

Лабораторная работа 4.1 Гадильшина Валентина БД-231м

1. Скачать [отсюда](#) и запустить Pentaho DI. Pentaho DI требует установку Java 8. Попробуйте скачать архив и распаковать его. Необходимо запустить spoon.sh для Linux/Mac и spoon.bat для Windows. Видео по установке Pentaho DI на примере Windows 10 +
2. Скачать [примеры Pentaho jobs](#) для Staging и Dimension Tables. +
3. Создайте еще одну трансформацию, в которой создать sales_fact таблицу. +

The screenshot displays the Pentaho Data Integration (PDI) interface. The top pane shows a transformation job named 'transformation_general' with the following steps:

- Orders Excel input → Sort orders
- Returns Excel input 3 → Sort returns
- People Excel input 2 → Sort region
- Sort orders → Merge join
- Sort returns → Merge join
- Sort region → Merge join
- Merge join → Sort people
- Sort people → Merge join 2
- Merge join 2 → sales_fact Select values
- sales_fact Select values → sales_fact file output scv

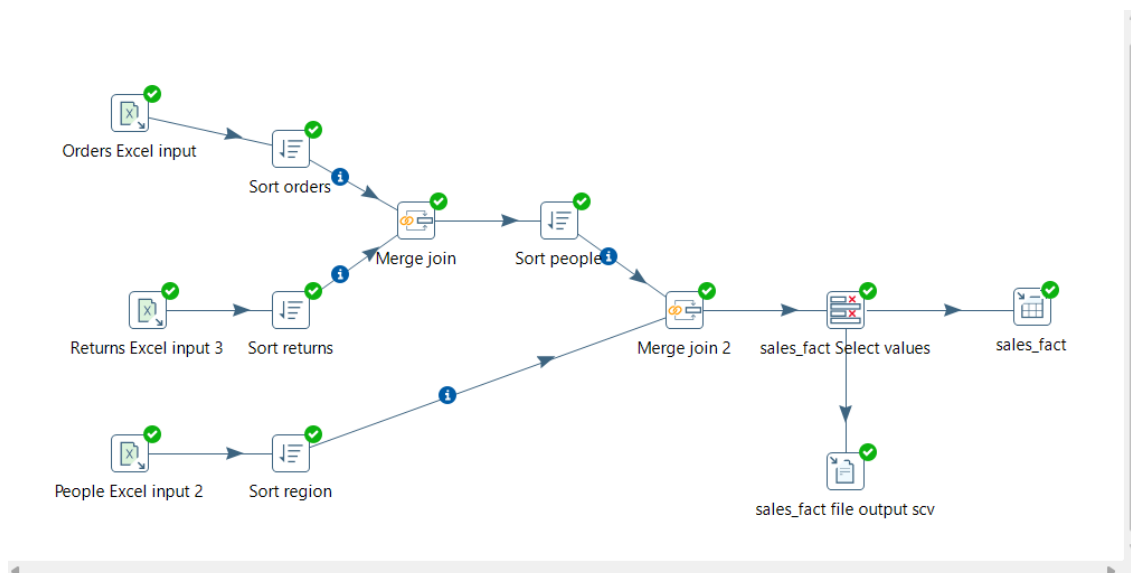
The bottom pane shows the 'Execution Results' for the 'sales_fact Select values' step. The data is as follows:

#	Row ID	Order ID	Sales	Quantity	Discount	Profit
1	6569	CA-2016-100678	2,688	2,0	0,2	1,008
2	6570	CA-2016-100678	317,058	3,0	0,3	-18,1176
3	6571	CA-2016-100678	149,352	3,0	0,2	50,4063
4	6572	CA-2016-100678	227,976	3,0	0,2	28,497
5	6315	CA-2016-100762	151,92	4,0	0,0	45,576








Below the execution results, the 'Database Connection' dialog is open, showing the 'General' tab. The connection name is empty. The connection type is 'PostgreSQL'. The settings are:




- Host Name: 95.131.149.21
- Database Name: dep3
- Port Number: 5432
- Username: m_03
- Password: [masked]

Buttons at the bottom include 'Test', 'Feature List', 'Explore', 'OK', and 'Cancel'.



4. Выявить 8-10 подсистем в ETL Pentaho DI и написать небольшой отчет, в котором приложить print screen компонента (ETL подсистемы) и написать про его свойства. Результат сохраните в Git. +

№	Компонент	Описание (свойства)
1	 Table input	Используется для чтения данных из таблицы базы данных. Он позволяет выполнить SQL-запрос к указанной таблице и получить результат в виде строк данных, которые затем могут быть обработаны дальше в трансформации. Table Input поддерживает множество баз данных, включая MySQL, PostgreSQL, Oracle, Microsoft SQL Server и многие другие.
2	 Microsoft Excel input	Используется для чтения данных из файлов формата Excel. Он предоставляет возможность извлечения данных из листов Excel и дальнейшей их обработки в рамках трансформации. Excel Input позволяет читать данные из файлов формата Excel (.xls или .xlsx). Также можно указать конкретный лист, из которого следует извлечь данные. И можно определить структуру данных, указав имена столбцов, их типы данных и другие параметры.
3	 Select values	Используется для выбора (в том числе для удаления лишних столбцов), переименования и преобразования столбцов данных в рамках трансформации.
4	 Sort rows	Используется для сортировки строк данных в рамках трансформации. В частности, используется перед выполнением операции слияния данных Merge Join.
5	 Merge join	Используется для объединения двух наборов данных из различных источников данных, таких как таблицы баз данных, файлы или другие источники данных, на основе общих значений столбцов. Можно настроить тип объединения (INNER, LEFT, RIGHT) и указать столбцы для сравнения при выполнении операции объединения. Также поддерживает обработку дубликатов.
6	 Text file output	Используется для записи данных в текстовый файл. Text File Output поддерживает различные форматы данных, такие как CSV, TSV и другие.
7	 Group by	Используется для группировки строк данных по определённому столбцу или набору столбцов и применения агрегирующих функций к данным внутри каждой группы. Group By часто используется для

		подготовки данных для создания отчетов или анализа, где требуется агрегация данных по определенным критериям.
8	 Calculator	Используется для выполнения различных математических операций, логических вычислений и преобразований значений в рамках трансформации данных. Также поддерживает вычисление выражений: используется для вычисления новых значений на основе выражений, включая формулы, функции или переменные.
9	 Data validator	<p>Используется для проверки качества данных и выполнения различных видов валидации данных в рамках трансформации. Data Validator может:</p> <ul style="list-style-type: none"> • проверять типы данных в столбцах данных и выявлять некорректные значения, несоответствующие ожидаемым типам данных. • проверять наличие значений в определенных столбцах данных и выявлять строки с недостающими данными. • поддерживает проверку формата данных, таких как даты, времена, числа или текстовые строки, на соответствие определенным форматам.
10	 Unique rows	Используется для удаления дубликатов строк из набора данных. Этот компонент позволяет получить уникальные строки данных на основе определенных критериев и удалить повторяющиеся строки.