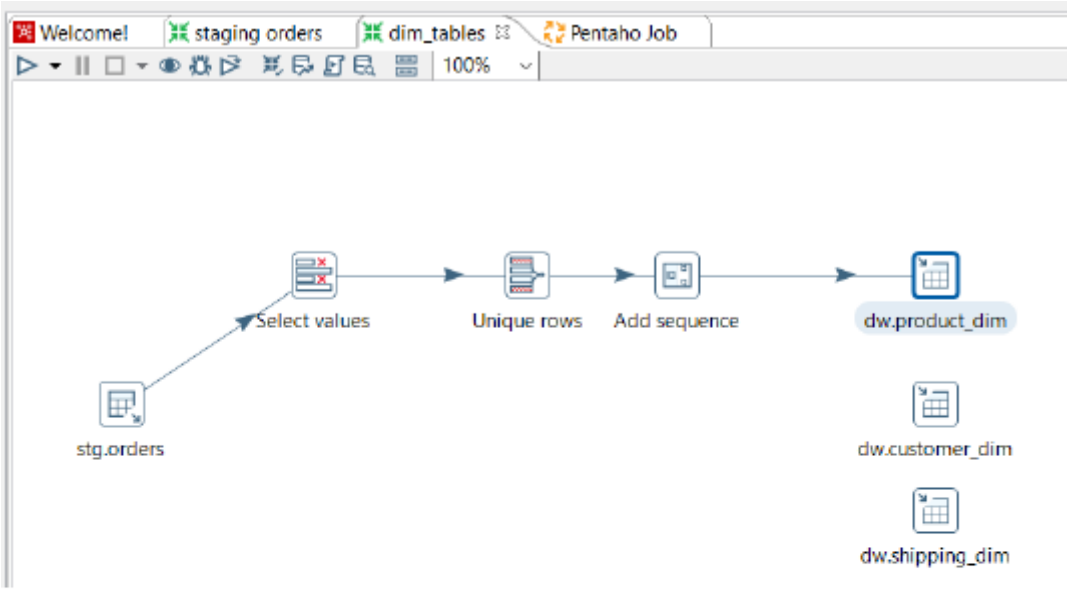
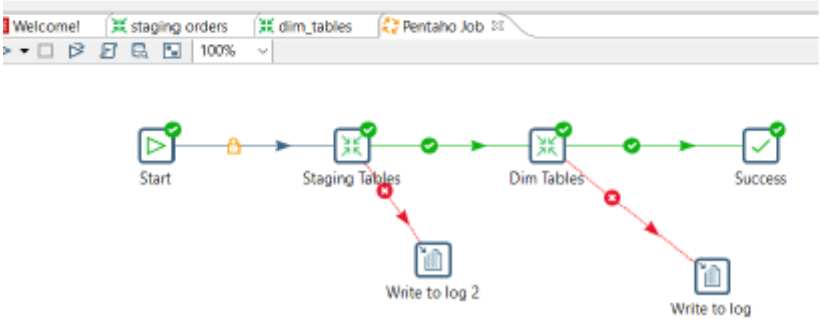
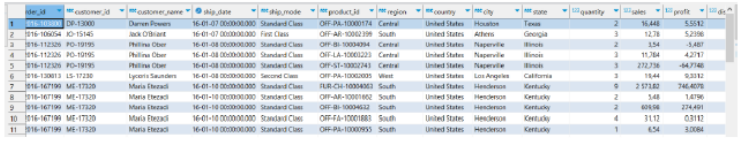
Лабораторная работа 4.1 Соловьев Михаил БД-231м

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. +



#	Table field	Stream field
1	order_id	order_id
2	customer_id	customer_id
3	customer_n...	customer_na...
4	ship_date	ship_date
5	ship_mode	ship_mode
6	product_id	product_id
7	region	region
8	country	country
9	city	city
1..	state	state
1..	quantity	quantity
1..	sales	sales
1..	profit	profit
1..	discount	discount
1..	sales_PK	sales_PK

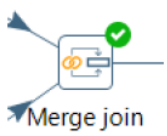
4.Выявить 8-10 подсистем в ETL Pentaho DI и написать небольшой отчет, в котором приложить print screen компонента (ETL подсистемы) и написать про его свойства. Результат сохраните в Git.

Table input - Используется для считывание данных их таблицы в базе данных. Необходимо настраивать подключение к базе.

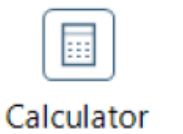
Microsoft Excel input - Используется для чтения данных из файлов формата Excel. Он предоставляет возможность извлечения данных из листов Excel и дальнейшей их обработки в рамках трансформации. Excel Input позволяет читать данные из файлов формата Excel (.xls или .xlsx). Также можно указать конкретный лист, из которого следует извлечь данные. И можно определить структуру данных, указав имена столбцов, их типы данных и тд.

Select values - Используется для выбора (в том числе для удаления лишних столбцов), переименования и преобразования столбцов данных в рамках трансформации.

Sort rows - Используется для сортировки строк данных в рамках трансформации. В частности, используется перед выполнением операции слияния данных Merge Join.



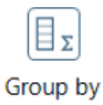
Merge join - Используется для объединения двух наборов данных из различных источников данных, таких как таблицы баз данных, файлы или другие источники данных, на основе общих значений столбцов. Можно настроить тип объединения (INNER, LEFT, RIGHT) и указать столбцы для сравнения при выполнении операции объединения. Также поддерживает обработку дубликатов.



Calculator - Используется для проверки качества данных и выполнения различных видов валидации данных в рамках трансформации. Data Validator может:

- 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.

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.



Group by - Позволяет группировать данные и использовать аггрегаты