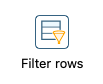
**Практика 3**

**Выявление подсистем Pentaho**

DATA CLEANING AND QUALITY HANDLER SYSTEM

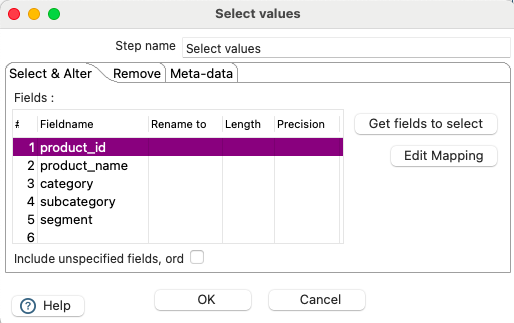
Задача системы подготовить данные для анализа, необходимо очистить данные и проверить их качество

1.  Этот компонент отвечает за фильтрацию записей. В его редакторе можно указать, значение, которые удовлетворяют условию.

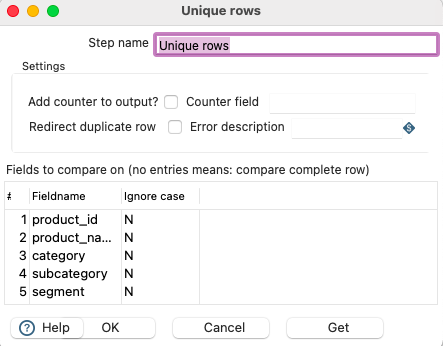
Изображение выглядит как текст

Автоматически созданное описание

1. Здесь мы можем поменять тип данных, изменить название, удалить объекты



1.  удаляет повторяющиеся строки из входного потока и фильтрует только уникальные строки в качестве входных данных для шага.



1. Изображение выглядит как текст

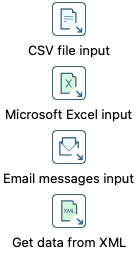
   Автоматически созданное описаниепозволяет управлять порядком, в котором будут считываться входные потоки. Если шаг имеет более одного входного потока - данные будут читаться одновременно из всех этих входов в непредсказуемом порядке. Однако, если вы хотите, чтобы один из входных потоков имел приоритет над другими, вы можете использовать шаг «Приоритет потоков».

Изображение выглядит как текст

Автоматически созданное описание

EXTRACTING SYSTEM

Задача системы понять источник и уметь подключаться к нему, чтобы забирать данные



Этот шаг обеспечивает возможность чтения данных из файлов с различными расширениями

CHANGE DATA CAPTURE

Система выявляет механизм изменения данных в исходных системах, чтобы загружать только свежие или изменённые данные в хранилище



1.

позволяет сравнить два потока строк. Это полезно для сравнения данных за два разных периода времени. Он часто используется в ситуациях, когда исходная система хранилища данных не содержит даты последнего обновления.

2. 

Этот шаг используется для чтения информации из базы данных с использованием соединения и SQL. Базовые операторы SQL можно сгенерировать автоматически, щелкнув Get SQL select.

DATA PROFILING SYSTEM

Система для сбора данных о наших данных

 Этот шаг позволяет просматривать строки вперед и назад.

 Шаг Data Validator позволяет определить простые правила для описания того, как должны выглядеть данные в поле. Это может быть диапазон значений, отдельный список значений или длины данных.

Этот шаг позволяет применять неограниченное количество правил проверки за один шаг к входящим данным.