АЛГОРИТМ ЗАГРУЗКИ МНОГОМЕРНЫХ ДАННЫХ В ДЕНОРМАЛИЗОВАННОЕ ХРАНИЛИЩЕ

Христофорова А. А., Гуринович А. Б.
Кафедра информационных технологий автоматизированных систем,
Белорусский государственный университет информатики и радиоэлектроники
Минск, Республика Беларусь
E-mail: nastafirik@gmail.com, gurinovich@bsuir.by

Работа посвящена обеспечению обработки многомерных данных в денормализованное хранилище данных. Описано основное назначение OLAP-систем. Проанализированы проблемы при создании хранилищ данных. Детально освещается процесс загрузки данных в хранилище, с использование технологии ETL, с последующим извлечением, преобразованием и загрузкой данных.

Введение

В современном мире компьютерные сети и вычислительные системы позволяют анализировать и обрабатывать большие массивы данных. Большой объем информации усложняет поиск решений, но дает возможность получить намного более точный расчет и последующий анализ полученных решений. Для этого существуют информационные системы, называемые системами поддержки принятия решений. В ходе работы будет рассмотрена эффективность использования таких подсистем. Прежде, чем приступать к анализу данных, необходимо выполнить обработку данных, цель которой — доведение данных до определенного уровня качества и информативности, а также организовать их интегрированное хранение в структурах, обеспечивающих их целостность, непротиворечивость, высокую скорость и гибкость выполнения аналитических запросов. Исследовался инструмент для (статистической) обработки данных, в основе которого лежит процесс ETL. Данный процесс представляет собой комплекс операций, реализующих процесс переноса первичных данных из различных источников в аналитическое приложение или поддерживающее его хранилище данных. Является составной частью этапа консолидации данных в анализе данных.

I. Основные задачи и проблематика построения хранилищ данных

Технология хранилищ данных предназначена для хранения и анализа больших объемов данных с целью дальнейшего обнаружения в них скрытых закономерностей. Термин «OLAP» неразрывно связан с термином «хранилище данных». Основное назначение OLAP-систем — поддержка аналитической деятельности, произвольных запросов пользователей-аналитиков. Целью OLAP-анализа является проверка возникающих гипотез. Основная проблематика при создании хранилищ данных заключается в следующем:

- интеграция разнородных данных;
- эффективное хранение и обработка больших объемов данных;
- организация многоуровневых справочников метаданных;
- обеспечение информационной безопасности хранилище данных.

Условия высокой конкуренции и растущей динамики внешней среды диктуют повышенные требования к системам управления предприятия. Развитие теории и практики управления сопровождались появлением новых методов, технологий и моделей, ориентированных на повышение эффективности деятельности. Методы и модели в свою очередь способствовали появлению аналитических систем.

II. Обзор процесса загрузки данных в хранилище

Извлечение данных из разнотипных источников и перенос их в хранилище данных с целью дальнейшей аналитической обработки связаны с рядом проблем:

- исходные данные расположены в источниках самых разнообразных типов и форматов, созданных в различных приложениях, и, кроме того, могут использовать различную кодировку;
- данные в источниках обычно излишне детализированы, тогда как для решения задач анализа в большинстве случаев требуются обобщенные данные;
- исходные данные, как правило, являются «грязными» (отсутствующие, неточные или бесполезные данные с точки зрения практического применения), что мешает их корректному анализу.

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

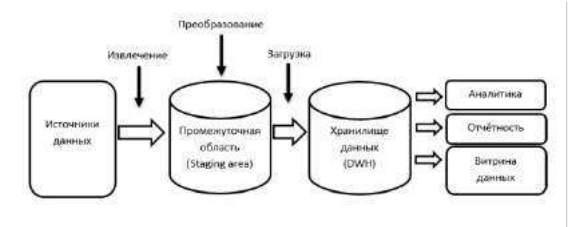

Рис. 1 – Процесс работы ETL-систем

На этапе извлечения данные извлекаются из одного или нескольких источников и подготавливаются к преобразованию. Для корректного представления данных после их загрузки в хранилище из источников должны извлекаться не только сами данные, но и информация, описывающая их структуру, из которой будут сформированы метаданные для хранилища. Процесс преобразования данных источников как правило включает в себя:

- преобразование типов данных;
- преобразования, связанные с нормализацией или денормализацией схемы данных;
- преобразования ключей, связанные с обеспечением соответствия бизнес-ключей суррогатным ключам;
- преобразования, связанные с обеспечением качества данных. Основная сложность на данном этапе заключается в получении требований о том, как именно данные должны быть преобразованы. Заключительный этап подразумевает быструю загрузку данных в хранилище данных.

Существует ряд особенностей данного этапа:

- загрузка данных, основанная на использовании команд обновления SQL, является медленной. Поэтому загрузка с помощью встроенных в СУБД средств импорта/экспорта является предпочтительной;
- индексы таблиц загружаются медленно. Во многих случаях целесообразно удалить индекс и построить его заново;
- следует максимально использовать параллелизм при загрузке данных.

Следует заметить, что при загрузке данных должна быть гарантирована ссылочная целостность данных, а агрегаты должны быть построены и загружены одновременно с подробными данными.

III. Заключение

Исследовалась актуальная область современных информационных технологий — системы анализа данных. Проанализирован основной инструмент аналитической обработки информации - OLAP - технологии. OLAP-технологии — это мощный инструмент обработки данных в реальном времени. Классифицированы основные проблемы создания хранилищ данных. Показано, что эффективным инструментом обработки и анализа данных задач является процесс ETL. Данный процесс представляет собой комплекс операций, реализующих процесс переноса первичных данных из различных источников в аналитическое приложение или поддерживающее его хранилище данных. Является составной частью этапа консолидации данных в анализе данных. При этом, увеличение эффективности обработки не влияет на целостность и безопасность данных.

IV. Список литературы

1. Использование MS SQL Server Analysis Services 2008 для построения хранилищ данных-2008 / В. В. Полубояров. – Национальный Открытый Университет "ИНТУИТ –2016. – 663 с.
2. Microsoft SQL Server 2005 Analysis Services. OLAP и многомерный анализ данных / Э. Меломед, В. П. Степаненко / БХВ-Петербург. – 2007. – 586 с. 5–13.
3. Импорт данных из Excel в SQL Server или базу данных [Электронный ресурс] / Microsoft. – Режим доступа: https://docs.microsoft.com/ru-ru/sql/relational-databases/import-export/import-data-from-excel-to-sql?view=sql-server-ver15. – Дата доступа: 17.10.2021.
4. Что такое хранилище данных? [Электронный ресурс] / PureStorage. – Режим доступа: https://www.purestorage.com/ru/knowledge/what-is-data-warehouse.html. – Дата доступа: 17.10.2021.
5. ETL: что такое, зачем и для кого [Электронный ресурс] / ЧЕРНОБРОВОВ АЛЕКСЕЙ АНАЛИТИК. – Режим доступа: https://chernobrovov.ru/articles/etl-chto-takoe-zachem-i-dlya-kogo.html. – Дата доступа: 17.10.2021.