МНОГОМЕРНЫЙ АНАЛИЗ ДАННЫХ. OLAP-кубы

САРАНЧУК Дориан, МАКИДОН Мария

Технический Университет Молдовы

***Аннотация**: В данной статье описываются ключевые моменты использования многомерных баз данных и OLAP модель обработки данных. Приведены правила Кодда для программных продуктов OLAP, а также тест FASMI. Описаны компоненты OLAP-систем. Рассмотрены особенности MDX запросов посредством SQL Server Analysis Services.*

***Ключевые слова:** многомерная база данных, OLAP, измерения, кортежи, MDX запрос.*

1. ВВЕДЕНИЕ

OLAP (On-Line Analytical Processing) — оперативная аналитическая обработка данных. Эта технология применяется в хранилищах данных для повышения эффективности анализа данных. Причём анализ данных происходит в режиме реального времени.

OLAP представляет собой инструмент для анализа больших объёмов данных. Взаимодействуя с OLAP-системой, пользователь может осуществлять гибкий просмотр информации, получать произвольные срезы данных и выполнять аналитические операции детализации, свёртки, сквозного распределения, сравнения во времени. Вся работа с OLAP-системой происходит в терминах предметной области.

OLAP-системы являются часть понятия Business Intelligence, которое включает в себя помимо традиционного OLAP-сервиса средства организации совместного использования документов, возникающих в процессе работы пользователей хранилища.

В основе концепции OLAP лежит принцип многомерного представления данных. По измерениям в многомерной модели выделяют факторы, влияющие на деятельность предприятия и получают гиперкуб, который затем наполняется показателями деятельности предприятия (цены, продажи, план, прибыли, убытки и т.п.). Наполнение это может вестись как реальными данными оперативных систем, так и прогнозируемыми на основе исторических данных, то есть данных, накопленных за определённый период времени.

Измерения гиперкуба могут носить сложный характер, быть иерархическими, между ними могут быть установлены отношения. В процессе анализа пользователь может менять точку зрения на данные, тем самым, просматривая данные в различных разделах и разрешая конкретные задачи. Над кубами могут выполняться различные операции, включая прогнозирование и условное планирование.

Важнейшим его элементом являются метаданные, то есть информация о структуре, размещении и трансформации данных. Благодаря им обеспечивается эффективное взаимодействие различных компонентов хранилища [1].

2. ТРЕБОВАНИЯ К OLAP-СИСТЕМАМ И СПОСОБЫ ХРАНЕНИЯ ДАННЫХ В НИХ

В 1993 году Е.Ф. Коддом — создателем концепции реляционных СУБД, а также OLAP — были сформулированы критерии OLAP. Они заключаются в недостатках реляционной модели и, в первую очередь, указывают на невозможность объединять, просматривать и анализировать данные с точки зрения множественности измерений, то есть понятным для корпоративных аналитиков способом. Общие требования к системам OLAP расширяют функциональность реляционных СУБД и включают многомерный анализ как одну из своих характеристик.

Кодд определил 12 правил для программного продукта OLAP:

1. Многомерное концептуальное представление данных. Концептуальное представление модели данных в продукте OLAP должно быть многомерным по своей природе, то есть позволять аналитикам выполнять интуитивные операции анализа вдоль и поперёк, вращения и размещения направлений консолидации.
2. Прозрачность. Пользователь не должен знать о том, какие конкретные средства используются для хранения и обработки данных, как данные организованы и откуда берутся.
3. Доступность. Аналитик должен иметь возможность выполнять анализ в рамках общей концептуальной схемы, но при этом данные могут оставаться под управлением оставшихся от старого наследства СУБД, будучи при этом привязанными к общей аналитической модели. Инструментарий OLAP должен накладывать свою логическую схему на физические массивы данных,

выполняя все преобразования, требующиеся для обеспечения единого, согласованного и целостного взгляда пользователя на информацию.

4. Устойчивая производительность. С увеличением числа измерений и размеров базы данных аналитики не должны столкнуться с каким бы то ни было уменьшением производительности. Устойчивая производительность необходима для поддержания простоты использования и свободы от усложнений, которые требуются для доведения OLAP до конечного пользователя.

5. Клиент – серверная архитектура. Главная идея работы в среде клиент – сервер — это то, что серверный компонент инструмента OLAP должен быть достаточно интеллектуальным и обладать способностью строить общую концептуальную схему на основе обобщения и консолидации различных логических и физических схем корпоративных баз данных для обеспечения эффекта прозрачности.

6. Равноправие измерений. Все измерения данных должны быть равноправными. Дополнительные характеристики могут быть предоставлены отдельным измерениям. Но поскольку все они симметричны, данная дополнительная функциональность может быть предоставлена любому измерению. Базовая структура данных, формулы и форматы отчётов не должны опираться на какое-то одно измерение.

7. Динамическая обработка разреженных матриц. Инструмент OLAP должен обеспечивать оптимальную обработку разреженных матриц. Скорость доступа должна сохраняться вне зависимости от расположения ячеек данных и быть постоянной величиной для моделей, имеющих разное число измерений и различную разреженность данных.

8. Поддержка многопользовательского режима. Зачастую несколько аналитиков имеют необходимость работать одновременно с одной аналитической моделью или создавать различные модели на основе одних корпоративных данных. Инструмент OLAP должен предоставлять им конкурентный доступ, обеспечивать целостность и защиту данных.

9. Неограниченная поддержка кроссмерных операций. Вычисления и манипуляция данными по любому числу измерений не должны запрещать или ограничивать любые отношения между ячейками данных. Преобразования, требующие произвольного определения, должны задаваться на функционально полном формульном языке.

10. Интуитивное манипулирование данными. Детализация данных в колонках и строках, агрегация и другие манипуляции, свойственные структуре иерархии, должны выполняться в максимально удобном, естественном и комфортном пользовательском интерфейсе.

11. Гибкий механизм генерации отчётов. Должны поддерживаться различные способы визуализации данных, то есть, отчёты должны представляться в любой возможности ориентации.

12. Неограниченное количество измерений и уровней агрегации. Рекомендуется допущение в каждом серьёзном OLAP инструменте как минимум пятнадцати измерений в аналитической модели. Более того, каждое из этих измерений должно допускать практически неограниченное количество определённых пользователем уровней агрегации.

Набор этих требований, послуживших фактическим определением OLAP, следует рассматривать как рекомендательный, а конкретные продукты оценивать по степени приближения к идеально полному соответствию всем требованиям.

Позднее все эти требования были переработаны в так называемый тест FASMI, который также определяет требования к продуктам OLAP. FASMI — это аббревиатура от названия пунктов теста:

- Fast (Быстрый). Приложение OLAP должно обеспечивать минимальное время доступа к аналитическим данным — в среднем порядка 5 секунд;
- Analysis (Анализ). Приложение OLAP должно давать пользователю возможность осуществлять числовой и статистический анализ;
- Shared (Разделяемый доступ). Приложение OLAP должно предоставлять возможность работы с информацией многим пользователям одновременно;
- Multidimensional (Многомерность). Приложение должно обеспечивать многомерное концептуальное представление данных, включая полную поддержку для иерархий.
- Information (Информация). Приложение OLAP должно давать пользователю возможность получать нужную информацию, в каком бы электронном хранилище данных она не находилась.

Данные могут храниться либо в реляционных, либо в многомерных структурах. В настоящее время применяются три способа хранения данных: MOLAP, ROLAP, HOLAP.

MOLAP (Multidimensional OLAP) —— исходные и агрегатные данные хранятся в многомерной базе данных. Хранение данных в многомерных структурах позволяет манипулировать данными как

многомерным массивом, благодаря чему скорость вычисления агрегатных значений одинакова для любого из измерений. Однако в этом случае многомерная база данных оказывается избыточной, так как многомерные данные полностью содержат исходные реляционные данные.

ROLAP (Relational OLAP) — исходные данные остаются в той же реляционной базе данных, где они изначально и находились. Агрегатные же данные помещают в специально созданные для их хранения служебные таблицы в той же базе данных.

HOLAP (Hybrid OLAP) — исходные данные остаются в той же реляционной базе данных, где они изначально находились, а агрегатные данные хранятся в многомерной базе данных.

Некоторые OLAP-средства поддерживают хранение данных только в реляционных структурах, некоторые — только в многомерных. Однако большинство современных серверных OLAP-средств поддерживают все способы хранения данных. Выбор способа хранения зависит от объема и структуры исходных данных, требований к скорости выполнения запросов и частоты обновления OLAP-кубов.

3. КОМПОНЕНТЫ OLAP

Таблица фактов является основной таблицей хранилища данных. Как правило, она содержит сведения об объектах или событиях, совокупность которых будет в дальнейшем анализироваться. Обычно говорят о четырех наиболее часто встречающихся типах фактов. К ним относятся:

– факты, связанные с транзакциями (Transaction facts). Они основаны на отдельных событиях (типичными примерамы - телефонный звонок или снятие денег со счета с помощью банкомата);

– факты, связанные с «моментальными снимками» (Snapshot facts). Основаны на состоянии объекта (например, банковского счета) в определенные моменты времени, например на конец дня или месяца. Примеры таких фактов - объем продаж за день или дневная выручка;

– факты, связанные с элементами документа (Line-item facts). Основаны на том или ином документе (например, счете за товар или услуги) и содержат подробную информацию об элементах этого документа (например, количестве, цене, проценте скидки);

– факты, связанные с событиями или состоянием объекта (Event or state facts). Представляют возникновение события без подробностей о нем (например, просто факт продажи или факт отсутствия таковой без иных подробностей).

Таблицы измерений содержат неизменяемые либо редко изменяемые данные. В подавляющем большинстве случаев эти данные представляют собой по одной записи для каждого члена нижнего уровня иерархии в измерении. Таблицы измерений также содержат как минимум одно описательное поле и, как правило, целочисленное ключевое поле для однозначной идентификации члена измерения.

Если будущее измерение, основанное на данной таблице измерений, содержит иерархию, то таблица измерений также может содержать поля, указывающие на «родителя» данного члена в этой иерархии. Нередко таблица измерений может содержать и поля, указывающие на «прародителей», и иных «предков», а также дополнительные атрибуты членов измерений, содержавшиеся в исходной оперативной базе данных.

Каждая таблица измерений должна находиться в отношении «один ко многим» с таблицей фактов. Пример схемы базы данных с таким типом связей приведен на рисунке 1.

4. ЯЗЫК MDX

Назначение языка MDX (Multidimensional Expressions) — предоставить разработчикам средство для простого и эффективного доступа к многомерным структурам данных.

Простейший вид запроса MDX выглядит следующим образом:

```
SELECT <множество1> ON COLUMNS,
<множество2> ON ROWS
FROM <куб>
WHERE <кортеж>
```

Формат инструкций SELECT, FROM и WHERE подобен языку структурированных запросов SQL. Кроме того, эти инструкции служат для тех же целей, что и в SQL. Однако, язык MDX позволяет выполнять более сложные операции.

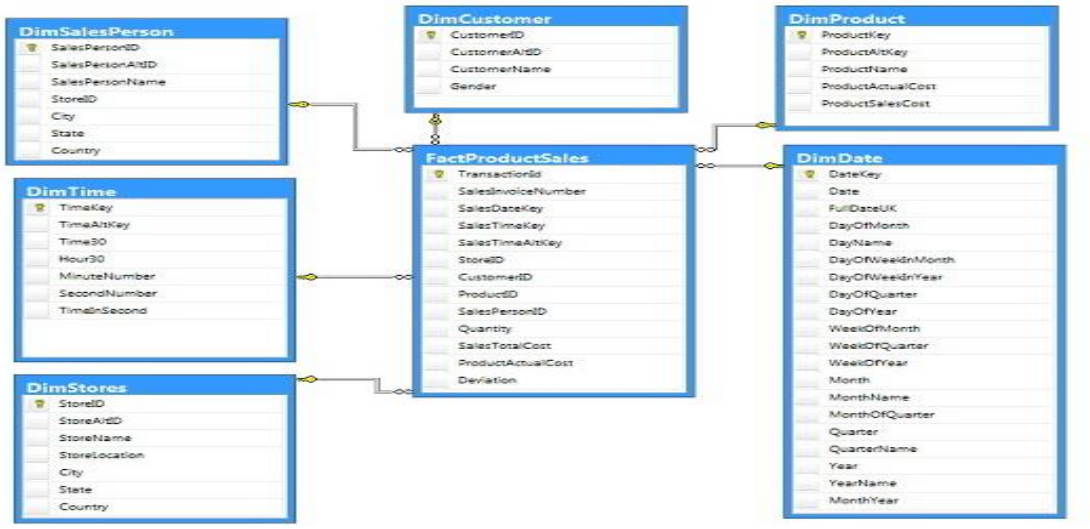

Рисунок 1 - Схемы базы данных

Запрос MDX возвращает единственное значение; кортеж, ссылающийся на него, использует для возвращения этого значения заданные по умолчанию члены измерений. Фактические данные содержатся в специальном измерении - Measures. Если приведенный выше запрос отправить к экземпляру Analysis Services, то в результате получим заданный по умолчанию член измерения Measures, который является одной из размерностей куба. Результатом данного запроса является значение, полученное путем агрегации значений всех относящихся к этой размерности ячеек куба для заданных по умолчанию значений каждого измерения куба.

Кортеж — это набор членов одного или нескольких разных измерений. Задавая кортеж, мы указываем на конкретную ячейку или набор ячеек внутри куба. Таким образом, кортеж – это декартово произведение множеств, определенных на различных измерениях куба.

Множество или набор — это совокупность (объединение) кортежей, определенных с использованием одинакового количества одних и тех же измерений.

Измерение среза (slicer dimension) создается при определении предложения WHERE; по сути, это фильтр, который исключает нежелательные измерения и члены.

Если в оси среза определено несколько кортежей, они будут обработаны как набор, а их значения — агрегированы с размерностью из запроса и функцией агрегации из этой размерности.

ЗАКЛЮЧЕНИЕ

В настоящее время аналитическая обработка информации привлекает возрастающее внимание в мире. Многомерный анализ данных превращает "сырые" данные в информацию и знание для конечных пользователей. Модули аналитической обработки присутствуют в составе большинства финансово-производственных приложений. Это обусловлено тем фактом, что качество информационной поддержки деятельности аналитиков и руководителей является одним из решающих факторов успеха предприятий.

Основными особенностями систем для многомерного анализа данных, накопленных в хранилище, являются выделение из большого объёма исторических данных содержательной информации (знаний) с использованием средств обработки информации на основе методов искусственного интеллекта, использование мощной вычислительной техники и специального хранилища данных, которое накапливает информацию из различных источников за длительный период времени, а также обеспечение оперативного доступа к данным.

ЛИТЕРАТУРА

1. Заботнев М.С. Методы представления информации в разреженных гиперкубах данных [Электронный ресурс]. — Режим доступа: http://www.olap.ru/basic/theory.asp
2. Барсегян А.А., Куприянов М.С., Степаненко В.В., Холод И.И. Методы и модели анализа данных: OLAP и Data Mining. - БХВ-Петербург, 2004.
3. Архипенков С. Я., Голубев Д. В., Максименко О. Б. Хранилища данных. - Диалог-МИФИ, 2002.