Учебный курс

Хранилища данных

Лекция 2

Технологии хранения данных

Лекции читает

Кандидат технических наук, доцент

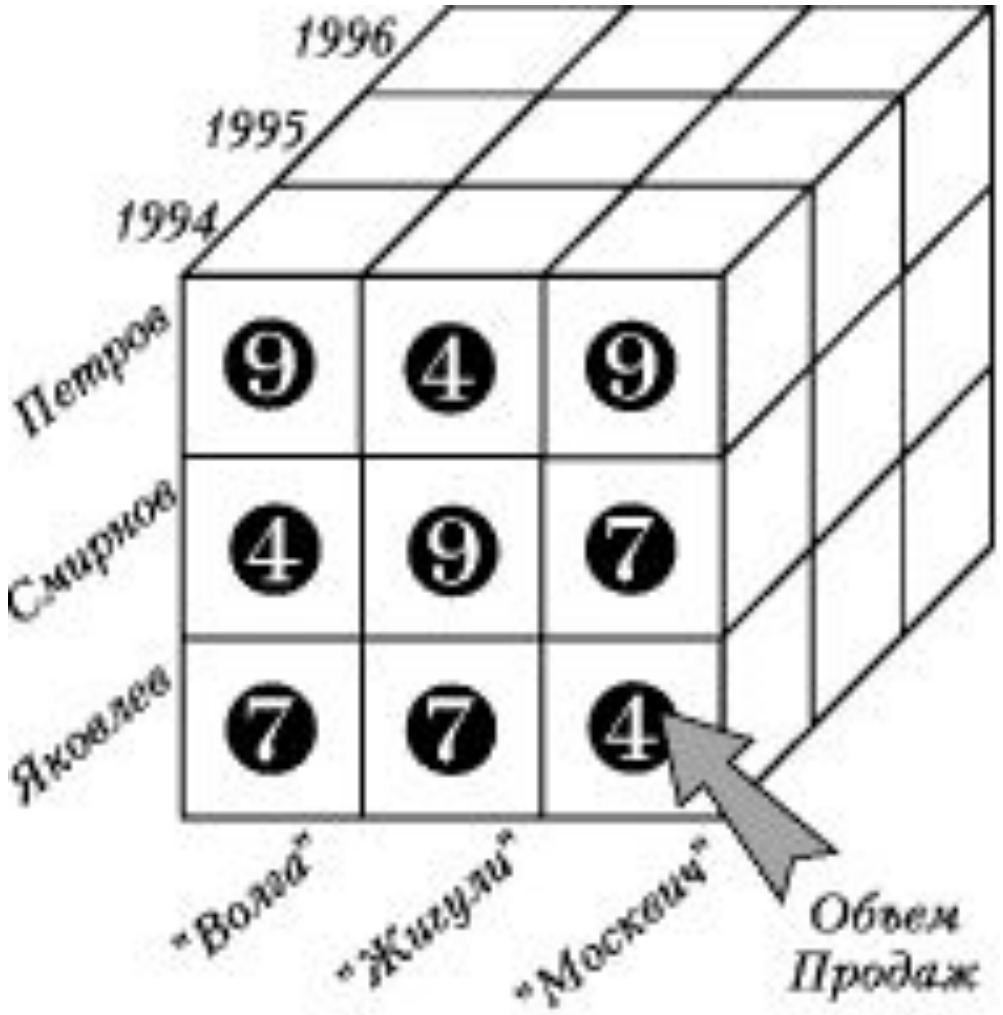
Перминов Геннадий Иванович

2. Кубы данных (многомерная модель данных)

Понятие о кубах

- Куб OLAP - это структура, в которой хранятся совокупности данных, полученные из базы данных OLAP путем всех возможных сочетаний измерений с фактами продаж в таблице фактов.
- Исходя из этого, создание окончательного отчета выполняется гораздо эффективнее, поскольку не требует выполнения никакого сложного запроса.

Вид трехмерного куба



Основными понятиями многомерной модели данных являются:

- **Показатель** - это величина (обычно числового типа), которая собственно и является предметом анализа. Один OLAP-куб может обладать одним или несколькими показателями.
- **Измерение** (dimension) - это множество объектов одного или нескольких типов, организованных в виде иерархической структуры и обеспечивающих информационный контекст числового показателя. Измерение принято визуализировать в виде ребра многомерного куба.
- Объекты, совокупность которых и образует измерение, называются **членами измерений** (members). Члены измерений визуализируют как точки или участи, откладываемые на осях гиперкуба. Например, временное измерение: Дни, Месяцы, Кварталы, Годы - наиболее часто используемые в анализе, могут содержать следующие члены: 8 мая 2002 года, май 2002 года, 2-ой квартал 2002 года и 2002 год.
  Как уже было сказано, объекты в измерениях могут быть различного типа, например "производители" - "марки автомобиля" или "годы" - "кварталы". Эти объекты должны быть организованы в иерархическую структуру так, чтобы объекты одного типа принадлежали только одному уровню иерархии.
- **Ячейка** (cell) - атомарная структура куба, соответствующая конкретному значению некоторого показателя. Ячейки при визуализации располагаются внутри куба и здесь же принято отображать соответствующее значение показателя.

Роль измерений в кубе

- Измерения играют роль индексов, используемых для идентификации значений показателей, находящихся в ячейках гиперкуба. Комбинация членов различных измерений играют роль координат, которые определяют значение определенного показателя. Поскольку для куба может быть определено несколько показателей, то комбинация членов всех измерения будет определять несколько ячеек со значениями каждого из показателей. Поэтому для однозначной идентификации ячейки необходимо указать комбинацию членов всех измерений и показатель.

Иерархии в измерениях необходимы для возможности агрегации и детализации значений показателей

Существуют следующие типы иерархий:

- **сбалансированные** (balanced);
- **несбалансированные** (unbalanced);
- **Неровные (balanced).**

Сбалансированные иерархии

- Это - иерархии, в которых число уровней определено её структурой и неизменно, и каждая ветвь иерархического дерева содержит объекты каждого из уровней. Каждому производителю автомобилей может соответствовать несколько марок автомобилей, а каждой марке - несколько моделей автомобилей, поэтому можно говорить о трёхуровневой иерархии этих объектов. В этом случае на первом уровне иерархии располагаются производители, на втором - марки, а на третьем - модели.
- Как видно, для формирования сбалансированной иерархии необходимо наличие связи "один-ко-многим" между объектами менее детального уровня по отношению к объектам более детального уровня. В принципе каждый уровень сбалансированной иерархии можно представить как отдельное простое измерение, но тогда эти измерения окажутся зависимыми, в значит неизбежно повышение разреженности куба.

Несбалансированные иерархии

- Это - иерархии, в которых число уровней может быть изменено, и каждая ветвь иерархического дерева может содержать объекты, принадлежащие не всем уровням, только нескольким первым.
- Необходимо заметить, что все объекты несбалансированной иерархии принадлежат одному типу.
- Типичный пример несбалансированной иерархии - иерархия типа "начальник-подчиненный", где все объекты имеют один и тот же тип - "Сотрудник".

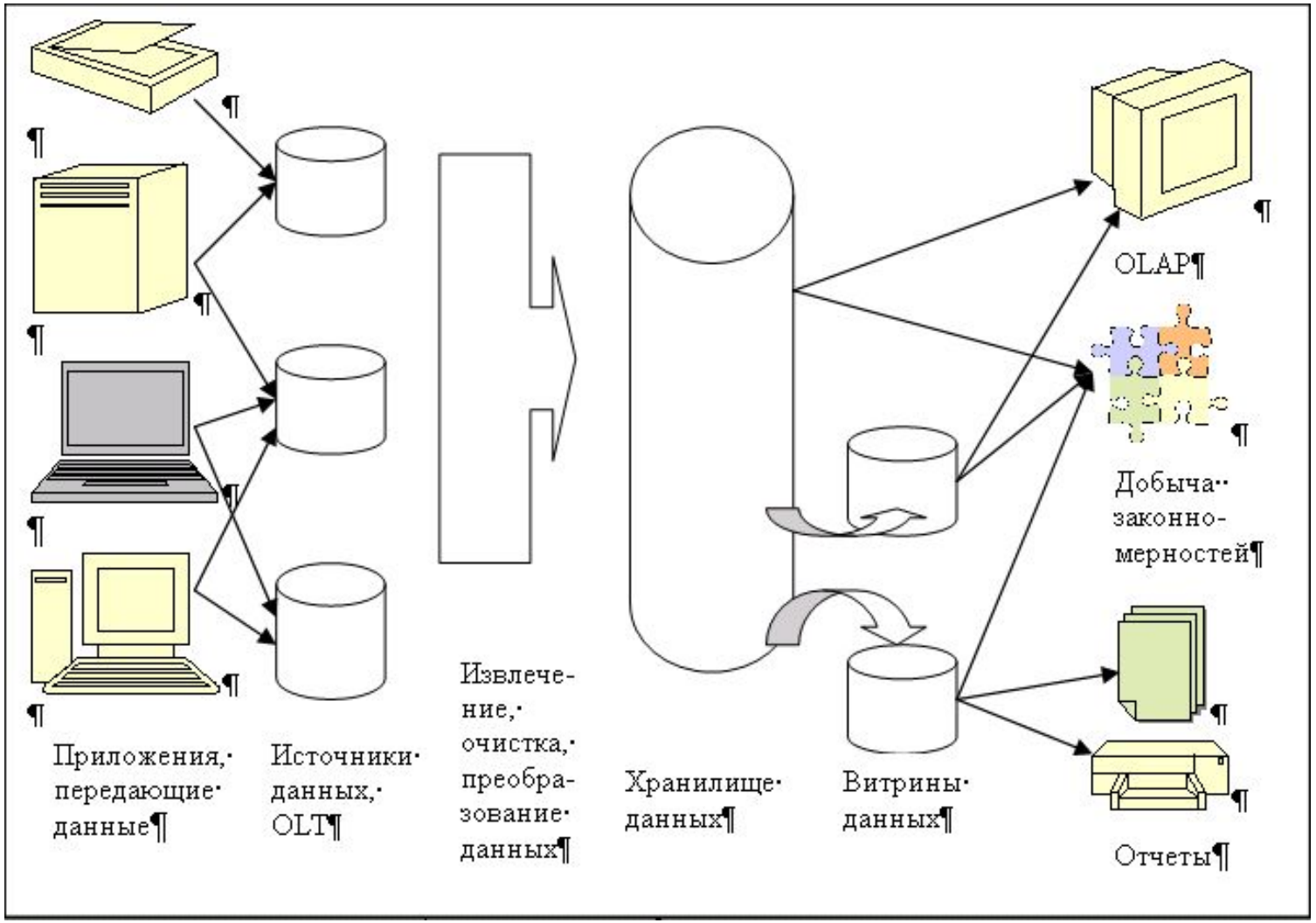
Неровные иерархии

- Это- иерархии, в которых число уровней определено её структурой и постоянно, однако в отличие от сбалансированной иерархии некоторые ветви иерархического дерева могут не содержать объекты какого-либо уровня.
- Иерархии такого вида содержат такие члены, логические "родители" которых не находятся на непосредственно вышестоящем уровне.
- Типичным примером является географическая иерархия, в которой есть уровни "Страны", "Штаты " и "Города", но при этом в наборе данных имеются страны, не имеющие штатов или регионов между уровнями "Страны" и "Города".

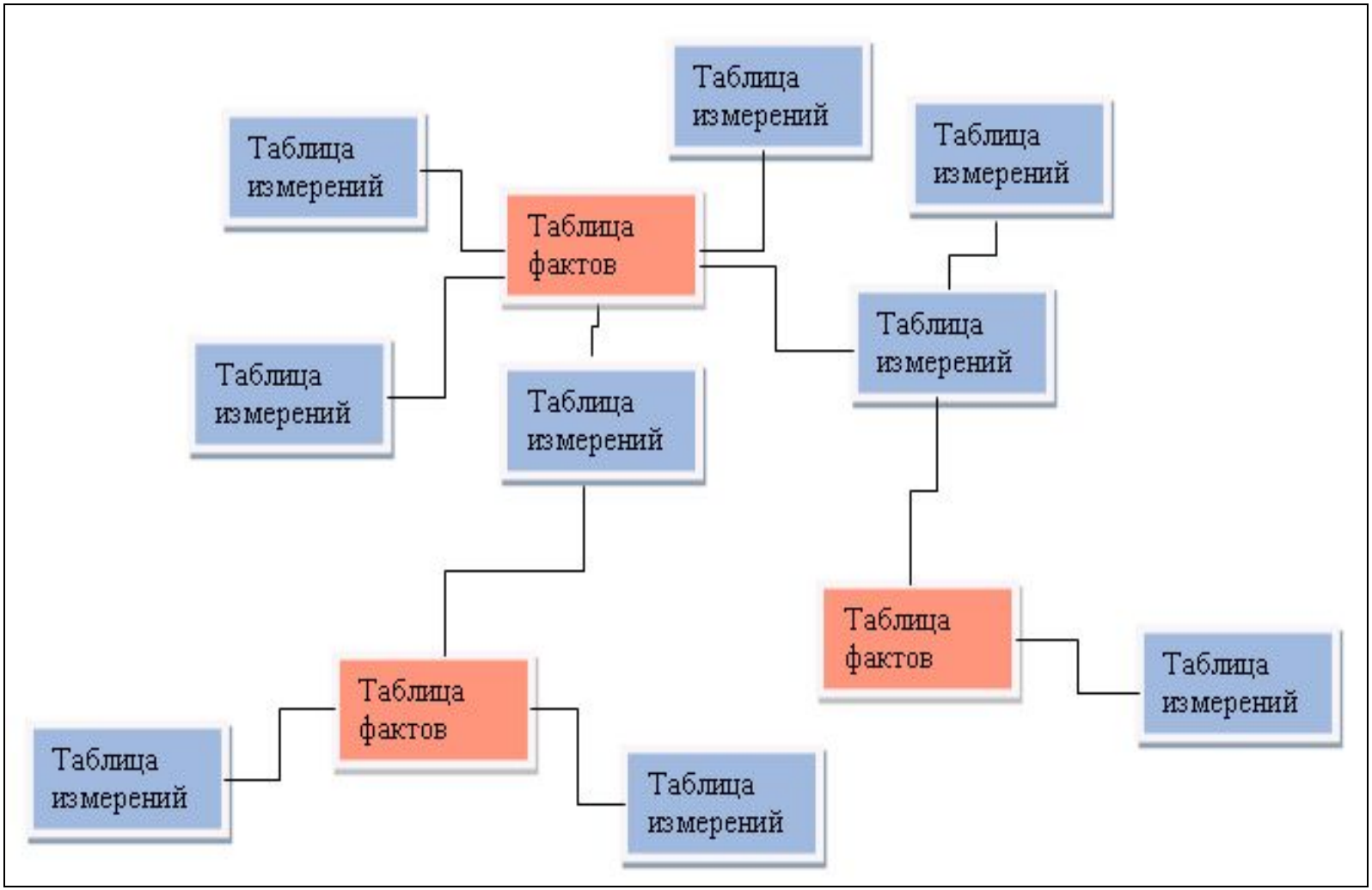
Агрегаты

- Агрегатами называют агрегированные по определенным условиям исходные значения показателей. Обычно под агрегацией понимается любая процедура формирования меньшего количества значений (агрегатов) на основании большего количества исходных значений. В дальнейшем под терминами агрегирование и агрегация будем понимать исключительно процесс суммирования данных.
- Заблаговременное формирование и сохранение агрегатов с целью уменьшения времени отклика на пользовательский запрос является основным свойством систем поддержки оперативного анализа.

DW с витринами данных



Многомерный куб с несколькими таблицами фактов



Варианты реализации хранилищ данных:

- Виртуальное хранилище данных
- Концепция CIF
- Концепция Data Warehouse Bus
- Гибридная многоуровневая архитектура хранилища данных

Виртуальное хранилище данных

- В данном случае в отличие от классического (физического) ХД данные из оперативных источников данных (ОИД) не копируются в единое хранилище.
- Они извлекаются, преобразуются и интегрируются непосредственно при выполнении аналитических запросов в оперативной памяти компьютера. Фактически такие запросы напрямую адресуются к ОИД

Основными достоинствами виртуального ХД являются:

- минимизация объема памяти, занимаемой на носителе информацией;
- работа с текущими, детализированными данными.

Недостатки технологии виртуального хранилища

- Время обработки запросов к виртуальному ХД значительно превышает соответствующие показатели для физического хранилища.
- Интегрированный взгляд на виртуальное хранилище возможен только при выполнении условия постоянной доступности всех ОИД. Таким образом, временная недоступность хотя бы одного из источников может привести либо к невыполнению аналитических запросов, либо к неверным результатам.
- Различные ОИД могут поддерживать разные форматы и кодировки данных. Часто на один и тот же вопрос может быть получено несколько вариантов ответа. Это может быть связано с несинхронностью моментов обновления данных в разных ОИД, отличиями в описании одинаковых объектов и событий предметной области, ошибками при вводе, утерей фрагментов архивов и т. д.
- Главным же недостатком виртуального хранилища следует признать практическую невозможность получения данных за долгий период времени. При отсутствии физического хранилища доступны только те данные, которые на момент запроса есть в ОИД. Основное назначение OLTP-систем — оперативная обработка текущих данных, поэтому они не ориентированы на хранение данных за длительный период времени. По мере устаревания данные выгружаются в архив и удаляются из оперативной БД.

Концепция Corporate Information Factory, (сокр. CIF) Билла Инмона

- Концепция CIF объединила оперативные приложения, накопители оперативных данных (Operational Data Store, ODS, OLTP-системы), центральное хранилище данных (DW), витрины данных (Data Mart) и системы интеллектуального анализа данных (Data Mining) в единый процесс выработки и потребления информации на предприятии.
- В CIF оперативные приложения служат для управления частными процессами. ODS накапливают в себе временные срезы различных процессов, происходящих на предприятии, и согласуют их между собой. ODS часто используется как оперативный источник информации. Как правило, ODS хранят значительно более детализированную информацию, чем хранилище, но за меньший период времени — от полугода до года, так как для доступа к данным в нем не используются предварительно рассчитываемые агрегаты.

Работа Хранилища CIF состоит из следующих этапов:

- скоординированное извлечение данных из источников.
- загрузка реляционной базы данных, состоящей из таблиц в третьей нормальной форме, содержащей атомарные данные.
- получившееся нормализованное Хранилище используется для того, чтобы наполнить информацией дополнительные репозитории презентационных данных, т.е. данных, подготовленных для анализа.
- Эти репозитории, в частности, включают специализированные Хранилища для изучения и "добычи" данных (Data Mining), а также витрины данных.

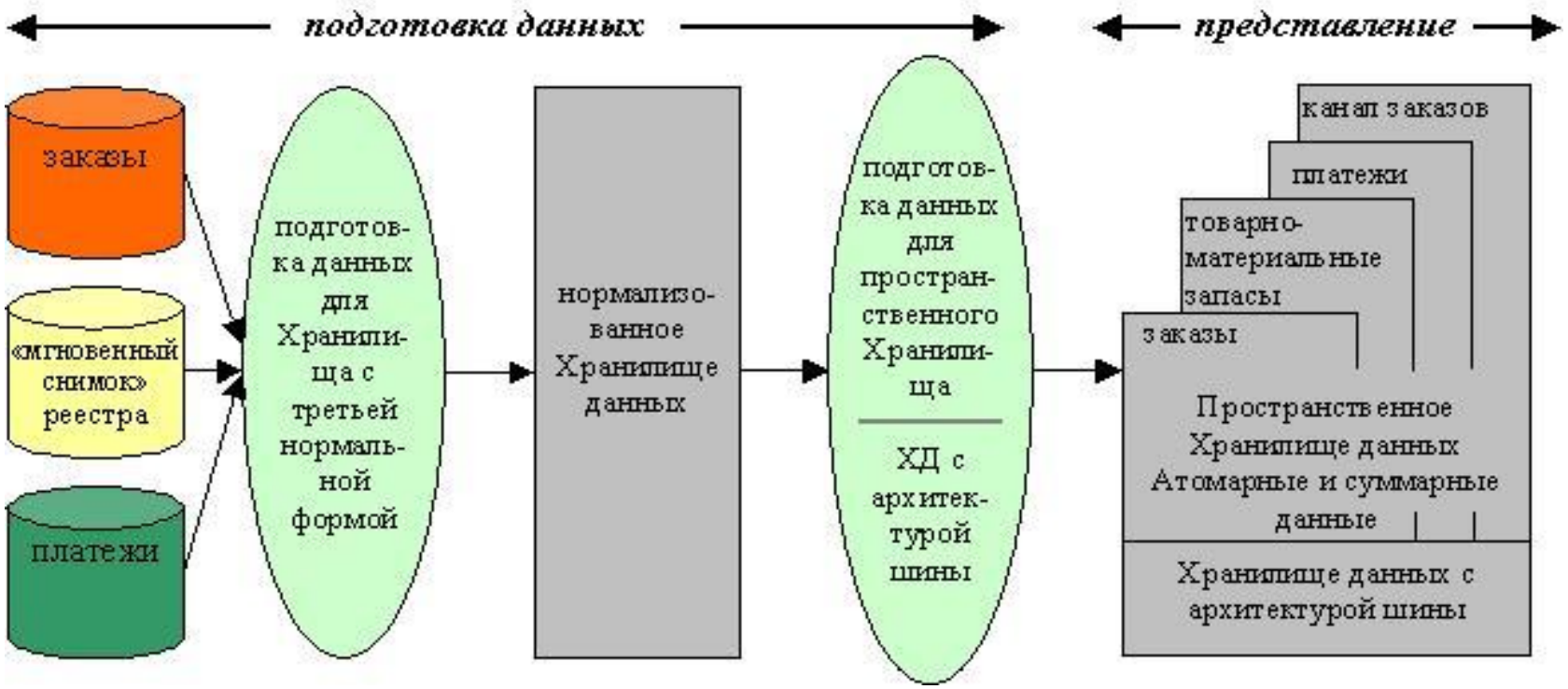
Концепция Data Warehouse Bus

- Использование пространственной модели организации данных с архитектурой "звезда" (star scheme).
- Использование двухуровневой архитектуры, которая включает стадию подготовки данных, недоступную для конечных пользователей, и Хранилище.
- В состав последнего входят несколько витрин атомарных данных, несколько витрин агрегированных данных и персональная витрина данных, но оно не содержит одного физически целостного или централизованного Хранилища данных.
- Хранилище Кимболла - скорее "виртуальный" объект. Это коллекция витрин данных, которые могут быть пространственно разобщенными.

Гибридное хранилище данных

- В последнее время все более популярной становится идея совместить концепции хранилища и витрины данных в одной реализации и использовать хранилище данных в качестве единственного источника интегрированных данных для всех витрин данных.
- Тогда естественной становится трехуровневая архитектура системы.

Гибрид нормализованного и пространственного Хранилищ данных



Первый уровень гибридного хранилища

- На **первом уровне** реализуется корпоративное хранилище данных на основе одной из развитых современных реляционных СУБД. Это хранилище интегрированных в основном детализированных данных. Реляционные СУБД обеспечивают эффективное хранение и управление данными очень большого объема, но не слишком хорошо соответствуют потребностям OLAP-систем, в частности, в связи с требованием многомерного представления данных.

Второй уровень гибридного хранилища

- На **втором уровне** поддерживаются витрины данных на основе многомерной системы управления базами данных (примером такой системы является Oracle Express Server). Такие СУБД почти идеально подходят для целей разработки OLAP-систем, но пока не позволяют хранить сверхбольшие объемы данных (предельный размер многомерной базы данных составляет 10-40 Гбайт). В данном случае это и не требуется, поскольку речь идет о витринах данных.
- Витрина данных не обязательно должна быть полностью сформирована. Она может содержать ссылки на хранилище данных и добирать оттуда информацию по мере поступления запросов. Конечно, это несколько увеличивает время отклика, но зато снимает проблему ограниченного объема многомерной базы данных.

Третий уровень гибридного хранилища

- На **третьем уровне** находятся клиентские рабочие места конечных пользователей, на которых устанавливаются средства оперативного анализа данных.

Форматы хранения данных в OLAP кубах

Данные форматы различаются методами хранения кубов данных

- многомерный OLAP-формат (Multi-dimensional OLAP - MOLAP);
- реляционный OLAP-формат (Relational OLAP - ROLAP);
- гибридный OLAP-формат (Hybrid OLAP - HOLAP).

MOLAP

- MOLAP является многомерным форматом хранения данных, который отличается высоким быстродействием. Помимо поддержки OLAP самих кубов данных при выборе данного формата данные будут храниться в многомерных структурах на OLAP-сервере (OLAP-структуры).
- MOLAP обеспечивает наилучшее быстродействие выполнения запросов, поскольку этот формат специально оптимизирован для многомерных запросов к данным.

Преимущества и недостатки MOLAP

- Поскольку MOLAP требует копирования и преобразования всех данных в надлежащий формат для многомерной структуры хранилища данных, MOLAP можно применять для небольших или средних объемов данных.
- Основное преимущество MOLAP заключается в превосходных свойствах индексации; ее недостаток — низкий коэффициент использования дискового пространства, особенно в случае разреженных данных.

Область применения MOLAP

- объем исходных данных для анализа не слишком велик (не более нескольких гигабайт), т.е. уровень агрегации данных достаточно высок;
- набор информационных измерений стабилен (поскольку любое изменение в их структуре почти всегда требует полной перестройки гиперкуба);
- время ответа системы на нерегламентированные запросы является наиболее критичным параметром;
- широкое использование сложных встроенных функций требуется для выполнения кроссмерных вычислений над ячейками гиперкуба, в том числе возможности написания пользовательских функций.

ROLAP

- Реляционные хранилища OLAP содержат данные, передаваемые в кубы данных, вместе с агрегациями данных куба, причем данные хранятся в реляционных таблицах, размещенных в реляционном ХД.

Преимущества ROLAP :

- в большинстве случаев корпоративные хранилища данных реализуются средствами реляционных СУБД, и инструменты ROLAP позволяют производить анализ непосредственно над ними. При этом размер хранилища не является таким критичным параметром, как в MOLAP;
- при переменной размерности задачи, когда изменения в структуру измерений приходится вносить достаточно часто, ROLAP-системы с динамическим представлением размерности являются оптимальным решением, так как в них такие модификации не требуют физической реорганизации БД;
- реляционные СУБД обеспечивают значительно более высокий уровень защиты данных и хорошие возможности разграничения прав доступа.

Недостатки ROLAP

- Главный недостаток ROLAP по сравнению с MOLAP — меньшая производительность.
- Для обеспечения производительности, сравнимой с многомерными базами данных, необходимо использовать звездообразные схемы. В этом случае производительность реляционных систем может быть приближена к производительности систем на основе MOLAP.

HOLAP

- Гибридная архитектура, которая объединяет технологии ROLAP и MOLAP. В отличие от MOLAP, которая работает лучше, когда данные более 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.
- При использовании данного формата OLAP-данные, передаваемые в куб данных, хранятся в реляционных базах данных подобно ROLAP. А агрегации данных (данные куба) записываются и представляются в многомерном формате.

Преимущества и недостатки HOLAP

- Преимуществом данной системы является обеспечение возможности связи с огромными наборами данных в реляционных таблицах и прирост производительности за счет использования многомерных хранилищ.
- Недостаток состоит том, что количество проводимых преобразований между ROLAP и MOLAP системами может существенно влиять на общую эффективность.

Сравнительные характеристики

Характеристика	OLTP	ROLAP	MOLAP
1	2	3	4
Типовая операция	Обновление	Отчет	Анализ
Уровень аналитических требований	Низкий	Средний	Высокий
Экраны	Неизменяемые	Определяемые пользователем	Определяемые пользователем
Объем данных на транзакцию	Небольшой	От малого до большого	Большой
Уровень данных	Детальные	Детальные и суммарные	Суммарные
Сроки хранения данных	Текущие	Исторические и текущие	Исторические, текущие и прогнозируемые
Структурные элементы	Записи	Записи	Массивы

Достоинства OLAP:

- простота использования и восприятия выходных таблиц;
- полнота аналитических данных;
- полная и легкая настройка отчета без программиста;
- возможность детализировать отчет в процессе анализа данных (от итогов к деталям);
- формирование отчетов в 10 раз быстрее;
- непротиворечивость данных в отчетах;
- консолидация информации из разных баз данных;
- повышенная защита данных;
- эквивалентность одного OLAP-отчета целому набору простых отчетов.

Недостатки OLAP:

- не ориентирован на получение форм отчетности с произвольным дизайном;
- некоторые пользователи визуально плохо воспринимают выходные таблицы;
- ограниченные возможности создания оперативных отчетов;
- основная проблема: необходимость разработки хранилищ данных.

Литература

- Перминов Г.И. УМК - «Системы интеллектуального анализа данных» (Business Intelligence). ГУ-ВШЭ, 2007.
- Microsoft SQL Server 2005. Analysis Services. Под ред. Горбач И. –С-Пб,: БХВ-Петербург, 2007
- Э. Спирли. Корпоративные хранилища данных. Планирование, разработка, реализация. Том. 1: Пер. с англ. - М.: "Вильямс", 2001.
- http://www.dw-institute.com/lessons