ПОДХОДЫ К ИНТЕГРАЦИИ ДАННЫХ НА ОСНОВЕ ETL-ПРОЦЕССОВ

Красников А.А.

Белорусский государственный университет информатики и радиоэлектроники
г. Минск, Республика Беларусь (магистрант)

Беляцкая Т.Н. – к.э.н., доцент

В работе представлены результаты исследования методов структурирования и совмещения данных в информационных системах для целей их анализа.

В начале восьмидесятых годов прошлого века, в период бурного развития регистрирующих информационных систем, возникло понимание ограниченности возможности их применения для целей анализа данных и построения на их основе систем поддержки и принятия решений. Регистрирующие системы создавались для автоматизации рутинных операций по ведению бизнеса – выписка счетов, оформление договоров, проверка состояния склада и т.д., и основными пользователями таких систем был линейный персонал. Основными требованиями к таким системам были обеспечение транзакционности вносимых изменений и максимизация скорости их выполнения. Именно эти требования определили выбор реляционных СУБД и модели представления данных "сущность-связь" в качестве основных используемых технических решений при построении регистрирующих систем.

Для менеджеров и аналитиков в свою очередь требовались системы, которые бы позволяли:
анализировать информацию во временном аспекте;

формировать произвольные запросы к системе;

обрабатывать большие объемы данных;

интегрировать данные из различных регистрирующих систем.

Регистрирующие системы не удовлетворяли ни одному из вышеуказанных требований. В регистрирующей системе информация актуальна только на момент обращения к базе данных, в следующий момент времени по тому же запросу вы можете получить совершенно другой результат. Интерфейс регистрирующих систем рассчитан на проведение жестко определенных операций и возможности получения результатов на нерегламентированный (ad-hoc) запрос сильно ограничены. Возможность обработки больших массивов данных также мала из-за настройки СУБД на выполнение коротких транзакций и неизбежного замедления работы остальных пользователей.

Ответом на возникшую потребность стало появление новой технологии организации баз данных – технологии хранилищ данных.

И благодаря этому появилась нужда в ETL процессах для загрузки данных в хранилища.

ETL завоевала популярность в 1970-х годах, когда организации начали использовать несколько репозиториев данных или базы данных для хранения различных типов деловой информации. Потребность в интеграции данных, которая распространялась по этим базам данных, быстро росла. ETL стал стандартным методом для получения данных из разрозненных источников и преобразования его перед загрузкой в целевой источник или пункт назначения[1].

В основе концепции хранилища данных лежат две основные идеи – интеграция разъединенных детализированных данных (детализированных в том смысле, что они описывают некоторые конкретные факты, свойства, события и т.д.) в едином хранилище и разделение наборов данных и приложений, используемых для оперативной обработки и применяемых для решения задач анализа. Определение понятия "хранилище данных(ХД)" первым дал Уильям Г. Инмон в своей монографии. В ней он определил хранилище данных как "предметно-ориентированную, интегрированную, содержащую исторические данные, не разрушаемую совокупность данных, предназначенную для поддержки принятия управленческих решений".

Данные из различных источников помещаются в ХД, а описания этих данных в репозиторий метаданных. Конечный пользователь, используя различные инструменты (средства визуализации, построения отчетов, статистической обработки и т.д.) и содержимое репозитория, анализирует данные в хранилище. Результатом его деятельности является информация в виде готовых отчетов, найденных скрытых закономерностей, каких-либо прогнозов. Так как средства работы конечного пользователя с хранилищем данных могут быть самыми разнообразными, то теоретически их выбор не должен влиять на его структуру и функции его поддержания в актуальном состоянии.

OLTP (англ. Online Transaction Processing), транзакционная система – обработка транзакций в реальном времени. Способ организации БД, при котором система работает с небольшими по размерам транзакциями, но идущими большим потоком, и при этом клиенту требуется от системы минимальное время отклика.

OLAP (англ. online analytical processing, интерактивная аналитическая обработка) – технология обработки данных, заключающаяся в подготовке суммарной (агрегированной) информации на основе больших массивов данных, структурированных по многомерному принципу. Реализации технологии OLAP являются компонентами программных решений класса Business Intelligence[2].

Любая транзакционная система, как правило, содержит два типа таблиц. Один из них отвечает за быстрые транзакции. Например, при продаже билетов необходимо обеспечить работу большого числа кассиров, которые обмениваются с системой короткими сообщениями. Действительно, вводимая и распечатываемая информация, касающаяся фамилии пассажира, даты вылета, рейса, места, пункта назначения, может быть оценена в 1000 байт. Таким образом, для обслуживания пассажиров необходима быстрая обработка коротких записей. Другой тип таблиц содержит итоговые данные о продажах за указанный срок, по направлениям, по категориям пассажиров. Эти таблицы используются аналитиками и финансовыми специалистами раз в месяц, или в конце года, когда необходимо подвести итоги деятельности компании. И если количество аналитиков в десятки раз меньше числа кассиров, то объемы данных, необходимых для анализа, превышают размер средней транзакции на несколько порядков величины. Естественно, что во время выполнения аналитических работ время отклика системы на запрос о наличии билета увеличивается. Создание систем с резервом вычислительной мощности может сгладить негативное воздействие аналитической нагрузки на транзакционную активность, но приводит к значительному удорожанию комплекса, при том, что избыточная мощность большую часть времени остается невостребованной. Вторым фактором, приведшим к разделению аналитических и транзакционных систем, являются разные требования, которые предъявляют аналитические и транзакционные системы к вычислительным комплексам.

История OLAP начинается в 1993, когда была опубликована статья 3 «Обеспечение OLAP (оперативной аналитической обработки) для пользователей – аналитиков». Первоначально

казалось, что разделения транзакционных и аналитических систем (OLTP – OLAP) вполне достаточно.

Однако вскоре выяснилось, что OLAP – системы очень плохо справляются с ролью посредника между различными транзакционными системами – источниками данных и клиентскими приложениями.

Стало ясно, что необходима среда хранения аналитических данных. И поначалу на эту роль претендовали единые базы данных, в которые предлагалось копировать исходную информацию из источников данных. Эта идея оказалась не вполне жизнеспособной, поскольку транзакционные системы разрабатывались, как правило, без единого плана, и содержали противоречивую и несогласованную информацию.

Так появились хранилища данных, предназначенные для надежного хранения информации, и системы извлечения, очистки и загрузки данных. OLAP-системы работали поверх хранилищ данных.

Вскоре выяснилось, что хранилища данных накапливают настолько важную для организации информацию, что всякий несанкционированный доступ в хранилище чреват серьезными финансовыми потерями. Кроме того, ориентированные на надежное хранение форматы данных плохо сочетаются с требованиями быстрого информационного обслуживания. Территориальная распределенность и организационная структура предприятия также требуют специфического подхода к информационному обслуживанию каждого подразделения. Решением является витрины данных, которые содержат необходимое подмножество информации из хранилища. Наполнение витрин из хранилища может происходить в часы спада активности пользователей. В случае сбоя информация может быть легко восстановлена из хранилища с минимальными потерями[3].

Витрины данных могут обслуживать задачи отчетности, статистического анализа, планирования, сценарных расчетов, и, в том числе, многомерного анализа (OLAP). Таким образом, системы OLAP, первоначально претендовавшие на роль чуть ли не половины вычислительного мира (отдавая вторую половину OLTP системам), в настоящее время занимают место аналитических средств уровня рабочих групп.

ХД строятся на основе многомерной модели данных. Многомерная модель данных подразумевает выделение отдельных измерений (время, география, клиент, счет) и фактов (объем продаж, доход, количество товара), которые анализируются по выбранным измерениям. Многомерная модель данных физически может быть реализована как в многомерных СУБД, так и в реляционных. В последнем случае она выполняется по схеме "звезда" или "снежинка". Данные схемы предполагают выделение таблиц фактов и таблиц измерений. Каждая таблица фактов содержит детальные данные и внешние ключи на таблицы измерений [4].

Таблица фактов – является основной таблицей хранилища данных. Как правило, она содержит сведения об объектах или событиях, совокупность которых будет в дальнейшем анализироваться.

Таблица фактов, как правило, содержит уникальный составной ключ, объединяющий первичные ключи таблиц измерений. Чаще всего это целочисленные значения либо значения типа “дата/время” в целочисленном формате — ведь таблица фактов может содержать сотни тысяч или даже миллионы записей, и хранить в ней повторяющиеся текстовые описания, как правило, не выгодно – лучше поместить их в меньшие по объему таблицы измерений. При этом как ключевые, так и некоторые не ключевые поля должны соответствовать будущим измерениям OLAP-куба. Помимо этого, таблица фактов содержит одно или несколько числовых полей, на основании которых в дальнейшем будут получены агрегатные данные[5].

Для многомерного анализа пригодны таблицы фактов, содержащие как можно более подробные данные (то есть соответствующие членам нижних уровней иерархии соответствующих измерений). Бывает предпочтительнее взять за основу факты продажи товаров отдельным заказчикам, а не суммы продаж для разных стран – последние все равно будут вычислены OLAP–средством, в случае использования такового. Исключение можно сделать, пожалуй, только для клиентских OLAP-средств, поскольку в силу ряда ограничений они не могут манипулировать большими объемами данных[6].

Таблица измерений - таблица в структуре многомерной базы данных, которая содержит атрибуты событий, сохраненных в таблице фактов. Атрибуты представляют собой текстовые или иные описания, логически объединенные в одно целое. Например, имя покупателя может являться атрибутом в таблице измерений покупателей, а наименование товара, - в таблице измерений товаров. В то время как сумма транзакции является величиной аддитивной, и ее значение должно храниться в таблице фактов.

Таблица фактов связана с таблицами измерений с помощью ключа[7].

ETL (от англ. Extract, Transform, Load – дословно «извлечение, преобразование, загрузка») – один из основных процессов в управлении хранилищами данных, который включает в себя:

Извлечение данных из внешних источников.

Трансформация и очистка, чтобы они соответствовали потребностям бизнес-модели.

Загрузка их в хранилище данных.

Это системы корпоративного класса, которые применяются, чтобы привести к одним справочникам и загрузить в DWH и EPM данные из нескольких разных учетных систем.

Для переноса исходных данных из различных источников в ХД следует использовать специальный инструментарий, который должен извлекать данные из источников различного формата, преобразовывать их в единый формат, поддерживаемый ХД, а при необходимости – производить очистку данных от факторов, мешающих корректно выполнять их аналитическую обработку.

ETL–процессы являются важной частью любого среднего и крупного юридического лица. Потребность в интеграции данных, которая распространялась по базам данных, быстро росла. ETL стал стандартным методом для получения данных из разрозненных источников и преобразования его перед загрузкой в целевой источник или пункт назначения.

Список использованных источников

1. Что такое ETL? [Электронный ресурс.] – Электронные данные. – Режим доступа : https://www.sas.com/en_us/insights/data-management/what-is-etl.html

2.] Беляцкая, Т.Н. Формирование и развитие национальной электронной экономической системы (теория, методология, управление) // Автореферат, – 2019 49 с.

3. Беляцкая, Т. Н. Формирование электронной экономики Беларуси: макроэкономические условия / Т. Н. Беляцкая // Наука и инновации. – 2018. – № 12. – С. 49–55

4.] Kimball R. The Data Warehouse Toolkit : The Definitive Guide to Dimensional Modeling / R. Kimball – USA: Wiley, 2016. - 601 с.

5. Farooq F. Real-Time Data Warehousing : A State-of-the-Art Survey / F. Farooq – USA : LAP Lambert Academic Publishing, 2011 – 112 с.

6. Хранилища данных [Электронный ресурс.] – Электронные данные. – Режим доступа : http://www.bipartner.ru/resources/dw_arch.html

7. Fact Tables and Dimension Tables [Электронный ресурс.] – Электронные данные. – Режим доступа : http://www.kimballgroup.com/2003/01/fact-tables-and-dimension-tables/

8. Fact Tables [Электронный ресурс.] – Электронные данные. – Режим доступа : http://www.kimballgroup.com/2008/11/fact-tables/

9. Таблицы фактов и измерений в хранилищах данных [Электронный ресурс.] – Электронные данные. – Режим доступа: http://bourabai.ru/tpoi/tables.htm