Построение размерностных хранилищ данных для систем анализа хозяйственной деятельности предприятий Республики Беларусь

Ахрамович А.В.; Чваркова И.Л.
Кафедра интеллектуальных систем, факультет радиофизики и компьютерных технологий
Белорусский Государственный Университет
Минск, Беларусь
e-mail: anton.ahramovich@gmail.com, iryna.chvarkova@googlemail.com

***Аннотация* - В данной работе предлагается подход к проектированию размерностных хранилищ данных для систем анализа хозяйственной деятельности предприятий в реальном масштабе времени, приводятся основные сведения о технологии анализа данных в реальном времени, а также рассматриваются и классифицируются источники данных для аналитических систем.**

Ключевые слова: размерностные хранилища данных, многомерные кубы данных, аналитическая обработка данных в реальном времени (OLAP).

I. Постановка задачи

В настоящее время решение проблем хранения структурированных данных больших объёмов, а также анализа этих данных и дальнейшее принятие управленческих решений приобретает особую значимость для успешного введения хозяйственной деятельности предприятия. Использование концепции хранилищ данных позволяет оптимальным образом решить перечисленные задачи. Существует несколько готовых решений сбора и анализа данных, однако, все они основаны на дополнительном использовании систем планирования ресурсов предприятия от компании производителя аналитического программного обеспечения, что увеличивает время и стоимость внедрения аналитических систем.

Для предприятий Республики Беларусь характерно использование программного обеспечения планирования производства разработанного собственными силами предприятия и полностью отражающего как его внутреннюю структуру, так и удовлетворяющего его потребностям в полной мере. Использование подхода индивидуального проектирования размерностных хранилищ данных для систем анализа хозяйственной деятельности предприятия позволяет снизить как временные так и финансовые затраты на внедрения аналитической системы. Наличие такого хранилища данных обеспечивает возможность обработки информации в реальном масштабе времени для интерактивного анализа многомерной структуры.

В данной работе рассматривается принципы проектирования размерностных хранилищ данных для построения аналитических систем, рассматриваются необходимые уровни таких аналитических систем и анализируются возможные сценарии их использования.

II. Общая архитектура аналитической системы

В зависимости от индивидуальных особенностей хозяйственной деятельности предприятия допустимы различные подходы к организации аналитических систем. В тоже время аналитическая система может быть представлена в виде абстрактной схемы состоящей из трёх уровней (Рис. 1). Первый уровень содержит разнородные источники данных, такие как существующие базы данных предприятия, плоские таблицы, документы и др. Второй уровень представляет собой размерностное хранилище данных. Третий уровень содержит инструменты для предоставления информации конечным пользователям и/или сценарии использования разработанного хранилища в виде нового источника данных для других систем.

Рис. 1. Абстрактное описание системы

III. Подготовка исходных данных

Для решения задачи сбора данных из разнородных источников используют специальные программные инструменты - ETL, предоставляющие возможность в автоматическом режиме аккумулировать данные из источников, осуществлять фильтрацию, преобразовывать и загружать преобразованные данные в единую базу данных. Единая база данных состоит из таблиц, структурированных в соответствии со схемой звезды или снежинки (Рис. 2). Обе схемы организации данных представлены централизованной таблицей фактов, которая в свою очередь связана с таблицами измерений. Отличием между этими схемами является то, что в схеме «снежинка» таблицы измерений нормализованы с рядом других связанных таблиц измерений, в то время как в схеме «звезда» таблицы измерений полностью денормализованы: каждое измерение представлено в виде единой таблицы [1]. Структура базы данных и выбор той или иной схемы напрямую связаны с индивидуальными особенностями хозяйственной деятельности предприятия. Также допускается комбинированный вариант проектирования базы данных, при котором одна таблица измерений могла быть связана сразу с несколькими таблицами фактов.

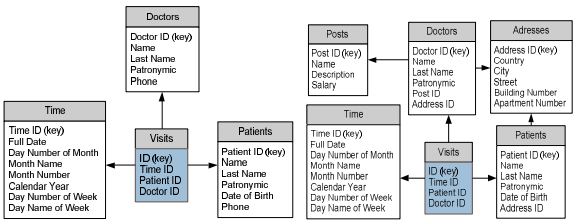

Рис. 2. Схема звезды и снежинки

IV. Проектирование хранилища данных

Проектирование хранилищ данных является ключевым фактором при создании системы анализа хозяйственной деятельности предприятий. Гибкость конечной системы, простота её использования и качество обработанной информации напрямую обусловлены качеством и продуманностью спроектированного хранилища данных. Существуют следующие общие подходы к проектированию хранилищ данных: “сверху вниз” и “снизу вверх”. При использовании подхода “сверху вниз” проектирование хранилища данных начинается с создания общей структуры. Такой подход применим в случаях, когда необходимые для разработки технологии и бизнес процессы анализируемой системы хорошо изучены и полностью понятны. В случае использования подхода “снизу вверх” проектирование начинается с экспериментов, предположений и разработки прототипов. Данный подход зачастую используется на ранних стадиях разработки аналитической системы и моделирования бизнес процессов [2].

Таким образом, процесс проектирования хранилища данных можно представить в виде следующих последовательных действий:

1. Определение набора бизнес процессов для последующего анализа (например, данные о продажах, складировании и т.д.)
2. Определение минимального «звена» бизнес модели. «Звено» является наименьшим уровнем представления данных в таблице фактов хранилища данных (например, транзакция, покупка и т.д.).
3. Определение необходимых измерений, позволяющих полностью описать состояние модели (например, измерения времени, заказчиков, поставщиков и т.д.)
4. Определение необходимых атрибутов, характеризующих каждую запись в таблице фактов. Зачастую это числовые данные, такие как количество проданных товаров, сумма продажи и т.п.
5. Выбор схемы проектируемого хранилища данных.

V. Сценарии использования хранилища данных

Разработанное хранилище данных может быть использовано в соответствии с различными сценариями. Во-первых, на основе данного хранилища существует возможность использования OLAP системы для анализа данных в реальном времени. Накопленная в базе данных информация объединяется в специальные структуры (OLAP-кубы), позволяющие осуществлять комплексный анализ за минимальное время. Базы данных OLAP содержат два основных типа данных: показатели, являющиеся числовыми данными, и признаки, являющиеся категориями, используемыми для организации этих показателей. OLAP системы обладают широким набором инструментов для работы с иерархичными данными (поддерживаются операции детализации данных, операции агрегирования, операции среза/поворота куба и многие другие). В зависимости от реализации хранилища данных выделяют реляционные, многомерные и гибридные системы обработки данных в реальном времени [3].

Во-вторых, разработанное хранилище данных может быть использовано в качестве источника данных как для уже разработанных внутренних систем, специализированых на выполнение определённых задач, так и для других хранилищ данных.

В-третьих, данные, накопленные в разработанном хранилище, могут быть предоставлены конечному пользователю при помощи специализированных программных инструментов. В зависимости от сложности используемых инструментов пользователю может быть предоставлена возможность просмотра данных как в виде статических, так и в виде сложных и интерактивных отчётов [4].

VI. Заключение

Таким образом, использование размерностных хранилищ для систем анализа хозяйственной деятельности предприятий обнаруживает следующие преимущества:

1. На базе хранилища данных существует возможность создания гибкой аналитической системы, позволяющей формировать отчёты любой сложности на основе большого количества данных.
2. Автоматизированный сбор данных для аналитической системы позволяет уменьшить количество сотрудников, осуществляющих подготовку, сортировку, загрузку данных.
3. Размерностные хранилища данных являются универсальным источником данных для сторонних систем анализа и позволяют легко предоставлять накопленные данные, как конечным пользователям, так и другим аналитическим системам.

[1] Microsoft SQL Server 2005 Analysis Services. OLAP и многомерный анализ данных / А. Б. Бергер [и др.]; под общ. ред. А. Б. Бергера, И. В. Горбач. – Санкт-Петербург: БХВ-Петербург, 2007. – 928 с.

[2] Pedersen, T.B. Multidimensional Database Technology / T.B. Pedersen, C. S. Jensen // IEEE Computer. – 2001. –Vol. 34, № 12. – P. 40 - 46.

[3] Холод, И.И. Методы и модели анализа данных: OLAP и Data Mining / И. И. Холод, М. С. Куприянов, В. В. Степаненко. – Санкт-Петербург: БХВ-Петербург, 2004. –336 с.

Ponniah, P. Data warehousing fundamentals for IT professionals/ P. Ponniah – John Wiley & Sons Inc., 2010 – 571 p.

[4] Mark I. Hwang, Hongjiang Xu The Effect of Implementation Factors on Data Warehousing Success: An Exploratory Study // Journal of Information, Information Technology, and Organizations Volume 2, 2007.

[5]