Лабораторная работа 4.1. Установка и настройка ETL-инструмента. Создание конвейеров данных

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.

Условие выполнения работы:

Работа выполняется в образе **Ubuntu 22.04 (.ova) для VirtualBox 7.0** https://disk.yandex.ru/d/gagWU_zn1erR8g, в котором предварительно установлены все необходимые компоненты для работы с Pentaho Data Integration, либо проводится установка окружения в ОС Linux.

Задачи

- **Настроить среду для работы с Pentaho Data Integration (PDI)**:
  - Запуск виртуальной машины с **Ubuntu 22.04** в **VirtualBox**.
  - Проверка установки Java и WebKitGTK.
  - 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.

Инструменты и технологии

Для выполнения лабораторной работы используются следующие инструменты:

Компонент	**Описание**
Ubuntu 22.04 (.ova)	Образ операционной системы для VirtualBox 7.0
VirtualBox 7.0	Виртуализация среды
Pentaho Data Integration 9.4	ETL-инструмент для работы с данными
MySQL/PostgreSQL	База данных для хранения обработанных данных
CSV-файлы	Исходные данные для обработки
Java 11	Необходима для работы Pentaho
libwebkitgtk-1.0-0	Библиотека для корректного запуска Spoon
SQL	Язык запросов для работы с базами данных

Установка и настройка окружения

Запуск образа Ubuntu 22.04 в VirtualBox

Установите **VirtualBox 7.0** с официального сайта.

Скачайте **готовый образ Ubuntu 22.04 (.ova)**.

Импортируйте его в VirtualBox:

File → Import Appliance → Выберите .ova файл → Import.

Запустите виртуальную машину.

Установка и запуск Pentaho Data Integration (в случае установки на новую ВМ)

Шаг 1. Установка и подготовка окружения

Открываем терминал в Linux и выполняем команды:

Распаковка архива

```
unzip pdi-ce-9.4.0.0-343.zip
```

Переход в каталог

```
cd Downloads/
cd data-integration/
```

Проверка установленной версии Java

```
java -version
```

Установка OpenJDK 11, если не установлено

```
sudo apt install openjdk-11-jdk -y
```

Проверка версии Java после установки

```
java -version
```

Шаг 2. Добавление репозиториев для установки WebKitGTK 1.0

Так как `libwebkitgtk-1.0-0` больше не поддерживается в новых версиях Ubuntu, добавим старый репозиторий Ubuntu Bionic:

Открываем файл источников пакетов

```
sudo nano /etc/apt/sources.list
```

Добавляем одну из следующих строк в конец файла:

```
deb http://cz.archive.ubuntu.com/ubuntu bionic main universe
```

ИЛИ

```
deb http://mirrors.kernel.org/ubuntu bionic main universe
```

Сохраняем изменения (Ctrl+X, Y, Enter).

Шаг 3. Обновление списка пакетов и установка WebKitGTK 1.0

Обновляем список пакетов

```
sudo apt-get update
```

Добавляем ключи для старого репозитория

```
sudo apt-key adv --keyserver keyserver.ubuntu.com --recv-keys 3B4FE6ACC0B21F32
```

Повторно обновляем пакеты после добавления ключей

```
sudo apt-get update
```

Устанавливаем устаревший WebKitGTK 1.0 для запуска Pentaho Spoon

```
sudo apt-get install libwebkitgtk-1.0-0 -y
```

Шаг 4. Запуск Pentaho Data Integration

Переход в каталог data-integration

```
cd ~/Downloads/data-integration/
```

Делаем исполняемым файл Spoon (если нужно)

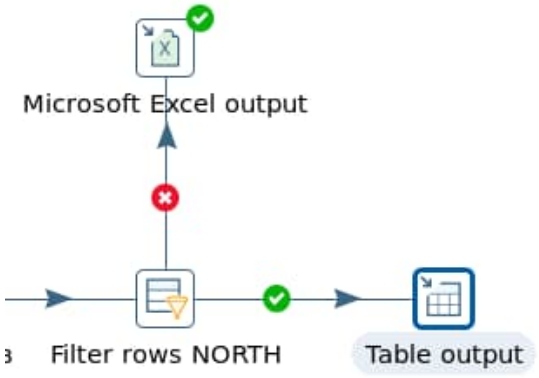
```
chmod +x spoon.sh
```

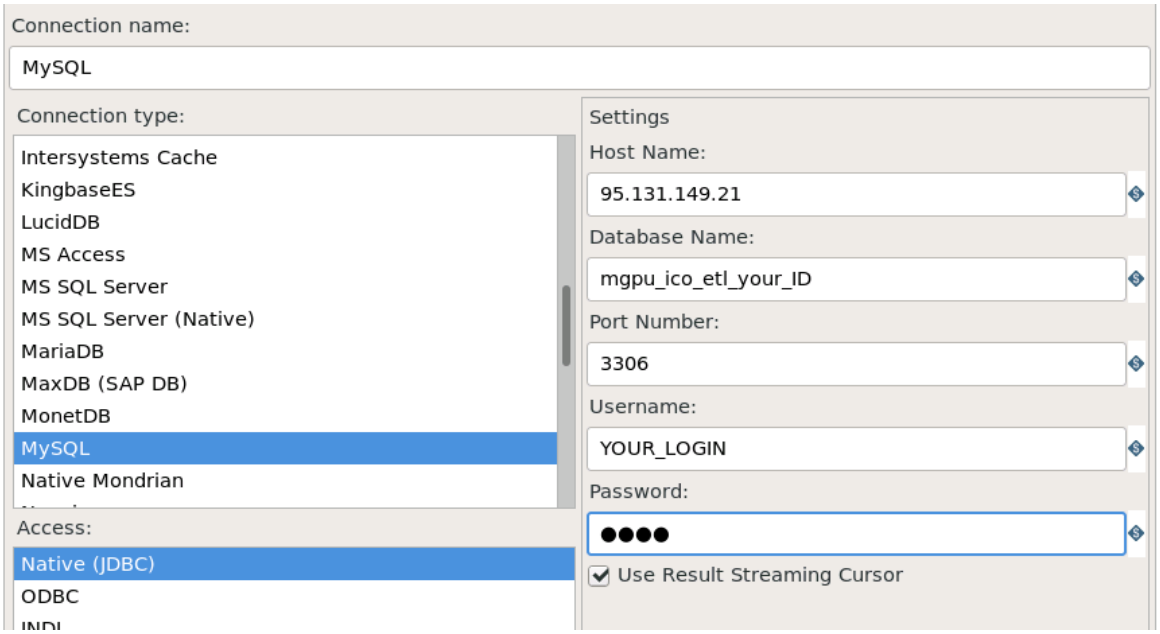
Запускаем Pentaho Spoon

```
./spoon.sh
```

Подготовка данных для загрузки в MySQL

Добавьте компонент **"Table output"**





Установка MySQL драйвера для Pentaho Data Integration

1. Загрузка драйвера

1. Скачайте MySQL Connector/J (JDBC driver) с официального сайта MySQL: https://dev.mysql.com/downloads/connector/j/

2. Выберите Platform Independent версию Ubuntu Linux 22.04 (Architecture Independent), DEB Package

3. Скачайте mysql-connector-j_9.2.0-1ubuntu22.04_all.deb

Установка deb пакета

sudo dpkg -i mysql-connector-j_9.2.0-1ubuntu22.04_all.deb

Проверка установленного пакета

dpkg -l | grep mysql-connector-j

Поиск установленного JAR файла

ls /usr/share/java/mysql-connector-j-9.2.0.jar

Правильная команда для копирования (используем путь относительно домашней директории)

sudo cp /usr/share/java/mysql-connector-j-9.2.0.jar ~/Downloads/data-integration/lib/

Проверяем, что файл скопировался

ls -l ~/Downloads/data-integration/lib/mysql-connector-j-9.2.0.jar

Устанавливаем правильные права доступа

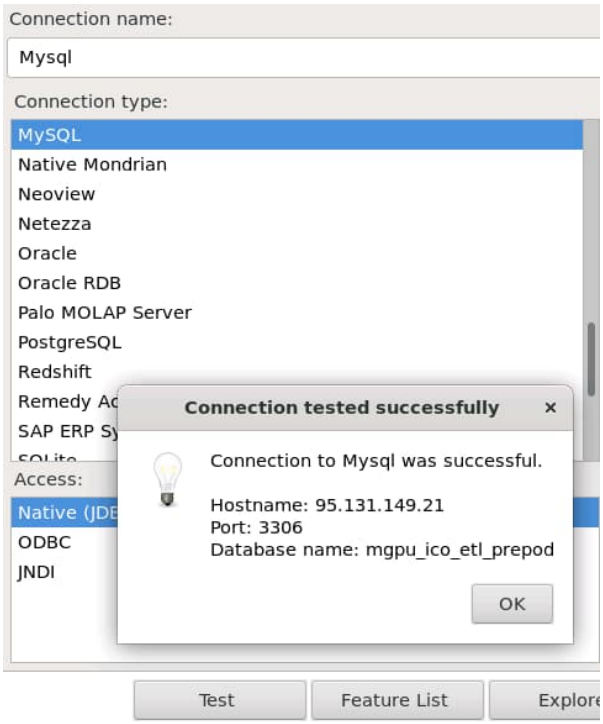
sudo chmod 644 ~/Downloads/data-integration/lib/mysql-connector-j-9.2.0.jar

Меняем владельца файла на текущего пользователя

sudo chown dba:dba ~/Downloads/data-integration/lib/mysql-connector-j-9.2.0.jar

После успешного копирования:

- Перезапустите Pentaho Data Integration.
- Проверьте подключение к MySQL.



Варианты индивидуальных заданий

Каждый студент выполняет одно из следующих заданий, используя Pentaho Data Integration для обработки данных. В каждой задаче выбрать по тематике Kaggle-датасет, CSV-файл.

№	Описание задания	Kaggle-датасет
1	Анализ розничных продаж: фильтрация транзакций, выявление аномалий, расчет метрик эффективности	Retail Sales Dataset (https://www.kaggle.com/datasets/mohammadtalib786/retail-sales-dataset)
2	Анализ электронной коммерции: очистка данных о заказах, сегментация клиентов	E-Commerce Dataset (https://www.kaggle.com/datasets/carrie1/ecommerce-data)
3	Финансовая аналитика: обработка данных о котировках, расчет показателей	Financial Market Data (https://www.kaggle.com/datasets/finnhub/financial-market-data)
4	HR-аналитика: анализ данных о сотрудниках, расчет KPI	Human Resources Dataset (https://www.kaggle.com/datasets/rhuebner/human-resources-data-set)
5	Анализ цепочек поставок: оптимизация логистических данных	Supply Chain Dataset (https://www.kaggle.com/datasets/shashwatwork/dataco-smart-supply-chain-dataset)
6	Маркетинговая аналитика: анализ эффективности рекламных кампаний	Marketing Campaign Dataset (https://www.kaggle.com/datasets/rodsaldanha/arketing-campaign)
7	Анализ клиентского сервиса: обработка данных обращений клиентов	Customer Support Dataset (https://www.kaggle.com/datasets/thoughtvector/customer-support-on-twitter)
8	Анализ веб-аналитики: обработка данных о посещаемости сайта	Web Analytics Dataset (https://www.kaggle.com/datasets/tunguz/ga-customer-revenue-prediction)
9	Прогнозирование спроса: анализ исторических данных продаж	Demand Forecasting Dataset (https://www.kaggle.com/datasets/felixzhao/productdemandforecasting)
10	Анализ банковских транзакций: выявление паттернов, сегментация	Banking Transactions Dataset (https://www.kaggle.com/datasets/apoorvwatsky/bank-transaction-data)
11	Анализ операционной эффективности: обработка производственных данных	Manufacturing Dataset (https://www.kaggle.com/datasets/inIT-OWL/production-plant-data)
12	Анализ рынка недвижимости: очистка и трансформация данных	Real Estate Dataset (https://www.kaggle.com/datasets/arslanali4343/real-estate-dataset)
13	Анализ социальных медиа: обработка данных активности	Social Media Dataset (https://www.kaggle.com/datasets/gokulrajkmv/social-media-sentiment-analysis)
14	Биржевая аналитика: обработка данных торгов	Stock Market Dataset (https://www.kaggle.com/datasets/borismarjanovic/price-volume-data-for-all-us-stocks-etfs)
15	Аналитика телекоммуникаций: анализ данных об использовании услуг	Telecom Dataset (https://www.kaggle.com/datasets/blastchar/telco-customer-churn)
16	Бюджетная аналитика: обработка финансовых показателей	Budget Analytics Dataset (https://www.kaggle.com/datasets/city-of-seattle/seattle-budget-data)

17	Анализ программ лояльности: обработка данных о бонусных программах	Loyalty Program Dataset (https://www.kaggle.com/datasets/arjunbhasin2013/ccdata)
18	Анализ цифровой рекламы: обработка данных рекламных кампаний	Digital Ads Dataset (https://www.kaggle.com/datasets/vidurpunj/facebook-ad-campaign)
19	Анализ бизнес-рисков: обработка данных страховых случаев	Insurance Risk Dataset (https://www.kaggle.com/datasets/mirichoi0218/insurance)
20	Аналитика инвестиций: обработка данных инвестиционного портфеля	Investment Portfolio Dataset (https://www.kaggle.com/datasets/stefanoleone992/mutual-funds-and-etfs)

Формат предоставления отчета по лабораторной работе:

1. Скачать CSV-датасет с Kaggle.
2. Создать Pentaho ETL-конвейер с фильтрацией, заменой и обработкой данных.
3. Выгрузить данные в MySQL/PostgreSQL.
4. Подготовить отчет с:
   - Описанием процесса;
   - Скриншотами Spoon;
   - SQL-запросами для проверки.
5. Загрузить отчет, CSV-датасет (или ссылку на github) **и lab_4_01.ktr** в LMS.