**Министерство образования и науки Российской Федерации**

**Федеральное государственное бюджетное образовательное учреждение**

**высшего образования**

**Псковский государственный университет**

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**Контрольная работа**

по дисциплине

**« Технологии обработки информации »**

тема:

**Хранилища данных (ХД)**

Выполнил студент

группы 1022-03

Ковалевский Р.А.

Проверил: Полетаева О.А.

Дата проверки:\_\_\_\_\_\_\_\_

**Псков**

**2018**

**Вариант №3**

1. **Какие факторы стимулировали появление концепции ХД?**
2. **В чем заключаются основные различия между ХД и обычными базами данных?**
3. **Какие функциональные требования предъявляются к ХД?**
4. **Какие принципы лежат в основе построения ХД?**
5. **Каковы цели использования концепции ХД в процессе поддержки принятия решений и интеллектуального анализа данных?**
6. **Зачем выполняется агрегирование данных?**
7. **Что такое метаданные и какова их роль в процессе функционирования ХД?**
8. **Какие виды метаданных вам известны?**
9. **Какие ХД называются кросс-платформенными?**
10. **Какие архитектуры ХД вам известны?**
11. **Какую цель преследует использование гибридной концепции ХД?**
12. **Каков принцип организации гибридных ХД?**
13. **В чем недостаток гибридных ХД?**
14. **Что такое витрина данных?**
15. **Каковы преимущества и недостатки использования витрин данных?**
16. **Имеет ли смысл использовать витрины данных на предприятии, не имеющем подразделений?**
17. **Что такое виртуальное ХД?**

**1.Какие факторы стимулировали появление концепции ХД**

В начале 80-х годов, в период бурного развития регистрирующих информационных систем, появилось осознание ограниченности их применения для анализа данных и построения систем поддержки и принятия решений.

Регистрирующие системы создавались для автоматизации рутинных операций: выписки счетов, оформления договоров, проверки состояния склада и т.д., и предназначались для линейного персонала. Основными требованиями к таким системам были обеспечение транзакционности вносимых изменений и максимизация скорости, что и определило тогда выбор реляционных СУБД и модели представления данных «сущность-связь» в качестве основных технических решений при построении регистрирующих систем.

Для менеджеров и аналитиков в свою очередь требовались системы, которые бы позволяли: анализировать информацию во временном аспекте, формировать произвольные запросы к системе, обрабатывать большие объемы данных, интегрировать данные из различных регистрирующих систем. Очевидно, что регистрирующие системы не удовлетворяли ни одному из этих требований — информация в такой системе актуальна только на момент обращения к базе данных, а в следующий момент времени к тому же запросу можно получить совершенно другой результат.

Интерфейс регистрирующих систем рассчитан на проведение жестко определенных операций и возможности получения результатов на нерегламентированный (ad-hoc) запрос сильно ограничены. Возможности обработки больших массивов данных также были невелики из-за настройки СУБД на выполнение коротких транзакций.

Ответом на возникшую потребность стало появление технологии хранилищ данных.

**2.В чем заключаются основные различия между ХД и обычными базами данных?**

Хранилище данных - предметно-ориентированный, интегрированный, зависимый от времени набор данных, предназначенный для поддержки принятия решений различными группами пользователей.

Так как хранилище носит предметно-ориентированный характер, его организация нацелена на содержательный анализ информации, а не на автоматизацию бизнес-процессов. Это свойство определяет архитектуру построения хранилища и принципы проектирования модели данных, отличные от тех, что применяются в оперативных системах.

ХД строится на базе клиент-серверной архитектуры, реляционной СУБД и утилит поддержки принятия решений.

К хранилищу присоединяются сторонние программные продукты, которые делают модели, основанные на интеллектуальном, а не статистическом анализе данных и получают скрытые закономерности. Но включают они генетические алгоритмы, нейронные сети, нелинейная динамика, кластеризация, гибридные системы - достаточно большое количество технологий для построения моделей в хранилище. Это необходимо тогда, когда количество данных таково, что прямой поиск и статистические методы анализа результата не дают.

Основные составляющие Хранилища данных:

1.Предметная ориентированность

Локальные базы данных содержат мегабайты информации, абсолютно не нужной для анализа (адреса, почтовые индексы, идентификаторы записей и др.). Подобная информация не заносится в хранилище, что ограничивает спектр рассматриваемых данных при принятии решения до минимума.

Для принятия решений требуется некоторая строго определенная совокупность данных, которая и вытаскивается из БД в ХД, второстепенные ненужные атрибуты отсеиваются.

2.Интегрированность (целостность и внутренняя взаимосвязь)

Несмотря на то что данные погружаются из различных источников, но они объединены едиными законами именования, способами измерения атрибутов и др. Это имеет большое значение для корпоративных организаций, в которых одновременно могут эксплуатироваться различные по своей архитектуре вычислительные системы, представляющие одинаковые данные по-разному. Например, могут использоваться несколько различных форматов представления дат или один и тот же показатель может называться по-разному. В процессе погружения подобные несоответствия устраняются автоматически. Это самая трудоемкая часть создания ХД.

3.Временная привязка

Оперативные системы охватывают небольшой интервал времени, что достигается за счет периодического архивирования данных. ХД, напротив, содержит данные, накопленные за большой интервал времени (от пяти-семи до десятков лет).

4.Неразрушаемая совокупность данных.

Модификация данных не производится, поскольку может привести к нарушению их целостности.

**Отличия:**

Если БД невелика, узкоспециализирована и имеется квалифицированный программист, способный писать нестандартные запросы, собирающие данные в массив и анализирующие эти данные, то вместо ХД можно использовать обычную БД. Хранилище же предназначено не для программиста - им должен суметь воспользоваться аналитик, руководитель, человек, не владеющий навыками написания сложных запросов.

Недостатки использования БД в поддержке принятия решений:

·недостоверность данных;

·низкая производительность при нестандартных запросах;

·невозможность преобразования разнородных данных, так как они часто не имеют меток времени;

·Проблемы при подготовке отчетов возникают из-за того, что:

- трудно понять, где находятся данные, необходимые для анализа и принятия решения;

- большинство БД ориентировано только на стандартные запросы;

- требуется привлекать программистов для выполнения нестандартных запросов.

Особенности хранилищ 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, поскольку требования к функциональности и производительности аналитических приложений отличаются от требований к транзакционным системам.

- Хранилища данных создаются специально для приложений поддержки принятия решений и предоставляют накопленные за определенное время, сводные и консолидированные данные, которые более приемлемы для анализа, чем детальные индивидуальные записи.

- Рабочая нагрузка состоит из нестандартных, сложных запросов, которые обращаются к миллионам записей и выполняют огромное количество операций сканирования, соединения и агрегирования. Время ответа на запрос в данном случае важнее, чем пропускная способность.

**3.Какие функциональные требования предъявляются к ХД?**

Хранилища данных – это процесс сбора, отсеивания и предварительной обработки данных с целью представления результирующей информации пользователям для статистического анализа и аналитических отчетов. Ральф Кинболл (автор концепции хранилищ данных) описывал хранилища данных как «место, где люди могут получить доступ к своим данным». Он же сформулировал основные требования к хранилищам данных:

– поддержка высокой скорости данных из хранилища;

– поддержка внутренней непротиворечивости данных;

– возможность получения и сравнения данных;

– наличие удобных утилит просмотра данных хранилища;

– полнота и достоверность хранимых данных;

– поддержка качественного процесса пополнения данных.

Всем перечисленным требованиям удовлетворять зачастую не удается, поэтому для реализации хранилищ данных используют несколько продуктов. Одни из которых представляют средства хранения данных, другие – средства их извлечения и просмотра, в-третьих – средства пополнения хранилищ данных. Типичное хранилище данных как правило отличается от реляционной базы данных: 1) Обычная база данных предназначена для того, чтобы помочь пользователям выполнять повседневную работу, тогда как хранилища данных предназначены для принятия решений; 2) Обычная база данных подвержена постоянным изменениям в процессе работы пользователей, а хранилища данных относительно стабильно; данные в нем обновляются согласно расписанию (например, ежечасно, ежедневно, ежемесячно), в идеале, процесс пополнения данными за определенный период времени без изменения прежней информации находящейся уже в хранилище. 3) Обычная база данных чаще всего является источником данных попадающих в хранилище, кроме того хранилище может пополняться за счет внешних источников (например, сжатия данных).

Требования:

• высокая скорость получения данных из хранилища;

• автоматическая поддержка внутренней непротиворечивости данных;

• возможность получения и сравнения срезов данных;

• наличие удобных средств для просмотра данных в хранилище;

• обеспечение целостности и достоверности хранящихся данных.

**4.Какие принципы лежат в основе построения ХД?**

Информация, которая загружается в хранилище, должна интегрироваться в целостную структуру, отвечающую целям анализа данных. При этом минимизируются несоответствия между данными из различных оперативных систем, в хранилище именуются и выражаются единым образом. Данные интегрированы на множестве уровней: на уровне ключа, атрибута, на описательном, структурном уровне и так далее. Общие данные и общая обработка данных консолидированы и являются единообразным для всех данных, которые подобны или схожи в хранилище данных. При этом информация структурируется по разным уровням детализации:

– высокая степень суммаризации;

– низкая степень суммаризации;

– текущая детальная информация.

Хранилища можно рассматривать как набор моментальных снимков состояния данных: можно восстановить картинку на любой момент времени. Атрибут времени всегда явно присутствует в структурах данных хранилища.

Попав однажды в хранилище, данные уже никогда не изменяются, а только пополняются новыми данными из оперативных систем, где данные постоянно меняются. Новые данные по мере поступления обобщаются с уже накопленной информацией в хранилище данных.

**5.Каковы цели использования концепции ХД в процессе поддержки принятия решений и интеллектуального анализа данных?**

Цель концепции ХД - прояснить отличия в характеристиках данных в операционных и аналитических системах, определить требования к данным помещаемым в целевую БД ХД, определить общие принципы и этапы её построения, основные источники данных, дать рекомендации по решению потенциальных проблем возникающих при их выгрузке, очистке, согласовании, транспортировке и загрузке в целевую БД.

Для правильного понимания данной концепции необходимо понимание следующих принципиальных моментов, заключающихся в том, что концепция ХД:

это не концепция анализа данных, скорее это концепция подготовки данных для анализа.

не предопределяет архитектуру целевой аналитической системы. Она говорит о том, какие процессы должны выполняться в системе, но не о том, где конкретно и как эти процессы должны выполняться.

предполагает не просто единый логический взгляд на данные организации (как иногда это трактуется). Она предполагает реализацию единого интегрированного источника данных.

Анализ данных в хранилищах базируется на технологиях интеллектуального анализа данных (НАД). Целью НАД является извлечение знаний из данных, т. е. обнаружение в исходных данных ранее неизвестных нетривиальных практически полезных и доступных для интерпретации знаний, необходимых для принятия решений в различных предметных областях.

Методы НАД реализуются в технологиях:

• интерактивной аналитической обработки данных (On-Line Analytical Processing — О LAP);

• глубинного анализа данных (Data Mining — DM);

• визуализации данных.

Технология OLAP и многомерные модели данных. Технология OLAP ориентирована на обработку нерегламентированных запросов к хранилищам данных. Создание хранилищ данных вызвано тем, что анализировать данные OLAP-систем напрямую невозможно или затруднительно, так как они являются разрозненными, хранятся в форматах различных СУБД и в разных сегментах корпоративной сети. Основная задача хранилища — представление данных для анализа в одном месте в рамках простой и понятной структуры.

**6.Зачем выполняется агрегирование данных**

Агрегирование данных (data aggregation): процесс сбора, обработки и представления информации в окончательном виде. Агрегирование данных в основном выполняется для формирования отчетов, выработки политики, управления здравоохранением, научных исследований, статистического анализа и изучения здоровья населения..."

Источник: информатизация здоровья. Требования к архитектуре электронного учета здоровья. Гост р исо/тс 18308-2008.

Процедура агрегировать данные позволяет преобразовать группы наблюдений в наблюдения, содержащие агрегированную информацию по соответствующей группе, и создавать новый - агрегированный - файл данных или создавать новые переменные в активном наборе данных, содержащие агрегированную информацию. Наблюдения агрегируются на основе значений нуля или большего числа группирующих переменных. Если группирующие переменные не заданы, то весь набор данных представляет собой единую группу для агрегирования.

Если вы создаете новый агрегированный файл данных, то новый файл будет содержать по одному наблюдению на каждую группу, определяемую группирующей переменной. Например, если имеется одна группирующая переменная с двумя значениями, то новый файл данных будет содержать только два наблюдения. Если не заданы группирующие переменные, то новый файл данных будет содержать одно наблюдение.

Если агрегированные переменные добавляются в активный набор данных, сам файл данных не агрегируется. Каждое наблюдение с одинаковыми значениями группирующих переменных получает одинаковые значения для новых агрегированных переменных. Например, если имеется единственная группирующая переменная gender , то все мужчины будут иметь одинаковое значение новой агрегированной переменной, представляющей средний возраст. Если не задано группирующих переменных, то все наблюдения будут получать одно и то же значение для новой агрегируемой переменной, представляющей средний возраст.

Иными словами, для достоверного описания предметной области использование данных с максимальным уровнем детализации не всегда целесообразно, поэтому наибольший интерес для анализа представляют данные, обобщенные по некоторому интервалу времени, по группе клиентов, товаров и т.д. Такие обобщенные данные называются агрегированными (иногда агрегатами), а сам процесс их вычисления — агрегированием.

В результате агрегирования большое количество записей о каждом событии в бизнес-процессе заменяется относительно небольшим количеством записей, содержащих агрегированные значения. Например, вместо информации о каждой из 365 ежедневных продаж в году в результате агрегирования будут храниться 52 записи с обобщением по неделям, 12 — по месяцам или 1 — за год. Если цель анализа — разработка прогноза продаж, то для краткосрочного оперативного прогноза достаточно использовать данные по неделям, а для долгосрочного стратегического прогноза — по месяцам или даже по годам.

**7.Что такое метаданные и какова их роль в процессе функционирования ХД?**

Метаданные — это любые данные о данных. Метаданные играют важную роль в построении Систем Поддержки Принятия Решений (СППР). Одновременно это один из наиболее сложных и недостаточно практически проработанных объектов. В общем случае можно выделить по крайней мере три аспекта метаданных, которые должны присутствовать в системе.

С точки зрения пользователей:

метаданные для бизнес-аналитиков,

метаданные для администраторов,

метаданные для разработчиков.

С точки зрения предметных областей:

структуры данных хранилища,

модели бизнес-процессов,

описания пользователей,

технологические и пр.

С точки зрения функциональности системы:

метаданные о процессах трансформации,

метаданные по администрированию системы,

метаданные о приложениях, метаданные о представлении данных

пользователям.

Присутствие трех перечисленных аспектов метаданных подразумевает, что, например, прикладные пользователи и разработчики системы будут иметь различное видение технологических аспектов трансформации данных из источников: прикладные пользователи - семантику, состав и периодичность пополнения хранилища данными из источника, разработчики - ER-диаграммы, правила трансформации и интерфейс доступа к данным источника.

В настоящее время отсутствует единая промышленная технология проектирования, создания и сопровождения метаданных. Поэтому вопросы, связанные с управлением метаданными, рассматриваются отдельно, применительно к каждому конкретному проекту построения СППР.

**8.Метаданные Хранилища данных**

Популярность Хранилищ данных в последние годы существенно возросла. Конкурентоспособные организации находятся на пути построения ХД либо расширения, перепроектирования и усовершенствования уже имеющихся. Метаданные считаются ключевым фактором успеха в проектах по внедрению Хранилищ. Они содержат всю информацию, необходимую для извлечения, преобразования и загрузки данных из исходных систем, а также для последующего использования и интерпретации содержимого ХД.

Метаданные систем Хранилищ данных иногда подразделяют на два типа:

служебные метаданные, используемые для функций извлечения, преобразования и загрузки, для переноса OLTP-данных (информации из транзакционных систем) в Хранилище; интерфейсные метаданные, использующиеся для описания экранов и создания отчетов.

**Функции метаданных в хранилище данных**

Роль метаданных для ХД значительно важнее, чем в системах операционной обработки данных. Если в системах операционной обработки данных интерфейс системы настроен на бизнес-процедуры обработки данных конкретными специалистами и понятен им после специального обучения, то интерфейс систем складирования данных конструируется таким образом, чтобы помимо всего прочего отвечать на непредопределенные вопросы (ad hoc). Как правило, такие вопросы формулируются в терминах предметной области и бизнес-процессов, к тому же специалистами, для которых ИТ-технологии не являются основной профессией: аналитиками, менеджерами среднего и высшего уровня.

Таким образом, одним из главных аспектов использования метаданных в ХД является их предметная ориентация. Основные вопросы, на которые должны ответить метаданные, — это какие данные представлены в системе и как их получить в нужном для анализа данных виде.

Первой основной функцией метаданных в ХД является представление соответствия данных источников и данных ХД. Как правило, это описание представляет собой фиксацию взаимосвязи атрибутов данных источника и атрибутов данных ХД, правила преобразования первых во вторые, изменение в наименовании данных, в их физических характеристиках и т. д.

Такая информация позволяет идентифицировать источники данных для ХД, правильность данных в ХД и их корректность.

Вторая основная функция метаданных в ХД — управление данными во времени. Время жизни данных в ХД, как правило, 5-10 лет, а то и более. А для систем операционной обработки данных время жизни данных — от нескольких дней до нескольких месяцев. Затем данные архивируются в случае необходимости.

Таким образом, временной горизонт данных в ХД гораздо больше, и это обстоятельство изменяет коренным образом некоторые принципы управления данными. Например, в системах операционной обработки данных в одно и то же время существует только одно корректное определение данных. Для ХД это не так.

Структура данных (схема или модель данных) в системах операционной обработки данных меняется во времени, т.е. данные в таких системах в разное время имеют различные формы представления. Хронология таких изменений должна быть зафиксирована в ХД.

Таким образом, в ХД в одно и то же время может существовать несколько схем данных, отвечающих различным периодам эволюции источников данных. Запись о таких структурных изменениях сохраняется в метаданных ХД. На основании записей аналитики получают ответы на вопросы, какими данными и за какие периоды они располагают.

Третья, и немаловажная, функция метаданных в ХД — это поддержка версионности. Эта функция тесно связана с управлением данными с большим временным горизонтом. Метаданные должны отражать изменения внутренней структуры данных источников и, следовательно, должны сами изменяться, для того чтобы обеспечить непрерывность истории изменения структуры данных ХД.

Таким образом, поддержка версионности метаданных позволяет в каждый момент времени в прошлом обеспечить правильное описание модели данных, а аналитики получают возможность знать, какие данные, когда и как попали в ХД.

Четвертая основная функция метаданных в ХД — это интерпретация данных в терминах бизнес-пользователей. Метаданные должны поддерживать в запросах понятную для пользователя терминологию, независимо от того, какие правила наименования атрибутов были использованы проектировщиком ХД.

Пятая основная функция метаданных — обеспечение открытости (доступности другим информационным системам) системы складирования данных для ее интеграции с другими аналитическими системами организации. Опрос метаданных ХД другой системой позволяет последней выяснить структуру данных ХД и поддерживать обмен данными между системами.

**9.Какие ХД называются кросс-платформенными**

Кроме того, все ХД можно разделить на одноплатформенные и кросс-платформенные. Одноплатформенные ХД строятся на базе только одной СУБД, а кросс-платформенные могут строиться на базе нескольких СУБД.

**10.Какие архитектуры ХД вам известны?**

Пять архитектур

На сегодняшний день предложено множество архитектур, опишем пять наиболее распространенных:

независимые витрины(independent data marts);

шина взаимосвязанных витрин (data-mart bus architecture with linked

dimensional data marts);

архитектура «звезда» (hub-and-spoke);

централизованное Хранилище (centralized data warehouse);

федеративная архитектура (federated architecture).

Независимые витрины (independent data marts)

Нередка ситуация, когда каждое подразделение компании разрабатывает свою собственную витрину данных. Все эти витрины удовлетворяют потребностям, для которых создавались, но при этом не зависят друг от друга и не обеспечивают единого представления о ситуации в компании. В них несогласованно заданы данные, используются разные измерения и показатели, а следовательно анализ данных между витринами затруднен.

Шина взаимосвязанных витрин данных

Предложена Ральфом Кимбэлом (Ralph Kimball).

Создание такой архитектуры начинается с анализа требований для конкретных бизнес-процессов, таких как заказы, клиенты, счета и проч. Первая витрина (data mart –DM) строится для одного бизнес-процесса с использованием измерений и показателей, которые в дальнейшем будут применяться в других компонентах.

Последующие DM разрабатываются с использованием этих измерений, что в результате приводит к созданию логически интегрированных витрин.

Архитектура «Звезда»

Продвигается известным экспертом в области Хранилищ Билом Инмоном (Bill Inmon). Представляет собой централизованное Хранилище с зависимыми витринами данных.

Эта архитектура разрабатывается на основе корпоративного анализа требований к данным. Тут важно обратить внимание на создание масштабируемой и поддерживаемой инфраструктуры. Используя корпоративное представление данных, выполняется итеративная разработка архитектуры, при этом вовлекается одна предметная область за другой. Детальные данные хранятся в нормализованной форме в Хранилище. Зависимые витрины получают данные из Хранилища.

Зависимые витрины разрабатываются для подразделений или конкретных функциональных областей, целей (например, для data mining) и могут быть как нормализованными, так и денормализованными, либо в виде любой агрегированной структуре данных. Большинство пользователей выполняет запросы на зависимых витринах данных.

Иногда архитектуру hub-and-spoke называют подходом «сверху вниз», а шину витрин - подходом «снизу вверх». В этом есть некоторый смысл, так как первая ориентирована на исходно заданную инфраструктуру и процессы, а шина витрин – на разработку проекта, который решает критические бизнес-задачи.

Подходы «снизу вверх» и «сверху вниз» со временем сближаются. Сторонники первого из них утверждают важность поэтапной разработки и достижения «маленьких побед». Приверженцы второй методологии считают важным наличие корпоративного плана для интеграции поэтапно создаваемых витрин данных.

Централизованное Хранилище (Без зависимых витрин)

Эта архитектура похожа на архитектуру «звезда», исключая отсутствие зависимых витрин. Хранилище содержит детальные данные, некоторое количество агрегированных данных и логические представления. Запросы и приложения выполняются как на реляционных данных, так и на многомерных представлениях.

Федеративная архитектура

Эта архитектура использует уже существующие структуры поддержки принятия решений (операционные системы, витрины данных и Хранилища). Данные извлекаются из перечисленных систем на основе бизнес-требований. Данные логически или физически интегрируются с помощью метаданных, распределенных запросов и других методов. Эта архитектура является практическим решением для компаний, которые уже пользуются аналитическими средствами и не хотят от них отказываться.

**11.Какую цель преследует использование гибридной концепции ХД**

**12.Каков принцип организации гибридных ХД**

**13.В чем недостаток гибридных ХД**

Многомерная и реляционная модели хранилищ данных имеют свои преимущества и недостатки. Например, многомерная модель позволяет быстрее получить ответ на запрос, но не дает возможности эффективно управлять такими же большими объемами данных, как реляционная модель.

Логично было бы использовать такую модель ХД, которая представляла бы собой комбинацию реляционной и многомерной моделей и позволяла бы сочетать высокую производительность, характерную для многомерной модели, и возможность хранить сколь угодно большие массивы данных, присущую реляционной модели. Такая модель, сочетающая в себе принципы реляционной и многомерной моделей, получила название гибридной, или HOLAP (Hybrid OLAP).

Хранилища данных, построенные на основе HOLAP, называются гибридными хранилищами данных (ГХД) (рис. 10).

Главным принципом построения ГХД является то, что детализированные данные хранятся в реляционной структуре (ROLAP), которая позволяет хранить большие объемы данных, а агрегированные — в многомерной (MOLAP), которая позволяет увеличить скорость выполнения запросов (поскольку при выполнении аналитических запросов уже не требуется вычислять агрегаты).

Пример:

В супермаркете, ежедневно обслуживающем десятки тысяч покупателей, установлена регистрирующая OLTP-система. При этом максимальному уровню детализации регистрируемых данных соответствует покупка по одному чеку, в котором указываются общая сумма покупки, наименования или коды приобретенных товаров и стоимость каждого товара. Оперативная информация, состоящая из детализированных данных, консолидируется в реляционной структуре ХД. С точки зрения анализа представляют интерес обобщенные данные, например, по группам товаров, отделам или некоторым интервалам дат. Поэтому исходные детализированные данные агрегируются, и вычисленные агрегаты сохраняются в многомерной структуре гибридного ХД.

Если данные, поступающие из OLTP-системы, имеют большой объем (несколько десятков тысяч записей в день и более) и высокую степень детализации, а для анализа используются в основном обобщенные данные, гибридная архитектура хранилища оказывается наиболее подходящей.

Недостатком гибридной модели является усложнение администрирования ХД из-за более сложного регламента его пополнения, поскольку при этом необходимо согласовывать изменения в реляционной и многомерной структурах.

Преимущества:

Построение OLAP-куба выполняется по запросу OLAP-средства на основе реляционных и многомерных данных. Такой подход позволяет избежать взрывного роста данных. При этом можно достичь оптимального времени исполнения клиентских запросов.

**14.Что такое витрина данных**

**15.Каковы преимущества и недостатки использования витрин данных**

**16.Имеет ли смысл использовать витрины данных на предприятии, не имеющем подразделений.**

Хранилище данных хранится информация по всем аспектам деятельности организации. Витрина же данных (data mart) – это специализированное хранилище данных, содержащее данные по одному из направлений деятельности предприятия. Витрины данных - это комплекс тематически связанных баз данных, относящихся к конкретным аспектам деятельности компании. В этом случае аналитики видят и работают не со всеми имеющимися в компании данными, а только с реально необходимыми данными. Это максимально приближает их к конечному пользователю.

Витрина данных представляет собой срез хранилища данных, представляющий собой массив тематической, узконаправленной информации, ориентированный на пользователей одной конкретной рабочей группы. Часто витрины еще называют киосками данных.

Т.к. Витрины Данных обычно содержат тематические подмножества заранее агрегированных данных, то их проще проектировать и настраивать. Витрина данных проектируются для ответов на конкретный ряд вопросов. Данные в витрине оптимизированы для использования определенными группами пользователей, что облегчает процедуры их наполнения, а также способствует повышению производительности

Т.к. конструирование хранилища данных — сложный процесс, который может занять несколько лет, некоторые организации вместо этого строят витрины данных, содержащие информацию для конкретных подразделений. Например, витрина данных отдела маркетинга может содержать только информацию о клиентах, продуктах и продажах и не включать в себя планы поставок. Существуют также витрина данных отдела продаж, витрина данных финансового отдела, витрина данных отдела анализа рисков и т.п. 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.

Достоинства Витрин данных:

· Витрина данных максимально приближена к пользователю - аналитики работают только с теми данными, которые им необходимы для принятия решения.

· Витрины Данных значительно меньше по размеру, чем Хранилища данных.

· Создавать Витрины данных просто, наполнять их и работать с ними также просто.

· Витрины Данных содержат агрегированные данные по определенным темам, что упрощает их проектирование.

· Витрины данных внедряются достаточно быстро.

· Витрины проектируются для ответов на конкретный ряд вопросов.

· Данные в витрине оптимизированы для использования определенными группами пользователей, что облегчает процедуры их наполнения, а также способствует повышению производительности.

Недостатки Витрин данных:

· сложно контролировать избыточность, целостность и непротиворечивость данных в витрине данных, т.к. одни и те же данные могут храниться сразу в нескольких витринах. Дублирование данных - данные хранятся многократно в различных витринах данных. Это приводит к увеличению расходов на хранение, а также к потенциальным проблемам, связанным с необходимостью поддержания непротиворечивости данных.

· сложно работать с витриной, если для наполнения витрины используется очень большое количество источников данных - Потенциально это очень сложный процесс. Наполнения витрин данных при большом количестве источников данных – процесс очень сложный (требуется команда профессионалов в области их построения и работы с ними).

· информация накапливается в различных витринах, но дальнейшее объединение информации (т.е. ее консолидация) не предусматривается. Данные не консолидируются на уровне предприятия, поэтому получить полную единую картину бизнеса нельзя.

· каждая витрина содержит информацию о конкретном аспекте деятельности компании, что не дает возможность руководству компании увидеть как работает его бизнес в целом.

В большинстве случае Витрина данных - это аналитическая структура, которая обычно поддерживает область работы одного приложения, бизнес-процесса или отдела. Сотрудники отдела обобщают требования к информации и приспосабливают каждую витрину к своим нуждам. Затем они обеспечивают персонал, работающий с информацией, средствами интерактивной отчетности (например, инструментами OLAP, средствами формирования незапланированных запросов или параметризованных отчетов).

**17.Что такое виртуальное ХД**

При всех положительных сторонах хранилища данных как отдельного консолидированного источника встречаются ситуации, когда эта идея не работает. Дело в том, что устоявшейся практикой является ночная загрузка собранных за день данных из OLTP-систем в ХД. Такой регламент позволяет уменьшить нагрузку на OLTP-систему в течение рабочего дня, то есть в период ее активного использования. Чаще чем один раз в сутки пополнять ХД смысла не имеет. Однако подобное положение вещей не обеспечивает возможности анализировать информацию в течение рабочего дня сразу по мере ее поступления. Иногда это очень критично, и приходится искать альтернативу традиционному (физическому) ХД.

Кроме того, неизбежной проблемой при использовании ХД в бизнес-аналитике является избыточность. Она снижает эффективность использования дискового пространства и оперативной памяти серверов, а при очень больших объемах хранящейся и обрабатываемой информации может вызвать снижение производительности, возрастание времени ожидания отклика на запрос и даже привести к полной неработоспособности системы. Избыточность в той или иной степени характерна как для реляционных, так и для многомерных хранилищ.

Ситуация усугубляется еще и тем, что ХД хранят историческую информацию и реализуют принцип неизменчивости данных. То есть в отличие от обычных систем оперативной обработки (OLTP-систем), где хранятся лишь актуальные данные, а данные, утратившие актуальность, уничтожаются, ХД могут только пополняться новыми данными, а удаление исторических данных не производится. Кроме того, часто требуется хранить большие объемы агрегированных данных. В совокупности эти факторы могут привести к «взрывному» росту объемов ХД. Хотя ради справедливости заметим, что непрекращающееся развитие информационных технологий приводит к снижению стоимости хранения информации на машинных носителях.

Преодолеть вышеперечисленные проблемы позволяет концепция виртуального хранилища данных (ВХД). В ее основе лежит принцип, в соответствии с которым данные из локальных источников, внешнего окружения, баз данных и учетных систем не консолидируются в единое ХД физически, а извлекаются, преобразуются и интегрируются непосредственно при выполнении запроса в оперативной памяти компьютера. Фактически запросы адресуются непосредственно к источникам данных.

Определение

Виртуальным хранилищем данных называется система, которая работает с разрозненными источниками данных и эмулирует работу обычного хранилища данных, извлекая, преобразуя и интегрируя данные непосредственно в процессе выполнения запроса.

Преимущества такого подхода очевидны.

Появляется возможность анализа данных в OLTP-системе сразу после их поступления без ожидания загрузки в хранилище.

**Список используемой литературы:**

1. Тартынский В. А. Реализация хранилищ данных в системах поддержки принятия решений // Молодой ученый. — 2009. — №9. — С. 31-34. — URL https://moluch.ru/archive/9/618/ (дата обращения: 29.05.2018).

2. В.В. Корнеев, А.Ф. Гареев, С.В. Васютин, В.В. Райх. Базы данных. Интеллектуальная обработка информации. – М.: Издатель Молгачев С.В., Издательство Нолидж, 2001. – 496 с.

3. Вирт, Н. Алгоритмы и структуры данных / Н. Вирт. - М.: [не указано], 1989. - 542 c

4. Нанда Oracle PL/SQL для администраторов баз данных / Нанда, др. А. и. - М.: Символ, 2008. - 496 c.

5. Энсор, Д. Oracle. Проектирования баз данных / Д. Энсор, Й. Стивенсон. - М.: Киев: BHV, 2000. - 560 c.

6. Архипенков С.Я., Голубев Д.В., Максименко О.Б. Хранилища данных. М.: Диалог-МИФИ, 2002.