УДК 005.591.6:004.62:004.383.2:004.63

В. Ю. Бурикин, Р. И. Мастерской

СРЕДСТВА АВТОМАТИЗАЦИИ ДЛЯ ПОЛУЧЕНИЯ И ОБРАБОТКИ ДАННЫХ С УДАЛЕННОГО СЕРВЕРА

В статье рассматривается способ применения современных информационных технологий для автоматизации процесса получения и обработки большого объема данных. Описывается один из подходов к созданию системы по получению и обработке статистической информации с использованием инструментов API и StreamSets. Результатом работы является актуальная автоматизированная система, которая обеспечивает оперативный доступ, получение и обработку информации с удалённого сервера.

Введение. Для прогнозирования демографических процессов и оценки результативности долгосрочных мер в области медицины используются современные технологии обработки статистических данных, базирующиеся на математических методах, значительную часть которых составляют методы математического моделирования. В качестве параметров моделирования, как правило, необходимы статистические данные, характеризующие динамику процесса смертности, которая представлена на медицинских и демографических сайтах в виде возрастных показателей общей смертности и возрастных показателей смертности по разным причинам. Объем этих данных и скорость их прироста требует применения современных методов и средств автоматизации, обеспечивающих быструю обработку информационных потоков.

Решение этой задачи возможно с применением технологий Big Data. Эта тема интересна как с теоретической, так и практической точек зрения. Сами технологии продолжают непрерывно развиваться, что позволяет наблюдать за процессом их распространения и совершенствования в режиме реального времени, а также самим участвовать в создании и усовершенствовании новых технологий для обработки огромных массивов данных.

В статье излагается один из подходов, разработанный для оперативного извлечения и обработки данных с сайта CDC (Center for Disease Control and Prevention) [1]. Для этого предлагается использовать API (Application Programming Interface) [2], с помощью которого реализовано получение данных с удалённого сервера в форматах CSV, JSON или XML. Обработка, структурирование данных реализовано с использованием ETL-платформы в виде системы StreamSets [3]. Результат обработки данных оформлен в виде базы данных, содержащей сведения о показателях повозрастной смертности по различным причинам для разных стран и временных интервалов.

Инструментарий для получения данных с удалённого сервера. Получение данных с сайта CDC реализовано в виде API – интерфейса программирования приложений, который позволяет различным сервисам кооперироваться, обмениваться данными, получая к ним доступ. Можно отметить следующие преимущества API. При разработке сервисов экономится много времени. Разработчик получает уже готовые и хорошие решения, ему не приходится тратить ресурсы на написание подходящего ему кода. Учитываются мелочи, на которые сторонний программист может не обратить внимания. Приложениям присуща системность и прогнозируемость. Например, при помощи API одинаковая функция в различных проектах может быть осуществлена таким образом, что будет интуитивно знакома и понятна разным пользователям. Предоставляется доступ к сервисам сторонним программистам.

Существует четыре основных вида API: открытые, у которых отсутствуют ограничения на доступ, так как они общедоступны; внутренние, которые, предназначены для использования внутри компании (организация использует этот вид API для разных внутренних команд, чтобы была возможность усовершенствовать свои услуги и продукты); партнерские, в которых для доступа нужны соответствующие лицензии или права; составные, которые объединяют разные API серверов и данных.

В работе использовалось REST API (Representable State Transfer Application Programming Interface), предоставляющие набор методов, которые используются для получения и отправки данных. Эти методы используют HTTP, поэтому любой язык программирования может работать с ними. С применением этих методов были получены данные с сайта. Далее неструктурированные данные в формате csv (рисунок 1), извлеченные с помощью API, передаются в StreamSets для их обработки.

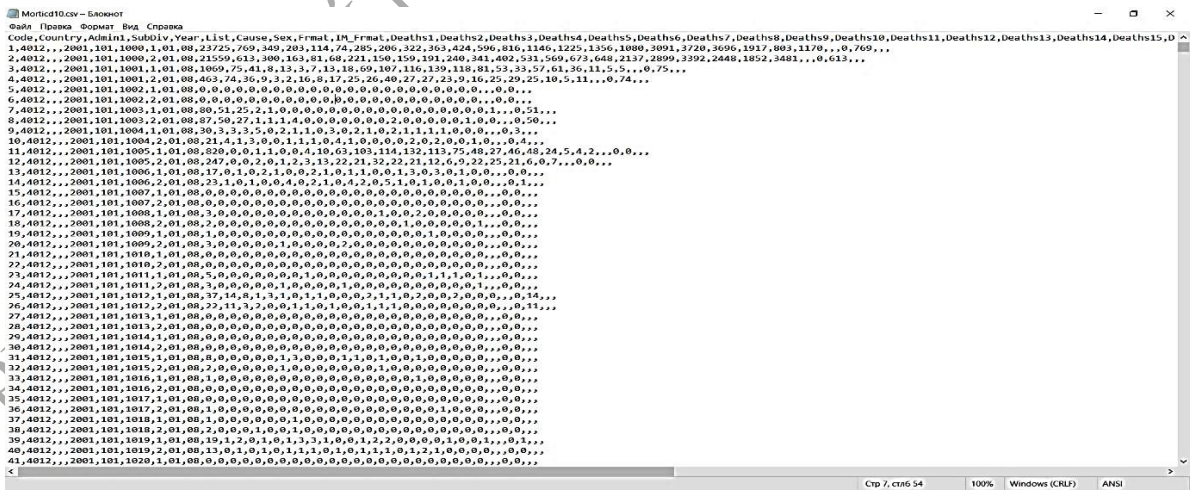

Рисунок 1 – Окно с данными в формате csv, полученными с сайта CDC

Инструментарий для обработки данных. Для работы с данными в StreamSets был создан пайплайн и в него были добавлены необходимые компоненты (рисунок 2).

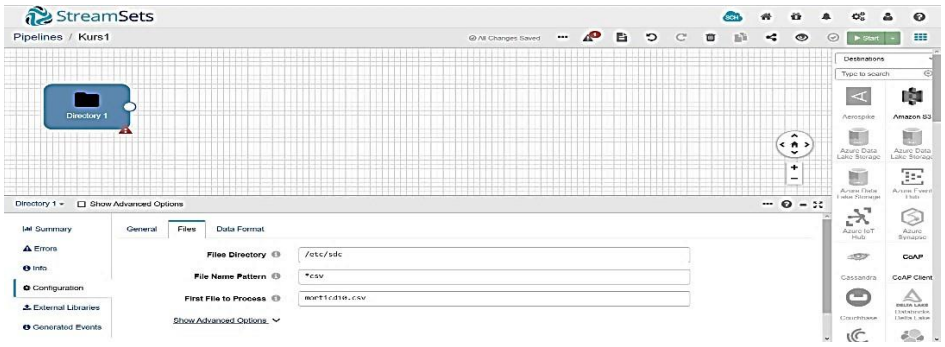

Рисунок 2 – Окно настройки параметров получения данных и определения их формата

Первым был добавлен компонент Directory. С его помощью указывается источник получения данных и указывается их формат. В поле Header Line выбирается опция With Header Line. Это позволяет использовать верхнюю строку csv файла в качестве заголовка, а в нашем случае – для названия полей.

Далее добавляется компонент Field Remover. С его помощью путем выбора нужных столбцов осуществляется фильтрация, выбираются данные из исходного файла, которые необходимо занести в базу данных (рисунок 3).

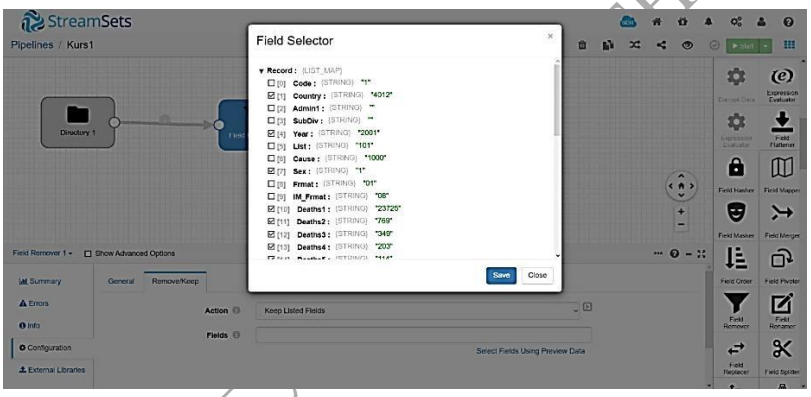

Рисунок 3 – Окно настройки параметров фильтрации

Для структурирования данных использовался компонент Field Remover (рисунок 4). Путем настройки полей Fields to Convert и Convert to Type были заданы поля, которые необходимо конвертировать и определен тип, в который было необходимо конвертировать данные выбранных полей.

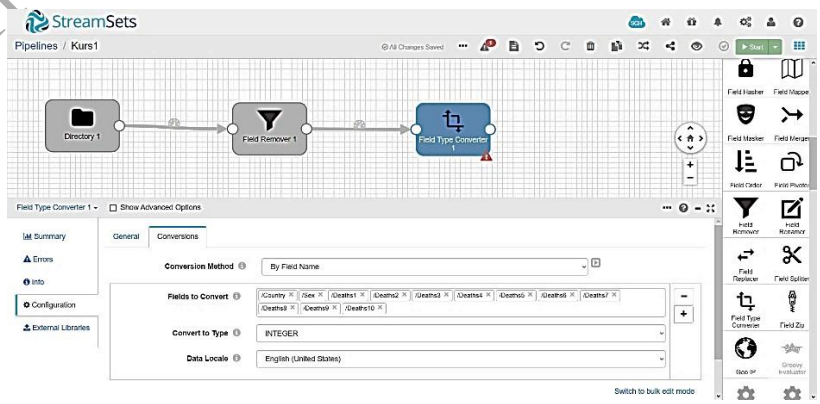

Рисунок 4 – Окно настройки преобразования типа данных к нужному формату

Для сохранения полученных данных в базе данных SQL использовался компонент JDBC Producer, в настройках которого были указаны логин и пароль от сервера, задано название таблицы (рисунок 5).

После выполнения спроектированного пайплана на сервере отображаются обработанные и структурированные данные (рисунок 6), которые могут использоваться в качестве исходных данных моделирования.

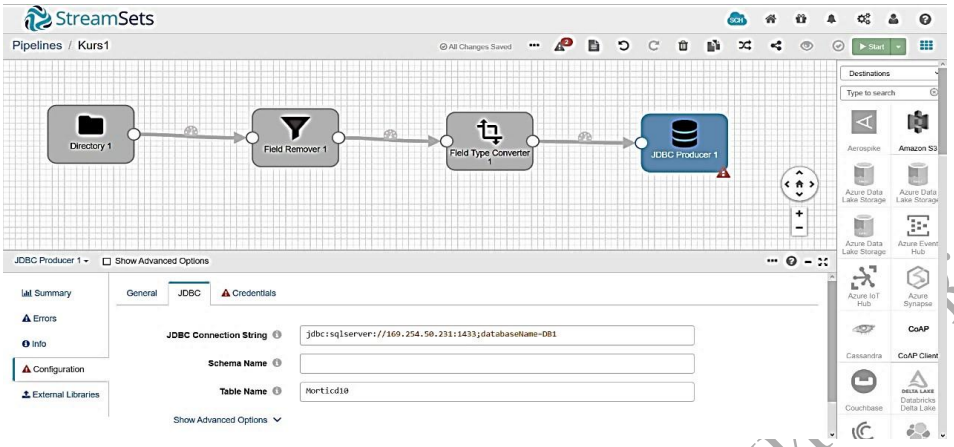

Рисунок 5 – Окно настройки доступа к месту хранения данных

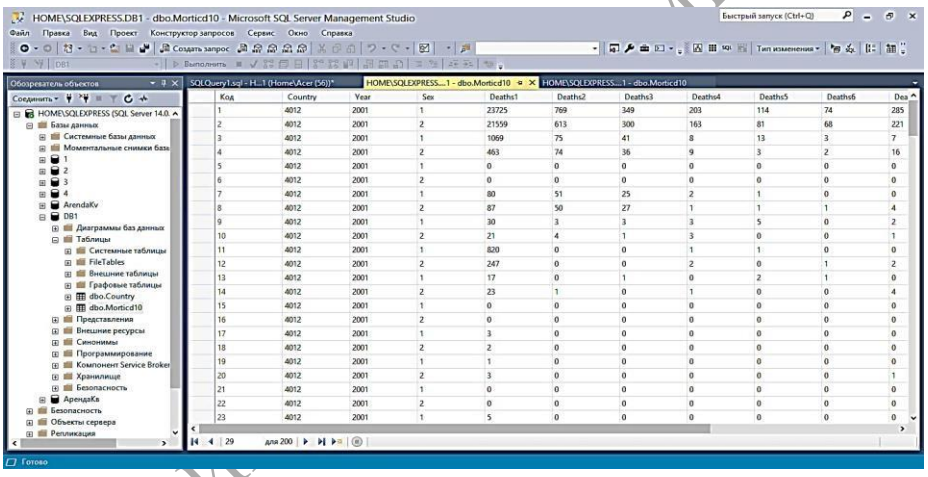

Рисунок 6 – Окно со структурированными данными в базе данных

Заключение. Одним из способов извлечения и обработки больших данных может быть применение современных технологий API и StreamSets. В статье приведены результаты разработки системы, автоматизирующей извлечение и обработку информации с сервера CDC. Способ является универсальным и может быть использован для получения данных с удаленных серверов.

Литература

1 Center for Disease Control and Prevention [Электронный ресурс]. – Режим доступа : http://www.cdc.gov/nchs/deaths.htm. – Дата доступа : 05.05.2021.

2 Арно, Л. Проектирование веб-API / Л. Арно. – Москва: ДМК-Пресс, 2020. – 440 с. 3 Система управления потоком данных StreamSets [Электронный ресурс]. –
Режим доступа : https://streamsets.com. – Дата доступа : 28.04.2021.