Технология баз данных в системах поддержки принятия решений

Лекция №6 для студентов 4-го курса специальности «Прикладная информатика»

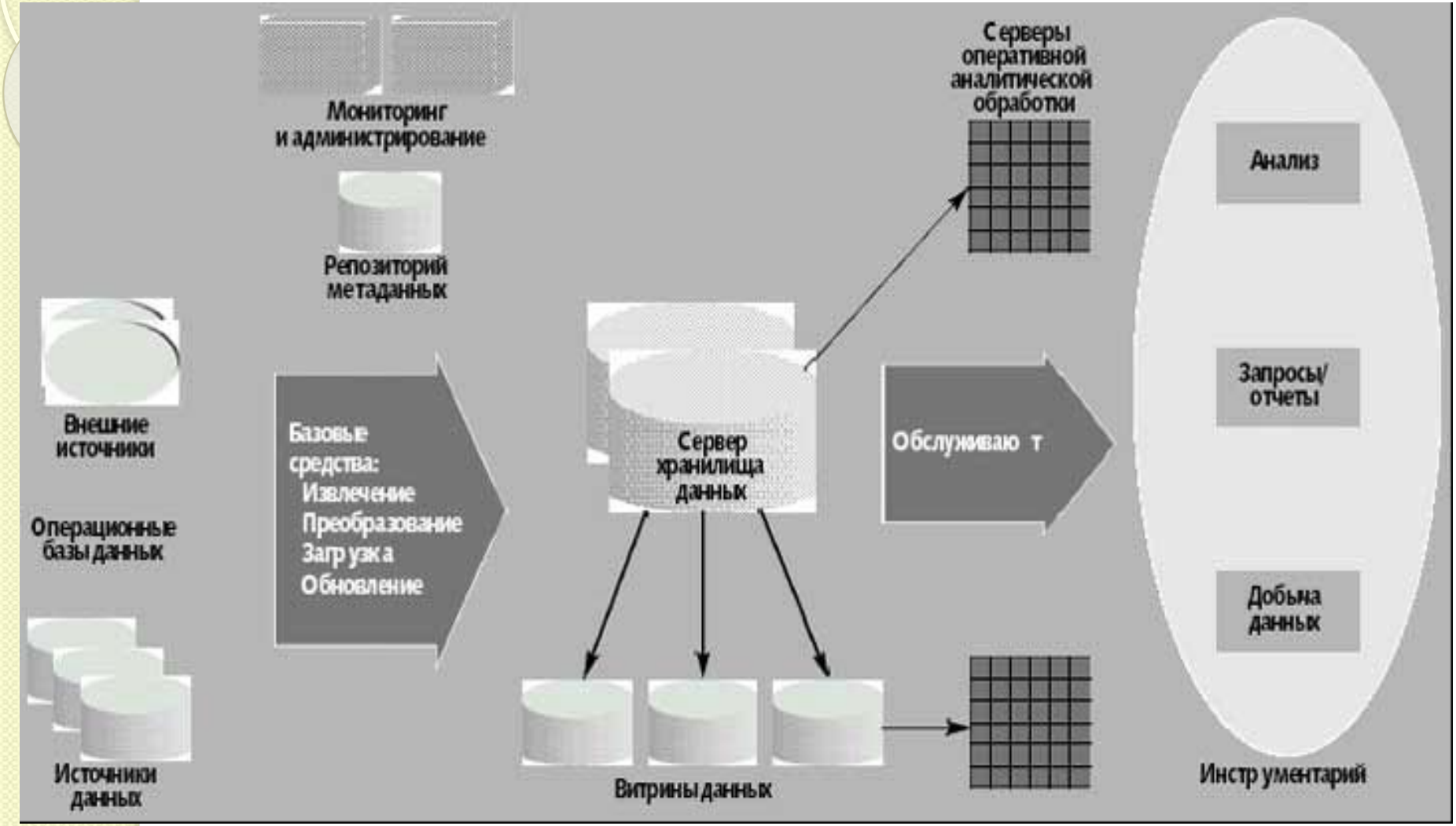
Вопросы

1) Компоненты систем поддержки принятия решений.
2) Структура хранилища данных.
3) Материализованные представления.
4) Оперативные аналитические приложения.

Системы поддержки принятия решений – основа ИТ-инфраструктуры различных компаний, поскольку эти системы дают возможность преобразовывать обширную бизнес-информацию в ясные и полезные выводы.

Сбор, обслуживание и анализ больших объемов данных, – это гигантские задачи, которые требуют преодоления серьезных технических трудностей, огромных затрат и адекватных организационных решений.

Системы оперативной обработки транзакций (online trans-action processing – OLTP) позволяют накапливать большие объемы данных, ежедневно поступающих из пунктов про-даж. Приложения OLTP, как правило, автоматизируют струк-турированные, повторяющиеся задачи обработки данных, такие как ввод заказов и банковские транзакции. Эти под-робные, актуальные данные из различных независимых то-чек ввода объединяются в одном месте, и затем аналитики смогут извлечь из них значимую информацию. Агрегирован-ные данные применяются для принятия каждодневных биз-нес-решений – от управления складом до координации рекламных рассылок.


Мониторинг
и администрирование
Репозиторий
метаданных
Внешние
источники
Операционные
базы данных
Источники
данных
Базовые
средства:
Извлечение
Преобразование
Загрузка
Обновление
Сервер
хранилища
данных
Витрины данных
Серверы
оперативной
аналитической
обработки
Обслуживают
Анализ
Запросы/
отчеты
Добыча
данных
Инструментарий

1 Компоненты систем поддержки принятия решений

Система поддержки принятия решений – сложная структура с многочисленными компонентами. Рассмотрим гипотетическую компанию Footwear Sellers Company, которая производит обувь и предлагает ее покупателям по каналам прямых продаж и через реселлеров.

Руководителям отдела маркетинга FSC необходимо извлечь следующую информацию из агрегированных бизнес-данных:

- пять штатов, сообщивших о самых больших за последний год тем-пах роста объема продаж в категории продуктов для молодежи;
- общий объем продаж обуви в Нью-Йорке за последний месяц по различным видам продуктов;
- 50 городов с самым большим количеством индивидуальных клиентов;
- один миллион клиентов, которые, скорее всего, приобретут новую модель обуви Walk-on-Air.

Прежде чем создавать систему, которая предоставит такую информацию, в FSC должны рассмотреть и решить три основных вопроса:

- какие данные накапливать и как на концептуальном уровне моделировать данные и управлять их хранением;
- как анализировать данные;
- как эффективно загрузить данные из нескольких независимых источников.

1.1 Хранилища данных

Хранилища данных содержат информацию, собранную из нескольких оперативных баз данных.

Хранилища, как правило, на порядок больше оперативных баз, зачастую имея объем от сотен гигабайт до нескольких терабайт. Как правило, хранилище данных поддерживается независимо от оперативных баз данных организации, пос-кольку требования к функциональности и производитель-ности аналитических приложений отличаются от требова-ний к транзакционным системам.

Хранилища данных создаются специально для приложений поддержки принятия решений и предоставляют накоплен-ные за определенное время, сводные и консолидирован-ные данные, которые более приемлемы для анализа, чем детальные индивидуальные записи. Рабочая нагрузка сос-тоит из нестандартных, сложных запросов, которые обра-щаются к миллионам записей и выполняют огромное коли-чество операций сканирования, соединения и агрегирова-ния. Время ответа на запрос в данном случае важнее, чем пропускная способность.

Поскольку конструирование хранилища данных – сложный процесс, который может занять несколько лет, некоторые организации вместо этого строят витрины данных (data mart), содержащие информацию для конкретных подразделений.

Например, маркетинговая витрина данных может содержать только информацию о клиентах, продуктах и продажах и не включать в себя планы поставок. Несколько витрин данных для подразделений могут сосуществовать с основным хранилищем данных, давая частичное представление о содержании хранилища.

Витрины данных строятся значительно быстрее, чем хранилище, но впоследствии могут возникнуть серьезные проблемы с интеграцией, если первоначальное планирование проводилось без учета полной бизнес-модели.

2 Структура хранилища данных

Большинство хранилищ используют технологию реляционных баз данных, поскольку она предлагает надежные, проверенные и эффективные средства хранения и управления большими объемами данных.

Важнейший вопрос, связанный с конструированием хранилищ данных, – архитектура базы данных, как логическая, так и физическая. Создание логической схемы корпоративного хранилища данных требует всеобъемлющего моделирования бизнеса.

2.1 Логическая архитектура базы данных

В архитектуре, основанной на схеме «звезда», база данных состоит из таблицы фактов, которая описывает все транзакции, и таблицы измерений для каждой из сущностей.

В примере с FSC каждая транзакция охватывает несколько сущностей – клиент, продавец, продукт, заказ, дата сделки и город, где сделка состоялась. Каждая сделка также имеет параметры – в нашем случае число проданных экземпляров продукта и общая сумма, которую заплатил покупатель.

Схема «снежинка» – усовершенствованная схема «звезда», в которой иерархия измерений представляется точным образом благодаря нормализации таблиц измерений. В «звезде» набор атрибутов описывает каждое измерение и может быть связан иерархией отношений.

2.2 Физическая архитектура базы данных

Системы баз данных используют избыточные структуры, такие как индексы и материализованные представления для эффективной обработки сложных запросов. Определение самого подходящего набора индексов и представлений – это сложная задача формирования физической архитектуры.

Хотя поиск в индексе и сканирование индекса могут быть эффективны для запросов, связанных с выбором данных, запросы, предполагающие интенсивную обработку данных, могут потребовать последовательного сканирования всей реляционной таблицы или ее вертикальных фрагментов.

Увеличение эффективности сканирования таблиц и использование распараллеливания для уменьшения времени ответа на запрос – важные моменты, которые следует учитывать при проектировании физической архитектуры.

2.3 Индексные структуры

Методы обработки запросов, которые используют операции пересечения и объединения индексов, полезны при ответе на запросы с множественными предикатами.

Пересечение индексов используется при выборке по нескольким условиям и может значительно снизить необходимость (или вообще устранить ее) в доступе к базовым таблицам, если все столбцы проекции можно получить посредством сканирования индексов.

Благодаря особой природе «звезды» соединение индексов детальных данных особенно удобно для систем поддержки принятия решений. Хотя индексы традиционно устанавливают соответствие значения в столбце списку строк с этим значением, соединенный индекс поддерживает связь между внешним ключом и соответствующими ему первичными ключами. В контексте схемы «звезда» соединенный индекс может связать значения одного или нескольких столбцов таблицы измерений с соответствующими строками таблицы фактов.

3 Материализованные представления

Многие хранилища данных используют запросы, которые требуют сводных данных и потому работают с агрегатами. Материализация сводных данных (т.е. их вычисление и сохранение) может ускорить обработку многих распространенных запросов.

Задачи, которые возникают при использовании материализованных представлений, аналогичны тем, которые возникают при работе с индексами:

- определение представлений, которые следует материализовывать;
- использование материализованных представлений для ответа на запросы;
- обновление материализованных представлений при загрузке новых данных.

4 Оперативные аналитические приложения

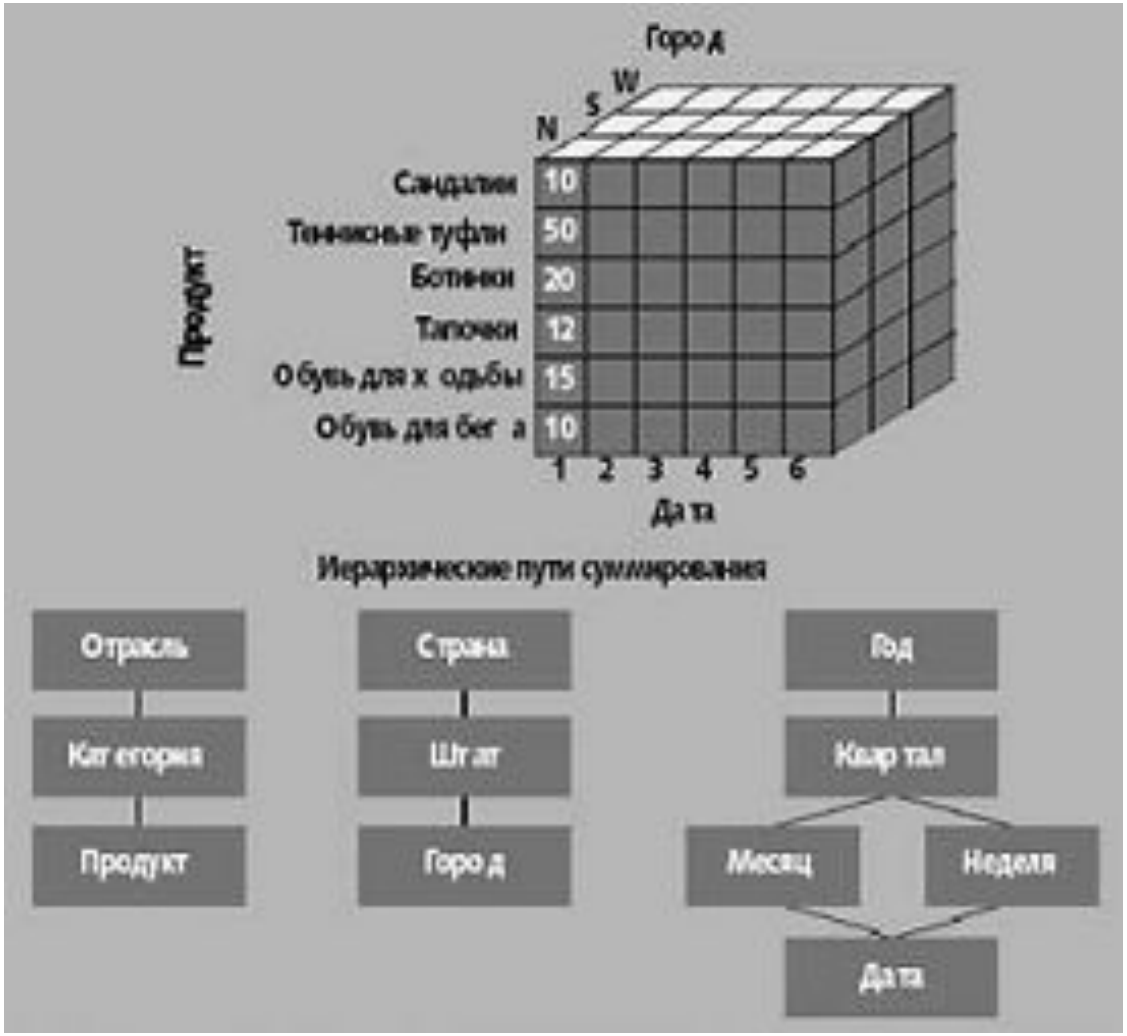
В типичном оперативном аналитическом приложении запрос агрегирует численные параметры более высоких уровней в иерархию измерений.

Пример – первый маркетинговый запрос FSC, для выполнения которого необходим набор агрегированных параметров – пять штатов, сообщивших о самом большом увеличении объема продаж в категории молодежных продуктов за последний год. «Штат» и «Год» – обобщения сущностей «Город» и «Дата».

Ключевые вопросы, касающиеся OLAP, связаны с концептуальной моделью данных и серверными архитектурами.

4.1 Концептуальная модель данных

Многомерная модель, изображенная на рисунке, использует численные параметры как объекты своего анализа.



Каждый численный параметр в концептуальной модели данных зависит от измерений, которые описывают сущности в транзакции. Например, измерения, связанные с продажами в примере FSC, – это клиент, продавец, город, название продукта и дата совершения сделки. Все вместе измерения уникальным образом определяют параметр, поэтому многомерная модель данных трактует параметр как значение в многомерном пространстве.

В многомерном представлении данных запросы drill-down и roll-up – это логические операции на кубе. Еще одна популярная операция – сравнить два параметра, которые агрегированы по одним и тем же измерениям, такими как продажи и бюджет.

OLAP-анализ может включать в себя более сложные статистические вычисления, нежели простые агрегаты, такие как сумма или среднее. Примером может служить такая функция, как изменение процента агрегата в определенный период по сравнению с различными периодами времени. Подобные дополнительные функции поддерживают многие коммерческие средства OLAP.

Измерение «Время» имеет особое значение для таких процессов поддержки принятия решений, как анализ тенденций. К примеру, аналитикам FSC может понадобиться проследить покупательскую активность в отношении спортивной обуви перед крупнейшими национальными легкоатлетическими соревнованиями или после них.

Развернутый анализ тенденций возможен, если база данных поддерживает встроенную информацию о календаре и ряд других характеристик «Времени». OLAP Council определил перечень операций для многомерных кубов.