

ETL - системы

13.04.2023



ETL - системы

ETL (от англ. Extract, Transform, Load – дословно «извлечение, преобразование, загрузка») представляет из себя один из основных процессов в управлении хранилищами данных. Это процесс извлечения данных из внешних источников, их дальнейшей трансформации и очистки в целях приведения в соответствие с потребностями бизнес-модели, а затем загрузки их в хранилище данных. Наша компания занимается разработкой ETL систем на заказ, а также внедрением, популярных решений, таких как Informatica.

Идея создания ETL родилась как способ решения проблемы бизнеса в получении достоверной отчетности из того не достаточно структурированного объема информации, который существует в данных любой ERP-системы. Данная неструктурированность имеет, как правило, две основные причины:

1. Случайные ошибки, возникшие на уровне ввода, переноса данных и т.п.;
2. Различия в справочниках и детализации данных между смежными IT-системами.

При этом, если первую причину можно попытаться ликвидировать,

то вторая может и не являться ошибкой, т.к. контролируемые различия в структуре данных – это нормальная оптимизация под цели конкретной системы. Отсюда можно вывести не одну, а две задачи ETL-системы:

1. Приведение всех данных к единой системе значений и детализации с попутным обеспечением их качества и надежности;
2. Обеспечение аудиторского следа при преобразовании данных, чтобы после завершения процесса было понятно из каких именно исходных данных и сумм собралась каждая строка преобразования данных.

Архитектура хранилища данных с точки зрения процесса ETL может быть представлена в виде трех компонентов:

- Источник данных: содержит структурированные данные (файл, таблица, совокупность таблиц);
- Промежуточная область: здесь содержатся временные вспомогательные таблицы, созданные исключительно для организации процесса выгрузки;
- Получатель данных: хранилище/база данных для помещения извлеченных данных. Подобное перемещение называют потоком данных.

Заметим, что ETL следует рассматривать не только как процесс переноса данных из одного приложения в другое, но и как инструмент подготовки данных к анализу.

Теперь разберем подробнее **этапы процесса ETL**.

Извлечение данных

Процедура извлечения записи из источников данных и подготовка их к процессу преобразования является начальным этапом процесса ETL. В первую очередь при разработке данной процедуры необходимо определить частоту выгрузки данных из OLTP-систем или отдельных источников. Время, которое занимает выгрузка данных, называется окном выгрузки.

Процедура извлечения данных может быть реализована двумя

способами:

- с помощью специализированных программных средств;
- средствами той системы, в которой они хранятся.

После извлечения данные помещаются в «промежуточную область», в которой для каждого источника данных создается своя таблица/отдельный файл либо и то, и другое.

Преобразование данных

Целью данного этапа является подготовка данных к размещению в хранилище данных, а затем приведение их к более удобному для последующего анализа виду. Здесь важно учитывать некоторые выдвигаемые аналитиком требования к уровню качества данных, а потому в процессе преобразования может быть задействован самый разнообразный инструментарий, от простейших средств ручного редактирования данных до систем, реализующих сложные методы обработки и очистки данных. В рамках ETL в процессе преобразования данных чаще всего выполняются такие операции, как преобразование структуры данных; агрегирование данных; перевод значений; создание новых данных; очистка данных.

Загрузка данных

Процесс загрузки представляет собой перенос данных из промежуточных таблиц в структуру хранилища данных. Во время очередной загрузки в хранилище данных с источников переносится только та информация, которая была изменена в течение промежуточного времени, прошедшего с предыдущей загрузки. При этом выделяют два потока:

- поток добавления (в хранилище данных передается ранее не существующая информация);
- поток обновления/дополнения (передается информация существующая ранее, но измененная или дополненная).

Чтобы распределить загружаемые данные на потоке используют средства данных, которые фиксируют состояния данных в некоторые моменты времени и определяют, какие данные были изменены или дополнены.

В отношении ETL-интеграции можно справедливо использовать фразу «дорого, но эффективно». Не смотря на то, что это одна из наиболее затратных составляющих проекта по построению аналитического хранилища данных, выгоды для бизнеса остаются не менее очевидными:

- единый взгляд на картину производства и продажи продуктов и услуг, их доходность и себестоимость позволяет минимизировать разногласия между производственными и финансовыми подразделениями компании и, в случае необходимости, перераспределить ресурсы в пользу более выгодных и эффективных направлений;
- вместо отдельных департаментов и сотрудников, собирающих данные в рамках своих бизнес-единиц, данные собираются в автоматизированном режиме и гораздо быстрее, при этом их качество несравнимо выше. Данное обстоятельство может привести к существенному снижению затрат;
- использование ETL-интеграции позволяет заказчику сосредоточить усилия на организационной, а не технической составляющей проекта.