OLAP. Процессы добычи данных.

Лекция №7 для студентов 4-го курса специальности «Прикладная информатика»

Вопросы

1) Архитектуры OLAP-серверов
2) Процессы добычи данных
3) Дополнительные вопросы OLAP и добычи данных

1 Архитектуры OLAP-серверов

Традиционные реляционные серверы не обеспечивают эффективное выполнение сложных OLAP-запросов и поддержку многомерных представлений данных. Но, тем не менее, три типа реляционных серверов баз данных:

- реляционной,
- многомерной и
- гибридной оперативной аналитической обработки

позволяют выполнять OLAP-операции в хранилищах данных, построенных с использованием систем управления реляционными базами данных.

1.1 ROLAP

Размещаются между основным реляционным сервером, где находится хранилище данных и клиентским инструментарием переднего плана.

Серверы ROLAP поддерживают многомерные OLAP-запросы и, как правило, оптимизированы для конкретных реляционных серверов. Они указывают, какие представления должны быть материализованы, возможные запросы пользователей в терминах соответствующих материализованных представлений, и генерируют сложные SQL-серверы для основного сервера.

Они также предусматривают дополнительные службы, такие как планирование запросов и распределение ресурсов. Серверы ROLAP наследуют возможности масштабирования и работы с транзакциями реляционных систем, однако существенные различия между запросами в стиле OLAP и SQL могут стать причиной низкой производительности.

Нехватка производительности становится менее острой, благодаря ориентированным на задачи OLAP расширениям SQL, реализованным в серверах реляционных баз данных наподобие Oracle, IBM DB2 и Microsoft SQL Server. Такие функции, как median, mode, rank, percentile дополняют агрегатные фун-кции. К другим дополнительным возможностям относятся аг-регатные вычисления на перемещающихся окнах, текущие сводные значения и точки прерывания для улучшенной под-держки формирования отчетов.

Многомерные электронные таблицы требуют группировки по различным наборам атрибутов. Для того чтобы удовлетворить эти требования Джим Грей и его коллеги предлагают расширить SQL двумя операторами — roll-up и cube. Свертка списка атрибутов, включающего продукт, год и город, помогает находить ответы на вопросы, в которых фигурируют:

- группировка по продуктам, годам и городам;
- группировка по продуктам и годам;
- группировка по продуктам.

1.2 MOLAP

Серверная архитектура напрямую поддерживает многомерные представления данных с помощью многомерного механизма хранения. MOLAP позволяет реализовывать многомерные запросы на уровне хранения путем установки прямого соответствия.

Основное **преимущество** заключается в превосходных свойствах индексации; ее **недостаток** – низкий коэффициент использования дискового пространства, особенно в случае разреженных данных.

Многие серверы MOLAP при работе с разреженными множествами данных используют двухуровневую организацию памяти и сжатие. При двухуровневой организации пользователь либо непосредственно, либо с помощью специальных инструментов проектирования, идентифицирует набор подмассивов. Индексировать эти массивы меньшего размера можно с помощью традиционных индексных структур. Многие из методик, разработанных для статистических баз данных, подходят и для MOLAP. Серверы MOLAP обладают хорошей производительностью и функциональностью, но не в состоянии должным образом масштабироваться в случае очень больших баз данных.

1.3 HOLAP

Гибридная архитектура, которая объединяет технологии ROLAP и MOLAP. В отличие от MOLAP, которая работает лучше, когда данные более менее плотные, серверы ROLAP лучше в тех случаях, когда данные довольно разрежены.

Серверы HOLAP применяют подход ROLAP для разреженных областей многомерного пространства и подход MOLAP – для плотных областей. Серверы HOLAP разделяют запрос на несколько подзапросов, направляют их к соответствующим фрагментам данных, комбинируют результаты, а затем предоставляют результат пользователю.

Материализация выборочных представлений в HOLAP, выборочное построение индексов, а также планирование запросов и ресурсов аналогично тому, как это реализовано в серверах MOLAP и ROLAP.

2 Процессы добычи данных

Обнаружение знаний (knowledge discovery) – процесс определения и достижения цели посредством итеративной добычи данных.

Состоит из трех этапов:

- подготовка данных;
- построение модели и ее оценка;
- применение модели.

2.1 Подготовка данных

На этапе подготовки данных аналитик готовит набор данных, содержащий достаточно информации, для того чтобы создать точные модели на последующих этапах. В случае с FSC, точная модель должна помочь прогнозировать, с какой вероятностью клиент купит продукты, рекламируемые в новом каталоге.

Как правило, добыча данных включает в себя итеративно создаваемые модели на основе подготовленного множества данных, а затем применение одной или нескольких моделей. Поскольку создание моделей на больших множествах данных может оказаться весьма дорогостоящим, аналитики часто сначала работают с несколькими выборками множества данных. Платформы добычи данных, таким образом, должны поддерживать вычисления на случайно выбранных экземплярах данных в сложных запросах.

2.2 Построение и оценка моделей

Только после того, как принято решение о том, какую модель применять, аналитик создает модель на всем подготовленном множестве данных.

Цель этого этапа **создания модели** – указать шаблоны, которые определяют целевой атрибут (target attribute). Пример целевого атрибута во множестве данных FSC: приобрел ли клиент хотя бы один продукт из предыдущего каталога?.

Предсказать как точно указанные, так и скрытые атрибуты помогают несколько классов моделей добычи данных.

На выбор модели влияют два важных фактора:

- точность модели,
- эффективность алгоритма для создания модели на больших множествах данных.

Многие коммерческие продукты создают модели для конкретных областей применения, но реальная база данных, на которой должна применяться такая модель, возможно, будет работать с другим сервером баз данных. Платформы добычи данных и серверы баз данных, таким образом, должны поддерживать взаимозаменяемость моделей.

Недавно рабочая группа Data Mining Group предложила воспользоваться Predictive Model Markup Language, стандартом на базе XML, для обмена рядом популярных классов моделей прогнозирования. Идея состоит в том, чтобы любая база данных, поддерживающая этот язык, могла импортировать и применять любую описанную на нем модель.

2.3 Применение модели

На этом этапе аналитики применяют выбранную модель к наборам данных, чтобы прогнозировать целевой атрибут с неизвестным значением.

Для каждого текущего набора клиентов в примере FSC, прогноз касается того, будут ли они приобретать продукты из нового каталога. Применение модели на входном наборе данных может породить другой набор данных. В примере FSC этап применения модели указывает подмножество клиентов, которым будет разослан каталог.

Когда входной набор данных очень большой, стратегия применения модели должна б ыть достаточно эффективной. В этом случае может потребоваться использование индексов на входной таблице для фильтрации кортежей, которые не будут входить в развертываемый результат, но это требует более тесной интеграции между системами управления ба-зами данных и применением модели.

3 Дополнительные вопросы OLAP и добычи данных

К другим важным вопросам в контексте OLAP и технологии добычи данных относятся

- пакетные приложения,
- платформы и их API-интерфейсы, влияние XML,
- приближенная обработка запросов,
- интеграция OLAP и добычи данных,
- добыча данных в Web.

3.1 Пакетные приложения

Пакетные приложения и средства формирования отчетов могут использовать знания о конкретной вертикальной отрасли для упрощения задачи анализа путем учета специфических для отрасли абстракций более высокого уровня. Data Warehousing Information Center и KDnuggets предлагают обширный список решений, ориентированных на конкретные отрасли.

Компании могут приобрести такие пакеты, а не разрабатывать свое собственное аналитическое решение, но пакеты, ориентированные на конкретную область применения, меняющиеся по мере развития бизнеса, ограничены по набору своих функций и потому не могут удовлетворить все потенциальные требования к анализу.

3.2 API-интерфейсы и влияние XML

Некоторые платформы OLAP и добычи данных предлагают API - интерфейсы, которые позволяют аналитикам создавать собственные решения. Однако поставщики решений, как правило, вынуждены писать специальные программы для различных платформ, чтобы предоставить не зависящее от платформ решение.

Новые ориентированные на XML службы на базе Web обеспечивают общий интерфейс для механизмов OLAP. Компании Microsoft и Hyperion опубликовали XML for Analysis, API-интерфейс, основанный на протоколе SOAP, предназначенный специально для стандартизации взаимодействий при доступе к данным между клиентским приложением и источником данных, работающими через Web. На основе этой XML-спецификации поставщики решений смогут писать программы с помощью одного API-интерфейса, а не использовать множество интерфейсов, ориентированных на решения разных производителей.

3.3 Приближенная обработка запросов

Обработка сложных агрегатных запросов, как правило, требует обращения к огромным объемам данных. Например, вычисление среднего объема продаж FSC в различных городах требует сканирования всех данных в хранилище. Во многих случаях достаточно точную оценку позволяет получить приближенная обработка запросов.

Идея состоит в том, чтобы на основе базовых данных максимально точно сформировать сводные данные, а затем получать ответы на агрегатные запросы с помощью этих сводных, а не полных данных. Дополнительную информацию по этому вопросу можно найти в описании проектов Approximate Query Processing и AQUA Project.

3.4 Интеграция OLAP и добычи данных

OLAP-инструментарий помогает аналитикам выявить актуальные порции данных, а модели добычи данных обогащают эту функциональность. Например, если темпы роста объема продаж FSC не соответствуют прогнозируемым, специалисты по маркетингу хотели бы знать аномальные регионы и категории продуктов, для которых не выполняются заданные показатели.

Пробный анализ, который выявляет аномалии, использует методику, позволяющую отметить агрегатный параметр на более высоком уровне в иерархии измерений с аномальным результатом. Аномальный результат определяет общее отклонение реальных агрегатных величин от соответствующих прогнозируемых значений над всеми своими потомками. Для вычисления прогнозируемых значений аналитики могут использовать такие средства добычи данных, как регрессионные модели.

3.5 Добыча данных в Web

Большинство крупных компаний поддерживают Web-сайты, где клиенты могут просмотреть информацию, запросить данные о товарах и приобрести их.

Поскольку каждый клиент имеет личный контакт с компанией через Web-сайт, компании могут персонифицировать работу с ним. Например, сайт может рекомендовать клиенту продукты, услуги или статьи, относящиеся к области его интересов.

При создании таких Web-систем возникают два важных вопроса:

- сбор данных,
- методы персонификации.