Министерство образования Республики Беларусь

Учреждение образования

БЕЛОРУССКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ИНФОРМАТИКИ И РАДИОЭЛЕКТРОНИКИ

Инженерно-экономический факультет

**ОТЧЕТ**

**по лабораторной работе №1**

**по предмету «СиТИОД»**

**на тему**

# «Консолидация данных и ETL»

Выполнила: cт.гр. 772303

Маркевич М.А.

Проверил: Кунцевич А.А.

Минск 2020

**ТЕОРЕТИЧЕСКИЕ ДАННЫЕ**

**Консолидация** — комплекс методов и процедур, направленных на извлечение данных из различных источников, обеспечение необходимого уровня их информативности и качества, преобразование в единый формат, в котором они могут быть загружены в хранилище данных или аналитическую систему.

**Основные критерии оптимальности с точки зрения консолидации данных:**

* обеспечение высокой скорости доступа к данным;
* компактность хранения;
* автоматическая поддержка целостности структуры данных;
* контроль непротиворечивости данных.

### **Основные задачи консолидации данных:**

* выбор источников данных;
* разработка стратегии консолидации;
* оценка качества данных;
* обогащение;
* очистка;
* перенос в хранилище данных.

Сначала осуществляется выбор источников, содержащих данные, которые могут иметь отношение к решаемой задаче, затем определяются тип источников и методика организации доступа к ним. В связи с этим можно выделить три основных подхода к организации хранения данных.

Данные, хранящиеся в отдельных (локальных) файлах, например в текстовых файлах с разделителями, документах Word, Excel и т.д. Такого рода источником может быть любой файл, данные в котором организованы в виде столбцов и записей. Столбцы должны быть типизированы, то есть содержать данные одного типа, например только текстовые или только числовые. Преимущество таких источников в том, что они могут создаваться и редактироваться с помощью простых и популярных офисных приложений, работа с которыми не требует от персонала специальной подготовки. К недостаткам следует отнести то, что они далеко не всегда оптимальны с точки зрения скорости доступа к ним, компактности представления данных и поддержки их структурной целостности. Например, ничто не мешает пользователю табличного процессора разместить в одном столбце данные различных типов (числовые и текстовые), что впоследствии обязательно приведет к проблемам при их обработке в аналитическом приложении.

Базы данных (БД) различных СУБД, таких как Oracle, SQL Server, Firebird, dBase, FoxPro, Access и т.д. Файлы БД лучше поддерживают целостность структуры данных, поскольку тип и свойства их полей жестко задаются при построении таблиц. Однако для создания и администрирования БД требуются специалисты с более высоким уровнем подготовки, чем для работы с популярными офисными приложениями.

Специализированные хранилища данных (ХД) являются наиболее предпочтительным решением, поскольку их структура и функционирование специально оптимизируются для работы с аналитической платформой. Большинство ХД обеспечивают высокую скорость обмена данными с аналитическими приложениями, автоматически поддерживают целостность и непротиворечивость данных. Главное преимущество ХД перед остальными типами источников данных — наличие семантического слоя, который дает пользователю возможность оперировать терминами предметной области для формирования аналитических запросов к хранилищу.

При разработке стратегии консолидации данных необходимо учитывать характер расположения источников данных — локальный, когда они размещены на том же ПК, что и аналитическое приложение, либо удаленный, если источники доступны только через локальную или Глобальную компьютерные сети. Характер расположения источников данных может существенно повлиять на качество собранных данных (потеря фрагментов, несогласованность во времени их обновления, противоречивость и т.д.).

Другой важной задачей, которую требуется решить в рамках консолидации, является оценка качества данных с точки зрения их пригодности для обработки с помощью различных аналитических алгоритмов и методов. В большинстве случаев исходные данные являются «грязными», то есть содержат факторы, не позволяющие их корректно анализировать, обнаруживать скрытые структуры и закономерности, устанавливать связи между элементами данных и выполнять другие действия, которые могут потребоваться для получения аналитического решения. К таким факторам относятся ошибки ввода, пропуски, аномальные значения, шумы, противоречия и т.д. Поэтому перед тем, как приступить к анализу данных, необходимо оценить их качество и соответствие требованиям, предъявляемым аналитической платформой. Если в процессе оценки качества будут выявлены факторы, которые не позволяют корректно применить к данным те или иные аналитические методы, необходимо выполнить соответствующую очистку данных.

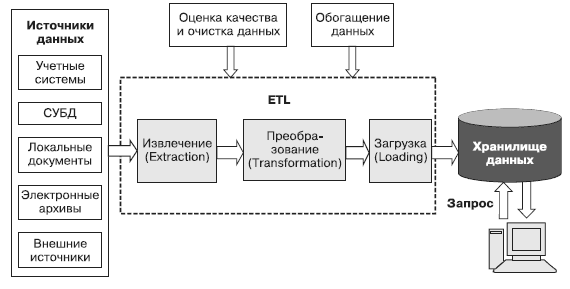
Очистка данных — комплекс методов и процедур, направленных на устранение причин, мешающих корректной обработке: аномалий, пропусков, дубликатов, противоречий, шумов и т.д.

Еще одной операцией, которая может понадобиться при консолидации данных, является их обогащение.

Обогащение — процесс дополнения данных некоторой информацией, позволяющей повысить эффективность решения аналитических задач.

Обогащение позволяет более эффективно использовать консолидированные данные. Его необходимо применять в тех случаях, когда данные содержат недостаточно информации для удовлетворительного решения определенной задачи анализа. Обогащение данных позволяет повысить их информационную насыщенность и, как следствие, значимость для решения аналитической задачи.

Место консолидации в общем процессе анализа данных может быть представлено в виде структурной схемы (рис. 1).

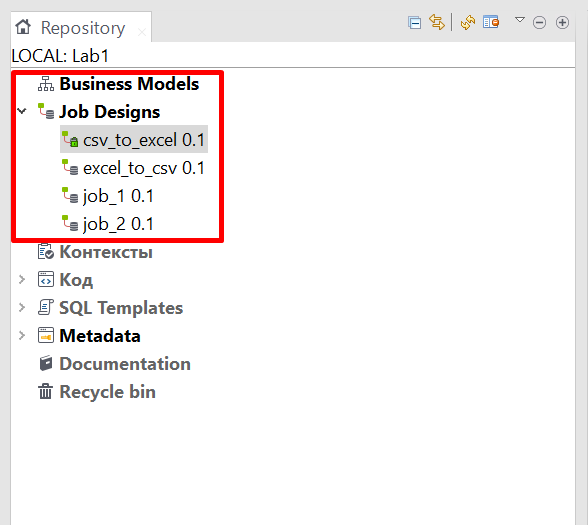
  
Рис. 1. Процесс консолидации данных

В основе процедуры консолидации лежит процесс **ETL (extraction, transformation, loading)**. Процесс ETL решает задачи извлечения данных из разнотипных источников, их преобразования к виду, пригодному для хранения в определенной структуре, а также загрузки в соответствующую базу или хранилище данных. Если у аналитика возникают сомнения в качестве и информативности исходных данных, то при необходимости он может задействовать процедуры оценки их качества, очистки или обогащения, которые также являются составными частями процесса консолидации данных.

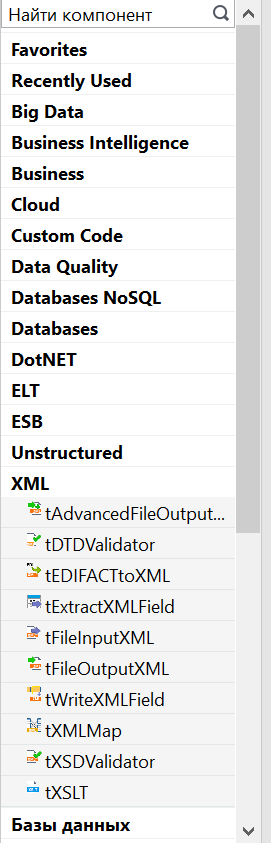
**ХОД РЕШЕНИЯ:**

**Были проведены действия аналогичные тем, что были показаны в методических указаниях.**

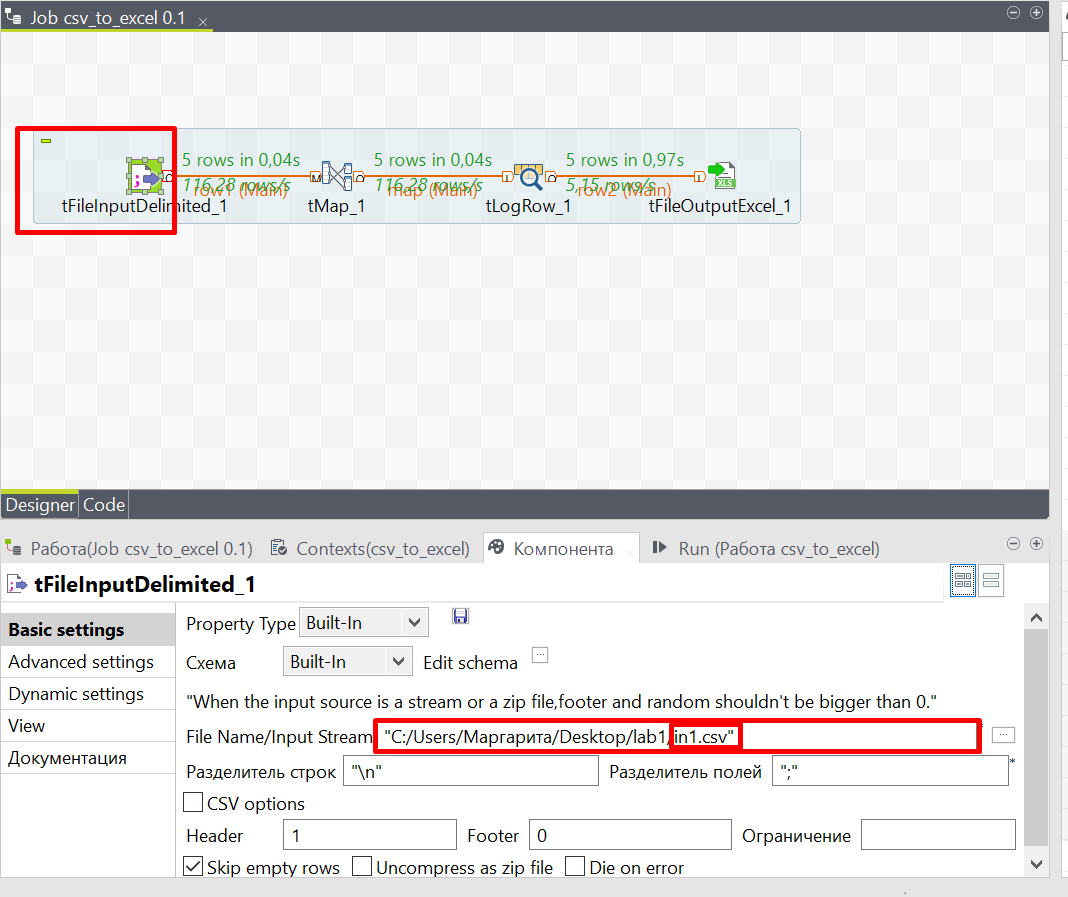
* 1. **В репозитории располагаются работы**

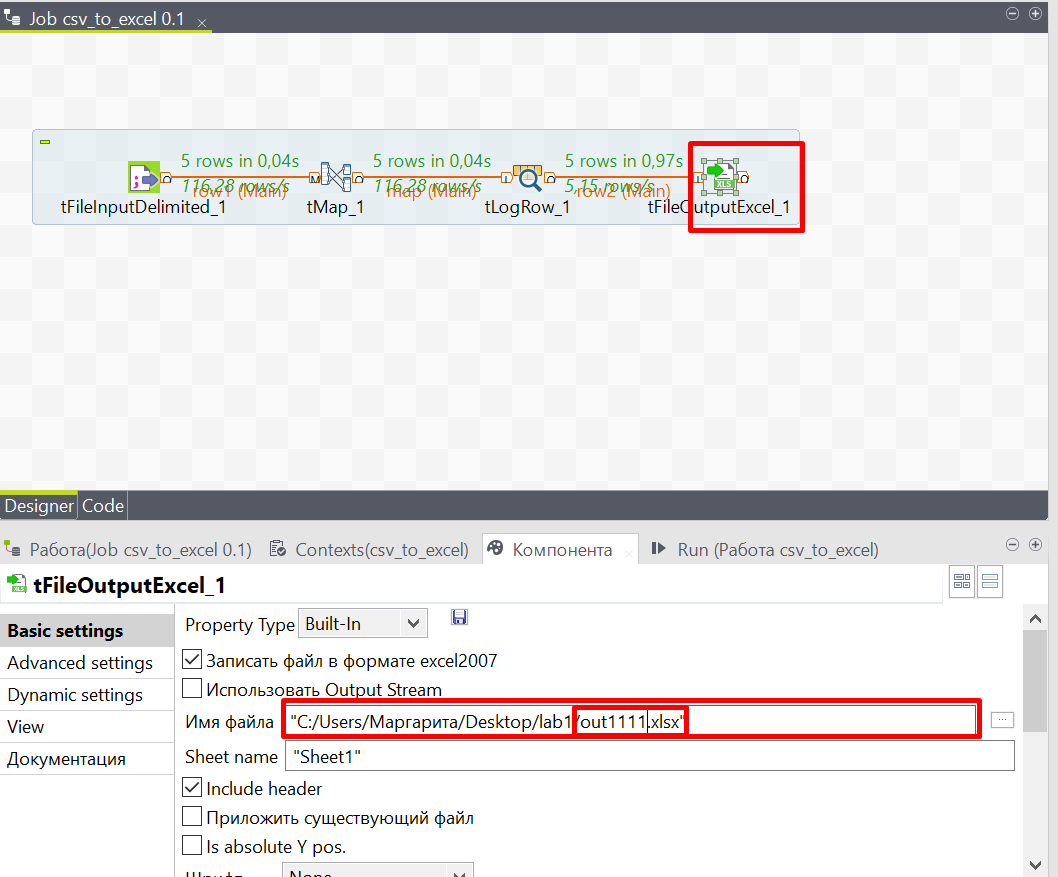


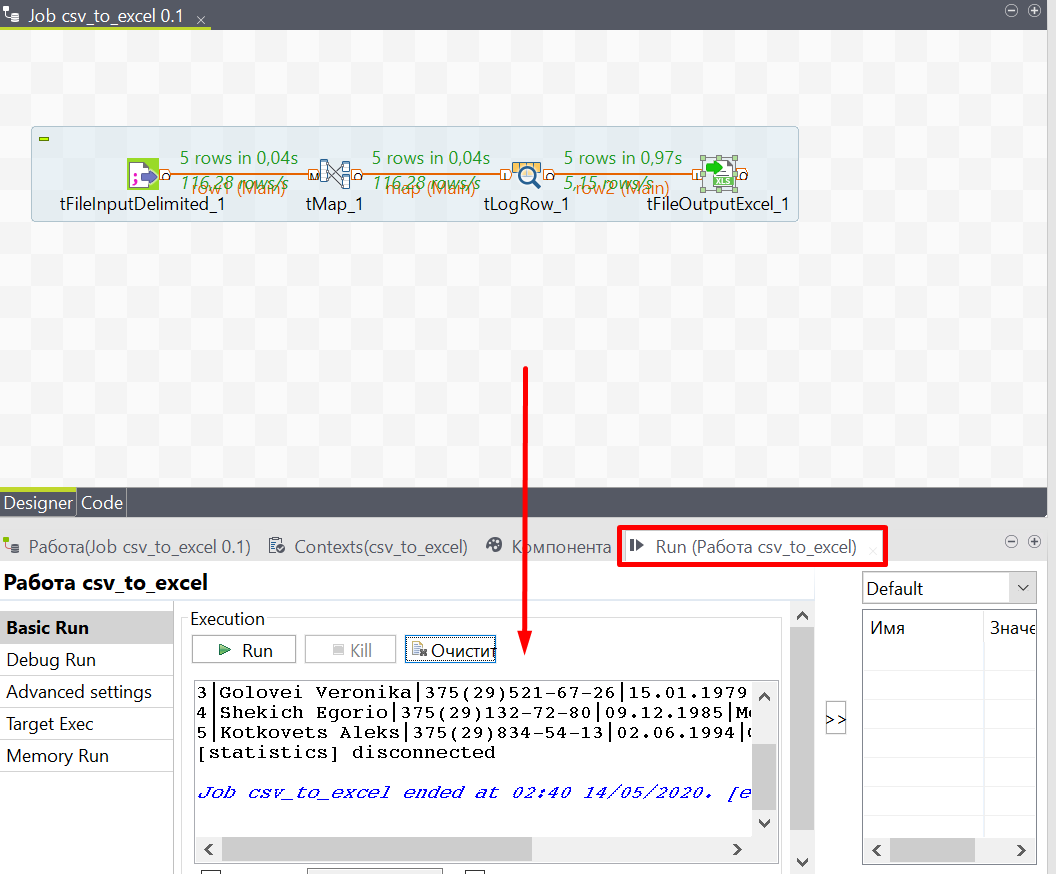
* 1. **Выбор компоненотов**



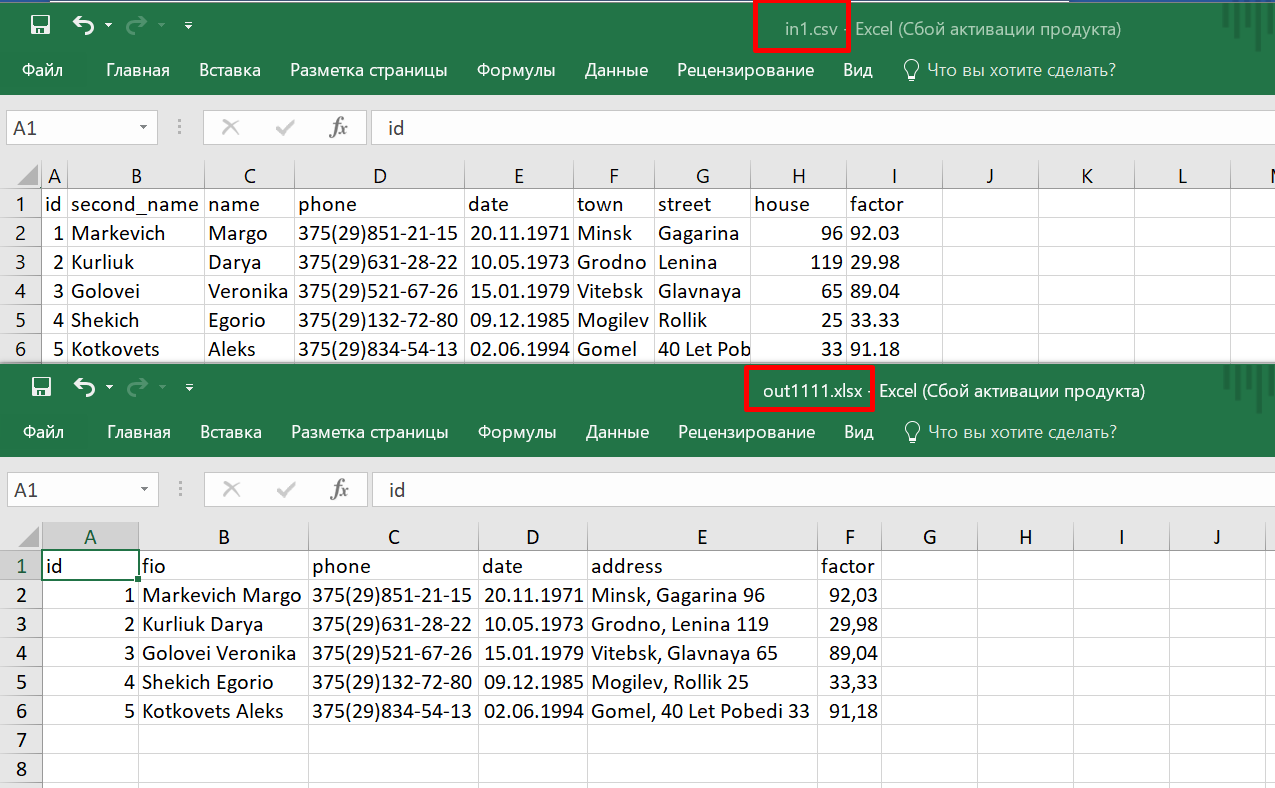
* 1. **Составленная схема, запуск и успешное выполнение**





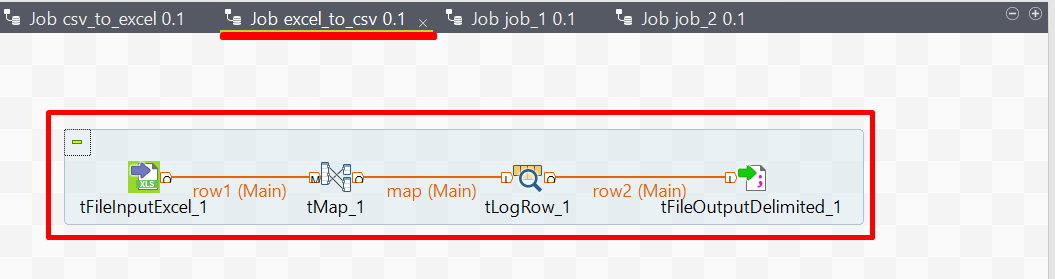


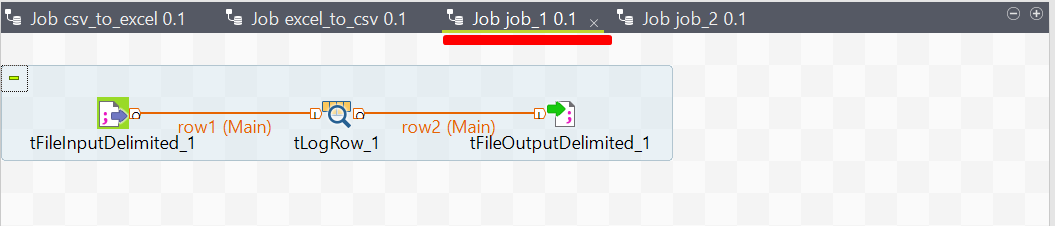
* 1. **Исходный и преобразованный файлы**

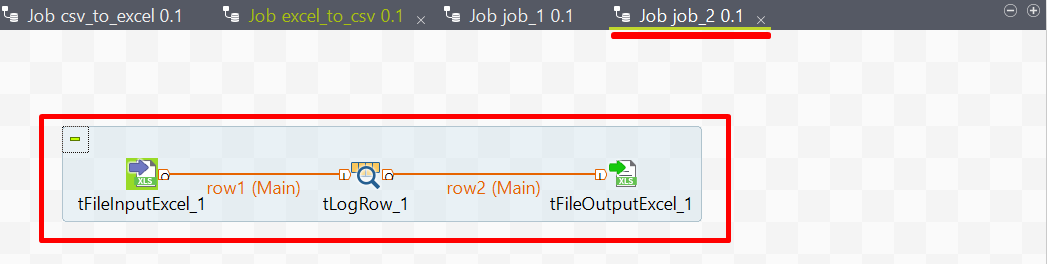


**Аналогично происходит работа с остальными работами и преобразованиями.**

**Схемы:**







**Успешное выполнение:**

