

**МИНИСТЕРСТВО ОБРАЗОВАНИЯ АЗЕРБАЙДЖАНСКОЙ
РЕСПУБЛИКИ
АЗЕРБАЙДЖАНСКИЙ ГОСУДАРСТВЕННЫЙ
ЭКОНОМИЧЕСКИЙ УНИВЕРСИТЕТ**

МАГИСТРАТУРА

На правах рукописи

Аббасов Аббас Аскер оглы

**ТЕМА: «РАЗРАБОТКА МЕТОДОЛОГИИ СБОРА,
УПОРЯДОЧЕНИЯ И ОБРАБОТКИ ДАННЫХ, ПОСТУПАЮЩИХ ИЗ
АВТОНОМНЫХ ИНФОРМАЦИОННЫХ СИСТЕМ ДЛЯ
КОНСТРУИРОВАНИЯ ХРАНИЛИЩ ДАННЫХ»**

МАГИСТЕРСКАЯ ДИССЕРТАЦИЯ

Название и шифр	ИМ 020000	Экономика
направления:		предприятия и управление
Название и шифр	ИМ 020004	Информационные
специальности:		технологии управления

Научный руководитель

Проф. А.К. Керимов

Заведующий кафедрой

Акад. А.М. Аббасов

БАКУ-2015

СОДЕРЖАНИЕ

РЕФЕРАТ.....	3-5
ВВЕДЕНИЕ.....	6-7
ГЛАВА I. КОНЦЕПЦИЯ СИСТЕМ СКЛАДИРОВАНИЯ ДАННЫХ	8-33
§1.1. Предпосылки создания концепции складирования данных...	8-15
§1.2. Системы поддержки и принятия решений и управленческие информационные системы.....	14-17
§ 1.3. Концепция хранилищ данных.....	18-33
ГЛАВА II. АРХИТЕКТУРА ХРАНИЛИЩ ДАННЫХ	34-68
§2.1. Типы хранилищ данных.....	35-39
§2.2. Выбор архитектуры ХД и методы его проектирования.....	40-48
§ 2.3. Модель типового проекта создания хранилища данных.....	49-57
§ 2.4. Разработка требований.....	58-68
ГЛАВА III. СИСТЕМЫ ДЕЛОВОЙ ОСВЕДОМЛЕННОСТИ И ХРАНИЛИЩА ДАННЫХ.....	69-83
§ 3.1. Системы бизнес-аналитики.....	70-80
§ 3.2. Интеллектуальная обработка данных (Data Mining).....	81-83
ЗАКЛЮЧЕНИЕ	84
ЛИТЕРАТУРА	85

РЕФЕРАТ

Актуальность темы. Хранилища данных — одна из самых актуальных тем в современной индустрии информационных технологий. Как и в случае с другими актуальными темами, поставщики стараются не отстать от моды и предлагают решения проблем, о существовании которых несколько лет назад мы даже не подозревали. Отсутствие интегрированных и непротиворечивых данных — не единственная проблема, стоящая перед пользователями и руководителями. Многие системы, которые проектировались с использованием традиционных методов и приемов, недостаточно хорошо оптимизированы для выполнения запросов к данным. В частности, нерегламентированные запросы, возможность появления которых не учтена при проектировании, могут выполняться плохо или не выполняться вовсе. Можно считать, что хранилище данных расположено в центре всех ориентированных на приложения систем организации. Хранилище регулярно получает данные из этих систем и формирует сводное представление. Данные могут быть простой копией транзакционных данных (в этом случае их называют *атомарными*) или же подвергаться на пути от источника к пункту назначения (хранилищу) трансформации либо агрегированию. При этом в хранилище может помещаться только какое-то их подмножество, или же данные могут подвергаться конвертированию для того, чтобы обеспечить их совместимость с данными из других источников.

Цель исследования. Цель, с которой написана данная работа, состоит в исследовании типовой архитектуры хранилища данных и ее основных компонентов. Особое внимание уделяется хранилищам данных с архитектурой шины данных и федеративным хранилищам данных. Для достижения этой цели проведен обзор типовых архитектур хранилищ данных и основных компонентов и корпоративных информационных фабрик.

Предмет исследования. Предметом разработки является процесс проектирования и разработки хранилищ данных, обладающих рядом характерных особенностей, связанных как с представлением предметной области в хранилище данных, так и с использованием встроенных механизмов программного обеспечения СУБД для воплощения хранилища данных в вычислительной среде.

Теоретической основой диссертации служат исследования специалистов и аналитиков в области, как экономической теории, интеллектуального анализа данных, так и других областей прикладной экономики, таких как, общей и экономической статистики, экономико-математических методов, технологии создания баз данных, баз знаний и экспертных систем, поскольку область извлечения знаний имеет мультидисциплинарный характер.

Методологической базой исследования являются современные работы отечественных и зарубежных специалистов поданной проблеме.

Приведенная методология и методические основы - дополнение известных научных разработок в области теории экономической информатики в части моделирования процессов проектирования информационных технологий управленческих решений по осуществлению функций бизнеса производственных систем.

Научная новизна. Научная новизна магистерской диссертации состоит в исследовании возможностей, которые предоставляют современные реляционные СУБД и CASE инструменты для проектирования хранилищ данных в системах бизнес – аналитики.

Теоретическая и практическая ценность исследования состоит из основных положений, рекомендаций и заключений по поводу организации работы информационных систем.

Структура и объем работы. Диссертация состоит из введения, трех глав, заключения и рекомендаций.

Первая глава работы посвящена разработке концепции систем складирования данных и хранилищ данных, исследованию основных причин ее возникновения и сферы применения, вводятся и обсуждаются базовые понятия, приведены примеры.

Во второй главе изучаются типовые архитектуры хранилищ данных, рассматриваются глобальное хранилище данных, централизованное хранилище данных, распределенное хранилище данных, киоски данных,

взаимосвязанные киоски данных, независимые киоски данных, корпоративная информационная фабрика, хранилище данных с архитектурой шины данных, федеральное хранилище данных. Обсуждаются основные типы хранилищ данных.

Третья глава описывает основные бизнес - функции процесса разработки хранилища данных и подробно излагаются бизнес - функции проектирования, вводится понятие систем деловой осведомленности или систем бизнес - аналитики, рассматриваются основные требования к таким системам и их архитектурные особенности. Подробно обсуждаются вопросы обеспечения информационной безопасности таких систем, проблемы их создания и возможные пути решения этих проблем. Показывается место хранилища данных при разработке систем бизнес – аналитики.

В конце работы даны заключение и литература.

ВВЕДЕНИЕ

В настоящее время хранилища данных являются не только востребованной ИТ, но быстро развивающейся отраслью компьютерной индустрии. По данным исследовательской группы RAMG, Inc объем мирового рынка систем складирования данных динамично развивается. Так что можно смело утверждать, что существует рынок систем бизнес - аналитики. Наличие такого рынка порождает и спрос на специалистов по системам бизнес - аналитики, и, в частности, проектировщиков хранилищ данных.

Процесс проектирования и разработки хранилищ данных обладает рядом характерных особенностей, связанных как с представлением предметной области в хранилище данных, так и с использованием встроенных механизмов программного обеспечения СУБД для воплощения хранилища данных в вычислительной среде. Как и любой процесс проектирования в информационной сфере, проектирование хранилища данных заключается в достижении компромиссов между функциональными, информационными, аппаратными, архитектурными и технологическими требованиями к хранилищу данных, и строится на информированном принятии решений по его структуре данных. Поэтому при изложении материала авторы предприняли попытку предоставить минимально достаточный набор понятий, навыков и умений для решения задачи проектирования хранилища данных.

Концепция баз данных (БД) как метод представления и накопления данных в электронном виде сформировалась к середине 60-х годов прошлого века в фирме IBM. В 1969 году была создана первая СУБД для управления и манипулирования данными как самостоятельными информационными объектами. В 1970 году была предложена реляционная модель данных для БД [4], и на ее основе начали создаваться популярные ныне реляционные СУБД. В рамках реляционной модели с единых позиций были решены многие проблемы операционной (транзакционной) обработки данных.

С середины 80-х годов прошлого столетия стали интенсивно накапливаться электронные информационные массивы данных организаций, корпораций, научно-исследовательских учреждений. Так, в начале 90-х годов прошлого века

только в области химических дисциплин было зарегистрировано более 7000 библиографических, фактографических и смешанных баз данных, ведущие мировые корпорации создали огромные электронные массивы конструкторской документации и документации по управлению производством. В это же время возникло четкое понимание, что сбор данных в электронном виде – не самоцель, накопленные информационные массивы могут быть полезны. Первыми осознали этот факт в области управления бизнесом и производством. В накопленных данных организации находится "информационный снимок" хронологии ее поведения на рынке. Анализ истории административно-хозяйственной деятельности организации позволил существенно увеличить эффективность ее управления, эффективно организовать взаимоотношения с клиентами, производство и сбыт.

Задачи анализа накопленных данных стали перелagаться "на плечи" компьютера и встраиваться в виде аналитических приложений в ИС с БД. Сейчас большинство исследователей сходятся к тому, что отправной точкой разработки концепции складирования данных явился ретроспективный (как иногда еще говорят, исторический) взгляд на данные, накопленные в организации как в электронном, так и в ином виде.

Отметим также, что использование технологий БД и ИС на уже разработанных моделях данных и методиках моделирования данных приводит к ряду проблем для аналитических приложений.

ГЛАВА 1. КОНЦЕПЦИЯ СИСТЕМ СКЛАДИРОВАНИЯ ДАННЫХ

Информационная технология складирования данных (data warehousing) родилась в недрах компании IBM [1] и была окончательно сформулирована Б. Инмоном и Р. Кимбаллом в 90-х годах прошлого столетия [2,3] как **метод решения информационно-аналитических задач в области принятия и поддержки решений**. Возникнув на стыке технологии баз данных (БД), систем поддержки принятия решений (СППР — DSS) и компьютерного анализа данных, в дальнейшем концепция складирования данных претерпела эволюцию, поскольку оказалась пригодной для широкого круга приложений в бизнесе, науке и технологии.

§1.1. Предпосылки создания концепции складирования данных

Основным посылом разработки концепции складирования данных явилось осознание руководством организаций потребности в анализе накопленных электронных массивов данных. На рис. 1.1 показана упрощенная принципиальная схема функционирования организации и место анализа непрерывно поступающей информации.



Рис. 1.1. Упрощенная принципиальная схема функционирования организации

Во всем мире организации накапливают или уже накопили в процессе своей административно-хозяйственной деятельности большие объемы данных, в том числе и в электронном виде. Эти коллекции данных хранят в себе большие

потенциальные возможности по извлечению новой аналитической информации, на основе которой можно и необходимо строить стратегию организации, выявлять тенденции развития рынка, находить новые решения, обуславливающие успешное развитие в условиях конкурентной борьбы. Для некоторых организаций такой анализ является неотъемлемой частью их повседневной деятельности, другие начинают активно приступать к такому анализу.

Системы, построенные на основе информационной технологии складирования данных, обладают рядом характерных особенностей, которые выделяют их как новый класс информационных систем (ИС). К таким особенностям относятся предметная ориентация системы, интегрированность хранимых в ней данных, собираемых из различных источников, инвариантность этих данных во времени, относительно высокая стабильность данных, необходимость поиска компромисса в избыточности данных.

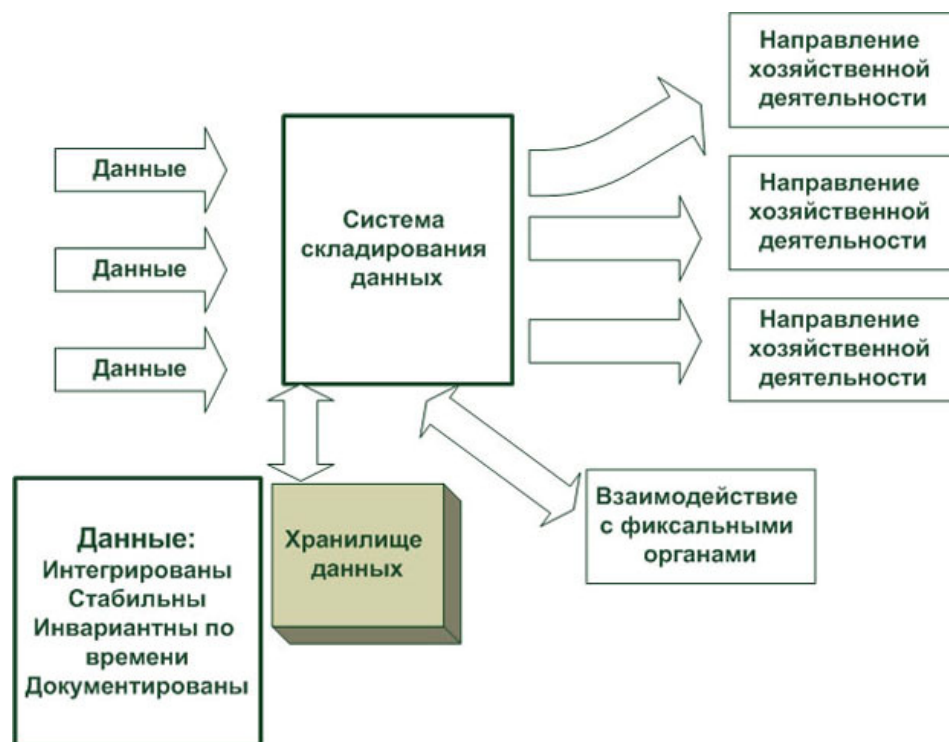


Рис. 1.2. Особенности систем складирования данных

Хранилище данных (ХД — data warehouse) является местом складирования собираемых в системе данных и информационным источником для решения задач анализа данных и принятия решений. Как правило, объем информации в ХД является достаточно большим. Упрощенно можно сказать, что хранилище данных **управляет данными**, которые были собраны как из

операционных систем организации (OLTP-систем — On-Line Transactions Processing), так и из внешних источников данных, и которые длительный период времени хранятся в системе. Более точное определение будет дано позже, после обсуждения истории создания концепции складирования данных.

Одной из главных целей создания систем складирования данных является их ориентация на анализ накопленных данных, т.е. структуризация данных в ХД должна быть выполнена таким образом, чтобы данные эффективно использовались в аналитических приложениях (analytical applications).

Заметим, что задачи анализа накопленных данных решали и до создания концепции складирования данных. В распоряжении аналитиков и сейчас имеется большой набор пакетов программ. Главным отличием использования концепции складирования данных является структуризация, систематизация, классификация, фильтрация и т. п. больших массивов электронной информации в виде, удобном для анализа, визуализации результатов анализа и производства корпоративной отчетности.

Автоматизированная информационная система (ИС) с БД, будучи средством удовлетворения потребностей пользователей в информации как производственном ресурсе, работает с потоками информации, выраженными в потоках данных и операциях с ними. Как было указано выше, основной акцент на ранних стадиях эксплуатации ИС с БД строился на операционной концепции работы с данными. ИС, грубо говоря, должна была быстро и адекватно "переварить" поток данных для решения поставленных перед ней задач с помощью унифицированного набора операций манипулирования данными. Обработка данных сводилась к операциям вставки, удаления и обновления.

Совместное действие этих операций в рамках ИС приводило к конфликтам в данных - потерям данных, ошибкам в обновлении и т.д. - так называемым аномалиям в данных. Предложив реляционную модель (которая является достаточно строго математической, а, следовательно, приемлемо контролируемой моделью), Е. Кодд в целом решил ряд проблем и задач операционной обработки данных [4,8-10]. Создание реляционных СУБД позволило достаточно грамотно (с учетом уровня компетентности разработчика) строить системы операционной

(или, как ее еще называют, транзакционной) обработки данных - OLTP (On-Line Transactions Processing).

На практике данные в операционных системах могут содержаться столь угодно долго, сколь в них имеется потребность. Несмотря на то, что производители жестких дисков постоянно увеличивают объемы этих дисков, хранить редко используемую информацию не имеет смысла по той простой причине, что производительность многих запросов с ростом объема данных начинает падать и совершенствование подсистем оптимизации запросов СУБД решает проблему ухудшения производительности запросов лишь отчасти. В целом с накоплением данных производительность обработки данных продолжает ухудшаться (эффект больших объемов).

Типичным организационным методом работы с редко используемыми данными является процедура архивизации. Во многих случаях процедура архивизации сводится к простому копированию данных на резервный носитель информации.

Таким образом, одной из проблем при решении задач анализа данных, помимо других скрытых проблем, в рамках операционных систем анализа данных является низкая производительность обработки запросов, которые готовят данные для последующего анализа. Такие запросы увеличивают нагрузку на процессоры ОС и в целом ухудшают обработку потока транзакций в БД, исходящего от систем операционной обработки данных.

Работа с архивом как чистой копией массива данных операционной системы обработки данных не решает проблему производительности. Отсюда простой практический ход - разделить решение задач обработки транзакций и задач анализа данных. В реляционных СУБД производительность запроса может быть улучшена за счет модификации модели данных. Архивные информационные массивы можно наделить структурой, отличной от структуры данных в несущей БД операционной ИС. Разработку таких структур данных можно связать с решением задач ретроспективного анализа данных, накопленных в системе. Это допустимо хотя бы потому, что в задачах анализа данных учитываются далеко не все функциональные зависимости, поддерживаемые в операционных БД. Поэтому

структуру данных архивов стали проектировать под задачи анализа данных, неявно породив тем самым новый класс приложений.

Фундаментальные требования к разработке операционных систем обработки данных и систем анализа данных различны: операционным системам нужна производительность, в то время как системам анализа данных нужны гибкость и широкие возможности для получения результата. Это противоречие в целевой направленности двух классов систем обработки данных явилось одной из основных предпосылок разработки концепции складирования данных (рис. 1.3).

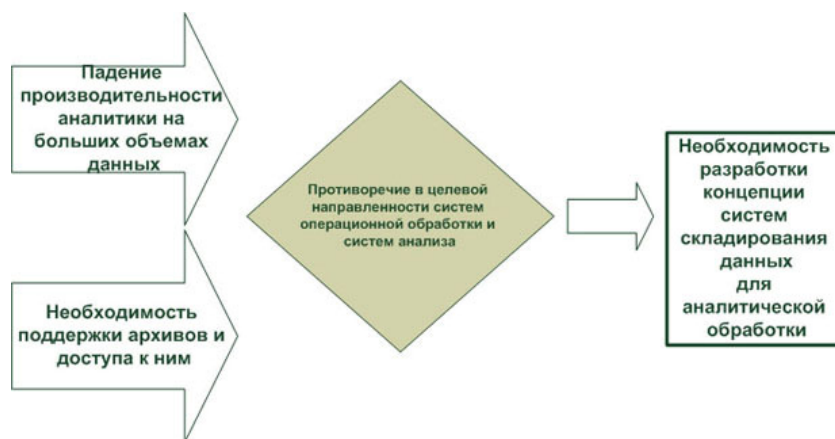


Рис. 1.3. Основной побудительный мотив разработки концепции систем складирования данных, следующий из опыта решения задач анализа на данных операционных систем обработки данных

Создание новой концепции потребовало пересмотра ряда традиционных подходов к обработке данных и перестройки технологических процедур. Поскольку перестройка технологических процедур является чрезвычайно затратным мероприятием, важно отметить те причины, которые явились дополнительными побудительными мотивами применения новой концепции на практике.

Одной из первых таких причин является работа с данными, доставшимися по наследству (legacy systems — система, доставшаяся по наследству). Средства вычислительной техники (ВТ) быстро эволюционировали. В 80-х годах прошлого века появились миникомпьютеры на платформах AS/400 и VAX/VMS. Конец восьмидесятых и начало девяностых сделали ОС UNIX популярной серверной платформой для повсеместного введения новой архитектуры клиент /сервер. Начиная с 90-х, быстро стало прогрессировать семейство ОС MS Windows.

В то же время, начиная с 70-х годов, большинство систем обработки данных в сфере бизнеса создавалось на мейнфреймах фирмы IBM. Несмотря на все изменения в операционных платформах, архитектуре вычислительных систем, инструментальных средствах разработки программ и информационных технологиях, значительное количество бизнес-приложений в Америке и Европе продолжало и продолжает работать на оборудовании такого класса, что, кстати, стимулировало новый виток в развитии информационных технологий на мейнфреймах.

За годы эксплуатации в этих системах накоплены огромные бизнес-знания, было зафиксировано значительное количество бизнес-правил. Этот огромный объем информации невероятно трудно перенести на новые аппаратно-программные платформы или в приложения.

Системы, обобщенно называемые системами, доставшимися по наследству (legacy systems), продолжают быть самым большим источником данных для систем анализа данных. Однако время, требуемое на получение результатов работы таких приложений, часто оказывается значительно больше того, которое может позволить себе для ожидания конечный пользователь (по большей части руководство организации) в условиях современного бизнеса.

Второй причиной стал персональный компьютер, который позволил перенести данные из централизованного вычислительного центра на рабочий стол пользователя (в частности бизнес-аналитика).

Всего за несколько лет персональный компьютер (ПК) прочно утвердился на рабочем столе руководителей бизнеса, аналитиков и финансистов. Такая популярность ПК повлекла за собой интенсивную разработку программного обеспечения, в том числе и для анализа данных бизнеса. Хорошо подготовленные пользователи могут использовать настольные базы данных, которые позволяют им хранить и работать с информацией, извлеченной из источников данных систем, доставшихся по наследству. Персональный компьютер и его программный инструментариум перенесли работу по анализу данных из больших вычислительных центров (ВЦ) на рабочий стол пользователя. Эффективность аналитической работы в особенно крупных организациях стала расти.

Вовлечение конечных пользователей для решения задач управления данными в условиях коллективного их использования не является выходом из создавшейся ситуации. Во-первых, это требует времени и усилий конечных пользователей (а следовательно, денег). Во-вторых, у них есть основная работа – анализ данных, которая им интересна и за которую им платят жалование. Вовлечение их в работу в сфере информационных технологий совершенно точно приведет к снижению эффективности их основной работы.

Широкое применение персональных компьютеров в анализе данных привело к другой проблеме. Отсутствие общих стандартов представления данных в организации, большая свобода в выборе представления данных конечным пользователем, сбрасывание со счетов требований коллективного использования данных приводит к анархии в работе с данными, и, как следствие, появляется опасная тенденция несогласованности коллективно используемых данных, которая может сказываться на качестве принятия стратегических решений.

§ 1.2. Системы поддержки и принятия решений и управленческие информационные системы

Еще одной причиной стало интенсивное использование систем поддержки и принятия решений (СППР — DSS) и управленческих информационных систем (ИСП — EIS, информационная система руководителя). СППР обычно фокусируются на более детальном представлении информации и ориентированы больше на менеджеров среднего уровня. ИСП обеспечивают более высокий уровень консолидации и многоаспектного (многомерного представления) взгляда на данные, поскольку руководители высокого уровня нуждаются в большем многообразии представления тех же самых данных для детального анализа.

Эти два схожих и перекрывающихся по функциям класса систем являются одной из главных предпосылок для создания концепции систем складирования данных. Отметим некоторые признаки, обычно связываемые с системами этого класса.

- В этих системах данные представлены в стандартных терминах бизнеса, а не в закодированной форме (имена полей в БД ИС). Наименования элементов данных и структуры данных в этих системах проектируются для

использования конечными пользователями с невысоким уровнем подготовки в области информационных систем.

- Данные в таких системах предварительно обрабатываются в контексте стандартных бизнес-правил, таких как размещения ассигнований по продуктам, производственным единицам и рынкам.
- Допускается консолидированное представление данных по таким категориям, как продукт, производитель и рынок. Хотя в таких системах время от времени допускается развертывание интегрированных данных, они способны обеспечить доступ ко всей детальной информации в одно и то же время.

В настоящее время системы складирования данных обеспечивают аналитические инструменты для решения таких задач, но их разработка строится не на специфических требованиях аналитиков или исполнителей, а основывается на структуре бизнеса организации. С этой точки зрения системы складирования данных дали новый виток в развитии СППР и ИСР.

Не следует забывать также о факторах, связанных с техническим прогрессом в области разработки аппаратного обеспечения ЭВМ и развитием компьютерных технологий в разработке программного обеспечения. Это обстоятельство привело к снижению цен на комплектующие с одновременным ростом их мощности, созданию дружественных интерфейсов для пользователей.

Наиболее важным фактором в развитии складирования данных стало увеличение мощности аппаратной платформы компьютеров, поскольку ХД хранят обычно очень большие объемы информации. Параллельно росла вычислительная мощность ПК и развитое программное обеспечение, которые позволили разработать и внедрить архитектуру клиент/сервер. Почти ко всем ХД можно обратиться с ПК, оснащенного развитыми инструментальными программными средствами. Эти средства изменяются от очень простых обработчиков запросов до мощных графических многомерных средств анализа данных. Создание серверных операционных систем, таких как Windows и Unix, повысило надежность в функционировании и дало мощные возможности распределенной вычислительной среде. Эти технологические факторы способствовали быстрому развитию систем складирования данных.

Создание и распространение Интернет/Интранет привело к тому, что бизнес стал перемещаться в Интернет. Сегодня одной из наиболее значительных областей компьютерной индустрии является разработка интранет-приложений. Грубо говоря, Интранет является совокупностью локальных компьютерных сетей, ориентированных на бизнес, которые основываются на стандартах сети Интернет, хотя проектируются для внутреннего использования в организации. ХД может быть доступно из любой точки сети, как локальной, так и глобальной, и стоимость доступа к нему значительно снижается по сравнению с обычной технологией. С другой стороны, использование технологии Интернет позволяет веб-серверу обеспечить обработку данных в узлах их размещения, что приводит к выполнению всех трудоемких процедур анализа до того, как его результаты представляются пользователю в его браузере.

Значительное влияние на формирование концепции складирования данных оказали фундаментальные изменения в организации бизнеса и изменения в его структуре в конце прошлого века. Появление ярко выраженной глобальной экономики изменило требования к информации и спрос на нее. Деятельность организаций пересекла границы своей страны и тем самым стала транснациональной.

Изменение экономических условий побудили большие корпорации к объединению (консолидации) своих усилий. Появление таких механизмов, как реинжиниринг бизнес-процессов (business process reengineering) и перестраиваемость бизнеса (downsizing), вынудило руководителей переоценить практику ведения бизнеса. Пересмотр процедур ведения бизнеса и изменения в финансовых потоках сыграли важную роль в развитии концепции складирования данных.

Глобализация экономики выдвигает не только требования непрерывного анализа потоков экономических данных, но и определенные требования к сбору и размещению деловой информации. Теперь процесс сбора и свертывания производственных и коммерческих данных от разбросанных по всему миру производственных подразделений оказывает сильное влияние на принятие решений в корпорациях. Глобализация делает процедуры размещения данных в централизованном хранилище данных более сложными. Колебания стоимости

валют или сезонные колебания в сбыте продукции в различных регионах мира добавили трудности к складированию данных и делают анализ данных более сложным.

Еще одним важным фактором, который повлиял на развитие ХД, явилось появление специализированных поставщиков решений в автоматизации бизнеса. Фирмы-разработчики ПО SAP AG, Baan, Oracle, Microsoft, IBM и др. предлагают быстро адаптируемые к бизнес-процессам программные продукты для управления бизнесом. Разработка комплексного ПО для управления бизнесом привела к интенсификации процессов стандартизации бизнеса и стандартизации программного обеспечения. Информация в ХД поступает в унифицированном виде из всех ИС управления бизнесом, а не только из систем, доставшихся по наследству. Следует также отметить тенденцию последних лет в разработке ПО данного класса приложений — обеспечение возможности собирать информацию в ХД из данных любых внешних источников (например, продукты компании SAS).

Один из наиболее важных результатов массивной инвестиции в технологию и создание высокопроизводительных ПК привел к созданию методов анализа, основанного на здравом смысле (*technology-savvy business analyst*). Даже если технология здравого смысла конечного пользователя не всегда выгодна для многих проектов, тенденция ее применения привела к созданию более сложных технологий анализа для сегодняшнего бизнеса. Однако именно на технологии здравого смысла была продемонстрирована выгода использования хранилищ данных и развития логических и физических моделей.

Текстовые редакторы и крупноформатные таблицы, столь популярные на первых ПК, оказали существенное влияние на представление данных в хранилищах данных.

Очень сильно влияют на тенденции развития информационных технологий требования к информации, предъявляемые средним и высшим звеном управляющего персонала. Информационные технологии стали производственным ресурсом компаний – это первый результат таких требований. Электронная почта, Интернет, мобильный телефон и карманный ПК вовлечены в процесс управления. Это также требования, приходящие от руководителей организаций и компаний.

Им нужен быстрый и качественный доступ к аналитической информации в любой момент времени и по любому виду каналов связи.

Как видно из вышесказанного, потребности бизнеса в новых экономических условиях, создание мощной программно-аппаратной платформы, распространение информационных технологий создали предпосылки рождения нового класса приложений — систем складирования данных и концепции ХД как информационного носителя для таких приложений.

На рис. 1.4 просуммированные основные факторы, способствующие созданию и развитию концепции систем складирования данных и хранилищ данных.



Рис. 1.4. Основные факторы, повлиявшие на разработки концепции хранилищ данных

§ 1.3. Концепция хранилищ данных

Отметим, что складирование данных — это развивающаяся технология. Как и для любой развивающейся технологии, определенная доля осторожности должна присутствовать при оценке действий производителей ПО ХД, пытающихся позиционировать себя среди конкурентов. Например, дискуссии о размерах ХД — с какого размера хранилище данных можно считать собственно хранилищем? С 50 ГБ? Заметим, что в некоторых областях исследования размер анализируемого массива может быть очень небольшим. Просто нет данных. А анализ такого массива возможен.

Рассмотрим основные элементы концепции складирования данных.

Извлечение данных из операционных систем

Главный элемент концепции складирования данных состоит в том, что к данным, сохраняемым для анализа, может быть обеспечен наиболее эффективный

доступ только при условии выделения их из операционной (транзакционной) системы, т.е. данные из операционной системы должны быть вынесены в отдельную систему складирования данных. Такой подход носит исторический характер. Из-за ограничений в аппаратном обеспечении и технологии, для того чтобы обеспечить производительность транзакционной системы, данные архивировались на магнитных лентах или носителях вне такой системы. Проблема доступа к ним требовала определенных технологических решений.

Нужно отметить, что с развитием концепции позиция отделения данных для анализа от данных в OLTP-системе претерпела мало изменений. Она стала более формальной и обогатилась за счет применения средств многомерного анализа данных. В настоящее время ХД можно строить и на существующей OLTP-системе, и над ней, и как самостоятельный объект. Это должно решаться руководителем ИТ-проекта в рамках выбора архитектуры ХД.

Необходимость интегрирования данных из нескольких OLTP-систем

Системы складирования данных наиболее полезны, когда данные могут быть извлечены более чем из одной OLTP-системы. Когда данные должны быть собраны от нескольких бизнес-приложений, естественно предположить, что это нужно сделать в месте, отличном от места локализации исходных приложений. Еще до создания структурированных ХД аналитики во многих случаях комбинировали данные, извлеченные из разных систем, в одну крупноформатную таблицу или базу данных. ХД может очень эффективно воедино собрать данные от конкретных приложений, таких как продажи, маркетинг, финансы, производство, с учетом их накопления, т.е. сохранить временные ряды основных показателей бизнеса — так называемые исторические данные.

Заметим, что одним из свойств данных, собранных из различных приложений и используемых аналитиками, является возможность делать перекрестные запросы к таким данным. Во многих ХД атрибут "время" является естественным критерием для фильтрации данных. Аналитиков интересует поведение временных рядов данных, характеризующих процессы бизнеса.

Целью многих систем складирования данных является обзор деятельности типа "год за годом". Например, можно сравнивать продажи в течение первого квартала этого года с продажами в течение первого квартала предшествующих

лет. Время в ХД — фундаментальный атрибут перекрестных запросов. Например, аналитик может попытаться оценить влияние новой компании маркетинга, проходящей в течение определенных периодов, рассматривая продажи в течение тех же самых периодов. Способность устанавливать и понимать корреляцию между деятельностью различных подразделений в организации часто приводится как один из самых главных аргументов о пользе систем складирования данных.

Система складирования данных не только может работать как эффективная платформа для консолидации данных из различных источников, но может также собирать многократные версии данных из одного приложения. Например, если организация перешла на новое программное обеспечение, то ХД сохранит необходимые данные из предыдущей системы. В этом отношении система складирования данных может служить средством интеграции наследуемых данных, сохраняя преемственность анализа при смене программно-аппаратной платформы OLTP-системы.

Различия между транзакционной и аналитической обработкой данных

Одной из наиболее важных причин отделения данных для анализа от данных OLTP-систем было потенциальное падение производительности обработки запросов при выполнении процессов анализа данных. Высокая производительность и небольшое время ответа — критические параметры OLTP-систем. Потерю производительности и объем накладных затрат, связанных с обработкой predetermined запросов, обычно легко оценить. С другой стороны, запросы для анализа данных в ХД трудно предсказать, и следовательно, для них сложно оценить время выполнения запроса.

OLTP-системы спроектированы для оптимального выполнения predetermined запросов в режиме работы, близком к режиму реального времени. Для таких систем обычно можно определить распределение нагрузки во времени, определить время пиковых нагрузок, оценить критические запросы и применить к ним процедуры оптимизации, поддерживаемые современными СУБД. Также относительно легко определить максимально допустимое время ответа на определенный запрос в системе. Стоимость времени ответа такого запроса может быть оценена на основе отношения стоимости выполнения операций ввода-вывода / стоимость затрат на трафик по сети. Например, для

системы обработки заказов можно задать число активных менеджеров по оформлению заказов и среднее число заказов в течение каждого часа работы.

Несмотря на то, что многие из запросов и отчетов в системе складирования данных predetermined, почти невозможно точно предсказать поведение показателей системы (время отклика, трафик сети и т.п.) при их выполнении. Процесс исследования данных в ХД происходит зачастую непредсказуемым путем. Руководители всех рангов умеют ставить неожиданные вопросы. В процессе анализа могут возникать непредetermined (ad hoc) запросы, которые вызваны неожиданными результатами или непониманием конечным пользователем используемой модели данных. Далее, многие из процессов анализа имеют тенденцию принимать во внимание многие аспекты деятельности организации, в то время как OLTP-системы хорошо сегментированы по видам деятельности. Пользователю может потребоваться более детальная информация, чем хранящаяся в итоговых таблицах. Это может привести к соединению двух или более огромных таблиц, что закончится созданием временной таблицы объемом, равным произведению числа строк в каждой таблице, и резко снизит производительность системы.

Данные в системах складирования данных остаются неизменными

Другое ключевое свойство данных в системе складирования данных состоит в том, что данные в ХД остаются неизменными. Это означает, что после того, как данные разместятся в ХД, они не могут быть изменены. Например, статус заказа не меняется, размер заказа не меняется, и т. д. Эта характеристика ХД имеет большое значение для отбора типов данных при размещении их в ХД, а также выбор момента времени, когда данные должны быть занесены в ХД. Последнее свойство называется гранулированностью данных.

Рассмотрим, что означает для данных быть неизменяемыми. В OLTP-системе объекты данных проходят через постоянные изменения своих атрибутов. Например, заказ может многократно изменять свой статус до того, как будет оформлен. Или, когда изделие собирается на сборочной линии, к нему применяется множество технологических операций. Вообще говоря, данные из OLTP-системы нужно загружать в ХД лишь тогда, когда обработка их в рамках бизнес-процессов будет полностью завершена. Это может означать завершение

заказа или цикла производства изделия. Как только заказ закончен и отправлен, он вряд ли поменяет свой статус. Или, как только изделие собрано и сдано на склад, оно вряд ли попадет на первую стадию сборочного цикла.

Другим хорошим примером может быть размещение в ХД снимка постоянно изменяющихся данных в определенные моменты времени. Модуль управления запасами в OLTP-системе может изменять запас почти в каждой транзакции; невозможно занести все эти изменения в ХД. Вы можете определить, что такой снимок состояния запаса следует вносить в ХД каждую неделю или ежедневно, так, как это будет принято для анализа в конкретной организации. Данные такого снимка, естественно, неизменяемы.

После того, как данные занесены в ХД, их модификация возможна в крайне редких случаях. Очень трудно (хотя такие попытки есть) поддерживать динамические данные в ХД. Задача синхронизации часто изменяемых данных в OLTP-системах и системах складирования данных еще далека от приемлемого решения. Здесь следует также упомянуть, что размещение динамично меняющихся данных в ХД в настоящее время является предметом интенсивных исследований. Например, разработка процедур поддержки в ХД медленно меняющихся таблиц измерений является задачей, которая уже находит свое решение на уровне ПО производителей решений в области ХД.

Данные в хранилище данных хранятся значительно более длительное время, чем в OLTP-системах

Данные в большинстве OLTP-систем архивируются сразу после того, как они становятся неактивными. Например, заказ может стать неактивным после того, как он выполнен; банковский счет может стать неактивным после того, как он был закрыт. Главная причина для архивирования неактивных данных — это производительность OLTP-системы (зачем хранить данные, если к ним не обращаются). Большие объемы таких данных могут заметно ухудшить производительность выполнения запросов в предположении, что обрабатываются только активные данные. Для обработки таких данных в СУБД предлагаются различные процедуры разбиения базовых таблиц на секции. С другой стороны, поскольку ХД предназначены, в частности, быть архивом для OLTP-данных, данные в них хранятся в течение очень длительного периода.

Фактически, проект системы складирования данных может начинаться и без любого определенного плана архивирования данных из ХД. Стоимость сопровождения данных после их загрузки в хранилище невысока. Наибольшие затраты при создании хранилища выпадают на трансформацию данных (data transfer) и их очистку (data scrubbing). Хранение данных в течение пяти и более лет типично для систем складирования данных. Поэтому процедурам архивизации данных из ХД на стадиях их создания и эксплуатации в начале периода можно не уделять много времени. Особенно если учесть снижение цен на аппаратные средства ЭВМ.

Иначе говоря, отделение данных OLTP-систем от данных систем анализа является фундаментальной концепцией складирования данных. Сейчас бизнес невозможен без принятия обоснованных решений. Такие решения могут быть построены на основе всестороннего анализа результатов выполнения бизнес-процессов в организации и деятельности организации на рынке товаров и услуг. Время принятия решений в современных условиях и потоках информации сокращается. Роль создания и поддержки систем анализа данных на основе новых информационных технологий возрастает. ХД является одним из основных звеньев применения таких технологий.

Можно выделить следующие причины для разделения данных систем складирования данных и систем операционной обработки данных (рис. 1.5).

- Различие целевых требований к системам складирования данных и OLTP-системам.
- Необходимость собирать данные в ХД из различных информационных источников, т.е. если данные генерируются в самой OLTP-системе, то для системы складирования данных в большинстве случаев данные генерируются вне ее.
- Данные, попадая в ХД, остаются в большинстве случаев неизменными.
- Данные в ХД сохраняются длительное время.



Рис. 1.5. Основные причины разделения данных для анализа и оперативной обработки

Логическое преобразование данных OLTP-систем и моделирование данных

Данные в ХД логически являются преобразованными, когда они перенесены в него из OLTP-системы или другого внешнего источника. Проблемы, связанные с логическим преобразованием данных при переносе их в ХД, могут потребовать значительного анализа и усилий проектировщиков. Архитектура системы складирования данных и модели ХД имеют огромное значение для успеха таких проектов. Ниже будут рассмотрены некоторые фундаментальные понятия реляционной теории БД, которые полностью не применимы к системам складирования данных. Даже при том, что большинство ХД развернуто на реляционных БД, некоторые основные принципы реляционных БД сознательно нарушаются при создании логической и физической модели ХД.

Модель данных ХД определяет его логическую и физическую структуру. В отличие от просто архивированных данных, в данном случае невозможно обойтись без процедур детального моделирования. Такое моделирование на ранних стадиях проекта системы складирования необходимо для создания эффективной системы, охватывающей данные всех бизнес-процессов и процедур организации.

Процесс моделирования данных должен структурировать данные в ХД в виде, не зависящем от реляционной модели данных системы, которая предоставляет эти данные. Как будет показано ниже, модель ХД, вероятно, будет менее нормализована, чем модель OLTP-системы – источника данных.

В OLTP-системах данные по разным подсистемам могут значительно перекрываться. Например, информация относительно разрабатываемых изделий используется в различных формах во многих подсистемах OLTP-системы. Система складирования данных должна объединить все такие данные в одной системе. Некоторые атрибуты объектов, которые являются существенными для OLTP-системы, окажутся ненужными для ХД. Могут появиться новые атрибуты, так как сущность (entity) в ХД изменяет свое качество. Основное требование — все данные в ХД должны участвовать в процессе анализа.

Модель данных ХД должна быть расширена и структурирована таким образом, чтобы данные от различных приложений могли быть добавлены. Проект ХД в большинстве случаев не может включать данные от всех возможных бизнес-приложений организации. Обычно объем данных в ХД увеличивается по принципу инкремента: данные экстрагируются из OLTP-систем и добавляются в ХД определенными порциями. Начинают с сохранения особенно существенных данных, затем планомерным образом наращивают по мере необходимости их объем.

Модель хранилища данных подстраивается под структуру бизнеса

Следующий важный момент состоит в том, что логическая модель ХД настраивается на структуру бизнеса (ориентирована на предметную область), а не на агрегацию логических моделей конкретных приложений. Сущности (объекты), поддерживаемые в ХД, аналогичны сущностям (объектам) бизнеса — таким как клиенты, продукция (товар), заказы и дистрибьютеры. В рамках конкретных подразделений организации может быть очень узкое представление об объектах бизнеса организации, например, о клиентах. Так, группа обслуживания ссуд в банке может знать что-либо о клиенте только в контексте одной или нескольких выданных ссуд. Другое подразделение того же банка может знать о том же клиенте в контексте депозитного счета. Представление данных о клиенте в ХД намного превышает аналогичное представление конкретного подразделения банка. Клиент в ХД представляет клиента банка во всех его взаимоотношениях с банком. С точки зрения реляционной теории меняется базисный набор функциональных зависимостей, поддерживаемых в БД.

ХД следует строить на атрибутах сущностей бизнеса (предметно ориентированно), собирая данные об этих сущностях из различных источников. Структура данных любого отдельного источника данных, вероятно, будет неадекватна для ХД. На структуру данных конкретного приложения оказывают влияние такие факторы, как:

- вид конкретного бизнес-процесса. Так, в автоматизированной подсистеме закупок структура данных может быть продиктована характером бизнес-процедур закупок на данном секторе рынка;
- влияние модели действующих систем. Например, исходное приложение может быть достаточно старым и учитывать развитие модели данных за счет изменения структуры БД приложения. Многие такие изменения могут быть плохо документированы;
- ограничения программно-аппаратной платформы. Логическая структура данных может не поддерживать некоторые логические взаимоотношения между данными или иметь ограничения, связанные с ограничениями программно-аппаратной платформы.

Модель ХД не связана с ограничениями моделей данных источников. Для нее должна быть разработана модель, которая отражает структуру бизнеса организации, а не структуру бизнес-процесса. Такая расширенная модель данных должна быть понятна как аналитикам, так и менеджерам. Таким образом, проектировщик ХД должен выполнить настройку объектов ХД к структуре бизнеса организации, с учетом ее бизнес-процессов и бизнес-процедур.

Преобразование информации, описывающей состояние объектов в OLTP-системе

Следующий важный момент состоит в том, что перед размещением данных в ХД они должны быть преобразованы. Большинство данных из OLTP-системы или иного внешнего источника не могут поддерживаться в ХД. Многие из атрибутов объектов в OLTP-системе очень динамичны и постоянно изменяются. Многие из этих атрибутов не загружаются в ХД, другие же атрибуты являются статичными во времени и загружаются в ХД. ХД вообще не должно содержать информации об объектах, которые являются динамическими и постоянно находятся в состоянии модификации.

Чтобы понять, что означает потеря информации, описывающей текущее состояние объекта, рассмотрим пример системы управления заказами, которая отслеживает состояния запасов при заполнении заказа. Сначала рассмотрим сущность "Заказ" в OLTP-системе. Заказ может пройти множество различных статусов или состояний, прежде чем он будет выполнен и обретет статус завершенного. Статус заказа может указывать, что он готов к заполнению, что заказ заполняется, возвращен обратно на доработку, готов к отгрузке и т.д. Конкретный заказ может пройти много состояний, которые отражаются в статусе заказа и определяются бизнес-процессами, которые применялись к нему. Практически невозможно перенести все атрибуты такого объекта в ХД. Система складирования данных, вероятно, должна содержать только один конечный снимок такого объекта, как заказ. Таким образом, объект "Заказ" должен быть преобразован для размещения в ХД. Тогда в ХД может быть собрана информация о многих объектах типа "заказ" и построен окончательный объект ХД – "Заказ".

Рассмотрим более сложный пример трансформации данных при управлении запасом товара в OLTP-системе. Запас может изменяться в каждой транзакции. Количество конкретного товара на складе может быть уменьшено транзакцией подсистемы заполнения заказа или увеличено при поступлении купленного товара. Если система обработки заказа выполняет десятки тысяч транзакций в день, то, вероятно, фактический уровень запаса в БД будет иметь много состояний и зафиксируется во многих снимках в течение этого дня. Невозможно зафиксировать все эти постоянные изменения в БД и перенести их в ХД. Отображение такого поведения объекта в системе – источнике данных по-прежнему является одной из нерешенных задач в системах складирования данных. Есть ряд подходов к решению этой проблемы. Например, можно периодически фиксировать снимки уровня запаса в ХД.

Этот подход может быть применим к очень большой части данных в OLTP-системах. В свою очередь, такое решение повлечет за собой ряд задач, связанных с выбором периода времени, объема снимаемых данных и т.д. Таким образом, большая часть данных о состоянии объектов в OLTP-системе не может быть непосредственно перенесена в ХД. Они должны быть преобразованы на логическом уровне.

Денормализация модели данных

Следующий момент в проектировании реляционных ХД состоит в решении вопроса о том, насколько важно в ХД соблюдать принципы реляционной теории, а именно: разрешить денормализацию модели, в частности, для увеличения производительности запросов. Прежде чем мы рассмотрим денормализацию модели данных в контексте складирования данных, давайте кратко вспомним основные моменты теории реляционных БД и процесса нормализации. Е.Ф. Кодд разработал реляционную теорию БД в конце 60-х прошлого века, когда он работал в исследовательском центре ИВМ. Сегодня большинство популярных платформ БД полностью следует этой модели. Реляционная модель БД — коллекция двумерных таблиц, состоящих из рядов и колонок. В терминологии реляционной модели эти таблицы, строки и колонки соответственно называются отношениями или сущностями, кортежами, атрибутами (attribute) и доменами (domain). Модель идентифицирует уникальные ключи (Key) для всех таблиц и описывает отношения между таблицами через значения атрибутов (ключей).

Нормализация (Normalization) является процессом моделирования реляционной БД, где отношения или таблицы разбиваются до тех пор, пока все атрибуты в отношении полностью не будут определяться его первичным ключом. Большинство проектировщиков пытаются достичь третьей нормальной формы (3НФ) на всех отношениях до того, как они будут денормализоваться по тем или иным причинам. Три последовательных этапа нормализации реляционных БД кратко описаны ниже.

- Первая нормальная форма 1NF (1НФ). Говорят, что отношение находится в первой нормальной форме, если оно описывает одну единственную сущность и не содержит в качестве атрибутов массивов или повторяющихся групп значений. Например, таблица заказов, включающая в себя позиции заказа, не будет находиться в первой нормальной форме, поскольку содержит повторяющиеся атрибуты для каждой позиции заказа.
- Вторая нормальная форма 2NF (2НФ). Говорят, что отношение находится во второй нормальной форме, если оно находится в первой

нормальной форме и все неключевые атрибуты функционально полно зависят от первичного ключа отношения.

- Третья нормальная форма 3NF (3НФ). Говорят, что отношение находится в третьей нормальной форме, если оно находится во второй нормальной форме и его неключевые атрибуты полностью независимы друг от друга.

Процесс нормализации приводит к разбиению исходного отношения на несколько независимых отношений. Каждому отношению в БД отвечает по крайней мере одна таблица. Несмотря на большую гибкость реляционной модели в представлении данных, она может быть более сложной и трудной для понимания. Кроме того, полностью нормализованная модель может быть очень неэффективной при реализации. Поэтому проектировщики БД при преобразовании нормализованной логической модели в физическую модель допускают значительную денормализацию.

Основное назначение денормализации состоит в ограничении межтабличных соединений в запросах.

Еще одной из причин денормализации модели ХД является, так же, как и для операционных систем, производительность и простота. Каждый запрос в реляционных БД имеет свою стоимость выполнения (cost performance). Стоимость выполнения запросов очень высока в ХД из-за количества обрабатываемых данных в запросе (и межтабличных соединений, число которых растет пропорционально размерности модели). Соединение трех маленьких таблиц в OLTP-системе может иметь приемлемую стоимость выполнения запроса, но в системе складирования данных выполнение такого соединения может занять очень много времени.

Статичность взаимосвязей в исторических данных

Денормализация является важным процессом в моделировании ХД: взаимосвязь между атрибутами не изменяется для исторических данных. Например, в OLTP-системах товар может быть частью другого товара группы "А" в этом месяце и частью товара группы "В" в следующем месяце. В нормализованной модели данных для отображения этого факта необходимо включить атрибут "группа товаров" в отношение (сущность) "товар", но не в

отношение (сущность) "заказ", которая формирует заказы на этот товар. В сущность "заказ" включается только идентификатор товара. Реляционная теория будет требовать соединения между таблицами "Заказ" и "Товар" для определения группы товаров и других атрибутов этого продукта. Этот факт (функциональная зависимость) не имеет значения для ХД, поскольку сохраняемые данные относятся к уже выполненным заказам, т.е. принадлежность товара группе уже зафиксирована (фактически указанная функциональная зависимость не поддерживается). Даже если товар принадлежал различным группам в разное время, взаимосвязь между группой товаров и товаром каждого отдельного заказа статична. Таким образом, это не является денормализацией для ХД. В данном случае функциональная зависимость OLTP-системы не используется в ХД.

В качестве другого примера возьмем цену товара. Цены в OLTP-системе могут изменяться постоянно. Некоторые изменения этих цен могут быть перенесены в ХД, как периодические снимки таблицы "Цена товара". В ХД история прайс-листа товара зафиксирована и уже привязана к заказам, т.е. не нужно динамически определять прайс-лист при обработке заказа, поскольку он уже был применен к сохраненному заказу. В реляционных БД проще поддерживать динамические взаимоотношения между сущностями бизнеса, в то время как ХД содержит взаимосвязи между сущностями предметной области в заданное время.

Концепция логического преобразования данных приложений-источников, рассмотренная выше, требует определенных усилий при реализации и очень полезна при разработке ХД.

Физическое преобразование данных приложений источников

Важным моментом в системах складирования данных является физическое преобразование данных. Эти процедуры в складировании данных известны как процессы очистки данных ("data scrubbing", "data staging" или "data purge"). Процесс очистки данных является наиболее интенсивным и трудоемким в любом проекте создания ХД. Физическое преобразование включает использование стандартных терминов предметной области ХД и стандартов данных. В течение процесса физического преобразования данные находятся в некотором промежуточном файле до того, как будут занесены в ХД. Когда данные

собираются из многих приложений, их целостность может быть проверена в течение процесса формирования преобразованных данных до загрузки в ХД.

Термины и имена атрибутов сущностей, используемые в OLTP-системах, в процессе преобразования данных для ХД преобразуются в универсальные, стандартные термины, принятые для данной сферы бизнеса. Приложения могут использовать сокращения или трудные для понимания термины по множеству различных причин. Программно-аппаратная платформа может ограничивать длину и формат имен, а бизнес-приложения могут применять в разных предметных областях общие термины. В ХД необходимо пользоваться стандартными бизнес-терминами, которые понятны сами по себе большинству пользователей.

Идентификатор клиента (покупателя) в OLTP-системе может быть назван "Покуп.", "покуп_ид" или "покуп_но". Далее, различные приложения таких систем могут использовать различные имена (синонимы) при ссылке к одному и тому же атрибуту сущности. Проектировщик ХД выбирает простой стандартный бизнес-термин, такой, как "Идентификатор клиента". Таким образом, имена атрибутов сущностей из подающих систем должны быть унифицированы для использования в ХД.

Различные подсистемы OLTP-систем и внешних источников данных могут использовать различное определение доменов атрибутов на физическом уровне представления данных. Так, атрибут типа "идентификатор продукта" в одной системе имеет длину от 12 символов, а в другой — 18 символов. С другой стороны, ПО одних существующих систем может иметь ограничения на определение длин имен атрибутов и бедный набор типов для определения доменов, а в других такие ограничения могут отсутствовать и может предоставляться широкий выбор типов атрибутов.

При определении атрибутов в физической модели ХД необходимо использовать такие длины и типы данных в определении домена атрибута, которые позволили бы учесть как требования предметной области, так и возможности систем — источников данных. Определение стандартов доменов для ХД является одной из важных задач проектировщиков ХД. Правила

преобразования доменов атрибутов систем — источников данных в домены атрибутов ХД следует фиксировать в метаданных ХД.

Все атрибуты в ХД должны согласованно использовать predetermined значения. В различных приложениях могут быть приняты различные соглашения по predetermined значениям атрибутов. К таким predetermined значениям относятся значения по умолчанию, значения, заменяющие null-значения, и т. п. Например, признак пола в различных системах может иметь различные значения: в одних это символьные значения "М" и "Ж", в других — цифровые значения 0 и 1. Более неприятным примером является случай, когда одно значение данных используется в приложении в нескольких целях, т.е. атрибут на самом деле представляет множественное значение. Например, когда в атрибуте "тип метода измерения" две первые цифры означают метод измерения, а две вторые — метод физического контроля измерения. Такие различные значения перед загрузкой в ХД должны быть преобразованы к принятому в ХД predetermined значению.

В некоторых системах — источниках данных могут отсутствовать значения (проблема пропущенных значений, "missing data") или преобразование для них не может быть выполнено ("corrupt data" — данные, для которых преобразование не может быть выполнено). Важно, чтобы в процессе преобразования такие данные принимали в ХД значения, которые позволяли бы пользователям интерпретировать их правильно. Одним атрибутам можно просто назначить разумное значение по умолчанию в случае отсутствия значения или конфликтов при преобразовании, а другим атрибутам — определить значения из значений прочих атрибутов. Например, пусть в сущности "Заказ" значение атрибута единицы измерения товара пропущено. Это значение может быть получено из соответствующего атрибута сущности "Товар" этой системы-источника. Для некоторых атрибутов не существует подходящих значений по умолчанию в случае, когда их значения отсутствуют. Для таких пропущенных значений в ХД следует также определять значение по умолчанию, например, как null-значение.

Таким образом, в процессе преобразования данных проектировщик ХД должен привести данные систем-источников к определенным стандартам, а именно:

- стандартизовать наименования атрибутов в ХД;
- определить одинаковые домены для одних и тех же атрибутов различных систем-источников;
- принять соглашения о значениях по умолчанию для пропущенных данных;
- принять соглашения о predetermined значениях атрибутов.

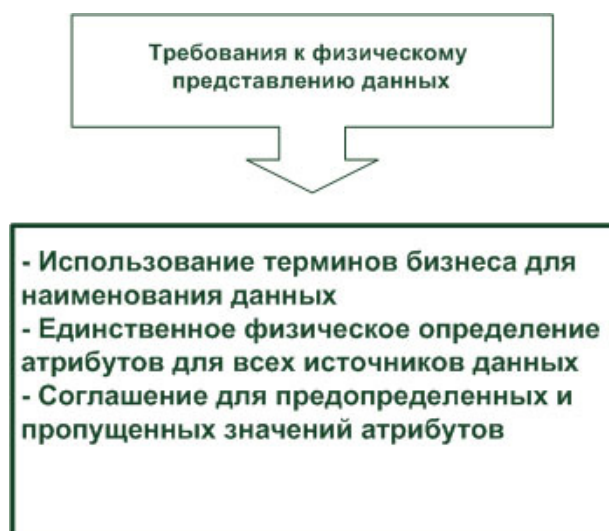


Рис. 1.6. Стандарты физического представления данных

ГЛАВА 2. АРХИТЕКТУРА ХРАНИЛИЩ ДАННЫХ

Концепция ХД развивалась по мере расширения сферы применения. Вначале под ХД понимался набор предметно-ориентированных, интегрированных, не меняющихся во времени исторических данных, предназначенных для принятия решений руководством.

Потом стало очевидным, что ХД обладают определенной внутренней структурой. Они содержат базовые данные, которые образуют единый источник для обработки данных во всех системах поддержки принятия решений (DSS). С помощью ХД можно выполнить согласование данных, несмотря на разногласие данных-источников. А элементарные данные, присутствующие в ХД, могут быть представлены в различной форме, отвечая не только известным требованиям, но и еще неизвестным.

ХД обычно имеют очень большой объем данных, поскольку в них содержатся исторические и детализированные данные, от нескольких терабайт и больше. По частоте использования данные в ХД подразделяются на два класса: активно и неактивно используемые данные. Большой объем неактивно используемых данных может значительно снизить производительность обработки запросов к ХД.

ХД содержат интегрированные данные. Они интегрированы на множестве уровней: на уровне ключа, атрибута, на описательном, структурном уровне и так далее. Общие данные и общая обработка данных консолидированы и являются единообразными для всех данных, которые обладают структурным сходством.

Несмотря на то, что указанные выше характеристики являются общими для всех ХД, в настоящее время довольно трудно типизировать и классифицировать всевозможные ХД. Можно предложить некоторую классификацию ХД в зависимости от характеристик предметной области, которые придают ХД индивидуальные особенности. Классификация архитектурных программно-аппаратных решений будет дана в следующей лекции.

Далее будут кратко описаны типы ХД по Б. Инмону. В основу его классификации положен отраслевой принцип применения ХД.

§ 2.1. Типы хранилищ данных

В большинстве случаев финансовые ХД организации строят в первую очередь. Создание **финансового ХД** — необходимый компонент финансовой инфраструктуры любой организации.

- Финансовые данные всегда находятся в центре внимания руководства организации. Поэтому привлечь интерес к созданию такой информационной системы данных очень легко.
- Финансовая активность большинства организаций (за исключением финансово-кредитных учреждений) невелика, поэтому объемы финансовых данных не очень большие, скорость поступления данных также невелика. Финансовые данные хорошо структурированы. Поэтому имеющиеся программно-аппаратные средства позволяют создать и поддерживать компактные финансовые ХД.
- Финансы охватывают все аспекты функционирования организации и имеют один общий знаменатель — деньги.
- Финансовые данные по своей природе имеют структуру, на которую напрямую влияет повседневная практика обработки финансовой информации.

По этим причинам финансы становятся самой предпочтительной областью построения корпоративного ХД. Однако финансовые ХД имеют серьезные, присущие только этому типу проблемы. Первая проблема заключается в следующем. Руководство организации ожидает, что сведения из финансовых ХД будут с точностью до одной копейки совпадать с данными существующей финансовой среды. Ожидание того, что информация в финансовом ХД должна точь-в-точь совпасть с цифрами из текущего финансового отчета, является глубоко ошибочным. Люди (то есть финансовые работники), которые так думают, просто не понимают, что, когда данные переходят из операционной среды в финансовое ХД, происходит их трансформация. А когда данные перетекают из мира приложений в реальный мир организации, их рассматривают в другом измерении. При такой трансформации происходит следующее.

- Меняются отчетные периоды. В операционной среде отчетный период завершается в конце месяца, а в ХД отчетный период заканчивается на корпоративном календаре, например, 15-го числа месяца.
- Меняются схемы группировки и кодирования счетов. В операционной среде данные рассчитываются в соответствии с планом бухгалтерских счетов, а в финансовой среде всей организации может быть совершенно другой набор схемы группировки и кодирования.
- Меняются классификации данных.
- Меняются валюты. Операционные денежные средства соответствуют той валюте, в которой они обращаются: рубли, доллары, евро, фунты и так далее. В глобальной среде деньги преобразуются к одной общей валюте.

Как видно, существует много причин, почему данные, находящиеся в ХД, отличаются от данных операционных систем. Финансовые работники думают иначе, и поэтому необходимо им разъяснять, что такое трансформация и что означают различные измерения данных.

ХД в области страхования, за некоторыми небольшими исключениями, похожи на все другие. Первое исключение (характерное для западных компаний) заключается в том, что продолжительность существования имеющихся ХД очень велика. Такие ХД содержат данные, которые являются очень старыми (до начала XX века)

Второе отличие этих ХД определяется датами, сведения о которых хранятся в этой сфере деятельности. Среда страхования — по каким бы то ни было причинам — отличается наличием огромного числа дат, связанных с бизнесом, большим, чем в каком-либо другом виде деятельности. Так, в сфере розничной торговли имеется несколько важных дат: дата продажи, дата появления на складе, возможно, дата производства. В банковском деле существенна дата транзакции. В телекоммуникации — дата телефонного звонка. В страховании же присутствуют даты всевозможных типов.

Наконец, третье отличие заключается в том, что эти ХД используют свой рабочий цикл деловой активности. Большинство организаций имеет весьма ограниченный и короткий экономический цикл. Так, в банках это — обналичивание чека. В торговле — покупка изделия. В телефонной компании —

звонок. В страховании им может быть заявка на страховое возмещение, которая может быть удовлетворена спустя пять лет, или закрытие полиса может сопровождаться двухмесячной отсрочкой. В итоге скорость, с которой функционирует страхование, отличается от скорости, характерной для других отраслей.

Эта разница в скорости отражается в ХД, как в зеркале. В других ХД транзакции просто собираются и обрабатываются. В области страхования транзакция может откладываться на неопределенный срок, а ее различные части могут отражаться в ХД, т.е. существенным становится не только сама операция, но и ее состояние. Результатом этого является совершенно особый подход при проектировании и внедрении таких ХД.

Хранилища данных для управления персоналом

ХД для управления людскими ресурсами имеют весьма существенные отличия от других ХД. Первое отличие — число предметных областей. Такое ХД неизбежно имеет одну важную предметную область — это работник. Практически все остальное подчинено этой области или занимает второстепенное положение. Большинство же других ХД имеют несколько базовых предметных областей.

Основное отличие ХД для управления людскими ресурсами состоит в том, что они используют очень мало транзакций. Так, имеется дата, когда субъект становится работником; дата, когда человек увольняется; годовые прибавки и повышения. Но, кроме транзакций фонда заработной платы и прочих редких, сгенерированных работником, транзакций, в таком ХД практически больше ничего и нет. Сравните сферу управления людскими ресурсами с коммуникацией или банковской средой, и разница в числе транзакций станет очевидной.

Эта разница в числе транзакций является причиной возникновения определенной сложности, которая заключается в том, что в области управления человеческими ресурсами наблюдается тенденция к объединению операционной обработки людских ресурсов и обработки людских ресурсов для систем принятия решения в одну среду. В других же отраслях соблазн совершить такую архитектурную ошибку весьма невелик.

Глобальные хранилища данных

Глобальные хранилища данных предназначены для глобального представления деятельности организации. Различают три типа таких ХД.

- Географически превалирующая обработка данных. Например, необходимо интегрировать бизнес в Гонконге с бизнесом в Париже, который, в свою очередь, следует интегрировать с Москвой, а тот — с Владивостоком.
- Функционально превалирующая обработка данных. Производственная деятельность должна быть интегрирована с поставками, которые необходимо интегрировать с продажами, а те — с исследованиями и так далее.
- Отраслевая превалирующая обработка данных. Например, требуется интегрировать печатное дело с консалтингом, который подлежит интеграции с бизнесом в сфере медицинского оборудования, а тот — со специализацией в области программного обеспечения.

Особенность глобального ХД заключается в том, что на глобальном уровне зачастую очень мало общих измерений. Единственное общее измерение — это деньги. И интеграция бизнеса может быть достигнута только с его помощью. Другие же измерения могут иметь или не иметь смысл на глобальном уровне. Так, клиент, продукт, поставщик, транзакция — все эти классические предметные области могут как присутствовать, так и отсутствовать в глобальной интегрированной сфере — глобальном ХД.

Помимо этого, глобальное ХД должно непрерывно реагировать на возможные изменения в бизнес-данных. В этом случае такие изменения, как правило, носят постоянный характер. Поэтому структура и технология, используемая для размещения и обслуживания глобального ХД, должна позволять поддерживать эти непрерывные перемены.

Хранилища данных с возможностями обнаружения новых данных (Data Mining)

ХД, поддерживающие технологию обнаружения новых данных (Data Mining), являются гибридом классических ХД. Они используются для выполнения мощной статистической обработки данных. Эти ХД являются:

- очень детальными;
- глубоко историческими;

- оптимизированными для статистического анализа.

Кроме того, для таких ХД характерна ориентация на какой-либо проект. Это означает, что, в отличие от всех других типов ХД, в большинстве случаев их перестают использовать сразу по завершении анализа, ради которого они создавались.

Еще одно важное отличие ХД с возможностями анализа заключается в том, что оно очень часто включает внешние данные. Такие данные очень полезны с точки зрения прогнозирования изменения бизнес-данных, которое не так легко увидеть без их участия.

Хранилища данных в области телекоммуникаций

Отличительная особенность этих ХД состоит в том, что они в значительной степени определяются данными, касающимися факта телефонных разговоров. Разумеется, в отрасли телекоммуникации присутствует множество других типов данных. Но ни одна другая область ХД не предопределяется в такой степени размером одной предметной области — деталями на уровне разговора.

Существуют несколько способов хранения подробностей на уровне телефонного разговора:

- хранение деталей на уровне разговора только за несколько месяцев;
- хранение множества деталей на уровне разговора, размещенных на различных носителях;
- резюмирование или агрегирование деталей на уровне разговора;
- хранение только отобранных деталей на уровне разговора, и так далее.

К сожалению, несмотря на разнообразие методов обработки, для данного ХД обработка может быть выполнена только над деталями на уровне разговора. А работа на итоговом или агрегированном уровне просто невозможна.

Характерные особенности различных типов ХД и доводы в пользу их внедрения

ХД — это логически интегрированный источник данных для систем принятия решений, информационных систем руководителей, систем анализа данных и систем обнаружения новых данных (Data Mining). ХД предназначено для информационной поддержки анализа данных, принятия решений, т.е.

информационной поддержки деятельности, а не собственно поддержки каждодневных бизнес-процедур организации, и поэтому многие принципы технологии БД утрачивают в ХД свое значение.

§ 2.2. Выбор архитектуры ХД и методы его проектирования

ХД ориентируется на определенную предметную область и организуется так, чтобы решать конкретные задачи анализа и информационной поддержки деятельности организации. Данные различных источников агрегируются ХД, приобретая при этом статус неизменяемых. Для ХД характерно массовое добавление данных и фактическое отсутствие операций обновления. Процесс пополнения данных включает в себя сложные процедуры очистки данных: устранения несоответствия типов, размеров и других свойств данных.

Основные отличия различных типов ХД состоят в следующем.

- Данные финансовых ХД — а именно их обычно создают в первую очередь — не будут с точностью до одной копейки совпадать с информацией в существующей финансовой среде.
- ХД в области страхования отличаются от других продолжительностью существования, а также разнообразием дат и продолжительностью экономического цикла.
- Для ХД управления человеческими ресурсами характерна только одна основная предметная область.
- ХД с возможностями обнаружения новых данных (Data Mining) и исследования данных (Exploration Data Warehouse), которые используются для выполнения мощной статистической обработки данных, являются гибридом классических ХД.
- Отличительная особенность ХД в области телекоммуникаций состоит в том, что они в значительной степени определяются данными, сгенерированными в одной предметной области.

Вне всяких сомнений, будут появляться и другие типы ХД, каждому из которых присущи свои отличительные особенности.

Технология хранения данных обеспечивает адекватную основу для информационной поддержки деятельности руководителей организаций в области

принятия решений и дает преимущества в тех областях деятельности, которые связаны с управлением и использованием долговременно хранимой информации, а именно:

- организация получает взгляд на данные как на единое целое. Например, это дает ответы на такие вопросы, как:
 - Сколько продуктов реально производится?
 - Что влияет на изменение спроса?
 - Какие товары или услуги приносят наибольший доход?
 - Каковы особенности и предпочтения ваших клиентов?
- среднее суммарное значение за три года возврата инвестиций, вложенных в создание ХД, составляет 400% (по результатам трехлетнего исследования опыта 62 корпораций, проведенного компанией IDC). При этом средняя стоимость организации хранилища данных составляет 2,2 млн долл., а среднее время самоокупаемости - 2,3 года. По оценкам, 90% участвовавших в этом проекте компаний получили более 40% прибыли на инвестированный капитал, а у 50% этот показатель составил более 160%;
- возрастает надежность данных для принятия решений. Данные, загружаемые в ХД, подвергаются очистке - согласуются, проверяются, уточняются;
- появляется возможность эффективного геопространственного анализа данных. Анализ такой информации имеет решающее значение в принятии решений по всем вопросам, связанным с географией бизнеса;
- исследование трендов и колебаний в бизнес-данных с помощью ХД позволяет достаточно надежно прогнозировать развитие бизнес-процессов организации во времени.

Одной из главных целей разработки ХД 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 ХД. 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.

ХД – это сложная компьютерная система. Под архитектурой ХД понимают совокупность программно-аппаратных компонент, совокупность технологических и организационных решений, предпринимаемых для создания, разработки и функционирования ХД, т.е. выбор аппаратного и программного обеспечения, выбор способов взаимодействия программно-аппаратных компонент, выбор способа решения проектной задачи по разработке и созданию ХД. 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.

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 ХД 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 рис. 2.1 приведена типовая обобщенная концептуальная схема для архитектуры ХД. В конкретных решениях по архитектуре ХД некоторые компоненты схемы могут отсутствовать.



Рис. 2.1. 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 малоприспособленными 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 рис. 2.2.

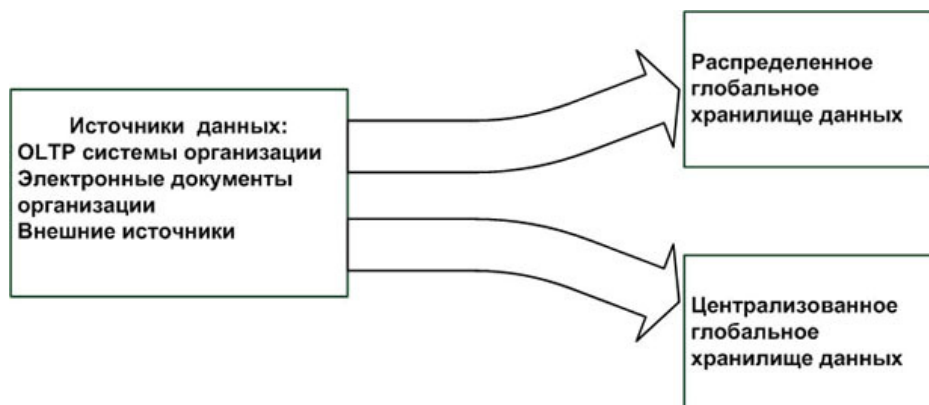


Рис. 2.2. Основные архитектурные решения для глобального ХД

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.

§ 2.3. Модель типового проекта создания хранилища данных

Сложные компьютерные системы, к которым относятся информационные системы с ХД (системы складирования данных, информационно-аналитические системы, BI-системы и системы бизнес-аналитики), создаются командами ИТ-специалистов. Работа команды является процессом, который требует тонкой психологической настройки, мобилизации усилий каждого члена команды для реализации поставленных в проекте целей. Каждый член такой команды должен знать, какова цель проекта, какие задачи решаются в рамках проекта, какие задачи и в какие сроки должен решить каждый член команды. Помимо решения своих непосредственных задач, проектировщик ХД должен четко представлять себе:

- от кого и какие данные он должен получить (оценить риск получения неполных данных);
- какие и кому передать результаты решения своих задач (оценить риск срыва сроков их решения);
- с кем и когда согласовывать свои проектные решения (оценить риск межличностных конфликтов в команде);
- в каких точках реализации проекта он может получить распоряжение пересмотреть результаты решения своих задач (оценить риск повторной работы) и т. д.

Плохо спроектированная структура ХД обычно приводит к сложности реализации ХД и зачастую к увеличению сроков проекта примерно в три раза. Плохо организованный проект создания ХД, даже при качественном проектировании, ведет в более чем 50% случаев к краху проекта. Нередко проекты создания ХД реализуются со второй и даже с шестой попыток.

Организация любого проекта разработки ХД (используемая бизнес-модель) во многом определяет его успешность. В основу организационной структуры проекта, как правило, закладывается некоторая модель жизненного цикла создаваемого объекта.

Жизненный цикл разработки хранилища данных

Главная цель создания ХД состоит в том, чтобы собрать вместе информацию из различных источников и представить эту информацию в формате, который является удобным для принятия решений по основным направлениям деятельности организации. Действия, необходимые для достижения поставленной цели, оказываются значительно более сложными, чем просто собрать данные и получать из них ответы на запланированные вопросы или получить ряд отчетов.

Подготовка данных для таких систем требует всесторонней экспертизы со стороны как технических специалистов, так специалистов предметной области, что обычно включает в себя:

- точную идентификацию бизнес-информации, которая должна храниться в ХД;
- идентификацию предметных областей (бизнес-направлений деятельности организации) и определение их приоритетов, которые составят в итоге набор предметных областей ХД;
- управление границами каждой предметной области, которая будет представлена в ХД на интерактивной основе;
- разработку масштабируемой архитектуры для технического обслуживания ХД и его приложений, точное определение и обоснованный выбор структурных компонент (аппаратной, программной, лингвистической и организационной), которые будут реализовать ХД;
- определение процедур экстрагирования, очистки, агрегации, преобразования и проверки данных для обеспечения гарантии их точности и согласованности;
- определение корректных уровней суммирования числовых показателей для поддержки бизнес-решений;

- определение и разработку процедур своевременного обновления данных, которые отвечают производственным потребностям, временным и жизненным циклам бизнес-информации;
- разработку дружественных интерфейсов и мощных инструментов аналитической обработки данных для конечных пользователей ХД;
- разработку словаря метаданных для обеспечения общения с ХД широких групп пользователей;
- обучение персонала для работы с приложениями ХД, всестороннее освещение круга решения задач, которые возможно решить в системе складирования данных;
- определение и регламентирование процесса сопровождения и обслуживания ХД и т.д.

Таким образом, ХД создаются для информационного обеспечения в решении стратегических задач по определенным направлениям бизнеса (для удовлетворения потребностей руководства и бизнес-аналитиков организации), а не для удовлетворения потребностей в автоматизации ежедневных процедур обработки бизнес-информации сотрудников структурных подразделений организации. В этом и состоит главное и коренное отличие систем складирования данных и систем бизнес-аналитики, использующих технологию ХД, от систем оперативной обработки информации (OLTP-систем, или On-Line Transactions Processing).

Факторы, влияющие на структуру проекта создания хранилища данных

В современных экономических условиях стратегический бизнес-план организации разрабатывается на два-три года. Бизнес-процедуры, как правило, также пересматриваются в организации каждые два-четыре года. Это приводит к пересмотру функций и реинжинирингу OLTP-систем организации. ХД создаются на пять и более лет, в них накапливается огромный объем информации. ХД предназначено жить, даже если половина бизнес направлений организации будет закрыта, а на их смену придут новые. Информация не теряется, она остается и будет доступна для анализа предыдущего опыта организации. Весьма расточительно планировать ХД так, чтобы через два года его переделывать.

Если язык общения системы бизнес-аналитики с конечными пользователями будет им не понятен, весь смысл и экономический эффект от разработки и внедрения такой системы будет сведен на ноль. Поэтому сбор и систематизация бизнес-терминологии, разработка открытой и гибкой структуры лингвистического обеспечения, ее внедрение в структуру ХД (а не в аналитические бизнес-приложения) является важной задачей проекта создания ХД.

Системы бизнес-аналитики создаются для лиц, принимающих решения, или приравненных к ним ведущих специалистов компаний. Эти системы обслуживают в первую очередь стратегические решения. Наивно будет полагать, что любой персонал способен принимать подобные решения для организации. Однако такая постановка вопроса не мешает пользоваться персоналу организации ХД как проверенным информационным источником накопленных компанией знаний.

Можно возразить, приводя примеры систем складирования данных в научных исследованиях. Ими пользуются многие участники научного сообщества — но в качестве источника информации, научного справочника. Однако можно смело утверждать, что научные ХД создаются для определенной группы ученых. Такая группа имеет научного руководителя, который и определяет направленность работы с информацией. В основном это получение новых знаний, разработка и апробирование эмпирических и математических моделей с целью получения новых научных знаний. В этом отношении научные ХД практически ничем не отличаются от ХД коммерческих организаций. Последним также нужны скрытые знания о своем бизнесе.

Таким образом, в проекте ХД вес разработки каждого структурного компонента приблизительно одинаков, ничто: ни аппаратные решения, ни программное, ни сетевое (включая middleware), ни лингвистическое, ни технологическое обеспечение — не должно быть обойдено вниманием. Недооценка этого факта приводит к краху проекта создания ХД. Поэтому тщательное планирование и организационная структура проекта создания ХД так важна.

Модель жизненного цикла хранилища данных

В основу организационной структуры проекта создания или разработки ХД закладывается определенная **модель** (или представление) жизненного цикла ХД как продукта. Под жизненным циклом продукта понимается набор определенным образом расположенных во времени фаз (этапов), которые проходит продукт от момента его создания до момента его утилизации.

В зависимости от квалификации руководителя и масштаба проекта создания ХД используются модели жизненного цикла различной степени детализации. В настоящей лекции мы не будем углубляться в анализ различных методологических подходов к построению моделей жизненного цикла ХД. Все они основаны на обоснованном выборе того или иного способа разбиения процесса разработки ХД на этапы (фазы), идентификации задач каждого этапа и расположения выбранных этапов в пространстве и во времени. Отметим, что в настоящее время различные исследователи предлагают разбиение процесса разработки ХД на три, четыре, пять и более этапов. Отметим также, что все модели жизненного цикла разработки ХД имеют общие (типовые) элементы, которые в разных моделях могут быть отнесены к различным этапам.

Рассмотрим типовые элементы процесса создания ХД, которые и определяют его жизненный цикл. При этом предлагаемая модель не является единственно верной, так как сам процесс разработки не является строго структурированной задачей, она отражает лишь общий взгляд на этот процесс.

После принятия решения о создании ХД, до начала выполнения этапа планирования формируется проектная команда, определяются роли и задачи ее участников. Поэтому проектировщику важно представлять, что должны делать другие члены команды, а для этого нужно понимать свое место и роль в процессе разработки ХД.

Процесс создания и разработки ХД (жизненный цикл разработки хранилища данных) в общем случае можно представить состоящим из следующих основных стадий:

- планирование;
- формулирование требований к системе складирования данных;
- анализ;
- проектирование;

- конструирование;
- внедрение;
- поддержка.

Этап поддержки ХД в жизнеспособном состоянии включен для полноты представления жизненного цикла разработки ХД. Как правило, поддержка ХД осуществляется ИТ-службами заказчика в рамках отдельного проекта и имеет свой жизненный цикл. Отметим также, что в настоящее время внедрение также стало рассматриваться как самостоятельный проект — совместный (корпоративный) проект разработчика и заказчика ХД.

Охарактеризуем кратко задачи каждого из этих этапов.

Планирование

На этапе планирования решаются следующие задачи: выбор стратегии реализации и методологии разработки, анализ задач, для решения которых создается ХД, анализ ресурсов разработки с технологической точки зрения и с точки зрения задач бизнеса, выбор архитектуры ХД, определяется бюджет проекта, разрабатываются возможные сценарии использования ХД, начинается сбор метаданных для ХД.



Рис. 2.3. Этап планирования разработки хранилища данных

Целью этого этапа является идентификация задач ХД, выбор способа решения этих задач, определения программно-технологического объекта и того, как, в какие сроки и за какие деньги этот объект будет реализоваться.

Выбор стратегии реализации определяет подход, который будет использован при создании ХД. Обычно используют следующие подходы: Top Down ("сверху вниз"), Bottom Up ("снизу вверх"), Middle In ("из середины") и комбинированный подход, который в последнее время становится все более популярным. Подход "сверху вниз" выбирается для вновь создаваемого ХД, т.е. когда "с нуля" принимаются все решения о технологической реализации объекта (аппаратура, программное обеспечение и т.д.). Подход "снизу вверх" используется, когда уже есть определенная вычислительная среда и объекты, из которых можно построить новый объект. Подход "из середины" предполагает эволюционное, поэтапное создание объекта, когда сначала разрабатывается так называемое ядро объекта, которое на следующих этапах наращивается новой функциональностью. Комбинированный подход применяет комбинацию выше перечисленных подходов.

Стратегия реализации определяет, какая концептуальная архитектура ХД будет использоваться и как будет строиться ХД с точки зрения последовательности реализации выбранной концептуальной архитектуры.

Выбор методологии создания ХД определяет, образно говоря, язык проекта, на котором будут разговаривать члены проектной команды, как будет оформлена техническая документация, какие принципы разработки будут использоваться. Это могут быть метод структурного анализа и проектирования (SADT), спиральный метод, методы, основанные на применении UML (язык объектно-ориентированного моделирования), и т.п.

Методология разработки основывается на использовании типичной для БД концепции трех схем и включает методы анализа спецификаций, методы концептуального, логического и физического проектирования.

Метод структурного анализа и проектирования является хорошо разработанной методикой и опирается на использование стандартов и методик IDEF или аналогичных им. Методика создания ХД по спирали реализует концепцию эволюционного подхода к созданию системы. Использование методов

объектно-ориентированного анализа для реляционных ХД потребует преобразования полученной объектной схемы несущей базы данных в реляционную схему.

Анализ задач ХД предполагает идентификацию и определение объектов бизнеса, информация о которых будет содержаться в ХД. Вот далеко не полный перечень вопросов, на которые следует получить ответ на этой стадии:

- Что является предметной областью для хранилища данных?
- Какие программно-аппаратные платформы используются или какие планируется использовать?
- Какие возможности планируются в терминах свойств, характеристик и функций?
- Что представляют собой источники данных, которые можно или нужно интегрировать в хранилище данных?
- Когда хранилище данных должно начать функционировать?

Определение предметной (тематической) ориентации ХД является одной из самых главных задач на этом этапе. В рамках каждого фрагмента предметной области определяется:

- количество и типы источников данных (подразделения организации);
- число выбранных источников данных;
- данные, которые будут храниться в ХД;
- цели использования данных;
- каким должен быть спектр вопросов, на который должно отвечать ХД;
- каков размер метамодели хранилища данных;
- каков размер данных в хранилище данных;
- каковы источники входных данных и их количество;
- насколько востребованными являются данные из источников данных;
- насколько хорошо документированы данные из источников данных;
- каковы уровни управления организации (IFC);
- какие инструментальные средства доступны для логического проектирования;
- доступна ли модель данных в масштабе организации;

- есть ли профессионально подготовленный персонал;
- будет ли хранилище данных реализоваться в рамках существующей программно-аппаратной платформы или на платформе, аналогичной существующей.

На этой стадии планирования решается вопрос, что должно быть реализовано сначала, что можно реализовать в дальнейшем и, самое главное, что не будет реализовано в этом проекте создания хранилища данных. Задача состоит в том, чтобы четко ограничить предметную область хранилища данных и понять (переформулировать) спецификации для хранилища данных.

Выбор архитектурных решений должен задать такие ограничения:

- степень использования данных систем оперативной обработки;
- только хранилище данных;
- только киоски данных;
- хранилище и киоски данных;
- инфраструктура секционирования данных;
- использование архитектуры клиент-сервер (двухзвенная или трехзвенная архитектура).

Разработка программы проекта и бюджетного плана имеет целью составить план проекта, построить бюджет проекта, оценить стоимость и обеспечение методов для оценки окупаемости ХД в терминах:

- оценка стоимости восстановления и стоимости хранения;
- оценка возможности организации;
- оценка доходности организации;
- оценка расширения рынка;
- оценка конкурентного преимущества;
- оценка удовлетворения потребностей клиентов.

На этом этапе разработки ХД должны работать в тесном взаимодействии с экспертами и экономистами.

Проработка сценариев типовой коммерческой деятельности для согласования с конечными пользователями является тонкой задачей, решение которой дает ответ на вопрос, как пользователи видят использование ХД. Решение задачи очень важно не только с точки зрения дальнейшей реализации

проекта, но с точки зрения снижения риска неоправданных ожиданий от внедрения ХД.

Обычно анализ сценариев строится по следующей схеме. Каждый сценарий состоит из:

- четко определенного конечного пользователя с точно определенной его ролью в бизнес-процессах;
- тематической области, которая представляется в хранилище или киоске данных и будет использоваться данным конечным пользователем;
- одного или более вопросов (задач) по ключевым проблемам тематической области, которые воплощены неудовлетворительно в существующих ИС и полезны для решения следующих задач:
 - построение адекватных критериев запросов для ХД;
 - идентификация метамодели ХД;
 - идентификация объема необходимой хронологической информации;
 - идентификация точек зрения пользователя на данные (измерений);
 - идентификация потребностей по данным киосков и ХД.

Цель настоящего этапа — очертить круг вопросов, на которые должна отвечать система складирования данных.

Цель следующей стадии, сбора метаданных, — сформировать спецификации по описанию данных и описать все необходимые элементы данных ХД в терминах бизнес-процессов и процедур бизнеса.

На этом этап планирования заканчивается, и команда проекта переходит к формированию каталога требований проекта. Без этого документа не имеет смысла продолжать проект, в намеченный срок он не закончится и в планируемый бюджет не вложится.

§ 2.4. Разработка требований

Этап разработки требований к ХД включает в себя следующие стадии:

- определение требований владельца ХД;
- определение требований конечных пользователей;
- определение технологических требований;
- определение архитектурных требований.



Рис. 2.4. Этап разработки требований к хранилищу данных

Требования владельца связаны с уточнением предметно-ориентированной направленности ХД, требованиями потенциального роста предметной области по:

- целям бизнеса;
- масштабу и целям хранилища данных/киосков данных, по клиентам и посредникам;
- требованиям клиентов;
- источникам данных;
- плану — бюджет, календарный план-график, ресурсы;
- влиянию на текущие инвестиции — люди, технология, обучение.

Часть требований бизнеса является спецификацией предметной области для ХД — спецификацией бизнес-функций направлений деятельности определенной категории потенциальных пользователей (идентификация интереса). Рассмотрим, например, отдел маркетинга компании. Отдел маркетинга может интересоваться одна из следующих тем (направлений деятельности):

- маркетинговые исследования;
- анализ конкурентоспособности;
- анализ поведения покупателя;
- сегментация (рынка) реализации продукта;
- ценовые и бюджетные решения;
- решения по продукции;
- решения по продвижению;
- решения по каналам сбыта;
- прогнозирование тенденций;
- сертификация (Benchmarking).

Анализ выбранных выше направлений имеет дело со спецификацией предметных областей для отдела маркетинга, которыми в данном случае будут являться "Заказы", "Рынки", "Продажи", "Продвижение продукции", "Временные циклы".

Здесь важен такой параметр, как **степень детализации** (granularity) – чем ниже степень детализации, тем больше количество деталей. Это ключевая информация для проектировщика ХД, поскольку она определяет потенциальный объем ХД, скорость его роста, потенциальную производительность обработки запросов и т.д.

На этой стадии идентифицируется, в каких разрезах будут исследоваться данные потенциальными пользователями ХД. Для примера с отделом маркетинга это может быть:

- временной цикл с детализацией "День", "Неделя", "Месяц", "Квартал", "Год";
- группы клиентов с детализацией "Покупатели", "Сегмент рынка", "Рынок", "Отрасль промышленности";
- семейство продуктов с детализацией "Продукция", "Семейство продуктов", "Линия продуктов";
- география и месторасположение с детализацией "Район", "Регион страны", "Страна", "Международный регион";
- структура организации с детализацией "Подразделение", "Отдел", "Филиал", "Организация".

Здесь может быть также учтена специфика как самой организации, так и сектора промышленности и рынка, на которых работает организация.

На стадии определения требований по архитектуре ХД проводится планирование архитектуры в масштабе предприятия и по его структурным подразделениям, определяется архитектура данных в рамках принятой модели (например, ER-модели и т.д.), определяется архитектура приложений ХД, разрабатывается каталог приложений и функции каждого из них. На этой стадии полезно получить CRUD- (Create, Read, Update, Delete — создание, чтение, обновление, удаление) матрицу [44-49], чтобы на стадии проектирования физической структуры оценить частоту использования данных приложениями.

На стадии определения требований разработки создаются:

- спецификация приложений, интерфейсов, компьютеров, баз данных, коммуникаций и форм для конечного пользователя;
- технологические требования;
- требования размещения ПО;
- требования по поддержке ХД в программных продуктах;
- требования для разработчиков и обслуживающего персонала по их квалификации.

На стадии определения требования конечных пользователей оценивается роль ХД в делопроизводстве, соответствие функциональности ХД ежедневной последовательности операций конечных пользователей, формулируются критерии запросов к данным, требования по корпоративной отчетности, требования анализа данных.

Например, классифицируются в операциях над ХД действия пользователей, такие как:

- разделение элементов данных по различным аспектам анализа (операции Slice and dice);
- развертывание и экспозиция элементов данных по уровням иерархии (операции Drill-down up);
- поиск скрытых закономерностей в данных (алгоритмы Data Mining);
- просмотр данных непредусмотренным образом (Datasurfing);
- загрузка и выполнение модификаций данных.

В заключение этой стадии разрабатывается также набор бизнес-моделей использования ХД, после чего принимаются решения по просмотру данных. Например, принимается решение о том, как будут просматриваться различные категории данных:

- с помощью стандартной визуализации в рамках реляционных СУБД (двумерные таблицы);
- с помощью многомерных кубов данных (OLAP);
- с помощью предусмотренных отчетов и диаграмм;
- на основе жизненных циклов элементов данных.

Главным результатом этого этапа является создание каталога требований.

Располагая каталогом требований, можно приступить к реализации **этапа анализа** —получения согласованных по источникам логической модели и определения набора инструментальных средств для работы с ХД. Стадии



Рис. 2.5. Этап анализа

Цель первой стадии этого этапа состоит в разработке логической модели данных для ХД и киосков данных. Кроме создания логической модели, необходимо фиксировать модели логической структуры баз данных подающих систем.

На следующей стадии определяются процессы, которые необходимы для связи источников данных и ХД, процедур очистки, преобразования и агрегации данных источников, производится выбор инструментальных средств для загрузки данных в ХД и выбор инструментальных средств конечного пользователя, которые должны быть размещены на клиентских компьютерах.

Проектирование

После завершения стадии анализа переходят к реализации **стадии проектирования**. Цель этого этапа — разработка физической модели ХД, проектирование процедур поступления данных в него и проектирование архитектуры приложений.

Проектирование ХД можно разложить на две равноценных стадии – проектирование архитектуры данных (логическое и физическое проектирование) и архитектуры приложений (анализ запросов и фиксация процессов взаимодействия хранилища данных с внешними источниками и пользователями).

Проектирование



Рис. 2.6. Этап проектирования хранилища данных

Детальное проектирование архитектуры данных включает в себя:

- **логическое проектирование.** Разработка логических моделей данных для ХД и киосков данных в рабочем пространстве базы данных. Отображение логических моделей данных источников данных в логические модели ХД и киосков данных;
- **физическое проектирование.** Отображение логических объектов в физическое описание ХД. Денормализация логической модели. Создание табличного пространства, секционирование, создание индексов и ограничений. Оценка объема физического ввода-вывода.

На стадии логического проектирования рассматриваются логические взаимоотношения между объектами предметной области. Особенностью логического проектирования ХД является тотальная ориентация на потребности и задачи конечного пользователя — бизнес-аналитиков и руководителей организации различных уровней. Они будут анализировать данные и работать с агрегированными данными, а не с данными конкретной бизнес-операции. Этот класс пользователей, как правило, обычно не знает точно, что им все-таки нужно от данных. И здесь важно оценить уровень претензий к данным и очертить круг задач, которые могут возникнуть.

В процессе логического проектирования выделяется набор объектов предметной области с их атрибутами. Объект представляет собой фрагмент

информации о предметной области. В реляционных базах данных объект отображается в таблицу, а его атрибуты отображаются в колонки такой таблицы. В логическом проектировании ХД для создания ER-диаграмм используется многомерное моделирование, которое, грубо говоря, сводится к идентификации информации об объекте в виде фактов (таблицы фактов) и идентификации информации, с помощью которой на эти факты можно посмотреть (таблицы измерений). Результатом стадии логического проектирования является логическая схема ХД и, возможно, отображение логической схемы источников данных (подающих систем) на логическую схему ХД.

На стадии физического проектирования рассматриваются задачи размещения данных в БД для эффективной их выборки. На этой стадии (для реляционных хранилищ данных) реально пишутся SQL-операторы. Данные, собранные на стадии логического проектирования, превращаются в описание физической БД, включая таблицы, уникальные идентификаторы объектов – в первичные ключи, ограничения по значениям данных, взаимоотношения между объектами – во внешние ключи, индексы, табличные пространства, разбиения и представления.

Назначение табличного пространства состоит в отделении таблиц от их индексов, маленьких таблиц от больших таблиц, т.е. является механизмом для решения задачи оптимального размещения данных (некоторые СУБД решают эту задачу сами, без вмешательства пользователя).

Секционирование больших таблиц необходимо для увеличения производительности обработки запросов, т.е. является одним из механизмов решения задачи оптимизации выборки, так же, как и создание индексов.

Ограничения в ХД отличаются от ограничений в OLTP-системах. В системах складирования данных целостность и проверка данных обеспечивается на стадии загрузки данных. Поэтому роль ограничений в ХД не столь уж велика. Типичным ограничением в ХД является ограничение **NOT NULL**.

Параллельно с проектированием архитектуры данных начинается стадия детального проектирования структуры приложений, которая включает в себя определение:

- процессов, которые являются внутренними для источников данных и связаны с процедурами очистки и частичной экстракции информации;
- процессов, которые связывают источники данных с ХД и киосками данных;
- процессов, которые являются внутренними по отношению к ХД и предназначены исключительно для обслуживания данных в хранилище;
- процессов, которые связывают ХД и киоски данных;
- процессов, которые являются внутренними по отношению к киоскам данных и предназначены исключительно для обслуживания данных в них;
- процессов, которые связывают хранилище и киоски данных со средствами конечного пользователя;
- процессов, которые являются внутренними на рабочих станциях конечных пользователей и используются для установки связи с ХД и запуском средств анализа;
- процессов для поддержки управления и администрирования внутренних задач ХД как системы.

Целью этой стадии является получение спецификаций всех приложений ХД.

Все результаты стадии проектирования тщательно документируются, поскольку являются исходными и рабочими документами команды проекта и постоянно используются в ходе дальнейшей работы над проектом.

Построение хранилища данных

Следующий этап проекта создания ХД — это **построение ХД**.



Рис. 2.7. Построение хранилища данных

Цель этого этапа состоит в разработке программ и собственно физической БД под ХД. Выполнение этого этапа проекта, помимо создания собственно ХД, включает в себя разработку и отладку приложений ХД, а именно следующих групп программ:

- программы, которые создают и модифицируют БД для ХД и киосков данных;
- программы, которые экстрагируют данные из источников данных;
- программы, которые выполняют преобразования данных, такие как интеграция, суммирование и агрегация;
- программы, которые выполняют обновление реляционных БД;
- программы, которые реализуют поиск в очень больших БД.

Результатом выполнения этого этапа является комплекс программ, работающих с ХД.

Внедрение

Следующим этапом реализации проекта создание является **внедрение** в опытную эксплуатацию. Это очень ответственный и трудоемкий этап. Начинается он со стадии начальной инсталляции, включающей начальную загрузку хранилища из источников данных, и тестирования процедур обновления и синхронизации данных.



Рис. 2.8. Этап внедрения хранилища данных

Далее составляются и утверждаются планы поэтапного тестирования, тренировки и обучения персонала для всех классов пользователей, реализации обновления несущих платформ для ХД, информационных потребностей и т. д.

Выполняется обеспечение администрирования системы и данных о пользователях, возможностей архивирования и резервного копирования, возможностей восстановления, управления доступом и безопасностью, полных возможностей и процесса управления системой и ее структурными компонентами, информационного каталога/директории. Проводится целостная интеграция хранилища данных в существующую инфраструктуру организации.

Результатом выполнения этого этапа является всесторонняя подготовка перехода ХД в промышленную эксплуатацию.

Поддержка

Этап поддержки ХД в работоспособном состоянии, как уже отмечалось выше, является самостоятельным проектом. Это последний этап жизненного цикла ХД. По его завершении происходит либо уничтожение ХД как продукта, либо его реинжиниринг. Это связано с быстрыми изменениями сферы бизнеса организации, которые приводят к необходимости пересмотра стратегических и тактических планов организации. ХД, так же, как и другие элементы инфраструктуры организации, должно соответствовать генеральному бизнес-плану организации. В зависимости от структуры генерального бизнес-плана организации, эксплуатирующей ХД, продолжительность этого этапа может составлять два-четыре года, реже пять и более лет.

На этом этапе ИТ-служба организации обеспечивает:

- поддержку работоспособности и масштабируемости программно-аппаратного обеспечения ХД;
- сбор, очистку, преобразование, загрузку и актуализацию данных в соответствии с установленными бизнес-процедурами;
- поддержку автоматизированных мест пользователей.

Этот этап может также включать в себя техническую поддержку со стороны разработчика ХД, в частности консультирование, обучение пользователей или абонентское обслуживание.

Этот этап является также испытанием ХД "на прочность". Конечные пользователи, как правило, не являются ИТ-профессионалами. Они решают свои задачи. Они будут использовать ХД только в том случае, если оно будет эффективно помогать им решать их задачи. Поэтому ИТ-служба организации должна постоянно отслеживать активность пользователей на ХД и вести постоянный аудит мнений пользователей о полезности ХД в их профессиональной деятельности, отсекая при этом "зерна от плевел".

Временные затраты на реализацию этапов

Таким образом, учитывая замечание о поддержке ХД как о самостоятельном проекте, в жизненном цикле создания ХД можно выделить шесть основных этапов. Независимо от того, как заказчик ХД понимает процесс его создания, эти этапы должны быть пройдены и заложены в бюджет проекта. Они обычно выполняются в последовательном порядке, хотя некоторые из них могут разрабатываться параллельно.

ГЛАВА 3. СИСТЕМЫ ДЕЛОВОЙ ОСВЕДОМЛЕННОСТИ И ХРАНИЛИЩА ДАННЫХ

Многозначность английского слова "Intelligence" приводит к неопределенности трактовки термина "Business Intelligence" как в российских, так и в зарубежных литературных источниках, посвященных тематике использования информационных технологий для аналитической поддержки бизнеса. Английское слово "Intelligence" означает способность узнавать и понимать, готовность к пониманию, знания, переданные или приобретенные путем обучения, исследования или опыта, действие или состояние в процессе познания, разведку, разведывательные данные. В русском языке слово "интеллект" означает мыслительную способность человека.

Термин **"Business Intelligence"** получил широкое распространение, когда был введен в обращение аналитиками компании Gartner Group в конце 80-х годов прошлого века как **"пользовательцентрический процесс, включающий доступ и исследование информации, ее анализ, выработку интуиции и понимания, которые ведут к улучшенному и неформальному принятию решений"**. Хотя ранее этот термин, например, использовался в компании IBM в качестве внутрикорпоративного термина.

К 1996 году содержание термина было уточнено, и "Business Intelligence" стал пониматься как **"инструменты для анализа данных, построения отчетов и запросов, которые могут помочь бизнес-пользователям преодолеть море данных для того, чтобы помочь синтезировать из них значимую информацию"**.

В русскоязычной литературе термин "Business Intelligence" переводится как "бизнес-интеллект", "интеллектуальный анализ данных", "деловая осведомленность" или вводится просто как аббревиатура BI. В настоящем курсе мы будем использовать термины "деловая осведомленность" и "бизнес-аналитика" как синонимы.

Однако все-таки в настоящее время не существует однозначного определения термина "деловая осведомленность" (BI). Отметим следующие важные аспекты трактовки содержания данного термина.

- Деловая осведомленность понимается как процесс, методы, технологии, средства извлечения и представления знаний.
- Деловая осведомленность понимается как знания о бизнесе и для бизнеса.

Таким образом, деловая осведомленность в широком смысле слова понимается как:

- **процесс превращения данных в информацию и знания о бизнесе для обеспечения принятия улучшенных и неформальных решений;**
- **информационные технологии сбора данных, консолидации информации и обеспечения доступа пользователей к бизнес-знаниям;**
- **знания о бизнесе, полученные в результате анализа данных и консолидированной информации.**

§ 3.1. Системы бизнес-аналитики

В современном мире успех компании на рынке напрямую зависит от того, как быстро менеджмент компании может распознать изменения динамики рынка и насколько своевременно может отреагировать на них с целью увеличения прибыли, исходя из существующих реалий рынка. Менеджеры компании должны отслеживать тенденции рынка, идентифицировать конкурентов и угрозы, оценивать риски, преобразовывать стратегию компании, оценивать свои ресурсы и т.д. Информация является необходимым производственным ресурсом для принятия эффективных управленческих решений.

Компании накопили значительные объемы данных и имеют доступ к еще большим объемам внешних данных. Менеджерам необходимо, чтобы эта информация была преобразована, предварительно обработана и соответствующим образом организована для быстрого доступа, анализа и принятия решений. Такой подход к данным есть, с одной стороны, создание конкурентного преимущества, а с другой стороны – требование к публикации данных для менеджеров компании. Публикация данных для менеджеров, обеспечивающая быстрый доступ к данным, выполнение анализа данных и информационную поддержку процесса принятия решений, является основной целью систем бизнес-аналитики. Бизнес-аналитика помогает компании создавать знания из всей доступной информации для

принятия эффективных управленческих решений и превращения этих решений в действие.

Таким образом, ключевую роль в управлении организацией в целом и ее отдельными производственными функциями играет информация. Данные, которые доступны менеджерам и аналитикам непосредственно из корпоративных информационных систем, не унифицированы, разрозненны и в общем случае не готовы для анализа. Системы деловой осведомленности или бизнес-аналитики являются тем классом информационных систем, который позволяет превратить данные корпоративных информационных систем и данные из внешних источников в полезные для бизнеса информацию и знания, используемые в управлении, на основе которых можно принимать решения.

Информационным фундаментом для бизнес-анализа и систем бизнес-аналитики является хранилище данных. Основное требование к хранилищу данных системы бизнес-анализа состоит в том, чтобы обеспечить структурированную и организованную для решения задач бизнеса информационную среду. Как правило, такую среду лаконично представляют в виде информационной пирамиды.



Рис. 3.1. Информационная пирамида

Информационная пирамида формируется из нескольких уровней.

- **Уровень оперативной информации.** На этом уровне ИТ обеспечивают работу с данными на уровне бизнес-процедур компании. Данные в автоматизированных системах являются хорошо структурированными и

детальными. С этими данными работают специалисты компании: бухгалтеры, менеджеры продаж, плановики и т.д.

- **Уровень тактической информации.** На этом уровне ИТ обеспечивают интеграцию данных на уровне бизнес-процессов оперативного управления производством в рамках подразделений компании. С этими данными работают руководители подразделений компании при выполнении ежедневных производственных заданий.

- **Уровень стратегической информации.** На этом уровне ИТ обеспечивают интеграцию данных на уровне бизнес-процессов по направлениям хозяйственной деятельности компании. С этими данными работают аналитики и руководители высшего звена компании, которые готовят стратегические решения развития и деятельности компании на рынке.

- **Уровень принятия решений.** На этом уровне ИТ обеспечивают интеграцию и агрегацию данных на уровне бизнес-процессов компании для руководителей высшего звена компании. Этот уровень обеспечивает информационную поддержку принятия решений.

Информационная пирамида описывает среду бизнес-аналитики, которую можно описать следующим образом. В информационную среду бизнес-аналитики поступает первичный материал — данные, которые затем перерабатываются в автоматизированных системах и информационных продуктах.

В процессе переработки происходит переход от данных к информации. ХД извлекает данные из множества транзакционных или оперативных систем, а затем интегрирует и хранит данные в специализированной БД. Например, в ХД могут приводиться в соответствие и объединяться пользовательские записи из четырех оперативных систем (приложений для обработки заказов, обслуживания, продаж и поставок). Такой процесс извлечения и интеграции преобразует данные в новый информационный продукт — информацию.

Затем пользователи, работающие с аналитическими инструментами (например, для создания запросов, отчетов, OLAP-анализа и выполнения операций интеллектуального анализа данных), обращаются к данным из ХД и анализируют ее. Таким образом, определяются тенденции, структуры и

исключения. Аналитические инструменты помогают пользователям преобразовать информацию в знания.

Теперь дадим определение систем бизнес-аналитики или систем деловой осведомленности.

К основным функциям системы бизнес-аналитики, как правило, относят следующие.

- Управление в реальном времени бизнес-знаниями в рамках всего предприятия.
- Простой доступ к информации для сотрудников компании различных уровней.
- Рост объема перерабатываемой информации, повышение конкурентоспособности.
- Проведение более эффективного анализа доходов и расходов.
- Предоставление исполнительным директорам комплексной и более наглядной картины предприятия по всем направлениям бизнеса.
- Отслеживание ситуации как на всем предприятии в целом, так и на каких-то конкретных проблемных участках в частности.

К основным технологическим средствам реализации функциональности систем бизнес-аналитики относят:

- отчеты и средства их создания;
- специализированные средства создания отчетов;
- генераторы отчетов, встроенные в средства разработки;
- нетрадиционные средства создания отчетов;
- OLAP-средства;
- клиентские OLAP-средства;
- серверные OLAP-средства;
- средства поиска закономерностей (Data Mining-средства) и т.д.

Система бизнес-аналитики является стержнем, вокруг которого формируются потоки стратегической бизнес-информации. Данный инструмент помогает компании принимать решения, которые будут основаны на корректной информации, полученной вовремя.

В условиях, когда рынок постоянно меняется, а конкуренция становится все жестче, руководителям крайне необходимо выявлять и анализировать имеющиеся у предприятия резервы, которые могут существенно расширить возможности бизнеса.

Предлагаемые решения в области бизнес-аналитики должны предоставлять возможность оперативно анализировать тенденции рынка, осознавать движущие силы бизнеса и, основываясь на объективной информации, быстро реагировать на изменения рыночной ситуации и принимать верные решения.

Например, одним из возможных решений может быть графический инструмент для экономического анализа, относящийся к категории OLAP-приложений (On-line Analytical Processing), который:

- позволяет регулярно проводить анализ выполнения производственной программы;
- позволяет проводить анализ отклонений финансовых показателей, таких как, например, анализ расходов по приобретению сырья и материалов;
- позволяет собирать, обобщать, анализировать и представлять данные в виде легко читаемых графиков и аналитических приложений. В сочетании с системой планирования и управления ресурсами предприятия ERP обеспечивает гибкий, многоплановый анализ бизнеса на основе операционных данных;
- позволяет создавать так называемые "информационные кубы" — виртуальные информационные центры, которые содержат аналитические данные, существенные для выявления тенденций рынка, анализа и принятия стратегических решений.

Эти многомерные "информационные кубы" собирают и хранят всю информацию о деятельности предприятия. С их помощью можно моделировать и анализировать критические аспекты бизнеса, учитывая информацию о продукции, поставщиках, потребителях, товарообороте, ценах и доходах. Анализ ведется интерактивно, в реальном времени, с помощью удобных визуальных инструментов, а не просто на основе многочисленных отчетов с тысячами страниц, таблиц и чисел.

"Кубы данных" должны быть настроены для решения ряда критически важных аспектов бизнеса, включая анализ продаж, запасов, финансов, каналов снабжения и производства.

Специальные возможности должны обеспечивать мгновенную детализацию данных и всестороннее исследование проблемы. Результирующее двух- или трехмерное представление удобно для быстрого изучения тенденций и анализа отклонений.

Типовой состав программного обеспечения систем бизнес-аналитики включает в себя не только саму систему, но и обучающие материалы, техническую документацию, а также возможность получения технической поддержки и профессиональных консультаций. Все это помогает быстро и в совершенстве освоить систему, получить максимум преимуществ ее использования:

- система для оперативного анализа;
- стандартный каталог;
- стандартные шаблоны и модели;
- стандартные отчеты;
- обучение, поддержка, документация, профессиональные консультации.

Система бизнес-аналитики должна:

- иметь единый графический интерфейс пользователя (GUI), который обеспечивает интуитивную навигацию и комфортность в работе;
- предоставлять гибкий анализ и отчетность сотрудникам различных подразделений для просмотра сводных данных, используя одни и те же разрезы деятельности в сопоставимых показателях;
- иметь открытую архитектуру и организационную масштабируемость, которые дают контролируемое, последовательное и быстрое развертывание аналитической системы во всех подразделениях предприятия;
- иметь мощную систему административного контроля, которая освобождает ИТ-службу от необходимости составления многих форм отчетности, но в то же время обеспечивает контроль доступа к базам данных, конфиденциальность и мониторинг изменений;

- обеспечивать быстрое внедрение, которое способствует скорейшему получению практических результатов и ускоряет отдачу от инвестиций в программное обеспечение.

Таким образом, системы бизнес-аналитики позволяют:

- лучше анализировать структуру покупательского спроса и потребности ваших клиентов;
- использовать гибкую стратегию продаж и целевой маркетинг;
- по-новому представить на рынке возможности вашей продукции;
- выявить слабые места в цепочке поставок;
- исключить неоправданные производственные и коммерческие расходы;
- выявить стратегические тенденции в изменении ценовой структуры и номенклатуры выпускаемой продукции.

Информационная безопасность систем бизнес-аналитики

Следует отметить, что многие компании не придают значения вопросам безопасности, игнорируя тот факт, что архитектурные компоненты систем бизнес-аналитики таят в себе определенную опасность. Обеспечение безопасности среды бизнес-аналитики – не менее важная задача, чем защита оперативных приложений.

Необходимость безопасности систем оперативной обработки транзакций (On-Line Transaction Processing, OLTP) осознается большинством компаний. Особенность реализации этой задачи для OLTP-приложений заключается в том, что она хорошо поддается структуризации и является статичной (определенные приложения каждый раз одинаковым образом обращаются к определенным данным). Круг пользователей весьма ограничен — это работники с определенными бизнес-функциями, они работают с приложениями и данными, которые касаются только их поля деятельности. Кроме того, физическая структура этих приложений также остается довольно постоянной. Инструментальные средства и базовая структура данных меняются нечасто.

Среда бизнес-аналитики и ХД, наоборот, характеризуется значительной динамичностью вкупе с широкой и часто меняющейся пользовательской аудиторией, причем пользователи могут быть как внутренними, так внешними. В

такой ситуации гораздо сложнее (а иногда и практически невозможно) распределить пользователей по подмножествам данных; особенно это касается аналитических приложений высокого уровня, таких как, например, решения управления эффективностью корпорации (corporate performance management), где окончательная информация формируется на основе изучения данных всего предприятия. Помимо этого, физическая структура этой среды часто является неясной: в нее устанавливается множество различных средств, а сами данные пребывают в постоянном движении (из ХД в витрины данных и на пользовательские машины в информационные панели). В результате мероприятия по обеспечению безопасности корпоративной информации обходят стороной приложения бизнес-аналитики и ХД.

Для того чтобы гарантировать защищенность среды бизнес-аналитики, компании должны в первую очередь выполнить задачи безопасности, возникающие на уровне отдельных ее компонентов.



Рис. 3.2. Оболочка решений бизнес-аналитики

Каждый из основных компонентов среды бизнес-аналитики имеет свою степень риска и для обеспечения безопасности каждого компонента потребуется реализовать различные подходы (и различные технологии). Это крайне непростая задача, и, пожалуй, наибольшую сложность представляют "пробелы" между компонентами. Ведь программная оболочка для бизнес-аналитики практически никогда не поставляется одним поставщиком или в форме одной ИТ-технологии. При этом бесшовная интеграция между компонентами невозможна. Более того,

именно то, как компоненты работают друг с другом, и то, как информация проходит между ними, и образует "точки риска".

Безопасность данных

Сама суть бизнес-аналитики подталкивает бизнес-пользователей к расширению доступа к данным и контролю над ними. Поэтому необходима жесткая политика по защите информации, которая должна помочь "залатать дыры", созданные многочисленными, слабо интегрированными технологическими компонентами, а также минимизировать огромный риск, присущий человеческому фактору.

Данные в ХД, витринах данных и операционных складах данных создают условия для осуществления всей бизнес-аналитики и, как правило, включают гигантские объемы детальных, транзакционных данных. Поскольку они часто отображают длительный отрезок времени, относящейся к истории существования компании, как, например, финансовая информация, обеспечение защищенности таких данных чрезвычайно важно. При рассмотрении задач безопасности данных следует задаться следующими вопросами:

- Кто располагает доступом к ХД, витрине данных, кубам и так далее?
- Каковы рамки их доступа: одна предметная область, множество предметных областей или все предметные области?
- Каким типом доступа они обладают, например, только чтение или возможность модификации?

Топология данных в среде бизнес-аналитики влияет на возможности доступа к данным и обеспечение безопасности. Во многих компаниях результаты запросов часто загружаются на индивидуальные машины с целью дальнейшей детализации и использования. Эти данные оказываются в витринах данных, настольных БД, информационных панелях или крупноформатных электронных таблицах и быстро оказываются вне пределов инфраструктуры безопасности ИТ-отдела, хотя по-прежнему сохраняют свою конфиденциальную сущность. При рассмотрении топологии данных с точки зрения безопасности необходимо изучить следующие вопросы:

- Насколько распределенной является архитектура данных, поддерживающая бизнес-аналитику?

- Имеется ли дополнительное распределение данных и если да, то каковы связанные с ним риски?
- Что делают пользователи с загружаемыми данными?
- Передают ли пользователи данные внешним партнерам?

Процесс сбора данных

Обычно процесс сбора и подготовки данных для среды бизнес-аналитики очень сложный и "непрочный". Огромное число источников данных и значительное разнообразие данных приводят к многоступенчатым процессам, в которых данные интерактивно собираются и преобразуются для загрузки в ХД. Данные, подвергающиеся как процессу сбора, так и преобразования, также образуют следующие "точки риска".

- Кто располагает доступом к средствам извлечения данных из операционных систем?
- Где находятся данные, пребывающие в процессе сбора, перед тем как оказаться в ХД, и кто имеет доступ к этой области?
- Какова логика преобразования, безопасность которой реализуется в средствах извлечения, преобразования и загрузки (ETL)?
- Если никакие средства не используются, то какова защита ETL-процессов, написанных пользователем, от несанкционированных модификаций?

Пользовательские средства формирования запросов и аналитические приложения

Программные инструменты бизнес-аналитики и аналитические приложения — это, в первую очередь, механизмы, предназначенные для доступа к данным в ХД. Такие средства часто приобретались в большом количестве с целью широкого и глубокого развертывания бизнес-аналитики по всему предприятию. Эти инструменты представляют особую ценность только для определенных пользователей и несут серьезную опасность, если попадают не в те руки.

- Кто располагает разрешением на использование средств формирования запроса и отчетности?
- Назначен ли каждому пользователю личный ID?

Появление и развитие аналитических приложений для электронной коммерции по схеме "бизнес-бизнес" (business-to-business) и "поставщик-покупатели" (business-to-consumer) усилили насущность вопросов безопасности.

- Насколько свободно ваши клиенты и поставщики обмениваются предоставленной им информацией в рамках своих предприятий?
- Предоставляют ли они ее своим внешним акционерам?
- Не может ли эта информация попасть в руки ваших конкурентов?

Политика информационной безопасности

Корпоративная политика информационной безопасности часто не затрагивает информации, которая хранится, анализируется и поставляется посредством аналитических приложений. Поскольку бизнес-аналитика расширяет доступ к информации, часто передавая ее в непосредственное распоряжение бизнес-пользователей, информация быстро оказывается вне пределов инфраструктуры безопасности ИТ-отдела. Поэтому при формировании корпоративной политики информационной безопасности необходимо рассмотреть следующие вопросы:

- Учитывает ли корпоративная политика безопасности специфику ИТ?
- Имеются ли области значительного риска, которые могут быть устранены посредством такой политики?
- Затрагивают ли правительственные постановления информацию, которая хранится, анализируется и представляется среде бизнес-аналитики?
- Не нарушают ли текущие или планируемые мероприятия по развертыванию среды бизнес-аналитики эти постановления?

Технологии обработки данных

Технология диалоговой аналитической обработки (OLAP)

Концепция многомерного представления данных предполагает, что элементы данных (фактическая информация) являются точками многомерного пространства, размерности которого представляют собой содержательное описание таких фактов (точку зрения на них). В приложениях обработки многомерных данных сохраняются все проблемы с визуализацией многомерных массивов данных. Самые продвинутые, дорогостоящие и элитарные решения пребывают в пределах своих узких предметных ниш.

Конечные пользователи не намерены каким-то образом продвигать за свой счет электронные таблицы в интерфейс с многомерными БД. Конечно, электронные таблицы, в силу их удобства и простоты, являются излюбленным инструментом конечных пользователей. Однако, как показывает опыт, электронные таблицы хороши лишь тогда, когда они "заточены" под многомерность конкретных предметных областей.

Приложения OLAP зачастую бывают весьма громоздкими, обычно их рентабельность отвечает использованию в составе корпоративных рабочих групп, например в аналитических службах. Вообще, для эффективного использования решений OLAP нужна поддержка корпоративной инфраструктуры.

Как показывает анализ, Web-архитектуры быстро вытесняют традиционные клиент-серверные приложения для целого ряда категорий программного обеспечения, и рынок корпоративных OLAP-решений здесь не исключение.

Это направление стремительно развивается за счет появления различных Web-OLAP инструментов на базе HTML- и Java-технологий от известных поставщиков и быстро растущих новых компаний. В связи с расширением контингента пользователей продукты Web-OLAP разрабатываются для выполнения несколько иного анализа, чем традиционные клиент-серверные инструменты. Происходит переход от инструментальных средств исследования данных, ориентированных на специалистов-аналитиков, к готовым аналитическим приложениям, доступным для более широкого круга пользователей.

§ 3.2. Интеллектуальная обработка данных (Data Mining)

Возникновение Data Mining (добыча данных) связано с наличием противоречия между теоретическими методами прикладной статистики и практикой решения реальных задач. Синонимами этого понятия являются обнаружение знаний в БД и интеллектуальный анализ данных.

Стимулом развития технологии Data Mining явился прорыв в технологиях электронного хранения больших объемов данных — деятельность любого предприятия сопровождается регистрацией и записью на электронные носители всех подробностей его деятельности.

Очевидно, что без технологии переработки потока этих "сырых" данных (raw data), последние образуют просто большую свалку.

Требования к технологии переработки:

- данные имеют неограниченный объем;
- данные являются разнородными по типу (количественными, качественными, текстовыми);
- результаты должны быть понятны и конкретны;
- инструменты для обработки сырых данных должны быть просты в использовании.

Традиционная прикладная статистика не справляется с поставленными задачами. Главная причина — она работает с фиктивными, средними величинами (концепция усреднения по выборке). Ее методы полезны при проверке заранее сформулированных гипотез (verification-driven data mining) и для грубого предварительного анализа, составляющего основу OLAP (online analytical processing).

В основу технологии Data Mining (discovery-driven data mining) положена концепция шаблонов (паттернов), отражающих фрагменты многоаспектных взаимоотношений в данных. Эти шаблоны представляют закономерности, свойственные подвыборкам данных, которые могут быть выражены в понятной человеку форме. Поиск шаблонов проводится методами, не ограниченными рамками априорных предположений о структуре выборки и виде распределений значений анализируемых показателей.

Понятно, что такие шаблоны должны быть нетривиальны (unexpected — неожиданные регулярности в данных, которые составляют так называемые скрытые знания, Hidden knowledge).

Есть понимание того, что "сырые" данные содержат глубинный пласт знаний и его нужно раскопать.

Основные бизнес-приложения Data Mining

1. Розничная торговля (предполагается, что собирается информация о каждой покупке)

- Анализ покупательской корзины (анализ сходства) — выявление товаров, которые покупаются вместе. Необходимо для рекламы, выработки стратегии создания запасов товаров и способов их раскладки в торговых залах
- Исследование временных шаблонов. Создание запасов
- Создание прогнозирующих моделей. Характер потребностей различных категорий покупателей с определенным поведением (приобретение товаров известных марок или распродажа). Разработка направленных мероприятий по продвижению товаров

2. Банковское дело

- Выявление мошенничества с кредитными карточками. Анализ транзакций, которые оказались мошенническими
- Сегментация клиентов. Маркетинговая политика банка становится более целенаправленной
- Прогнозирование изменения клиентуры, модели ценности своих клиентов и соответствующее обслуживание каждой категории

3. Телекоммуникации

- Анализ записей о подробных характеристиках вызовов — выявление категорий клиентов с похожими стереотипами поведения и разработка привлекательных наборов цен и услуг
- Выявление лояльности клиентов. Определение характеристик клиентов, которые однажды воспользовавшись услугами компании, и с большей долей вероятности воспользуются еще раз

4. Страхование

- Выявление мошенничества. Определенные стереотипы в заявлениях о выплате страховки, характеризующие взаимоотношения между юристами, врачами и заявителями
- Анализ риска. Путем выявления факторов, связанных с оплаченными заявлениями, можно уменьшить свои потери

ЗАКЛЮЧЕНИЕ

Необходимость автоматизированного интеллектуального анализа данных стала очевидной, в первую очередь, из-за огромных массивов исторической и вновь собираемой информации. Трудно даже приблизительно оценить объем ежедневных данных, накапливаемых различными компаниями, государственными, научными и медицинскими организациями. По мнению исследовательского центра компании GTE, только научные институты собирают ежедневно около терабайта новых данных.

Другой причиной роста популярности интеллектуального анализа данных (data mining) является объективность получаемых результатов. Человеку-аналитику, в отличие от машины, всегда присущ субъективизм: он в той или иной степени является заложником уже сложившихся представлений. Иногда это полезно, но чаще приносит большой вред.

И, наконец, интеллектуальный анализ данных дешевле. Оказывается, что выгоднее инвестировать деньги в решения интеллектуального анализа данных, чем постоянно содержать целую армию высококвалифицированных и дорогих профессиональных статистиков. Интеллектуальный анализ данных вовсе не исключает полностью человеческую роль, но значительно упрощает процесс поиска знаний, делая его доступным для более широкого круга аналитиков, не являющихся специалистами в статистике, математике или программировании.

ЛИТЕРАТУРА

1. Devlin B.A., Murphy P.T. An Architecture For A Business And Information System IBM Systems Journal, 1988, Vol 17, No 1 – pp. 60-80
2. Inmon W. H. Building the Data Warehouse, Third Edition. John Wiley & Sons, Inc. New York, 2002 – 428 p
3. Kimbell R., Ross M. The Data Warehouse Toolkit: The Complete Guide to Dimensional Data Warehouses J. Willey & Sons. Second Edition, 2002 – 447 p
4. Codd E.F. A relational model of data for large shared data banks Comm. ACM, 1970, V. 13, N 6 – pp. 377-387
5. R&D Database Handbook. A Worldwide Guide to Key Scientific and Technical Databases. Ed. Davidson L. Technical Insights, Inc., 1984 –p. 194
6. Codd E.F. A data base sublanguage founded on the relational calculus Proc. ACM-SIGFIDET 1971, Workshop, San Diego, Calif., Nov. 1971 – pp. 35-68
7. Codd E.F. Recent investigations in relational data base systems. Proc. IFIP Congress 1974, North-Holland Pub. Co., Amsterdam – pp. 1017-1021
8. Туманов В. Data Warehouse – информационная среда для принятия решений. С чего начать. PCWEEK/RE, 1998, № 29 – С. 15-16
9. Gupta V. R. An Introduction to Data Warehousing. System Services corporation, Chicago, Illinois, 1997 – p. 19
10. Инмон Б. Типы хранилищ данных. Перевод Intersoftlab, 2001
11. Rainardi V. Building a Data Warehouse: With Examples in SQL Server 2008, APRESS – p. 541
12. Туманов В.Е. Системы складирования данных. Архитектура, продукты и подходы к реализации Машиностроитель. – 2003. –№ 8. – С. 58-65
13. Артемьев В., Кузнецов С. Обзор возможностей применения ведущих СУБД для построения хранилищ данных (Data Warehouse)
14. Ballard C., Bell R., Herreman D., Kim E., Schau D., Valencic A. Data Modeling Techniques for Data Warehousing International Technical Support Organization, IBM, 1998 – p. 216 Business Intelligence Tutorial. Хранилища данных
15. Data Warehouse Center Application Integration Guide
16. Oracle9i Data Warehousing Guide Release 1 (9.0.1) Part Number A90237-01

17. Oracle8i Data Warehousing Guide Release 2 (8.1.6) A76994-01
18. Сайт компании NCR
19. Sybase Warehouse WORKS Системы управления базами данных. 1998, № 4-5
20. Data Warehousing Alliance
21. Технологии хранилищ данных
22. Грачева Т.В. ADABAS — основа технологий SOFTWARE AG Системы управления базами данных. 1995, № 2
23. Архипенков С. Аналитические системы на базе ORACLE Express OLAP. Проектирование, создание, сопровождение М.: Диалог-МИФИ, 2000. – 320 с
24. Архипенков С., Голубев Д., Максименко О. Хранилища данных. От концепции до внедрения. М.: Диалог-МИФИ, 2002. – 528 с
25. Лоуенд Ш., Хилсон С., Хоббс Л. Oracle9iR2. Разработка и эксплуатация хранилищ баз данных. М: КУДИЦ-ОБРАЗ, 2004. – 592 с
26. Елманова Н., Федоров А. Введение в OLAP-технологии Microsoft. М.: Диалог-МИФИ, 2002. – 268 с
27. Microsoft Training and Certification. Designing and Implementing a Data Warehouse Using Microsoft® SQL Server 7.0. Delivery Guide Course Number: 1502B. Microsoft,. 1999 – p. 542
28. Маклаков С.В., Матвеев Д.В. Анализ данных. Генератор отчетов Crystal Reports СПб.: БХВ-Петербург, 2003. – 496 с
29. Маклаков С.В., Матвеев Д.В. Введение в Crystal Reports. М.: Богородский печатник, 2001. – 212с. – ISBN 5-89589-027-X
30. Туманов В.Е. Управление информацией на основе консолидации данных — важнейшее конкурентное преимущество современного предприятия Машиностроитель, 2006, № 1. – С. 32-40
31. Туманов В.Е. Корпоративная информационная фабрика и решения компании Oracle Машиностроитель, 2005, № 9. – С. 10-15
32. Маклаков С.В., Туманов В.Е. Проектирование реляционных хранилищ данных М.: Диалог-МИФИ, 2007. – с. 333

33. Хендерсон К. Профессиональное руководство по Transact SQL. 2005. СПб.: «Техническая книга». – с. 555
34. Лазунский М., Лапиров А. Создание «корпоративной информационной фабрики» в банке. Банковские технологии, 2005, № 7. – С. 43-47
35. Hancock J.C., Toren R. Toren R. Practical Business Intelligence with SQL Server 2005 Addison Wesley Professional, 2006, – p. 432
36. Langit L. Foundations of SQL Server 2005 Business Intelligence APRESS, 2007, – p. 415
37. Бергер А.Б., Горбач И.В., Меломед Э.Л., Степаненко В.П., Щербинин В.А. Microsoft SQL Server 2005. Analysis Services. OLAP и многомерный анализ данных СПб.: БХВ-Петербург, 2007. – с. 928
38. Туманов В.Е. Хранилища данных: Жизненный цикл разработки. Машиностроитель, 2005, № 8. – С. 22-30
39. Маклаков С.В. Моделирование бизнес-процессов с AllFusion Process Modeler (BPwin 4.1) М.: Диалог-МИФИ, 2003. – с. 240. – ISBN 5-86404-179-3
40. Маклаков С.В. Создание информационных систем с AllFusion Modeling Suite. М.: Диалог-МИФИ, 2003. – с. 432. – ISBN 5-86404-181-5
41. Маклаков С.В. Bpwin и Erwin. CASE-средства разработки информационных систем. М.: Диалог-МИФИ», 1999. – с. 256
42. Маклаков С.В. Моделирование бизнес-процессов с BPwin 4.0 М.: Диалог-МИФИ», 2002. – с. 224. – ISBN 5-86404-165-3
43. Стивенсон Й., Энсор Д. Oracle. Проектирование баз данных. К.: Издательская группа BHV, 1999. – с. 560
44. Мартин Дж. Планирование развития автоматизированных систем М.: Финансы и статистика, 1984. – с. 196

Xülasə

Bu işin yazılmasının məqsədi verilənlər ambarının tipik arxitekturasının və onun əsas komponentlərinin tədqiqindən ibarətdir. Xüsusi diqqət verilənlər şini ilə ambarlara və verilənlərin federativ ambarlarına verilmişdir, verilənlər ambarının qurulması üçün muxtar sistemlərdən daxil olan verilənlərin yığılması, nizamlanması və emalı metodologiyası işlənilmişdir.

Summary

The purpose for which this paper was written, is to study the typical architecture of the data warehouse and its main components. Particular attention is paid to the data warehouse bus architecture data and federated data stores. Developed methodology for collecting, collating and processing data from autonomous information systems for the construction of data warehouses