А.В. Свирновский,
магистрант 1 курса,
e-mail: ***vladimirovich1998@gmail.com,***
науч. рук.: ***В.Ф. Алексеев,***
к.т.н., доц.,
БГУИР,
г. Минск, Беларусь

ОСНОВНЫЕ ПРИНЦИПЫ РАБОТЫ ETL СИСТЕМ

Аннотация: в данной статье рассматриваются основные принципы работы ETL систем.

Ключевые слова: хранилище данных, данные, процесс загрузки

ETL (от англ. Extract, Transform, Load – дословно «извлечение, преобразование, загрузка») – комплекс методов, реализующих процесс переноса исходных данных из различных источников в аналитическое приложение или поддерживающее его хранилище данных.

Проблема, из-за которой родилась необходимость использовать решения ETL, заключается в потребностях бизнеса в получении достоверной отчетности из огромного количества разрозненных данных, которые находятся в любой ERP-системе.

Извлечение данных из разнотипных источников и перенос их в хранилище данных с целью дальнейшей аналитической обработки связаны с рядом проблем, основными из которых являются нижеследующие:

1 Исходные данные расположены в источниках самых разнообразных типов и форматов, созданных в различных приложениях, и, кроме того, могут использовать различную кодировку, в то время как для решения задач анализа данные должны быть преобразованы в единый универсальный формат, который поддерживается ХД и аналитическим приложением.

2 Данные в источниках обычно излишне детализированы, тогда как для решения задач анализа в большинстве случаев требуются обобщенные данные.

3 Исходные данные, как правило, являются «грязными»,

то есть содержат различные факторы, которые мешают их корректному анализу [1].

Принципы работы ETL-систем.

Все основные функции ETL системы показаны на рисунке 1.

Рисунок 1 – Функции ETL-системы

В разрезе потока данных это несколько систем-источников (обычно OLTP) и система приемник (обычно OLAP), а также пять стадий преобразования между ними, детализация процесса представлена на рисунке 2:

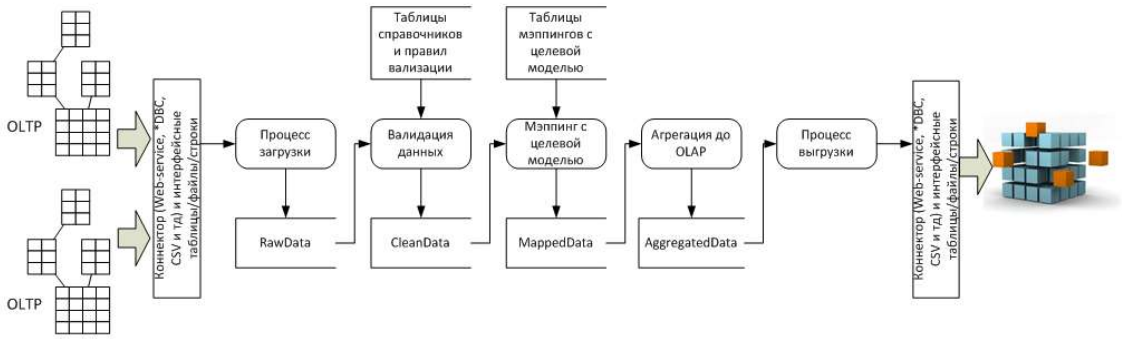

Рисунок 2 – Как работает ETL-система

Подробнее рассмотрим стадии преобразования данных между системой-источником и системой-приемником:

1 Процесс загрузки. Его задача загрузить 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.

2 Процесс валидации данных. На этом этапе данные последовательно проверяются на корректность и полноту, составляется отчет об ошибках для исправления.

3 Процесс мэппинга данных с целевой моделью. На этом этапе к валидированной таблице пристраивается некоторое количество столбцов по количеству справочников целевой модели данных, далее, по таблицам мэппингов в каждой пристроенной ячейке, в каждой строке проставляются значения целевых справочников. Значения могут проставляться как 1 к 1, так и многие к 1, так и 1 ко многим и многие ко многим, для настройки последних двух вариантов используют формулы и скрипты мэппинга, реализованные в ETL-инструменте.

4 Процесс агрегации данных. Этот процесс необходим из-за разности детализации данных в OLTP и OLAP системах. OLAP-системы – это полностью денормализованная таблица фактов и окружающие ее таблицы справочников (звездочка/снежинка), максимальная детализация сумм OLAP – это количество перестановок всех элементов всех справочников. OLTP система может содержать несколько сумм для одного и того же набора элементов справочников. Если удалять OLTP-детализацию на входе в ETL – происходит потеря «аудиторского следа». Этот след нужен для построения дрилл-даун отчета, который показывает – из каких строк OLTP, сформировалась сумма в ячейке OLAP-системы. В связи с этим сначала делается мэппинг на детализации OLTP, а потом в отдельной таблице данные агрегируют для загрузки в OLAP.

5 Выгрузка в целевую систему. Это технический процесс использования коннектора и передачи данных в целевую систему [2].

Для осуществления ETL-процесса допустимо использовать почти любой современный язык программирования. Однако, если требуется не разовая конвертация, а постоянно выполнять интеграцию данных, то целесообразно рассмотреть специализированное ПО. При этом стоит учитывать скорость, расширяемость и масштабируемость выбранного инструмента. Среди лидеров на рынке ETL-инструментов выделяются IBM DataStage, Informatica и Pentaho Data Integration. Обычно системы, созданные указанными компаниями, перекрывают потребности большинства компаний в области ETL. Исходя из этого целесообразно выбирать ETL-инструмент основываясь на поставленных задачах, а также

существующей платформе компании [3].

Литература и примечания:

[1] Интеллект [Электронный ресурс]. – Режим доступа: https://intellect.icu/osnovnye-funktsii-etl-sistem-6824 – Дата доступа: 25.01.2020.

[2] Основные функции ETL-систем [Электронный ресурс]. – Режим доступа: https://habr.com/ru/post/248231/ – Дата доступа: 27.01.2020.

[3] Молодой учёный [Электронный ресурс]. – Режим доступа: https://moluch.ru/archive/239/55368/ – Дата доступа: 30.01.2020.

© А.В. Свирновский, 2020