

Акимова Фаина Валерьевна

студентка

Манько Алёна Максимовна

студентка

Чубарова Олеся Викторовна

канд. техн. наук, доцент

Институт космических и информационных технологий

ФГАОУ ВО «Сибирский федеральный университет»

г. Красноярск, Красноярский край

ХРАНИЛИЩА ДАННЫХ: ОСНОВНЫЕ АРХИТЕКТУРЫ И ПРИНЦИПЫ ПОСТРОЕНИЯ В РЕЛЯЦИОННЫХ СУБД

Аннотация: в данной статье рассматриваются архитектуры хранилищ данных, рассмотрены некоторые общие принципы их построения. Отмечено, что в основе концепции хранилища данных лежат две основные идеи – интеграция разъединенных детализированных в едином хранилище и разделение наборов данных и приложений, используемых для оперативной обработки и применяемых для решения задач анализа.

Ключевые слова: база данных, хранилище данных, виртуальное хранилище данных, двухуровневая архитектура.

В основе концепции хранилища данных лежат две основные идеи – интеграция разъединенных детализированных в едином хранилище и разделение наборов данных и приложений, используемых для оперативной обработки и применяемых для решения задач анализа. Хранилище данных подразумевает под собой предметно-ориентированную, интегрированную, содержащую исторические данные, не разрушаемую совокупность данных, предназначенную для поддержки принятия управленческих решений.

Физическая реализация приведенной концептуальной схемы может быть самой разнообразной. Ниже приводятся наиболее часто встречающиеся подходы.

Виртуальное хранилище данных – это система, представляющая интерфейсы и методы доступа к регистрирующей системе, которые эмулируют работу

с данными в этой системе, как с хранилищем данных. Виртуальное хранилище данных можно организовать, создав ряд представлений (view) в базе данных, либо применив специальные средства доступа, например, продукты класса Desktop OLAP, к которым относятся, например, BusinessObjects, Brio Enterprise и другие [14].

Главными достоинствами такого подхода являются:

- простота и малая стоимость реализации;
- единая платформа с источником информации;
- отсутствие сетевых соединений между источником информации и хранилища данных.

Однако недостатков у него гораздо больше, чем достоинств. Создавая виртуальное хранилище данных, Вы создаете не хранилище как таковое, а иллюзию его существования. Структура хранения данных и само хранение данных не претерпевает изменений, и остаются проблемы:

- производительности;
- трансформации данных;
- интеграции данных с другими источниками;
- отсутствия истории;
- чистоты данных;
- зависимость от доступности основной БД;
- зависимость от структуры основной БД.

Двухуровневая архитектура хранилища данных подразумевает построение витрин данных (data mart) без создания центрального хранилища, при этом информация поступает из небольшого количества регистрирующих систем и ограничена конкретной предметной областью. При построении витрин данных используются основные принципы построения хранилищ данных, о которых пойдет речь ниже, поэтому их можно считать хранилищами данных в миниатюре. Плюсами витрин данных являются:

- простота и малая стоимость реализации;

- высокая производительность за счет физического разделения регистрирующих и аналитических систем, выделения загрузки и трансформации данных в отдельный процесс, оптимизированной под анализ структурой хранения данных;
- поддержка истории;
- возможность добавления метаданных.

Построение полноценного корпоративного хранилища данных обычно выполняется в трехуровневой архитектуре (следует отметить, что здесь под трехуровневой архитектурой понимается не структура «БД – Сервер приложений – клиент»). На первом уровне расположены разнообразные источники данных – внутренние регистрирующие системы, справочные системы, внешние источники (данные информационных агентств, макроэкономические показатели). Второй уровень содержит центральное хранилище данных, куда стекается информация от всех источников с первого уровня, и, возможно, оперативный склад данных (ОСД). Оперативный склад не содержит исторических данных и выполняет две основные функции. Во-первых, он является источником аналитической информации для оперативного управления и, во-вторых, здесь подготавливаются данные для последующей загрузки в центральное хранилище. Под подготовкой данных понимают их преобразование и осуществление определенных проверок. Наличие ОСД просто необходимо при различном регламенте поступления информации из источников. Третий уровень в описываемой архитектуре представляет собой набор предметно-ориентированных витрин данных, источником информации для которых является центральное хранилище данных. Именно с витринами данных и работает большинство конечных пользователей.

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.

При реализации проектов по построению хранилищ данных возникает ряд общих задач, независимых от предметной области обрабатываемой информации.

К числу таких задач можно отнести:

- проектирование структуры иерархических измерений;
- проектирование структуры медленно меняющихся измерений;
- проектирование и актуализация агрегатных значений.

В данной статье были рассмотрены возможные решения этих задач. В частности, были приведены способы реализации иерархических измерений с помощью введения дополнительных атрибутов (левая и правая граница), а также с помощью введения дополнительной таблицы – «helper-table». Однако во всех рассмотренных задачах существуют нерешенные вопросы, требующие дальнейших исследований. В частности, сложным для реализации является случай иерархических измерений с необходимостью поддержания истории изменений, которые имеют связи с какими-либо другими справочниками. В данную статью не вошли вопросы, касающиеся методов очистки данных и алгоритмов загрузки данных в хранилище. Эти темы требуют отдельного рассмотрения.

Список литературы

1. Joerg Reinschmidt, Allison Franchoise // Business Intelligence Certification Guide. IBM Red books.
2. Inmon W. Building the Data Warehouse. – New York: John Willey & Sons, 1992.
3. Спирли Эрик. Корпоративные хранилища данных. Планирование, разработка, реализация. – Том. 1: Пер. с англ. – М.: Вильямс, 2001.
4. Joe Celko. Trees in SQL: Intelligent Enterprise, October 20, 2000.
5. Дональд Э. Кнут. Искусство программирования. – Том 1 // Основные алгоритмы. – 3-е изд. – М.: Вильямс, 2000.