Лекция 15. Технологии хранилищ данных и добычи данных

Одной из **главных целей разработки хранилищ данных является информационное обеспечение компьютерной поддержки принятия решений по всем или основным видам деятельности организации.** Каждый вид деятельности организации является отдельной задачей, решение которой может быть, а может и не быть увязано с решением других задач в рамках организации. Вид деятельности организации или направление бизнеса совместно со спектром соответствующих ему бизнес-задач определяют предметную область хранилищ данных. Например, компания производит и продает оборудование для добычи газа, а с другой стороны, та же компания имеет *подразделения*, которые занимаются производством услуг в области автоматизации предприятий, в том числе и газодобывающих. Источники прибыли в этих случаях различны. Это два направления бизнеса компании (две предметных области). Общими задачами анализа данных для этих направлений бизнеса являются *прибыль* и бюджет.

Хранилище данных – это сложная компьютерная система. **Под *архитектурой хранилища данных* понимают совокупность программно-аппаратных компонент, совокупность технологических и организационных решений, предпринимаемых для создания, разработки и функционирования хранилищ данных, т.е. выбор аппаратного и программного обеспечения, выбор способов взаимодействия программно-аппаратных компонент, выбор способа решения проектной задачи по разработке и созданию хранилищ данных**. Как правило, *архитектуру ХД* составляют следующие компоненты:

- средства извлечения данных из различных БД OLTP-систем, унаследованных систем и других внешних источников данных;
- средства трансформации и *очистки данных*. Точность существующих данных доставляет немало хлопот организации. Поэтому перед тем как поместить данные в хранилище их необходимо привести в порядок, иначе говоря — очистить;
- программное обеспечение БД. Как правило, это высокопроизводительная РСУБД, используемая для структуризации и хранения информации;
- средства для соединения источников данных с хранилищем и клиентов с сервером.

Кроме этого, необходимы специальные *программные средства* проектирования хранилища, средства работы с репозиторием *метаданных* и собственно средства оперативной аналитики, или *OLAP*-средства.

Все это – сложное специальное *программное обеспечение, стоимость* которого также может исчисляться десятками и сотнями тысяч долларов.

Характер и масштаб решаемых задач анализа данных организации оказывает решающее *значение* на выбор *архитектуры ХД* и методы его проектирования. Проектировщик должен помнить, что, с одной стороны, ХД создается для решения конкретных, строго определенных задач анализа и воспроизводства новых данных, с другой — ХД должно обеспечивать корпоративную отчетность в рамках всей организации. Таким образом, определяющим моментом в построении ХД являются задачи обработки и анализа данных, производства и доставки отчетов.

Характер и масштаб решаемых задач анализа данных определяет и подходы к выбору архитектуры и проектированию ХД.

Желательно, чтобы выбор *архитектуры ХД* был сделан до начала его реализации, однако на практике не всегда следуют этому правилу. Задержка с выбором *архитектуры ХД* обычно приводит к пересмотру проделанной работы в свете новых принятых решений и, как правило, к увеличению объема работы.

Выбор *архитектуры ХД* относится к сфере компетенции руководителя ИТ-проекта по созданию *системы складирования данных*. На такой выбор влияют несколько различных факторов: *инфраструктура* организации, производственная и *информационная среда* организации, управление и *контроль*, масштабы проекта, возможности аппаратно-технологического обеспечения, готовность персонала и имеющиеся ресурсы.

Выбор подхода к конкретной реализации хранилищ данных также лежит в области влияния руководителя ИТ-проекта. Правильный выбор *архитектуры ХД* обычно определяет успех конкретного проекта по созданию *системы складирования данных*.

Существует несколько факторов, влияющих на принятие решений о выборе способа реализации: **время**, отведенное на проект, **возврат инвестиций, скорость ввода ХД в эксплуатацию, потребности пользователей, потенциальные угрозы по переделке, требования к ресурсам**, необходимым в определенный момент времени, выбранная *архитектура ХД*, **совокупная стоимость владения ХД**.

Проектировщик ХД должен знать, какие возможные решения могут быть приняты по *архитектуре ХД* и какой объем *работ* по проектированию ХД они повлекут. Выбор архитектуры будет определять, где ХД и/или *киоски данных* будут расположены и как ими будут организационно-технологически управлять. Например, данные могут быть расположены в центральном офисе организации, т.е. будут поддерживаться централизованно. Данные могут быть распределены по офисам организации или располагаться в филиалах организации, и могут поддерживаться как централизованно, так и независимо друг от друга.

Далее приводится краткий обзор типовых архитектур систем складирования данных и программных продуктов, наиболее часто используемых для реализации систем складирования данных.

Основные типы программно-аппаратной архитектуры хранилища данных

На рисунке приведена *типовая* обобщенная концептуальная схема для *архитектуры ХД*. В конкретных решениях по *архитектуре хранилищ данных* некоторые компоненты схемы могут отсутствовать.

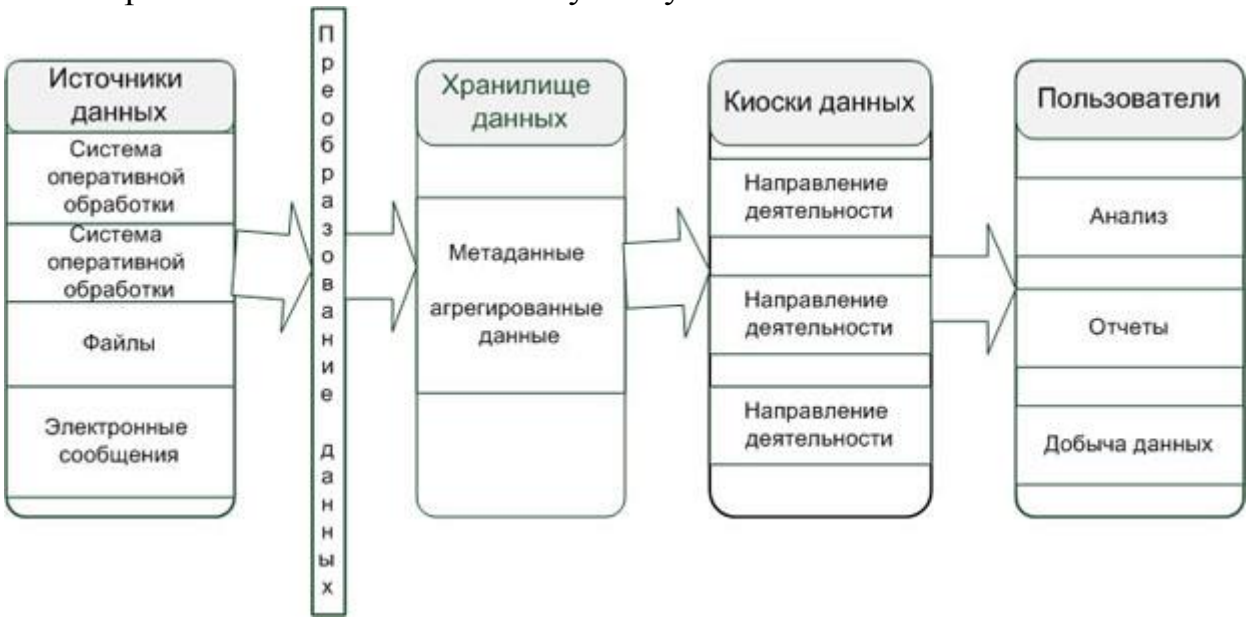

Рисунок - Типовая обобщенная концептуальная схема для архитектуры ХД

Компоненты типовой *архитектуры хранилища данных.*

- **Программное обеспечение промежуточного слоя**. Основное назначение этих компонент состоит в обеспечении доступа к сети и доступа к данным. Сюда можно отнести сетевые и коммуникационные протоколы, драйверы, системы обмена сообщениями и т.д. Поддержка такого программного обеспечения обычно выполняется информационными службами организации.
- **Базы данных систем оперативной обработки данных (OLTP) и данные внешних источников**. Для OLTP-систем характерна целевая направленность на эффективную обработку структур данных в рамках относительно небольшого числа четко определенных типовых транзакций. Количество таких транзакций может быть очень большим, число их типов незначительно. Направленность на быстрое выполнение транзакций делает такие системы малопригодными для решения аналитических задач. Транзакции для построения аналитических выборок по своей природе отличаются от транзакций OLTP-систем. В OLTP-системах выполнение таких выборок может приводить к снижению производительности.
- **Предварительная обработка и загрузка данных.** Предварительная обработка, связанная с фильтрацией, очисткой и преобразованием данных из OLTP-систем и внешних источников, обычно

выполняется в некотором промежуточном файле, который называется иногда загрузочной секцией. После обработки данные загружаются в ХД. Эта компонента включает в себя набор программных средств для выполнения указанных выше функций.

- **Хранилище данных**. Представляет собой ядро *системы складирования данных*. Это могут быть один или несколько серверов БД для поддержки ХД.
- **Метаданные**. Метаданные представляют собой репозиторий, который играет роль справочника о данных. Он включает терминологию предметной области, сведения об источниках данных, описание источников исходных данных, сведения об алгоритмах обработки исходных данных и т.д.
- **Уровень доступа к данным**. Этот компонент включает в себя программное обеспечение, которое обеспечивает взаимодействие конечных пользователей с данным ХД. В настоящее время универсальным средством общения служат SQL и его расширения.
- **Уровень информационного доступа**. Обеспечивает непосредственное общение пользователя с ХД. В качестве таких средств могут выступать стандартные пакеты MS Office, Lotus Notes или специальные программные продукты.
- **Уровень администрирования**. Компоненты этого уровня отслеживают выполнение процедур обновления ХД, включающих процедуры подкачки данных, обновления индексов, суммирования и агрегации данных, репликацию данных в распределенной вычислительной среде, авторизацию пользователя и разграничение доступа.

Типовыми архитектурами для систем складирования данных принято считать следующие:

- системы с *глобальным ХД*;
- системы с *независимыми киосками данных* ;
- системы с интегрированными *киосками данных* ;
- системы, разработанные на основе комбинации из вышеперечисленных архитектур.

Глобальное хранилище данных (*Global data warehouse*), или *хранилище данных* масштаба организации, — это такое ХД, в котором будут поддерживаться все данные организации или большая их часть. Это наиболее полное интегрированное ХД с высокой степенью интенсивности доступа к консолидированным данным и использованием его всеми подразделениями организации или руководством организации в рамках основных направлений деятельности организации. Таким образом,*глобальное ХД* проектируется и конструируется на основе потребностей аналитической информационной поддержки организации в целом. Его можно рассматривать как *общий репозиторий* для данных, обеспечивающих принятие решений.

Глобальное ХД необязательно должно быть реализовано физически как централизованное. Термин "глобальное" используется для отражения масштаба использования и доступа к данным в рамках всей организации. *Глобальное ХД* может быть физически как централизованным, так и распределенным.

Централизованное глобальное ХД характерно для организаций, расположенных территориально в одном здании. Оно поддерживается отделом информационных систем организации. **Распределенное глобальное ХД** также может быть использовано в рамках организации в целом. Оно физически распределяется по подразделениям организации и также поддерживается отделом информационных систем.

Поддержка ХД отделом информационных систем вовсе не означает, что именно эта служба управляет ХД. Например, отдельные части *распределенного ХД* могут управляться в рамках подразделений или направлений бизнеса.

Управление ХД определяет, кто решает:

- какие данные должны поступать в ХД;
- когда данные должны поступать в ХД;
- когда данные должны обновляться;
- кому разрешен доступ к данным в ХД.

Таким образом, для *глобального ХД* существуют два основных архитектурных решения, как показано на рисунке.

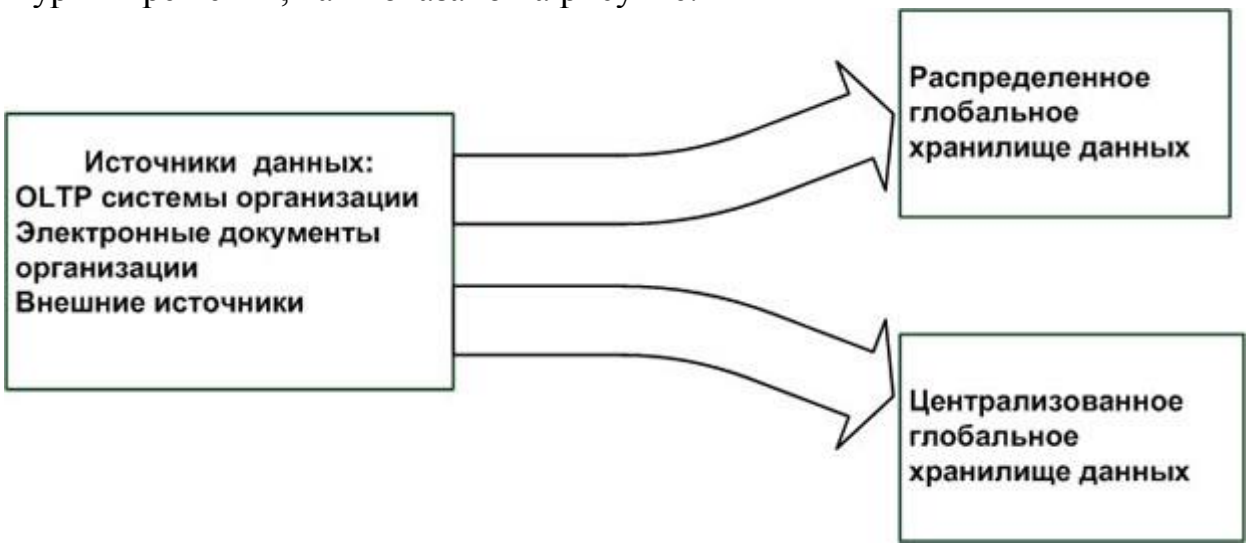

Рисунок - Основные архитектурные решения для глобального ХД

Данные для ХД обычно извлекаются из OLTP-систем организации, электронных документов организации и внешних источников данных. После фильтрации, очистки и преобразования они помещаются в ХД. Затем пользователи получают *доступ* к этим данным в соответствии с правилами управления доступом к данным, принятыми в организации.

Преимуществом *глобального ХД* является предоставление конечным пользователям доступа к информации в масштабах предприятия, недостатком — высокие *затраты* на реализацию, в том числе *затраты* времени на создание ХД.

Независимые киоски данных **включают в себя автономные или независимые киоски данных (Stand-alone Data Marts), которые управляются рабочими группами, отделами или направлениями бизнеса и разрабатываются исключительно для реализации аналитических потребностей последних**. Вполне возможно, что при этом не существует никакой связи между ними. Например, данные для таких *киосков данных* могут генерироваться непосредственно в самих подразделениях организации. Данные могут извлекаться из OLTP-систем, в частности, при помощи информационных служб организации. Информационные службы могут поддерживать вычислительную среду для *киосков данных*, но не управляют информацией в них. Данные в киоски могут поступать и из *глобального ХД*.

Для организации *независимых киосков данных* требуются некоторые профессиональные и технические навыки. Как правило, для их создания выделяются ресурсы и персонал в рамках того *подразделения*, для которого они создаются. Такой тип реализации ХД оказывает минимальное влияние на информационные ресурсы организации и может быть выполнен очень быстро. В то же время максимальная независимость и минимальная *интеграция*, а также отсутствие *глобального представления* о данных организации могут стать ограничением такой архитектуры.

Киоски данных могут быть взаимозависимы или взаимосвязаны (так называемые **связанные киоски данных**). Такая *архитектура ХД* **включает в себя совокупность киосков данных, которые управляются рабочими группами, отделами или направлениями бизнеса, но разрабатываются в рамках единой для организации схемы удовлетворения информационных и аналитических потребностей**. Для *взаимосвязанных киосков данных* типична распределенная *архитектура* реализации. Несмотря на то, что отдельные *киоски данных* реализуются в рамках рабочих групп, подразделений и направлений бизнеса, они могут быть интегрированы, т.е. взаимосвязаны, для того чтобы обеспечить представления данных в рамках организации в целом. Фактически, на наиболее высоком уровне интеграции, они могут стать *глобальным ХД*. В такой архитектуре пользователи одних подразделений могут получать *доступ* к данным других подразделений в рамках своих полномочий.

Требования интеграции данных в рамках архитектуры *взаимосвязанных киосков данных* делают реализацию ХД более сложной по сравнению с *независимыми киосками данных*. Например, необходимо решить вопрос, кто будет управлять данными в *киосках данных* и кто будет поддерживать вычислительную среду. Важным

становится вопрос о том, что делать с данными, которые являются общими для нескольких *киосков данных*, а также как разработать схему разграничения доступа пользователей к *киоскам данных* в рамках всей организации.

Главным достоинством создания ХД такой архитектуры является более глобальное *представление данных*. *Взаимосвязанные киоски данных* могут управляться в рамках того *подразделения*, в котором они создаются.

Реализация такой архитектуры не выдвигает высоких требований к программно-аппаратному обеспечению, и *стоимость* ее может быть невысокой. Однако время реализации будет больше по сравнению с *независимыми киосками данных*. Возрастают также сложность и *стоимость* процедур проектирования.

В заключение следует отметить, что развитие программно-вычислительных средств позволяет создавать так называемые виртуальные ХД, которые работают над OLTP-системами, ХД с многоуровневой архитектурой и так называемые встроенные ХД, которые встраиваются в существующую систему обработки данных организации.

Подходы в организации работ по созданию хранилища данных

Так же, как и для реализации любых типов информационных систем с базами данных, к ХД применимы следующие основные методологические подходы:

- "сверху вниз" (Top down design);
- "снизу вверх" (Bottom down design);
- "из середины" (Middle of design).

На выбор подхода к реализации ХД оказывают влияние следующие факторы:

- состояние текущей информационной инфраструктуры организации;
- имеющиеся в наличии ресурсы;
- требования по возврату инвестиций;
- потребности организации в интегрированном представлении данных о своей деятельности;
- скорость реализации.

Выбор методологического подхода к реализации ХД влияет на объем и тщательность проектирования.

Подход "сверху вниз". Подход "сверху вниз" требует детального планирования и проектирования ХД в рамках ИТ-проекта до начала выполнения проекта. Это связано с тем, что необходимо привлекать всех потенциальных пользователей ХД для выяснения их информационных потребностей в аналитической обработке данных, принимать решения об источниках данных, безопасности, структурах данных, качестве данных, стандартах данных. Все эти работы должны быть документированы и

согласованы. При этом подходе модель ХД должна быть разработана до начала реализации.

Обычно такой подход практикуют при создании *глобального ХД*. Если *киоски данных* включаются в конфигурацию, то они могут быть построены позже.

Достоинством такого подхода является получение более согласованных определений данных и бизнес-правил организации в самом начале работы над созданием ХД. *Стоимость* начального планирования и проектирования может оказаться достаточно высокой. Для этого подхода характерны большие *затраты* времени, что откладывает начало реализации и задерживает возврат инвестиций. Подход "сверху вниз" хорошо применять в организациях с четко организованной информационно-вычислительной структурой, когда программно-аппаратная платформа определена и существуют слаженно работающие источники данных.

Подход "снизу вверх". При использовании подхода "снизу вверх" начинают с планирования и проектирования *киосков данных*подразделений без предварительной разработки глобальной информационно-вычислительной инфраструктуры организации. Это не означает, что такая глобальная *инфраструктура* не будет разработана позже. Такой подход является более приемлемым во многих случаях, поскольку он быстрее приводит к конечным результатам. У него есть и недостатки: данные могут дублироваться и быть несогласованными в разных *киосках данных*. Чтобы избежать этого, необходимо тщательное планирование и проектирование.

Подход "проектирование из середины". Подходы "снизу вверх" и "сверху вниз" могут комбинироваться в зависимости от поставленных перед руководителем проекта по созданию ХД целей. Подход "проектирование из середины" представляет собой комбинацию вышеперечисленных подходов, которые применяются как бы по спирали. Сначала создается *ядро* системы (подход "сверху вниз"), а затем оно поэтапно наращивается за счет добавления новой или дополнительной функциональности (подход "снизу вверх"). Таким образом, на каждом витке спирали может быть использован каждый из двух указанных выше подходов.

Существуют и другие комбинации. Выбор подхода к реализации ХД наряду с выбором *архитектуры ХД* определяет тактические решения в проектировании и управлении проектом создания *системы складирования данных*. К таким решениям относятся планирование реализацией и *управление проектом*.

Характеристика решений ведущих производителей

В настоящем разделе дается краткий обзор решений основных производителей программного обеспечения для разработки ХД. При изложении материала используется, по возможности, следующая схема:

- название проекта компании и его цель;
- архитектурные решения;

- СУБД и используемая модель данных;
- возможности *языка обработки данных*;
- степень охвата жизненного цикла (анализ — проектирование — реализация — поддержка);
- возможные конкурентные преимущества.

IBM. Решение компании *IBM* называется *Data Warehouse* Plus. Целью компании в области разработки и поддержки систем складирования данных является обеспечение пользователя интегрированным набором программных продуктов и сервисов в рамках единой архитектуры.

IBM предлагает встроенную поддержку трех типов архитектурных решений для ХД:

- *независимый киоск данных* ;
- *взаимосвязанные киоски данных* ;
- *глобальное ХД*.

Несущая *СУБД* для ХД — семейство объектно-реляционных *СУБД DB2*. Язык манипулирования данными — *SQL*.

Преимущество решений *IBM* проявляется, когда и системы оперативной обработки данных, и ХД находятся на программном обеспечении *IBM*, т.е. предлагается так называемое замкнутое типовое решение.

С приобретением *компании Informix Software IBM* взяла под свое крыло ряд удачных решений этой компании в области систем складирования данных.

Oracle. Решения, предлагаемые компанией, преследуют две основные цели: предоставление пользователям широкого ассортимента программных продуктов самой компании и *деятельность* партнеров в рамках программы Warehouse *Technology*Initiative.

Компания Oracle не предлагает поддержку каких-либо встроенных архитектурных решений для ХД.

Несущая *СУБД* для ХД — семейство объектно-реляционных *СУБД Oracle* 11g/10g. Язык манипулирования данными — *SQL*. Начиная с версии 8i, диалект *SQL* существенно дополнен набором функций для аналитической обработки данных, вплоть до построения линейной регрессии.

Компания выпускает специальный CASE-инструментарий для проектирования ХД.

Конкурентные возможности *Oracle* определяются следующими факторами:

- имеется набор готовых приложений для разработки ХД, обеспечивающий полный жизненный цикл;
- компания является одним из лидеров по продажам в области анализа данных;

- совместимость с продуктами, производимыми другими компаниями.

NCR. Решение этой компании в области складирования данных ориентировано на организации, у которых имеются потребности в системах *DSS* (система поддержки и *принятия решений*) и системах *OLAP*. Предлагаемая *архитектура* называется *EnterpriseInformation Factory* (*виртуа льное* предприятие).

Несущая *СУБД* для ХД — реляционная *СУБД* Teradata.

Конкурентным преимуществом решений компании является большой *опыт* применения *СУБД* Teradata и связанных с ней методов*параллельной обработки данных*.

SAS Institute. Компания считает себя поставщиком полного решения для организации ХД. Компания предлагает методологию Rapid *Data* Warehousing для быстрого создания и наполнения ХД. В основу этой методологии положено:

- обеспечение доступа к данным в ХД с возможностью их извлечения из разнообразных источников данных (*интероперабельность*);
- преобразование и манипулирование данными в рамках *4GL* (Data Step);
- наличие у компании сервера многомерных БД;
- большой набор программных продуктов компании для аналитической обработки данных и статистического анализа.

Конкурентным преимуществом компании является наличие у нее длинной линейки программных продуктов для статистического и сравнительного анализа данных, который интегрирован в ее методологию построения и использования ХД.

Sybase. Стратегия компании в области ХД основывается на разработанной архитектуре Warehouse WORKS.

Несущая *СУБД* для ХД — реляционная *СУБД* Sybase *System* 11, средство подключения к базам данных OmniCONNECT. Язык манипулирования данными — *SQL* и средства быстрой разработки приложений.

Компания выпускает специальный CASE-инструментарий для проектирования ХД.

Конкурентным преимуществом компании является наличие набора программных продуктов для обеспечения полного жизненного *цикла* разработки ХД.

Microsoft. Компания сравнительно недавно стала активно предлагать комплексные решения в области ХД. Целью корпорации Microsoft является создание инструментальной и технологической среды, которая позволила бы минимизировать *затраты* на создание ХД и сделала бы этот процесс доступным для массового пользователя. Акцент предлагаемых компанией

решений в области складирования данных концентрируется на развитии инструментальных средств *OLAP*.

Корпорация предлагает спецификации среды Microsoft *Data* Warehousing Framework для создания и использования ХД. *Открытость среды* Microsoft *Data* Warehousing Framework обеспечила ее поддержку многими производителями программного обеспечения.

Цель Microsoft *Data* Warehousing Framework состоит в том, чтобы упростить разработку, внедрение и *администрирование* решений на основе ХД. Эта спецификация призвана обеспечить:

- открытую архитектуру, которая интегрируется и расширяется третьими фирмами;
- экспорт и импорт гетерогенных данных наряду с их проверкой, очисткой и ведением истории накопления;
- доступ к *разделяемым метаданным* со стороны процессов разработки ХД.

Несущая *СУБД* для ХД — реляционная *СУБД* MS *SQL Server* 2005/2008. Язык манипулирования данными — *SQL* со встроенными средствами обработки многомерных кубов.

Конкурентным преимуществом компании является наличие у нее набора программных продуктов для обеспечения разработки и поддержки ХД, в том числе для *очистки данных*, при невысокой цене на эти продукты. Ориентация продукции компании на средний и малый бизнес позволяет ей увеличить свои конкурентные преимущества.

Software AG. *Деятельность* компании в области ХД происходит в рамках программы *Open Data Warehouse* Initiative.

Несущая *СУБД* для ХД — сетевая *СУБД* ADABAS. Язык манипулирования данными — *Natural 4GL*.

У компании имеются собственные средства извлечения и анализа данных, а также *программный продукт* управления ХД SourcePoint.

Компания имеет сложившийся круг пользователей и долгое время не проявляла инициативы по переходу на распределенные архитектуры, основанные на компьютерах средней мощности. Компания обладает высоким потенциалом в области систем складирования данных и в последнее время компания наращивает свое участие в этом сегменте рынка.

Типовые программно-аппаратные решения (технологические решения)

Общие типовые решения

Из предыдущих разделов настоящей лекции следует, что существуют несколько вариантов реализации ХД в рамках типовой архитектуры. Рассмотрим некоторые из них.

- **Виртуальное хранилище данных**. Архитектура обеспечивает доступ к "живым" данным в режиме реального времени через *программное обеспечение промежуточного слоя*. В основе такого решения лежит

репозиторий метаданных, который описывает источники данных, процедуры их предварительной обработки и форматы представления информации конечному пользователю. Недостатки такого решения — интенсивный сетевой трафик, снижение производительности несущей системы, угроза нарушения целостности данных в случае неудачных действий пользователей ХД.

- ***Киоски данных*** Архитектура представляет собой облегченный вариант ХД тематической направленности. Бывают *киоски данных*, связанные с интегрированным ХД или несвязанные (автономные).
- **Глобальное хранилище данных**. Архитектура представляет собой единый источник интегрированных данных организации.
- **Хранилища данных с многоуровневой (в основном трехзвенной) архитектурой, или корпоративные ХД**. Архитектура является разновидностью *глобального ХД*, в которую технологически реализуются три уровня (рисунок 3). На первом уровне располагается корпоративное ХД организации. На втором уровне поддерживаются связанные *киоски данных* тематической направленности на основе многомерной СУБД. На третьем уровне находятся клиентские приложения пользователей с установленными на них средствами анализа данных.
- **Встроенные (комбинированные) хранилища данных.** Архитектура представляет собой ХД, которые органически встраиваются в виртуальное предприятие (Enterprise Information Factory, EIF) или используются как компонент аналитической поддержки в информационной реализации бизнес-функций.
- ***Корпоративная информационная фабрика*** (Corporate Information Factory, *CIF*). Эта архитектура является развитием архитектуры корпоративного ХД (*enterprise data* warehouse, EDW). Ее использование предполагает скоординированное извлечение данных из источников, загрузку их в реляционную БД со структурой в *третьей нормальной форме*, использование построенного ХД для наполнения дополнительных репозиториев презентационных данных.
- ***Хранилище данных с архитектурой шины данных*** (*Data Warehouse* Bus). В этой *архитектуре ХД* не является единым физическим репозиторием (в отличие от *CIF*). Это "виртуальное" ХД, представляющее коллекцию витрин данных, каждая из которых имеет архитектуру типа "звезда".
- **Объединенное (федеративное) ХД**. В этой *архитектуре ХД* состоит из ряда экземпляров ХД, которые функционируют на полуавтономной основе и, как правило, организационно или географически разнесены, однако могут рассматриваться и управляться как одно большое ХД.

Рисунок - Хранилища данных с многоуровневой архитектурой ХД

Существенные различия в программном обеспечении у различных производителей определяются следующими факторами: 1) используемая модель данных; 2) степень охвата жизненного цикла; 3) встроенная поддержка различных архитектур; 3) возможности *языка обработки данных*. Можно обратить внимание на следующие две основные тенденции.

- **Производители предлагают комплексные решения по созданию хранилищ данных**. Ведущие производители программного обеспечения в области проектирования и разработки информационных систем с базами данных стараются иметь свои собственные программы по *системам складирования данных* и обеспечивать полный жизненный цикл разработки и сопровождения таких систем.
- **Производители начинают предлагать готовые встроенные архитектурные решения для хранилищ данных**. Это обстоятельство позволяет значительно сокращать время на проектирование и разработку ХД.

С точки зрения применения программно-аппаратных платформ решения в области создания систем складирования данных можно условно разбить на три класса.

1. Комбинация готовых продуктов (решений) разных фирм без непосредственного программирования.
2. Использование полной замкнутой цепочки продуктов (решений) одной фирмы-поставщика.

3. Использование контура продуктов (решений) одной фирмы поставщика с дополнением до замкнутой цепочки совместимыми продуктами третьих фирм.

Простое масштабируемое решение

Пример простого *масштабируемого решения* можно предложить, основываясь на использовании Crystal Enterprise и Crystal Reports (фирма Business Objects) как инструментов конечного пользователя. Подробнее о возможностях Crystal Enterprise и Crystal Reports можно прочитать в литературе к курсу настоящих лекций.

ХД реализуется на СУБД Oracle, DB2, MS SQL Server или других, имеющих ODBC-интерфейс или интерфейс прямого доступа с Crystal Enterprise. Обычно применяется классическая *архитектура ХД* без *киосков данных*. Для этого решения большое значение имеет тщательное проектирование структуры ХД и запросов. Необходимо разработать и создать приложения для *очистки данных*(или воспользоваться имеющими у поставщиков средствами).

Преимущества

- Сводится к минимуму объем программирования, т.к. все стадии покрываются готовыми коробочными продуктами.
- Сокращается время разработки и создания ХД (за счет исключения трудоемкого процесса написания программ).
- Время разработки типового запроса — от 2-х до 6-ти часов, время разработки типового отчета – 1-2 дня.
- Такое решение хорошо для создания прототипов ХД, поскольку в данном случае отрабатываются практически все необходимые запросы и отчеты.
- Создается прекрасная инструментальная среда для использования нетиповых запросов.
- Такое решение прекрасно подходит и для создания виртуальных ХД.

Недостатки

- Разработка сложных *перекрестных запросов* может занять много времени.
- Это решение не подходит для сложной аналитической обработки данных, требующей разработки специальных приложений для анализа.

Замкнутое типовое решение

Замкнутое типовое решение можно предложить на основе использования замкнутой цепочки продуктов одной фирмы-поставщика, например Microsoft (рисунок 4), Oracle (рисунок 5), SAS или Sybase.

Рисунок - Типовое решение на основе продуктов Microsoft

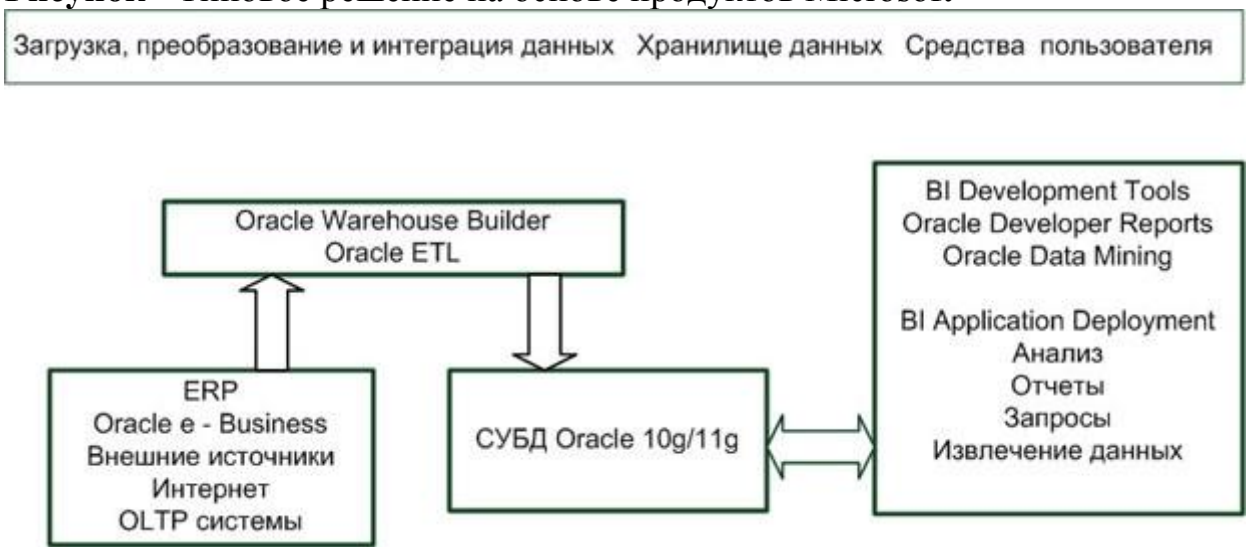

Рисунок 5 - Типовое решение на основе продуктов Oracle

Преимущества

- Как правило, все бизнес-направления поддерживаются за счет готовых сервисов.
- Время разработки и создания ХД поддается строгому описанию и достаточно точной оценке.
- Такое решение хорошо для создания ХД, которые предполагается использовать в организации длительное время.
- Такие решения подходят для сложной аналитической обработки данных, требующей разработки специальных приложений для анализа.

Недостатки

- Главным недостатком является высокий уровень затрат на разработку и создание, который при правильной организации проекта окупается.

- Кадровый вопрос: необходимо нанимать высококвалифицированные кадры, умеющие работать с набором продуктов выбранной компании. Как правило, обучение своих сотрудников по всем направлениям работы с ХД малоэффективно, хотя и привлекательно.

Области применения технологии хранилищ данных

Концепция хранилищ данных находит применение во многих сферах бизнеса, науки и управления. Рассмотрим типовые решения для бизнеса. Такие типовые решения использования технологии складирования данных в бизнесе можно разделить на следующие основные группы.

1. Разработка основы для создания аналитических подсистем сопровождения бизнеса.
2. Разработка ХД как составной части виртуального предприятия.
3. Разработка ХД для цифровых (электронных) библиотек и мультимедиа.

Основные сферы применения технологии складирования данных приведены в таблице 1. Имеется тенденция расширения проникновения концепции в те сферы бизнеса, где необходимо выполнять, с одной стороны, сравнительный анализ, искать зависимости в данных, выявлять тренды в рядах динамики, а с другой – использовать *системы складирования данных* в связке с системами операционной обработки.

Таблица 1 - Области применения концепции складирования данных

№	Сфера деятельности	Комментарий
1	Сегментация рынка	CRM
2	Планирование продаж, прогнозирование и управление	CRM, *SCM*
3	Опека клиентов	CRM
4	Схемы лояльности	
5	Проектирование и разработка продуктов	*MRP*/ERP
6	Интеграция цепочки поставок	*SCM*, ERP/*MRP*, SCP, SCE, DRP, JIT
7	Инновации и новые возможности	
8	Новые возможности приложений с использованием Интернет/Интранет	eBusiness, TMP
9	Приложения, основанные на агентах программного обеспечения	
10	Приложения для *извлечения знаний* и кибер-организация	EIF, виртуальное предприятие
11	Распространение DW из области стратегического планирования в область бизнес операций	VDW
12	Приложения для вертикальных секторов	CRM, TMP

	индустрии	
13	Готовые DW (off-the-shalf)	
14	Автоматизация принятия решений	DSS, *EIS*
15	Новые категории оперативных приложений, ориентированные на клиента	OLAP
16	Сбор и анализ экспериментальных данных в химии, физике, биологии	EDW
17	Хранение мультимедийной информации в DW	DL

Сокращения, использованные в колонке "Комментарий" таблицы 1 и не поясненные ранее в тексте, имеют следующие значения:

- CRM (Customer Relationship Management) – управление взаимоотношениями с клиентами;
- *SCM* (Supply Chain Management) – управление цепочкой поставок;
- SCP (Supply Chain Planing — планирование управления цепочкой поставок;
- SCE (Supply Chain Executing) — реализация управления цепочкой поставок;
- DRP (Distribution Resource Planing) — планирование потребностей распределения;
- JIT (Just-in-Time) — точно в срок;
- *MRP* (Manufacturing Resource Planing) – планирование материальных затрат;
- VDW (Virtual *Data Warehouse*) – виртуальные хранилища данных;
- DL (Digital Library) – цифровые библиотеки;
- ERP (Enterprise Resource Planing) – системы планирования масштаба предприятия;
- TMP (Trading Partner Management) – управление деловыми партнерами;
- EIF (Enterprise Information Factory) – виртуальное предприятие.

Рассмотрим несколько примеров применения технологии складирования данных в области создания аналитических подсистем информационного сопровождения бизнеса.

Аналитические CRM-системы

Оперативные системы CRM содержат следующие компоненты: центры обработки мобильных сообщений, данные по обслуживанию клиентов, данные из отдела продаж, данные о продажах через интернет-магазины, данные ERP систем, данные из *ИСР* (*EIS*) и других внешних источников. Эти системы выступают источниками данных для аналитических CRM. Типовая структура аналитического ХД CRM-системы приведена на рисунке 6.

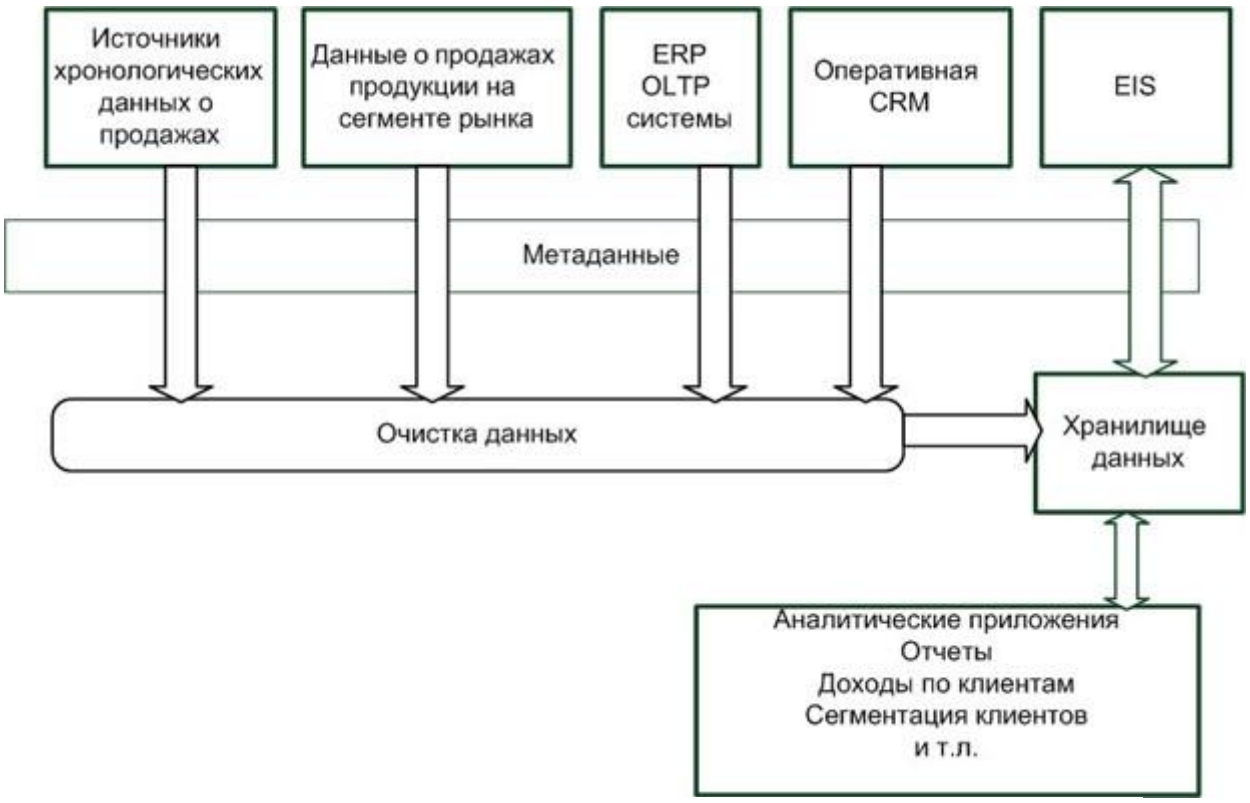

Рисунок - Архитектура аналитической CRM-системы

Внедрение такого решения позволяет оптимизировать цепочки работы с клиентами, провести персонализацию обслуживания клиентов, повысить доходы от продаж, а также позволяют разрабатывать стратегии расширения рынка за счет привлечения клиентов на основе индивидуального подхода.

Наиболее известное работающее решение в области аналитических CRM в телекоммуникациях имеет компания SAS Institute (US WEST Communications).

Аналитические SRM-системы

Аналитические *SRM* (Supply Relationship Management) системы занимаются управлением взаимоотношениями с поставщиками. Пример типовой архитектуры для ХД аналитических *SRM* систем приведен на рисунке 7.

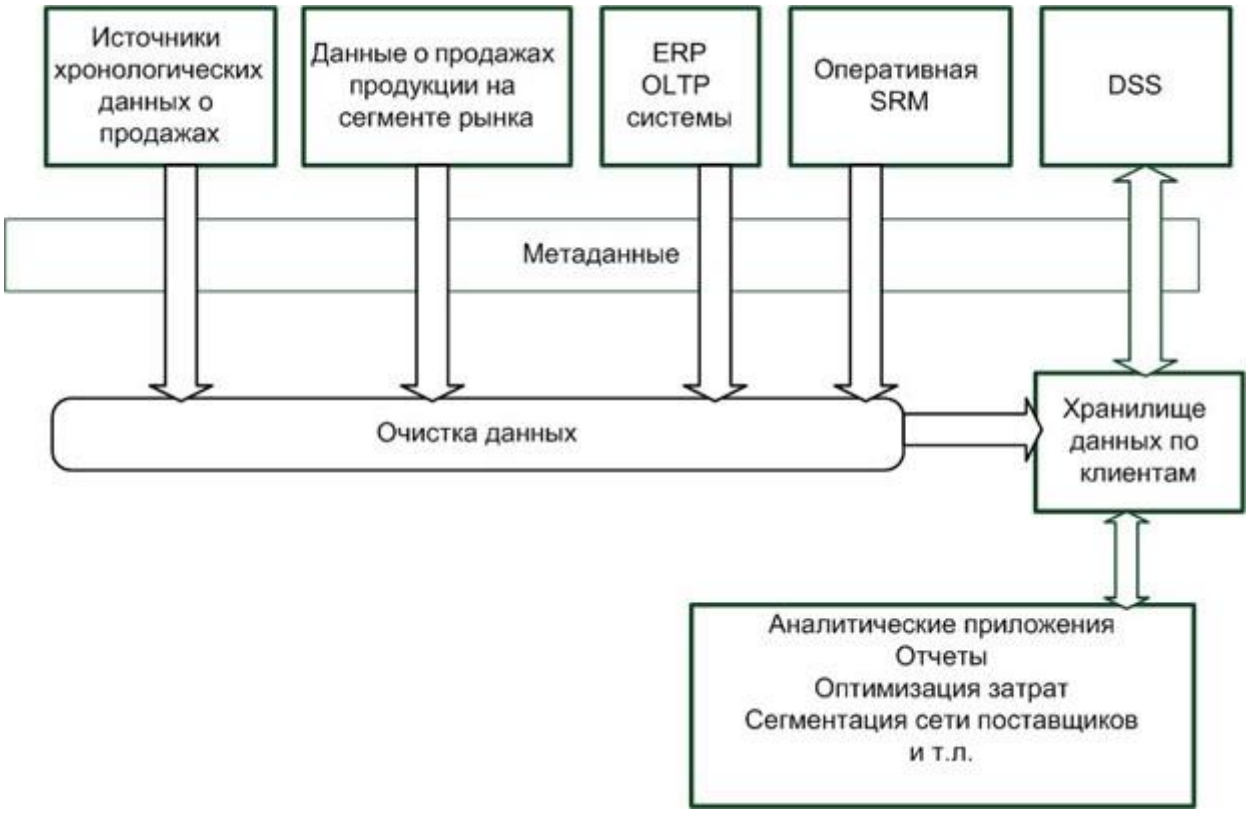

Рисунок - Архитектура аналитической SRM-системы

Конкурентные преимущества

- Снижение затрат (от 5 до 15%), потока сырья, планирования, исполнения и контроля прохождения.
- Повышение эффективности стратегии бизнеса в области управления финансовыми, материальными и информационными потоками
- Создание оптимальных циклов поставок.
- Оптимизация бизнес процессов на уровне работы с поставщиками.
- Сокращение времени поставок.
- Увеличение прибыли (от 5 до 15%)

Сопутствующие проблемы

- При использовании отдельных *SRM*-решений возможен конфликт с другими решениями.
- Возникает ряд сложностей с обучением персонала.
- Сопротивление поставщиков и дистрибьютеров.

Наиболее известное решение в области создания аналитических *SRM*-систем разработано компанией SAS Institute.

Аналитические SCM-системы

Аналитические *SCM*-системы, не встроенные в ERP-системы, представляют собой информационные системы для решения задач анализа и оптимизации в управлении жизненным циклом продукции. Пример типовой архитектуры для ХД аналитической *SCM*-системы приведен на рис. 2.8.

Рисунок - Архитектура аналитической SCM-системы

Достоинства использования SCM-решений

- Минимизация издержек сети сбыта.
- Снижение затрат, оптимизация потоков сырья, материалов, незавершенного производства, готовой продукции и услуг в результате планирования, исполнения и контроля от точки зарождения заявки до полного удовлетворения требований клиента.
- Повышение эффективности стратегии бизнеса в области управления финансовыми, материальными и информационными потоками
- Создание оптимальных жизненных циклов производства.
- Оптимизация бизнес-процессов на всех уровнях предприятия, начиная с поставки.

- Сокращение времени внедрения новых производственных технологий.

Сопутствующие проблемы

- При использовании *SCM*-решений возможен конфликт с другими решениями.
- Возникает ряд сложностей с обучением персонала.
- Сопротивление поставщиков и дистрибъютеров.

Конкурентные преимущества

- Уменьшение стоимости и времени обработки заказов (от 20 до 40%).
- Сокращение времени выхода на рынок (от 15 до 30%).
- Сокращение закупочных издержек (от 5 до 15%).
- Уменьшение складских запасов (от 20 до 40%).
- Сокращение производственных затрат (от 5 до 15%).
- Увеличение прибыли (от 5 до 15%).

По уровню использования *SCM*-решений телекоммуникации занимают второе место в мире (после нефти и газа). Перечень наиболее удачных решений в области оперативных *SCM*-систем приведен в таблице 2.

Таблица 2 - Решения в области оперативных *SCM*

Компания	**Программные продукты**
IBM	WebSphere (for e-Business), интеграция с ERP
SAP	Business Information WareHouse, SAP Advanced Planer & Optimizer *Logistics* Execution System
BAAN	IBAAN с совокупностью модулей в архитектуре ПО BAAN, в том числе и использованием хранилища данных

Виртуальные предприятия

Одной из перспективных областей применения систем складирования данных является разработка ХД как составной части виртуального предприятия. В этом случае ХД рассматривается как часть интегрированной информационной структуры организации, которая имеет типовую архитектуру, показанную на рисунке 9.

Рисунок - Место хранилища данных в виртуальном предприятии

Мультимедийные хранилища данных

Очень перспективным в последнее время становится разработка ХД для цифровых (электронных) библиотек и мультимедиа. Современные СУБД имеют ряд встроенных возможностей для хранения и выборки мультимедийных данных (например СУБД *Pilot*). Однако большинство решений по созданию мультимедийных баз данных реализуется на реляционных СУБД, обладающих возможностью работы с BLOB-данными и имеющими поддержку очень больших БД. Типичными представителями таких СУБД являются СУБД Oracle (имеет специальные средства выборки визуальной информации — VIR и интернет-систему обработки файлов iFS), DB2 и Informix (теперь IBM).

Примерами мультимедийных ХД являются разрабатываемые во всем мире электронные хранилища музейных данных (образы картин и других экспонатов).

Обсудим особенности типового решения создания мультимедийных ХД на основе реляционных СУБД. Следует отметить следующие свойства медиаданных:

- неструктурированная форма с точки зрения *теории реляционных баз данных*;
- размер элемента медиаданных очень большой;

- данные не имеют фиксированного максимального размера;
- внутренний формат для представления таких данных не может быть выражен простым типом данных реляционных СУБД;
- поиск данных затруднен или просто невозможен стандартными средствами СУБД.

С точки зрения разработки хранилищ мультимедийных данных следует отметить одно важное обстоятельство: измерения, в большинстве практических случаев, выражаются через простые типы данных, что значительно облегчает разработку хранилищ таких данных.

В этом отношении хранилище мультимедийных данных имеет типовую архитектуру, в которой медиаданные быстро извлекаются и визуализируются. Задачи сравнительного анализа медиаданных зависят от предметной ориентации ХД и требуют обычно специально разработанных процедур.

Преимущество

- Медиаданные классифицируются по иерархическим категориям и вводятся в ХД, что увеличивает скорость их выборки.

Сопутствующие проблемы

- Высокие требования к аппаратным решениям.
- Разработка систем классификации медиаданных.
- Разработка процедур и программ поиска медиаданных и их анализа.

Корпоративные информационные фабрики

В настоящее время в кругу бизнес-пользователей информационных технологий обсуждается предложенная Биллом Инмоном концепция так называемой *корпоративной информационной фабрики* (*Corporate Information Factory, CIF*) как одной из основополагающих вычислительных архитектур для производства информационных продуктов предприятия. Для любого предприятия реализацию такой концепции можно рассматривать как важную *перспективную задачу*, решение которой не только позволит повысить качество управления взаимоотношениями с внешними организациями (налоговыми и финансовыми государственными структурами) и партнерами, но и значительно увеличить *производительность* его подразделений, поставляющих информацию, необходимую для принятия стратегических решений.

Рассмотрим более подробно концепцию *CIF*.

Производству данных — свою технологию

Корпоративная информационная фабрика — это логическая архитектура программно-аппаратного решения по производству, складированию, управлению и доставке данных для поддержки принятия стратегических и тактических решений в масштабе организации. Концепция *CIF*, предложенная классиком в области теории хранилищ

данных Биллом Инмоном в серии его *работ*, подразумевала системно организованное взаимодействие репозиториев оперативных данных (*Operational Data Store*), центрального ХД, витрин данных и системы *интеллектуального анализа данных* (*Data Mining*) за счет создания технологических цепочек переработки и доставки данных.

В абстрактной форме процесс производства информации в *CIF* был представлен в аналогии с производством некоторого продукта. В соответствии с этим были выделены основные стадии производства информации (новых данных): получение исходных данных (сырья), их преобразование (производство отдельных деталей), складирование данных, создание информационных продуктов (из деталей готовой продукции) и доставка данных их потребителям (распределение конечной продукции).

Основная идея, положенная в основу концепции *CIF*, состоит в выделении элементов информационной архитектуры на основе их функционального назначения и регламентирования технологических процедур обработки данных.

Краеугольным камнем правильно спроектированной *CIF* являются, безусловно, *метаданные*. Задача этого слоя — описать в рамках единой терминологической базы (*метаданные* бизнес-пользователя) всю совокупность объектов управления средой *CIF*(*метаданные* администрирования). Только подход "от метаданных" позволяет из гетерогенного потока *входной* информации получить однородное описание среды и *предметной области*, что дает возможность одинаково легко обращаться к измерениям, кубам, отчетам и бизнес-объектам на основе произвольных выборок. Таким образом, обеспечивается высокое качество циркулирующей в *CIF* информации.

Структурные компоненты CIF

В основе *CIF* лежит модель функционального разделения процессов производства новых данных (информационных продуктов) и доставки информационных продуктов их потребителям, а также управления этими процессами.

Производители информационного продукта собирают данные из доступных источников (чаще всего из оперативных систем ввода и обработки данных), преобразуют и интегрируют их, размещая в *системе складирования данных* в унифицированном регламентированном формате. Потребители информационных продуктов извлекают необходимые тематические выборки из*системы складирования данных* (через специализированные предварительно настроенные интерфейсы — витрины данных) и затем используют их в процессе принятия решений.

Логическая структура *CIF* включает в себя несколько типовых архитектурных элементов (таблица 3).

Таблица 3 - Типовые архитектурные элементы логической структуры *CIF*

Элемент	Характеристика
Системы, доставшиеся "по наследству" (*Legacy Systems*)	Поддерживают бизнес-функции, которые были созданы в организации ранее. В таких системах обычно компоненты, обеспечивающие формирование отчетов и ввод и передачу данных, реализуются в рамках единого программного блока, что затрудняет решение задач по интеграции и преобразованию данных в соответствии с новыми требованиями бизнеса
Приложения оперативного управления организацией (OLTP)	Обеспечивают быструю обработку данных в рамках бизнес-направлений деятельности организации. Как правило, такие системы приобретаются у компании-разработчика, которая осуществляет их техническую поддержку
Оперативные склады данных (Operational*Data Store* - ODS)	Этот элемент наделяется свойствами как оперативных, так и аналитических систем. Основное его назначение - обеспечить осуществление анализа информации практически сразу после ее обновления в оперативных системах
Компоненты преобразования данных (*ETL*-tools, Staging Area, Near-line Storage)	Служат для перегрузки данных из одних программных компонентов в другие (с промежуточной очисткой и согласованием данных, получаемых из различных источников)
Корпоративное хранилище данных (*Enterprise Data*Warehouse)	Здесь накапливается детальная информация, необходимая для выполнения анализа. Данные перегружаются в корпоративное хранилище из оперативных элементов - унаследованных систем, автоматизированных *банковских систем* или оперативных складов данных. Как правило, обновление информации в EDW происходит с большой задержкой. Для разрешения этой проблемы используются ODS-элементы
Витрины данных (*Data Marts*)	Предназначены для хранения аналитической информации уровня подразделения или направления бизнеса
Приложения поддержки принятия решений (DSS) и приложения анализа данных (DM)	DSS, примером функционала которых могут быть системы анализа клиентской базы банка, обеспечивают поддержку принятия решений. Разнообразный статистический анализ выполняется в DM
Инфраструктура	Обеспечивает публикацию данных в сети Интранет

сетевых коммуникаций	(Интернет), а также обработку результатов ввода информации пользователями

CIF на предприятии

На предприятии производственные и финансовые потоки тесно взаимосвязаны с потоками информационными, которые отражают их динамические показатели и текущее состояние. Кроме того, такие информационные потоки являются источником данных для анализа при определении трендов изменений и их количественных характеристик.

Описанная выше в общих чертах схема превращения данных в информационные продукты и составляет суть концепции *CIF* на любом предприятии (рисунок 10).

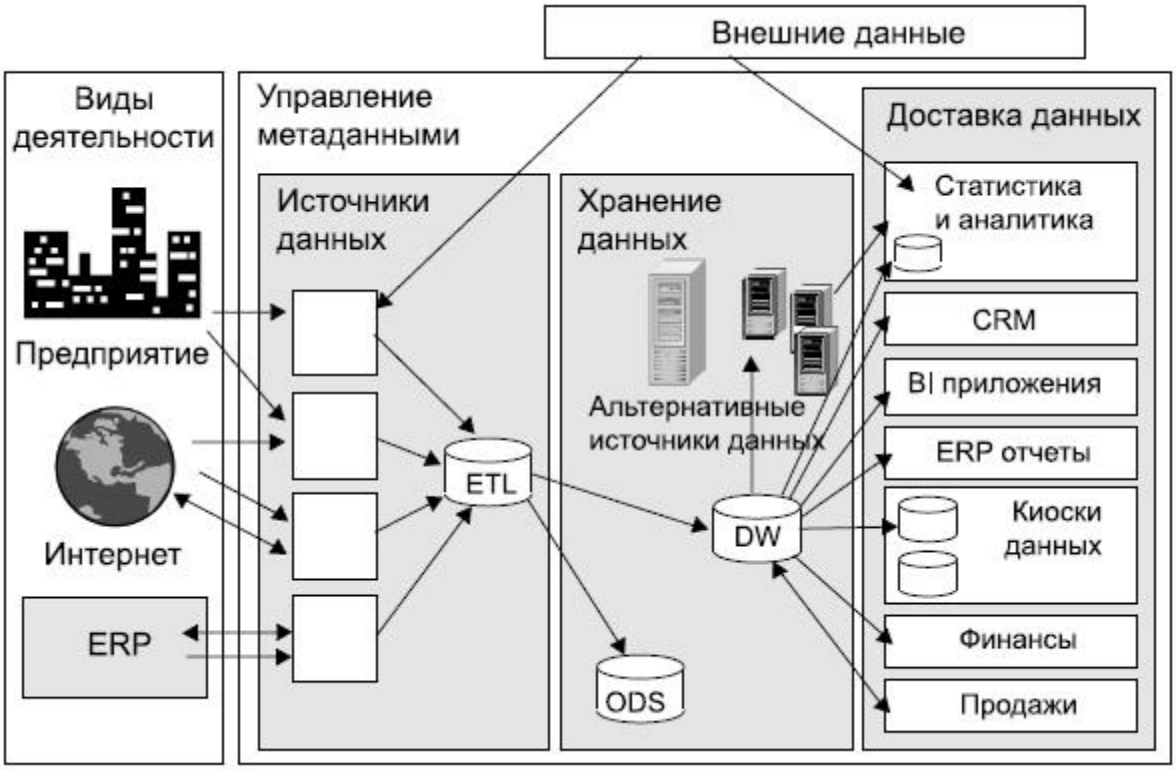

Рисунок - "Корпоративная информационная фабрика"

Хранилище данных — фундамент CIF предприятия

Складирование данных — это технология, с помощью которой можно оперативно собрать данные и на их основе решать разнообразные задачи по финансовому планированию, бюджетированию, риск-менеджменту, анализу взаимоотношений с партнерами, маркетинговому анализу и т.д. Однако самое главное преимущество отлаженной архитектуры *CIF* в другом: она позволяет адаптировать вычислительную среду как под четко определенные информационные потоки небольшого предприятия, так и под сложные схемы консолидации, которые характерны для предприятий с развитой филиальной

структурой и входящих в состав холдингов и отраслевых объединений предприятий.

Рассмотрим подробнее, как "фабрика управленческих данных" функционирует на предприятии.

ERP/MRP II системы как источники данных для CIF

Первоначальное наполнение корпоративного ХД и постоянное поддержание его в актуальном состоянии — это отнюдь не тривиальные задачи. Особые требования здесь предъявляются к качеству информации, кроме того, высока степень риска — ошибочные решения на основе неверных исходных посылок могут обернуться серьезными потерями.

На предприятиях основными источниками данных являются ERP-системы. Они представляют собой семейство оперативных приложений, обеспечивающих обработку производственных и финансовых данных, включая выполнение бухгалтерских проводок, логистических операций, генерацию текущей оперативной отчетности. Модули ERP ориентированы на те информационные продукты, которые они сопровождают или поддерживают. Разумеется, ERP не предназначены для обработки информации в историческом аспекте и не имеют развитого инструментария для агрегации и систематизации данных предприятия. Из-за строгой предметной направленности у подсистем ERP, как правило, слабо развиты взаимосвязи на уровне данных: обычно у них информационный обмен осуществляется небольшими объемами.

Таким образом, на первом шаге построения *CIF*-системы источники данных накапливают информацию в масштабе предприятия в "сыром" виде: она не подготовлена для анализа и компиляции аналитической отчетности.

Интеграция и преобразование данных

Организация процесса интеграции является еще одним фактором успеха в создании *CIF*: информация извлекается из разнородной вычислительной среды ERP, преобразуется с целью повышения ее качества и складируется. Все это делается для того, чтобы системы поддержки и принятия решений могли в дальнейшем ее активно использовать.

Для наполнения корпоративного ХД в нем обычно предусматриваются инструментальные средства:

- для извлечения и доставки из различных оперативных БД и внешних источников;
- для очистки, преобразования и интеграции;
- для загрузки;
- для актуализации.

Хранилище данных

ХД — это предметно-ориентированная, интегрированная, неизменяемая и поддерживающая хронологию коллекция данных, используемая для поддержки принятия решений. С позиций *CIF* хранилище является отправной точкой при преобразовании данных в информационные

продукты (аналитические отчеты и пр.). Оно всегда предоставляет своим потребителям проверенные и согласованные данные по всей организации в целом, независимо от источника их происхождения.

Управление данными

Процесс управления данными предусматривает комплекс процедур, отвечающих за прохождение информации в *CIF*. Он включает в себя архивацию и восстановление данных, секционирование, управление перемещением данных в системе, агрегацию и т.д.

Инструментальные средства для производства информационных продуктов

В конечном итоге, как мы помним, информация должна попасть к потребителю в заданном виде, чтобы послужить базисом для принятия взвешенных управленческих решений. Логично на выходе *CIF* применять:

- средства для многомерного представления данных и манипулирования ими;
- средства для формирования отчетов;
- систему информационных запросов.

В качестве отличительных характеристик подхода Билла Инмона к *архитектуре ХД* можно назвать следующие.

1. Использование реляционной модели организации атомарных данных и многомерной модели — для организации суммарных данных.
2. Использование подхода "проектирование из середины" при создании больших ХД, что позволяет создавать ХД поэтапно.
3. Использование *третьей нормальной формы* для организации атомарных данных, что обеспечивает высокую степень детальности интегрированных данных и, соответственно, предоставляет корпорациям широкие возможности для манипулирования ими и изменения формата и способа представления данных по мере необходимости.
4. ХД — это проект корпоративного масштаба, охватывающий все отделы и обслуживающий нужды всех пользователей корпорации.
5. ХД — это не механическая коллекция витрин данных, а физически целостный объект.

Хранилища данных с архитектурой шины данных

В данной *архитектуре ХД* с архитектурой шины данных, предложенной Ральфом Кимболлом, первичные данные преобразуются в необходимые структуры на стадии подготовки данных. При этом обязательно принимаются во внимание требования к скорости обработки информации и качеству данных. Подготовка данных начинается со скоординированного извлечения их из источников. Ряд операций совершается централизованно,

например, поддержание и хранение общих справочных данных, другие действия могут быть распределенными.

ХД с архитектурой шины данных изначально ориентированы на использование многомерной модели данных (см. следующие лекции). Поэтому, как правило, данные в его структуре денормализованы, чтобы оптимизировать выполнение запросов. Запросы в процессе выполнения обращаются к все более низкому уровню детализации без дополнительного перепрограммирования со стороны пользователей или разработчиков приложения.

В отличие от *корпоративной информационной фабрики*, в *ХД с архитектурой шины данных* чаще используются связанные *киоски данных*, которые разрабатываются для обслуживания бизнес-процессов (бизнес-показателей или бизнес-событий), а не направлений бизнеса. Например, данные о заказах, которые должны быть доступны для общекорпоративного использования, вносятся в ХД только один раз, в отличие от *CIF*, в котором их пришлось бы трижды копировать в витрины данных отделов маркетинга, продаж и финансов. После того, как в ХД появляется *информация* об основных бизнес-процессах, консолидированные *киоски данных* могут выдавать их перекрестные характеристики. *Матрица* шины данных корпоративного ХД с архитектурой шины выявляет и усиливает связи между показателями бизнес-процессов (фактами) и описательными атрибутами (измерениями).

ХД с архитектурой шины данных состоит из набора *взаимосвязанных киосков данных*, которые созданы для обслуживания бизнес-процессов организации (См. рисунок 2.11).

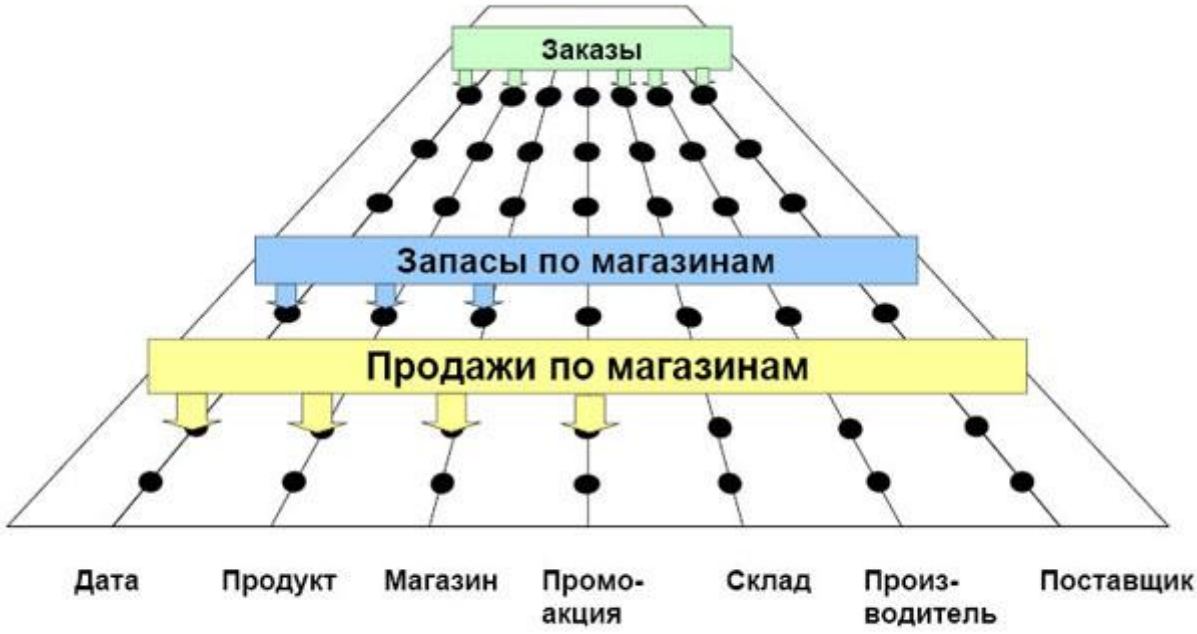

Рисунок - Хранилище данных с архитектурой шины данных

Суммируя все вышесказанное, можно отметить типичные характеристики *ХД с архитектурой шины данных*.

1. Использование многомерной модели организации данных с архитектурой "звезда" (star scheme).
2. Использование двухуровневой архитектуры, которая включает стадию подготовки данных, недоступную для конечных пользователей, и собственно ХД с архитектурой шины. В состав последнего входят несколько киосков атомарных данных, несколько киосков агрегированных данных и персональный *киоск данных*, но оно не содержит одного физически целостного или *централизованного ХД*.
3. ХД не является единым физическим репозиторием (в отличие от *CIF*). Это "виртуальное" ХД, представляющее коллекцию витрин данных, каждая из которых имеет архитектуру типа "звезда".

Отметим, что и *корпоративная информационная фабрика*, и *ХД с архитектурой шины данных* имеют своей целью создание корпоративного ХД. Соответственно, единство конечного объекта означает общность требований, которым должен удовлетворять любой подход для достижения искомого конечного результата, а это, в свою *очередь*, указывает на то, что и в самой архитектуре должны быть общие черты.

Обе эти архитектуры отличаются в основном способами представления данных. В *CIF*, они, как правило, нормализованы, а в *ХД с архитектурой шины данных* — нет.

Объединенное (федеративное) ХД

Для любой организации, особенно многофилиальной, наличие согласованной управленческой информации, необходимой для четкого понимания того, как функционирует бизнес, является одной из актуальных задач.

Обычной подход к улучшению информированности о бизнес-операциях — проведение стандартизации "сверху вниз" как структуры отчетности, так и модели данных. Однако с практической точки зрения стандартизация бизнес-структур оказывается для большинства организаций малоэффективной — требуется слишком много средств и времени.

В качестве одного из подходов для решения указанной проблемы может бытъ использована *архитектура* **федеративного ХД** (рисунок 12). В этой архитектуре на основе иерархии связанных ХД можно обмениваться данными, бизнес-моделями и структурами отчетности, благодаря чему возможно, с одной стороны, осуществлять общий *контроль* и предусмотреть определенную степень стандартизации, а с другой — позволить региональным отделениям сохранить автономность и учесть местную специфику.

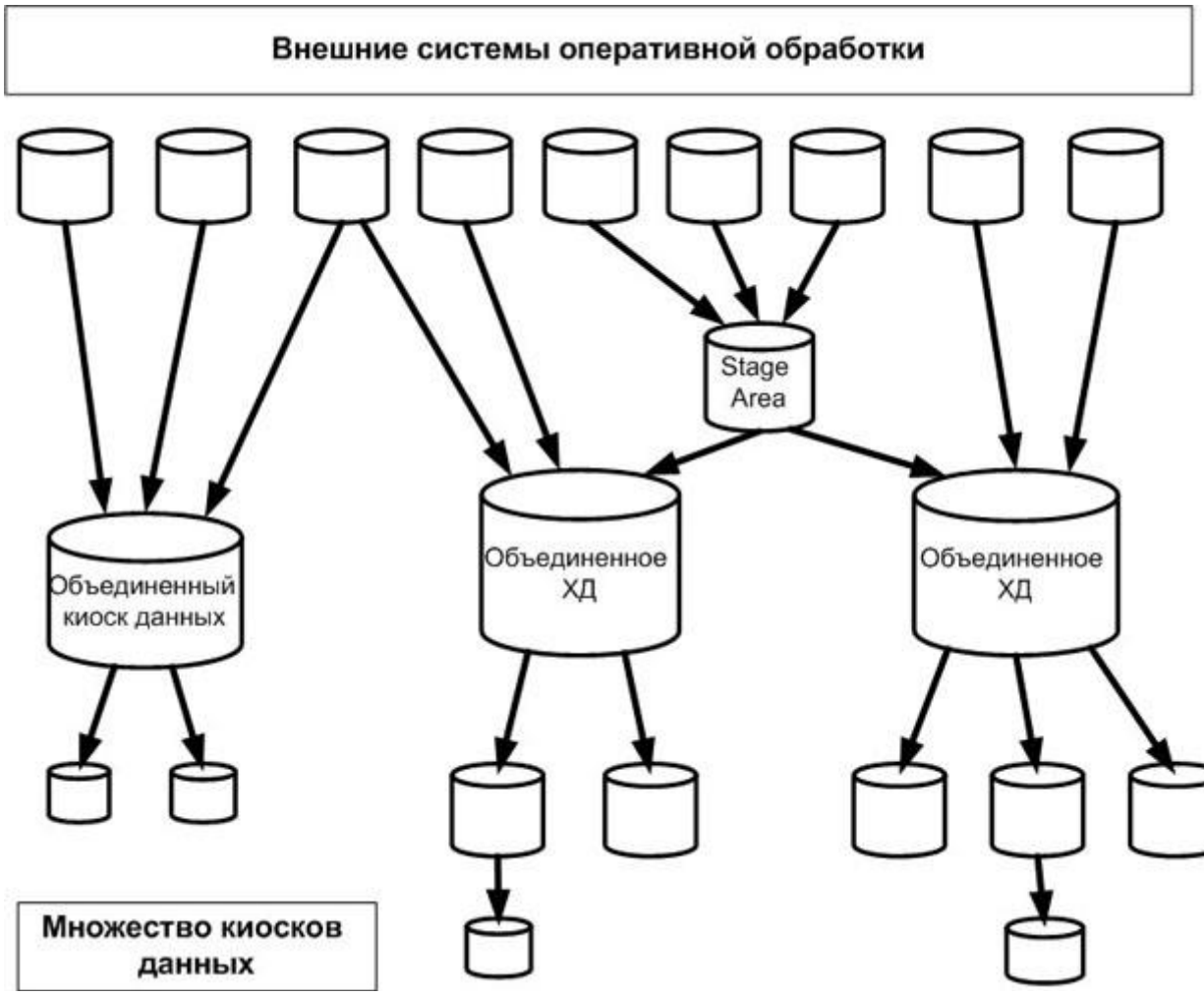

Рисунок - Федеративное хранилище данных

Система объединенных ХД характеризуется совместным использованием общих информационных точек, что устраняет, таким образом, *избыточность* и гарантирует *достоверность* информации по всей организации (рисунок 2.12). **Федеративное ХД состоит из ряда экземпляров ХД, которые функционируют на полуавтономной основе и, как правило, организационно или географически разнесены, однако могут рассматриваться и управляться как одно большое ХД**. Применение такой архитектуры снижает риск неудачи при глобальном развертывании системы, поскольку каждое локальное ХД меньше по масштабу, отвечает местным требованиям бизнеса и может управляться сотрудниками регионального *подразделения*.

Каждый из экземпляров федеративного ХД хранит копию базовой бизнес-модели и общие основные данные (common master dat), причем каждое ХД более высокого уровня содержит итоговые транзакционные данные более низкого уровня. Общие основные данные — например, схема организационной структуры компании — отправляются "вниз", т.е. из

корпоративного (глобального) ХД, а суммарные данные о транзакциях отправляются "верх", т.е. из локального ХД. Таким образом, "федерация" ХД может предоставить местным отделениям необходимую гибкость, а также обеспечить общий *контроль* и согласованность; при этом каждое ХД функционирует независимо от всех других остальных.

Для федеративных ХД характерны общая *семантика* и бизнес-правила, стандартизованный набор процессов извлечения из (о существовании бизнес-правил как таковых было сказано строкой выше) бизнес-правил, децентрализованные ресурсы и управление, параллельная разработка.

При этом следует учитывать, что важна необходимость в координировании *работ*, требуется согласованность среди различных отделов по вопросам архитектуры, бизнес-правил и семантики, сложная технологическая информационно-вычислительная среда.

Резюме

Компонентами типовой *архитектуры ХД* являются:

- *программное обеспечение промежуточного слоя*. Основное назначение этих компонент состоит в обеспечении доступа к сети и доступа к данным;
- БД OLTP систем и данные внешних источников;
- предварительная обработка и загрузка данных;
- ХД, реализованное средствами СУБД;
- метаданные, которые играют роль справочника о данных;
- уровень доступа к данным — программное обеспечение, которое обеспечивает взаимодействие конечных пользователей с данными ХД;
- уровень информационного доступа, который обеспечивает непосредственное общение пользователя с ХД;
- уровень администрирования.

Отметим, что в последнее время возрастает практический интерес к использованию ХД при формировании информационной инфраструктуры организаций. Преимущества, которые получает организация от внедрения хранилищ данных, следующие.

- *Взгляд на данные организации, как на единое целое*. Это ответы на такие вопросы: сколько продуктов реально производится? Что влияет на изменение спроса? Какие товары или услуги приносят наибольший доход? А также возможность учитывать особенности и предпочтения клиентов.
- *400% возврат инвестиций, вложенный в создание хранилища данных* (по результатам трехлетнего исследования опыта 62-х корпораций, проведенного IDC). Сроки исполнения — от 6 месяцев до 2-х лет в зависимости от объема хранилища данных, при следующем распределении затрат: для небольшого подразделения — $ 400000-600000, для большого подразделения — $800000-1500000, для большой корпорации — $15000000.

- *Возрастает надежность данных для принятия решений*. Данные, загружаемые в хранилище данных, подвергаются очистке — согласуются, проверяются, уточняются.
- *Геопространственный анализ данных*. Анализ такой информации имеет решающее значение в принятии решений по всем вопросам, связанным с географией бизнеса.
- *Исследование трендов и колебаний в бизнес-данных*. Позволяет достаточно надежно прогнозировать развитие бизнес-процессов организации во времени.