ETL (EXTRACT TRANSFORM AND LOAD) СИСТЕМЫ

ТОПАЛ Егор

Технический Университет Молдовы

***Аннотация**: Статья посвящена понятию ETL процесса и системам, его реализующим. Показано назначение данных систем. Дано краткое описание тому, какие функции должны выполнять ETL-системы.*

***Ключевые слова:** СУБД, ETL, Extract Transform Load, Data Migration, Big Data.*

1. Введение

Представьте себе ситуацию: вы – большая организация и у вас большое количество отделений, находящихся в разных районах, городах и, возможно, даже странах, но предоставляющих одни и те же услуги. При этом, каждое из этих отделений генерирует и сохраняет огромные объемы информации схожей по содержанию, но имеющей различный от отделения к отделению формат хранения. Вам же необходимо иметь возможность эту информацию обрабатывать: собирать статистику, производить анализ, создавать отчеты о состоянии организации в целом, а не какого-то конкретного отделения. Единственный способ это осуществить – каким-либо образом собирать данные со всех отделений в единую, как правило, OLAP-систему и уже там производить весь анализ. Процесс этот бывает достаточно сложен и чувствителен ко времени и ресурсам, а потому существует множество различных инструментов для осуществления подобного рода «миграций». Называется такой процесс ETL – Extract Transform and Load (рисунок 1).

Итак, ETL-процесс (что расшифровывается как ExtractTransformandLoad) – это комплекс методов, реализующих процесс переноса исходных данных из различных источников в аналитическое приложение или поддерживающее его хранилище данных.ETL-процесс состоит из трех этапов:

- Extract – извлечение данных из внешних источников
- Transform – преобразование и очистка данных в соответствии с бизнес-моделью
- Load – загрузка данных в целевое хранилище

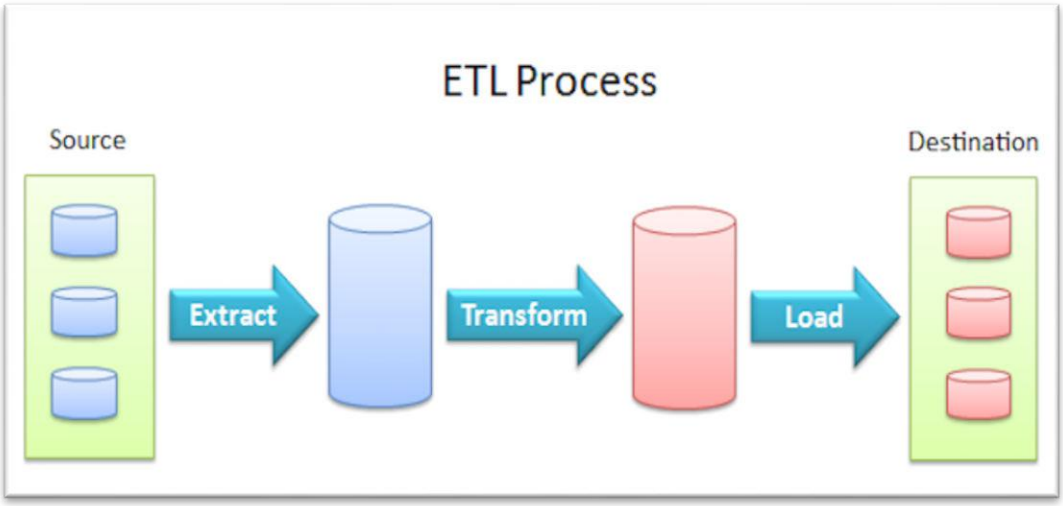

Рис. 1. Визуализации ETL процесса на высоком уровне

Стоит отметить существование сегодня очень похожего процесса, осуществляющего те же самые цели – ELT (ExtractLoadTransform). Отличие его состоит в том, что этап трансформации осуществляется уже после загрузки всех данных в так называемые DataLake базы. Хранение может быть осуществлено, например, следующим образом - для каждого источника создается отдельная схема, хранящая все необходимые для обработки таблицы. После этого, когда все данные загружены, выполняется трансформация. Такой подход обычно бывает более дорогим, однако становится все популярнее в условиях постоянно дешевеющих hardware-ресурсов.

2. Зачем нужны ETL системы?

Самостоятельный перенос с использованием тривиальных подходов, то есть копирования данных из одной базы вручную, зачастую трудно осуществим (так как в случае ETL речь часто идет об очень больших объемах данных), а особенно трудно осуществим контроль качества итогового и промежуточного результатов, а потому компании прибегают к использованию специальных ETL систем.

Коротко говоря, любая ETL система должна выполнять следующие задачи:

– Минимизировать влияние человеческого фактора на процесс путем автоматизации большинства рутинных действий;

– Обеспечить единый интерфейс доступа к различным типам источников данных (как правило, баз данных);

– Привести все данные к единой системе значений и детализации, попутно обеспечив их качество и надежность;

– Обеспечить аудиторский след при преобразовании (transform) данных, чтобы после преобразования можно было понять, из каких именно исходных данных и сумм собралась каждая строчка преобразованных данных.

3. Как работает ETL система

– Процесс загрузки – Его задача затянуть в ETL данные произвольного качества для дальнейшей обработки, на этом этапе важно сверить суммы пришедших строк, если в исходной системе больше строк, чем в RawData то значит — загрузка прошла с ошибкой;

– Процесс валидации данных – на этом этапе данные последовательно проверяются на корректность и полноту, составляется отчет об ошибках для исправления;

– Процесс мэппинга данных с целевой моделью – на этом этапе происходит трансформация данных в соответствии с существующими маппинг-таблицами (то есть схемы соответствий между таблицами источника для получения необходимого вида в хранилище данных)

– Выгрузка в целевую систему — это технический процесс использования коннектора и передачи данных в целевую систему (рисунок 2).

–

Рис. 2. Визуализации ETL процесса на более низком уровне

4. Реализация ETL на примере InformaticaBDMtool

ETL: Extract:

– PowerCentercчитывает данные, строку за строкой, из таблицы (или группы таблиц) или файла во внутреннюю базу данных

– Структура таблицы из источника сохраняется в SourceDefinitionObject

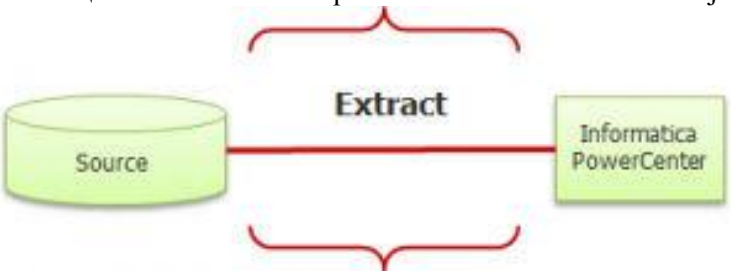

Рис. 3. Выгрузка данных в InformaticaBDM

ETL: Transform

– InformaticaPowerCenter конвертирует строки в формат, поддерживаемый хранилищем в «таргете»,

– Логика этих преобразований хранится в transformationobjects,

– InformaticaPowerCenter применяет к таблицам необходимую логику и сохраняет результат (рисунок 4).

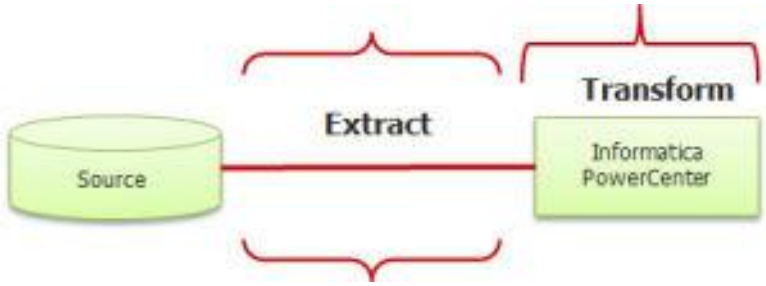

Рис. 4. Трансформация данных в Informatica BDM

ETL: Load

– InformaticaPowerCenter записывает данные, строку за строкой, в указанную таблицу, группу таблиц в базу данных или файл

– Эта база данных рассматривается как целевая

– Структура целевой базы данных хранится в targetdefinitionobject (рисунок 5).

Рис. 5. Загрузка данных в Informatica BDM

5. Заключение

Миграция данных между различными системами хранения данных – это сложный процесс, требующий особого внимания со стороны разработчиков и тщательного выбора инструментов. Использование уже спроектированных ETL-систем, таких как Informatica, позволяет облегчить процесс миграции и предупредить большинство возникающих ошибок.

Литература

1. Основные функции ETL-систем. [Электронный ресурс]. - Режим доступа:https://habrahabr.ru/post/248231/
2. Введение в ETL. [Электронный ресурс].-Режим доступа: https://bourabai.ru/tpoi/olap01-9.htm
3. Общая информация об ETL. [Электронный ресурс]. - Режим доступа: https://ru.wikipedia.org/wiki/ETL
4. What is Informatica ETL tool. [Электронный ресурс]. - Режим доступа: https://www.edureka.co/blog/what-is-informatica/