQL

Table output

Step name: Table output

Connection: MySQL

Target schema: mgpu_ico_etl_10

Target table: bank_segment

Commit size: 1000

Truncate table: ☐

ignore insert errors: ☐

Specify database fields: ☒

Main options | Database fields

Fields to insert:

Table field	Stream field
1 Account_No	Account_No
2 DATE	DATE
3 DEPOSIT_AMT	DEPOSIT_AMT
4 BALANCE_AMT	BALANCE_AMT
5 Balance_Sum	Balance_Sum

Get fields

Enter field mapping

Рисунок 12. Настройка выгрузки данных в бд



Рисунок 13. Этапы трансформации

Information_schema

- mgpu_ico_etl_10
 - New
 - bank_segment
 - Columns
 - New
 - Account_No (varchar)
 - BALANCE_AMT (varchar)
 - Balance_Sum (varchar)
 - DATE (varchar, nullable)
 - DEPOSIT_AMT (varchar)
- customers
- customers_new
- orders
- orders_new
- products
- performance_schema

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

Showing rows 0 - 24 (1283 total, Query took 0.0689 seconds.)

```

SELECT DATE(DATE) AS date_only, COUNT(DISTINCT Account_No) AS client_count FROM `bank_segment` GROUP BY DATE(DATE) ORDER BY client_count DESC;
  
```

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

1 > >> Number of rows: 25 Filter rows: Search this table

Extra options

date_only	client_count
2018-10-09	9
2018-11-27	9
2018-07-23	8
2018-08-10	8
2018-09-14	8
2018-07-09	8
2018-08-31	8
2017-11-27	8

Рисунок 14. Тенденция совершения банковских операций

Как видно, согласно выполненному запросу, большее количество клиентов обращалось в 2018 году в осенний период, приходящийся на октябрь и ноябрь.

Showing rows 0 - 0 (1 total, Query took 0.0316 seconds.) [DEPOSIT_AMT: 999999... - 999999...]

```
SELECT Account_No, DEPOSIT_AMT FROM bank_segment ORDER BY DEPOSIT_AMT DESC LIMIT 1;
```

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

Extra options

Account_No	DEPOSIT_AMT
1196428'	9999999

Рисунок 15. Клиент, у которого максимальный депозит

Showing rows 0 - 4 (5 total, Query took 0.0475 seconds.)

```
SELECT Account_No, SUM(BALANCE_AMT) AS total_balance FROM bank_segment GROUP BY Account_No ORDER BY `total_balance` DESC LIMIT 5
```

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

Extra options

Account_No	total_balance
409000493201'	540952848
409000611074'	486158787
409000425051'	-1276273266
409000405747'	-10944130644
409000493210'	-428937549830

Рисунок 16. Топ 5 клиентов из премиум сегмента, имеющие самый большой баланс

Execution Results

Logging Execution History Step Metrics Performance Graph Metrics Preview data

Stepname	Copynr	Read	Written	Input	Output	Updated	Rejected	Errors	Active
1 CSV file input	0	0	116201	116202	0	0	0	0	Finished
2 Select values	0	116201	116201	0	0	0	0	0	Finished
3 Value mapper	0	116201	116201	0	0	0	0	0	Finished
4 Filter rows	0	116201	62652	0	0	0	0	0	Finished
5 Memory group by	0	62652	62620	0	0	0	0	0	Finished
6 Table output	0	62620	62620	0	62620	0	0	0	Finished

Рисунок 17. Результаты трансформации

Согласно результату, видно, что после группировки количество прочитанных и записанных строк сократилось почти в 2 раза.

Execution Results

Logging Execution History Step Metrics Performance Graph Metrics Preview data

Stepname	Read	Written	Input	Output	Updated	Rejected	Errors	Active
Execute a transformation - Transformation 1_Morozova V : 1088016ms	116201	62620	116202	62620	0	0	0	Finished
Initialize a transformation - Transformation 1_Morozova V : 212ms	0	0	0	0	0	0	0	Finished
Initialize a step - CSV file input : 1ms	0	0	0	0	0	0	0	Finished
Initialize a step - Select values : 0ms	0	0	0	0	0	0	0	Finished
Initialize a step - Value mapper : 0ms	0	0	0	0	0	0	0	Finished
Initialize a step - Filter rows : 1ms	0	0	0	0	0	0	0	Finished
Initialize a step - Memory group by : 0ms	0	0	0	0	0	0	0	Finished
Initialize a step - Table output : 198ms	0	0	0	0	0	0	0	Finished
Connect to database - MySQL : 180ms	0	0	0	0	0	0	0	Finished
Execute a step - CSV file input : 346ms	116201	116201	116202	0	0	0	0	Finished
Execute a step - Select values : 533ms	116201	116201	0	0	0	0	0	Finished
Execute a step - Value mapper : 549ms	116201	116201	0	0	0	0	0	Finished
Execute a step - Filter rows : 586ms	116201	62652	0	0	0	0	0	Finished
Execute a step - Memory group by : 920879ms	62652	62620	0	0	0	0	0	Finished
Execute a step - Table output : 1087795ms	62620	62620	0	62620	0	0	0	Finished
Get DB metadata - MySQL : 0ms	0	0	0	0	0	0	0	Finished

Рисунок 18. Результат выполнения каждого степа по времени в мс

Выводы:

1. Было развернуто Pentaho Data Integration.

2. Создан ETL-конвейер:

- загружены данные из CSV-файла
- очищены, преобразованы и отфильтрованы данные
- выполнена замена значений
- выгружены обработанные и сгруппированные данные в MySQL.

3. Проверена корректность обработки.

Выполнены SQL-запросы для проверки результата и сегментации клиентов согласно частоте банковских транзакций.

Согласно результатам, выполнение трансформации заняло 1088016 ms, из которых большую часть заняли процессы группировки и выгрузки в MySQL (920879 и 1087795 ms соответственно).