

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:

Инструменты для хранения и обработки больших данных

Практическая работа 3

Тема:

Архитектура хранилищ данных: традиционная и облачная

Выполнила: Соколова М. С., группа: АДЭУ-201

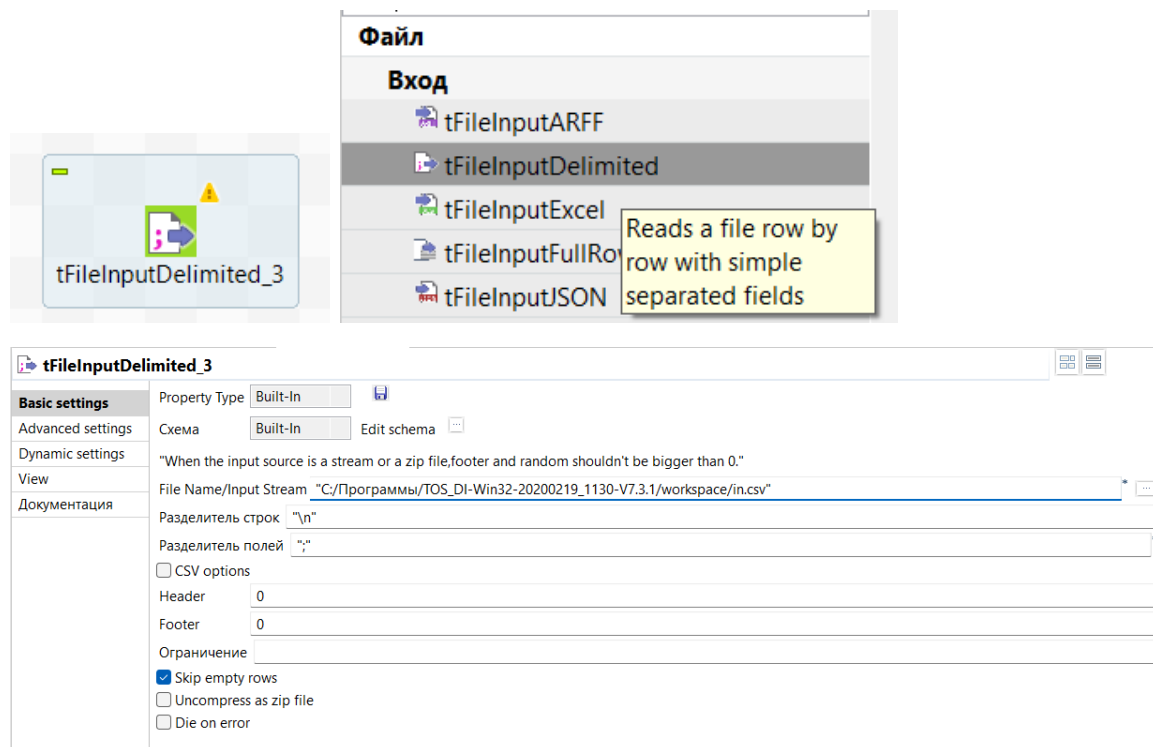
Преподаватель: Босенко Т. М.

Москва

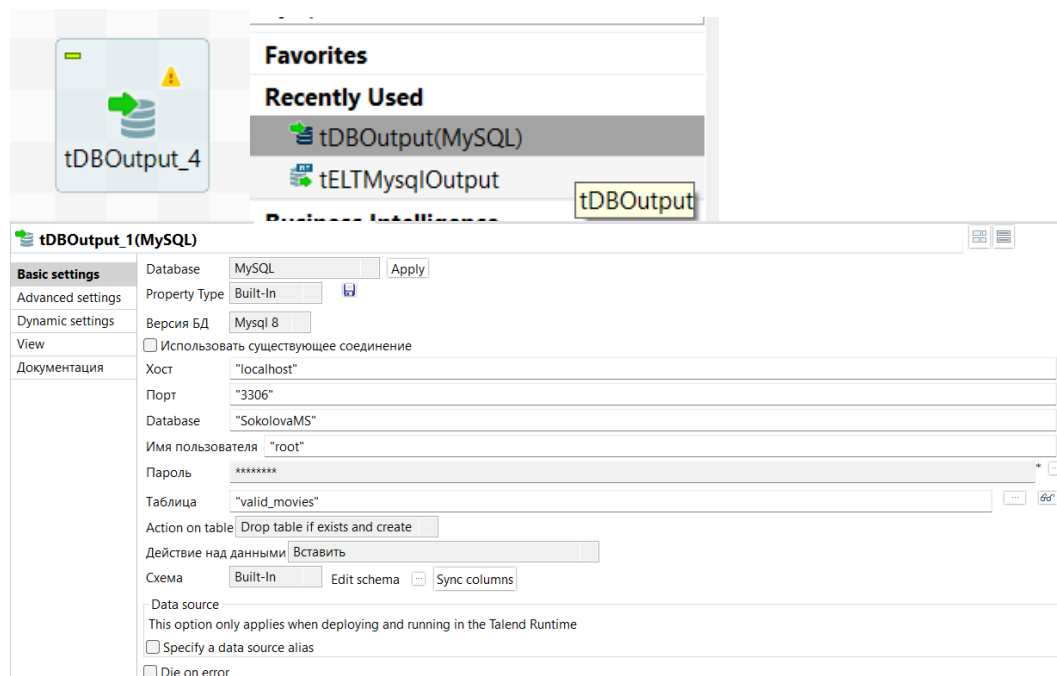
2023

Talend

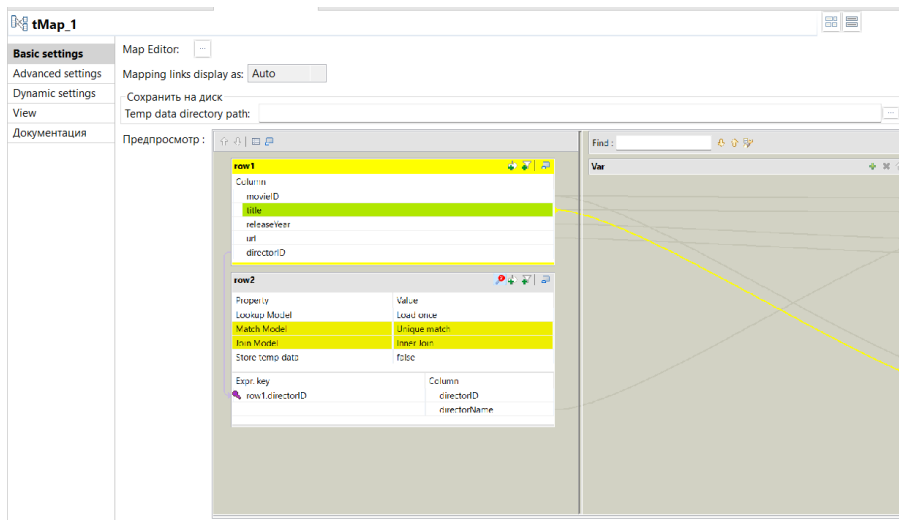
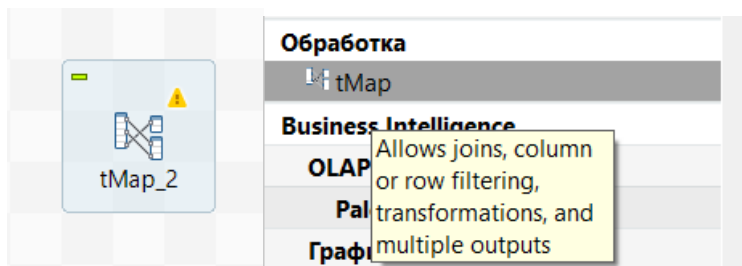
tFileInputDelimited – загрузка и считывание файлов формата csv или txt с разделителем. В атрибуты указывается путь и разделитель.



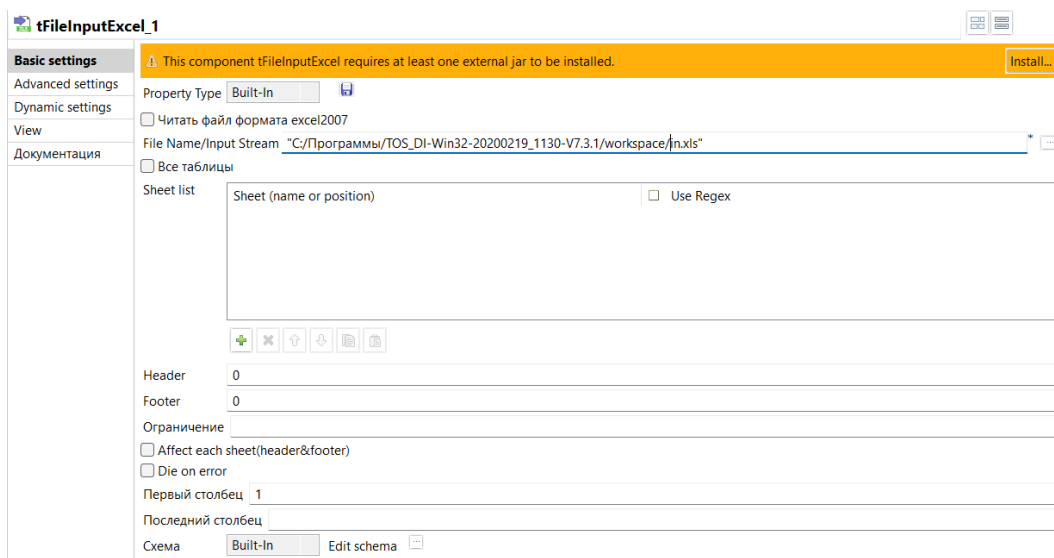
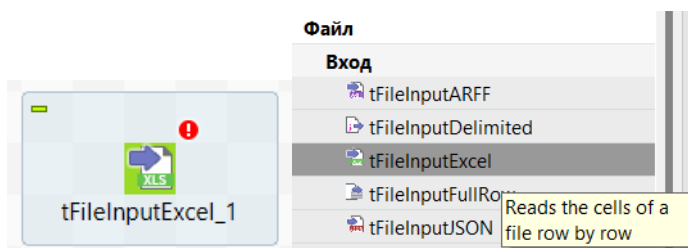
tMysqlOutput – Используется для записи, обновления, модификации данных в базе данных MySQL.



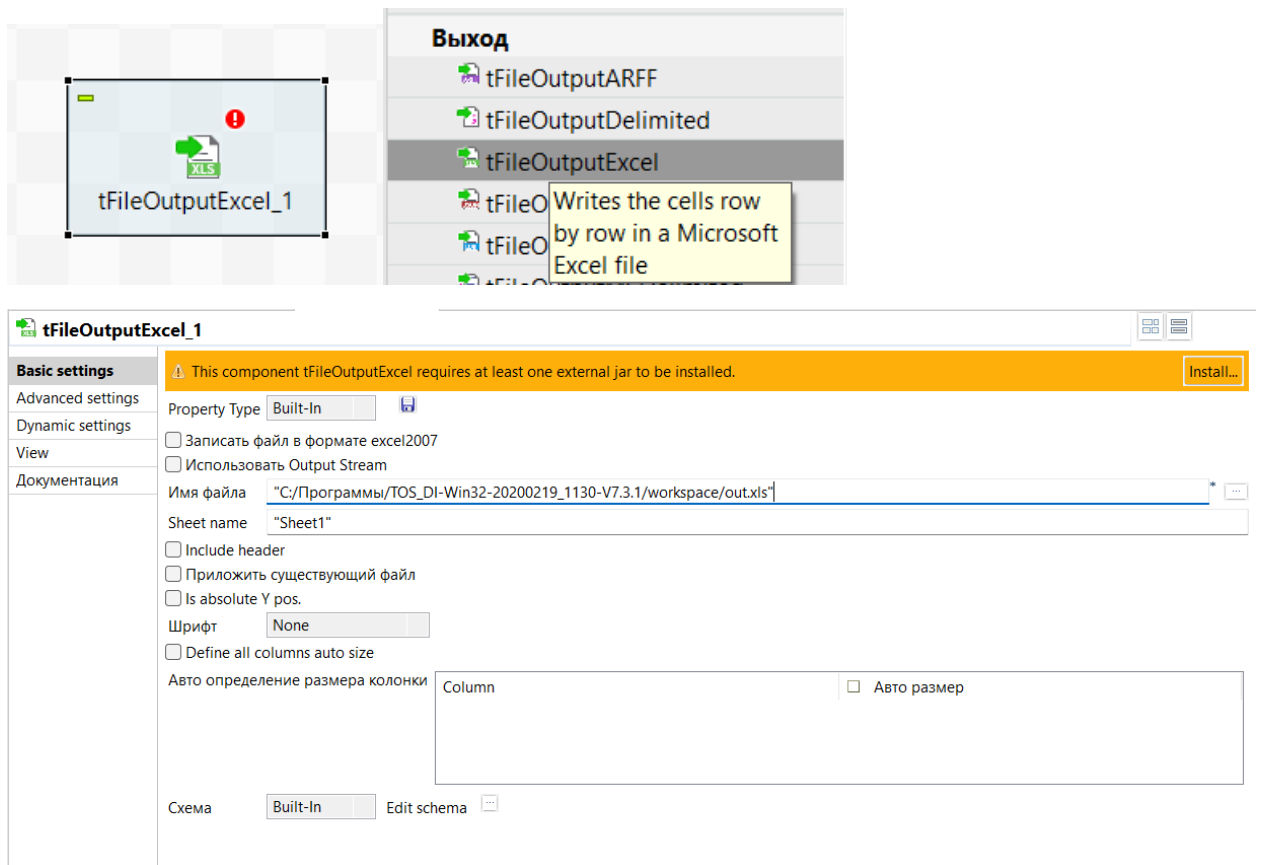
tMap – компонент, используемый для выполнения операций манипулирования данными с данными, таких как слияние данных и фильтрация данных. Основной функцией компонента tMap является сопоставление входного компонента с выходным компонентом.



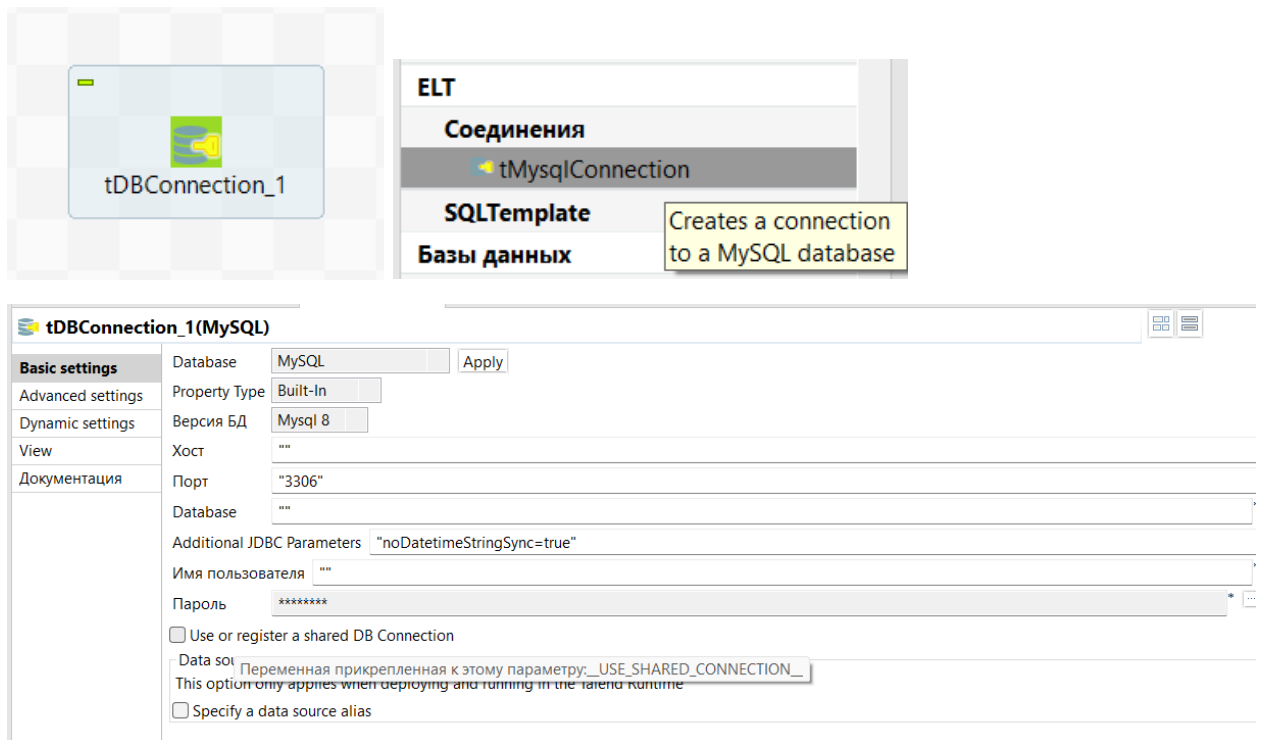
tFileInputExcel – считывает файл Excel строка за строкой, разбивает их на отдельные поля и передает его следующему компоненту.



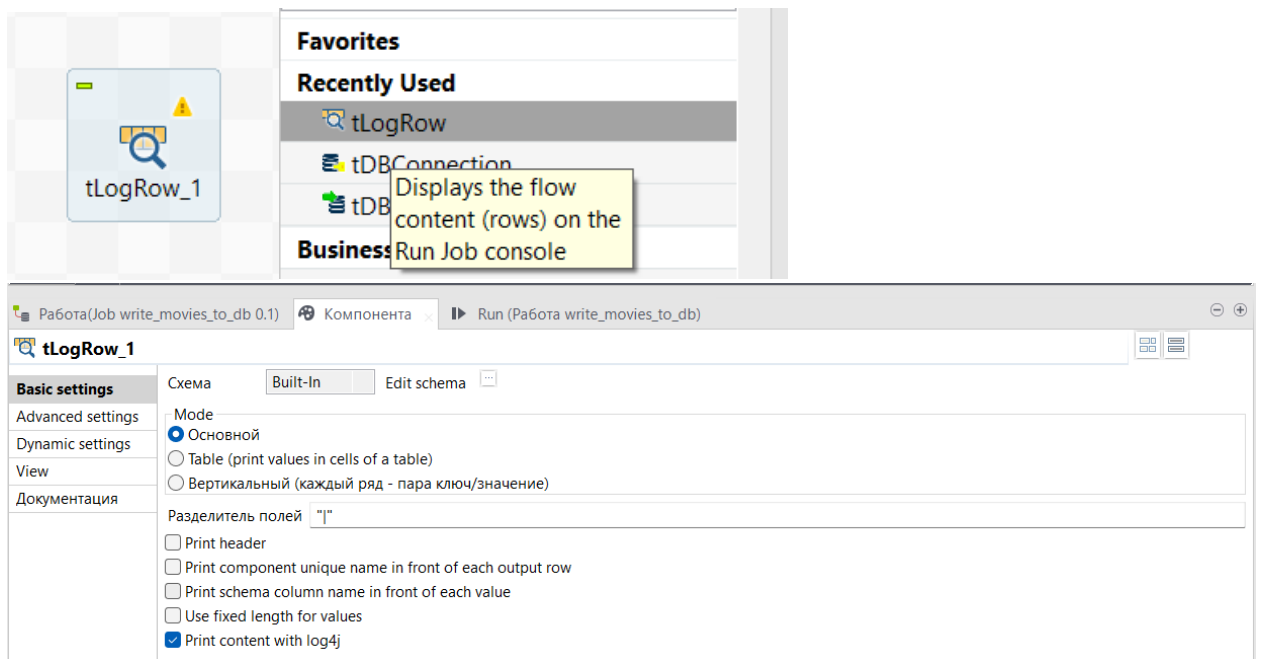
tFileOutput Excel – компонент выхода, который выгружает файл в формате Excel.



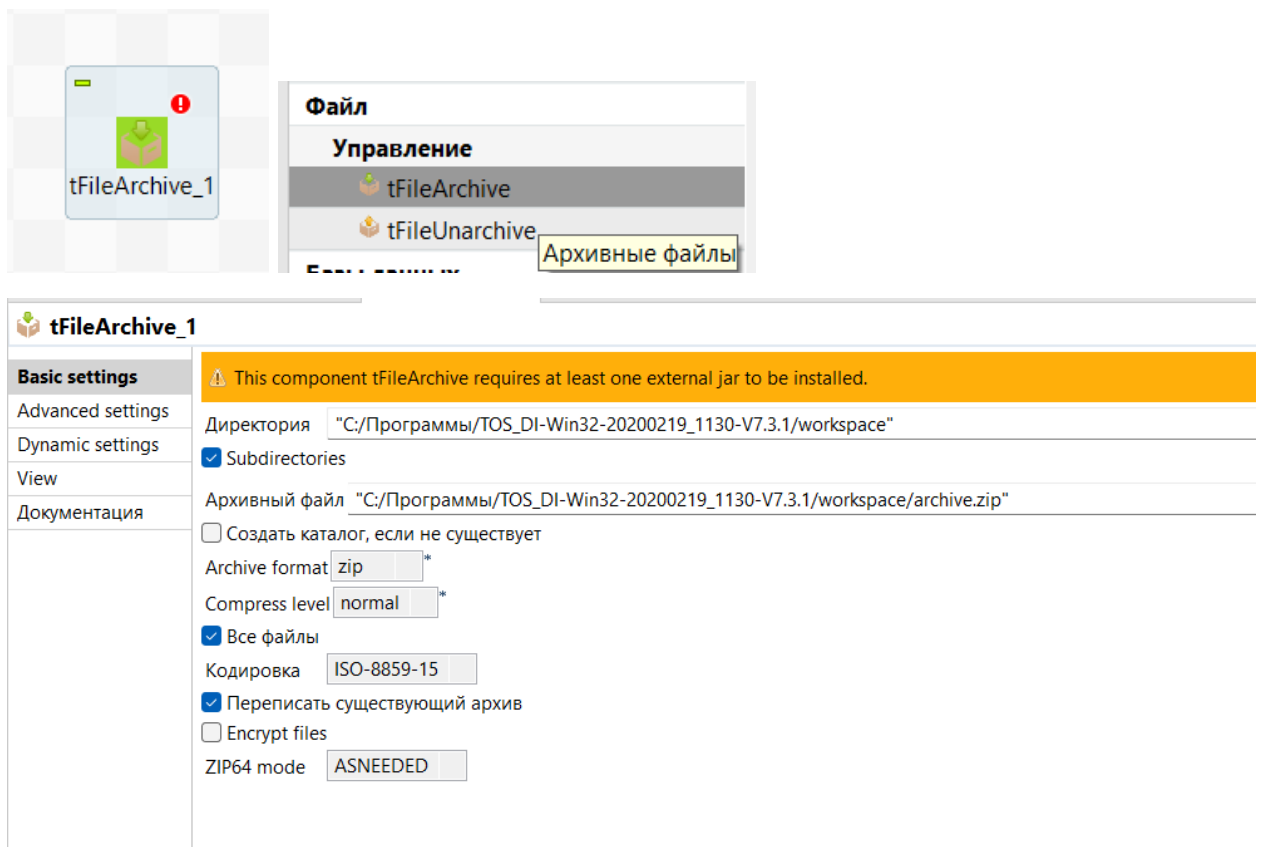
tMysqlConnection — подключается к базе данных MySQL, определенной в компоненте.



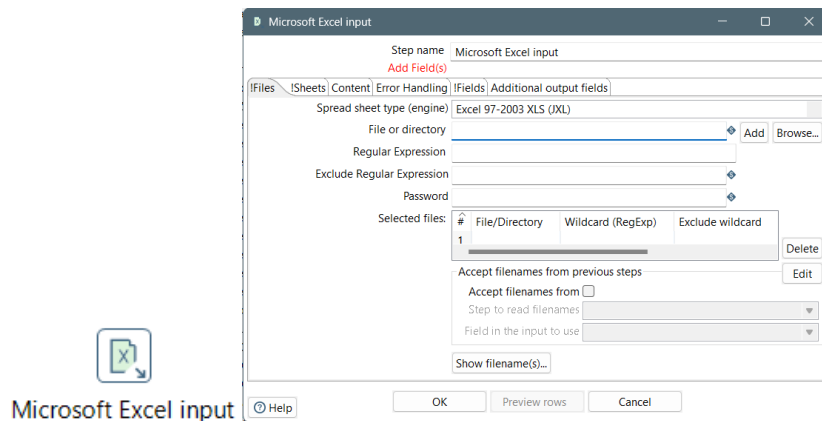
tLogRow — отслеживает обработку данных. Он отображает данные / выходные данные в консоли запуска.



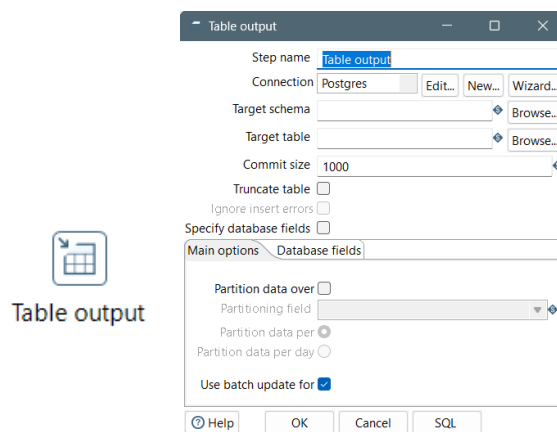
tFileArchive – сжимает набор файлов или папок в zip, gzip или tar.gz архивный файл.



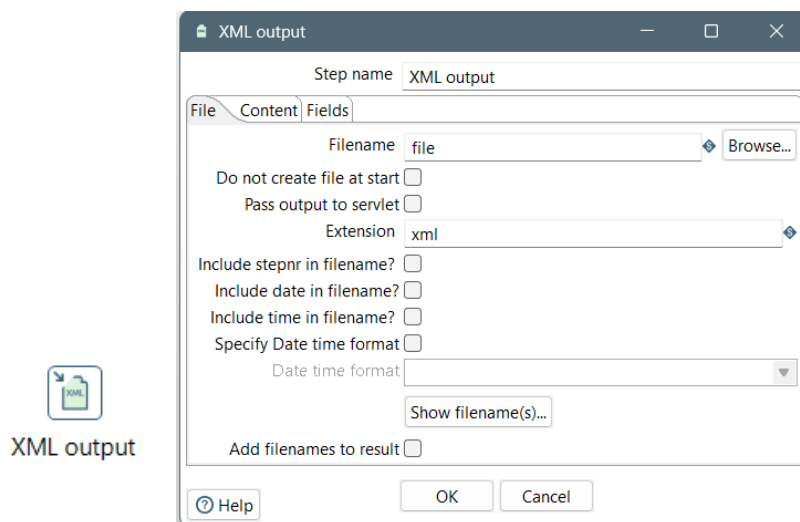
Pentaho



Компонент "Excel Input" используется для чтения данных из файлов формата Excel.

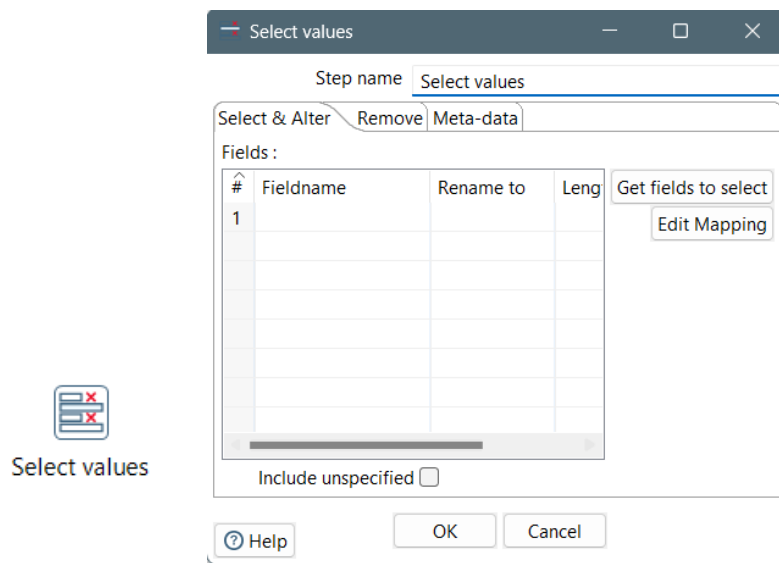


Output - Компонент "Table output" используется для записи данных в таблицы базы данных. Элемент "Connection" отвечает за указание параметров подключения к базе данных.

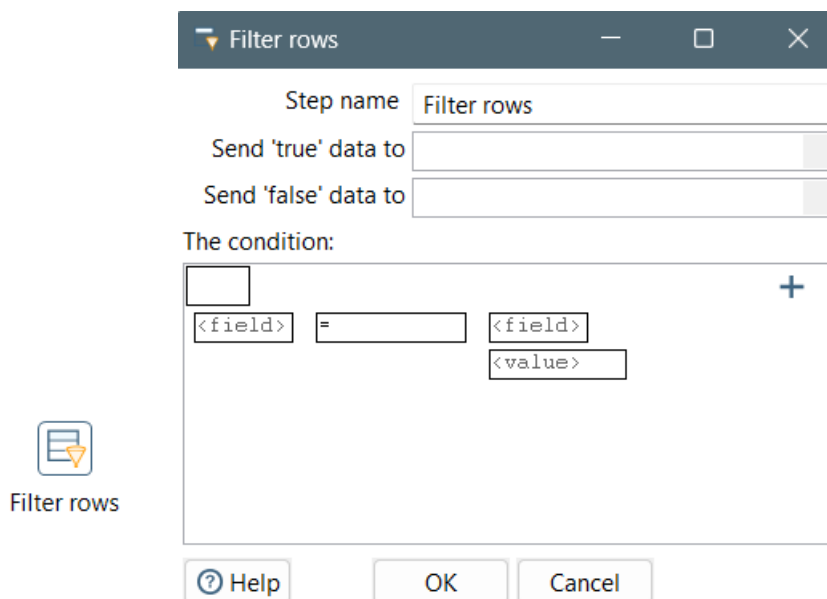


Для генерации XML-вывода вы можете использовать шаг "XML Output" (XML-вывод). Шаг "XML Output" позволяет преобразовать данные,

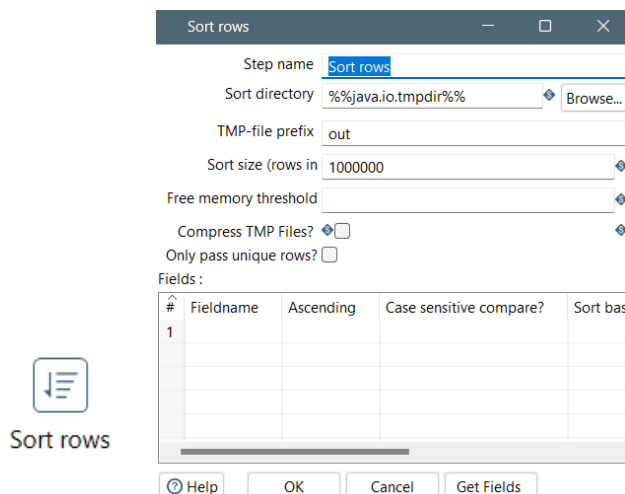
полученные на предыдущих шагах ETL-процесса, в элементы XML документа.



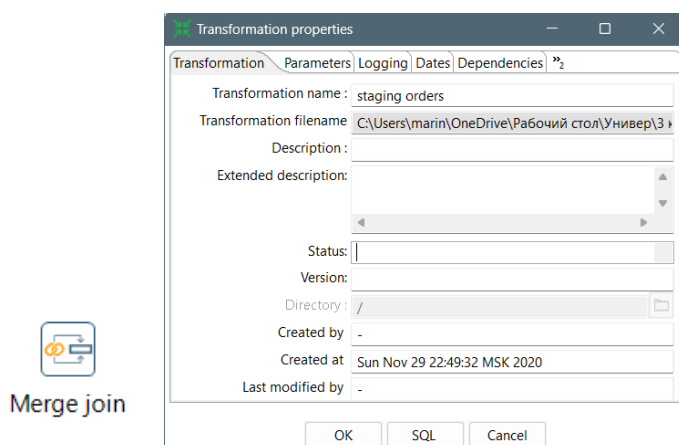
Компонент "Select values" используется для выборки определенных значений из исходного набора данных на основе заданных условий.



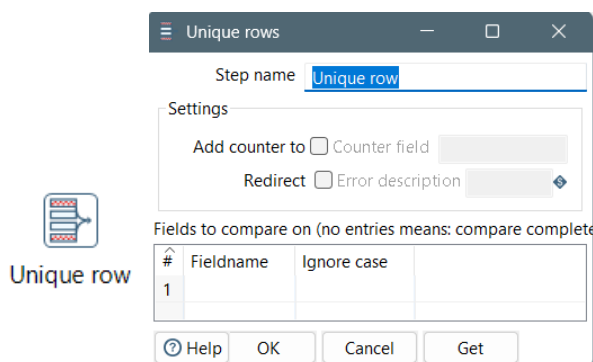
Filter - Компонент "Filter rows" используется для фильтрации данных на основе заданных условий. Один из ключевых элементов - "Condition", позволяющий задать условие фильтрации.



Sort - Компонент "Sort rows" используется для сортировки данных по заданным полям. Один из ключевых элементов - "Fields", позволяющий выбрать поля, по которым будет выполняться сортировка.



Merge Join - Компонент "Merge join" используется для соединения двух потоков данных на основе общего поля. Один из ключевых элементов - "Join type", позволяющий выбрать тип соединения данных.



Unique Rows - это шаг, который удаляет дубликаты строк из потока данных. Этот шаг может быть использован для удаления повторяющихся значений из любого набора данных, что делает его очень полезным при обработке больших объемов информации.