5_Оперативное хранилище данных, витрина данных: структура, связь с конкретными целями бизнеса.

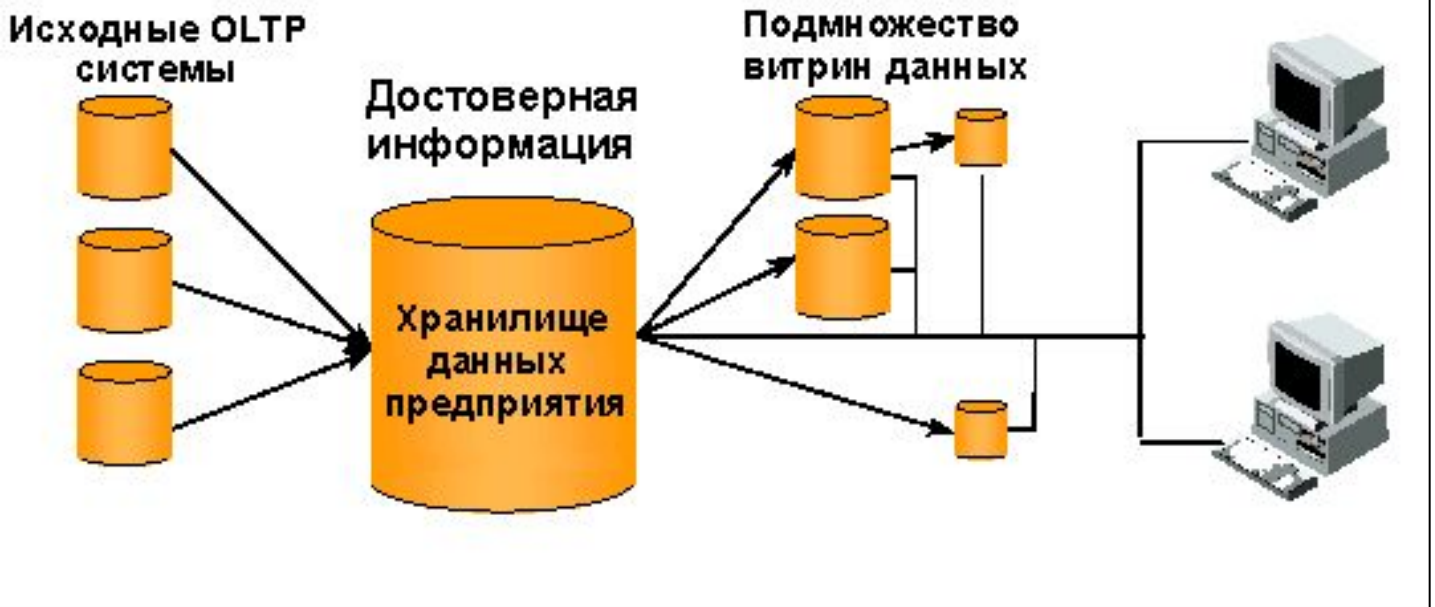
Хранилища данных - это новое технологическое решение, которое стало использоваться в начале 90-х годов 20-го века, после того как Билл Инмон (Bill Inmon), опубликовал книгу по этой теме.

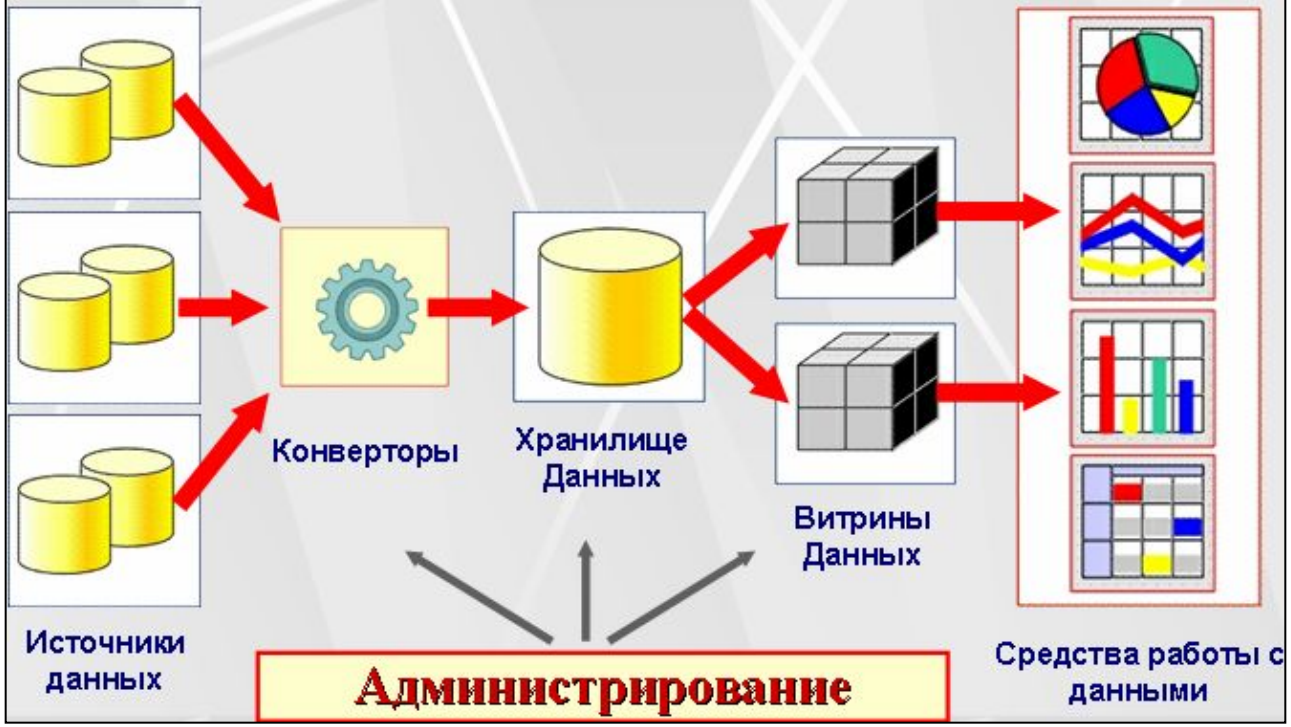
Любая транзакционная система, как правило, содержит два типа таблиц. Один из них отвечает за быстрые транзакции. Например, при продаже билетов необходимо обеспечить работу большого числа кассиров, которые обмениваются с системой короткими сообщениями (фамилия пассажира, даты вылета, рейса, места, пункта назначения). Другой тип таблиц содержит итоговые данные о продажах за указанный срок, по направлениям, по категориям пассажиров, что увеличивает время отклика системы на запрос. Стала необходима среда хранения аналитических данных. Поначалу предлагалось просто копировать исходную информацию из источников данных в единые БД. Но разные РБД содержали противоречивую и несогласованную информацию. Был нужен единый репозиторий – хранилище данных и системы извлечения, очистки и загрузки данных, а OLAP-системы работали поверх таких хранилищ данных.

Вскоре выяснилось, что форматы данных плохо сочетаются с требованиями быстрого информационного обслуживания. Территориальная распределенность и организационная структура предприятия также требуют специфического подхода к информационному обслуживанию каждого подразделения. Решением является витрины данных, которые содержат необходимое подмножество информации из хранилища. Наполнение витрин из хранилища может происходить в часы спада активности пользователей. Витрины данных могут обслуживать задачи отчетности, статистического

Эволюция понимания места OLAP

Классическое Хранилище данных





Хранилище - базис для глобальной интеграции всех бизнес-процессов, т.к. все отчеты строятся из одного легко доступного источника, поэтому данные во всех отчетах согласованы и непротиворечивы

Кроме БД хранилище включает в себя сложную инфраструктуру:

- средства изменения и расширения БД;
- технологию регулярного сбора данных;
- инструменты проверки и очистки данных;
- технологию ввода и изменения аналитических признаков;
- технологию расчета показателей, формульный язык;
- технологию агрегации и консолидации данных;
- инструменты выполнения запросов, создания отчетов и анализа в режиме реального времени;
- технологию построения сложных отчетов;
- средства разграничения прав доступа и др.

Билл Инмон: “Хранилище данных - это предметно-ориентированное, привязанное ко времени и неизменяемое собрание данных для поддержки процесса принятия управляющих решений”. Данные в хранилище попадают из оперативных систем (OLTP-систем), которые предназначены для автоматизации бизнес-процессов. Кроме того, хранилище может пополняться за счет внешних источников, например статистических отчетов.

В любом хранилище данных - и в обычном, и в многомерном - наряду с детальными данными, извлекаемыми из ИС, хранятся и суммарные показатели (агрегированные показатели, агрегаты), такие, как суммы объемов продаж по месяцам, по категориям товаров и т. п. Агрегаты хранятся в явном виде с единственной целью - ускорить выполнение запросов.

Разработчики инструментов генерирования Хранилищ - компании SAS, IBM, Informatica.

Разработчики инструментов управления Хранилищем - Oracle, IBM, Microsoft, Teradata.

Поставщики устройства для Хранилищ данных - Netezza, DATAllegro, IBM, Sun.

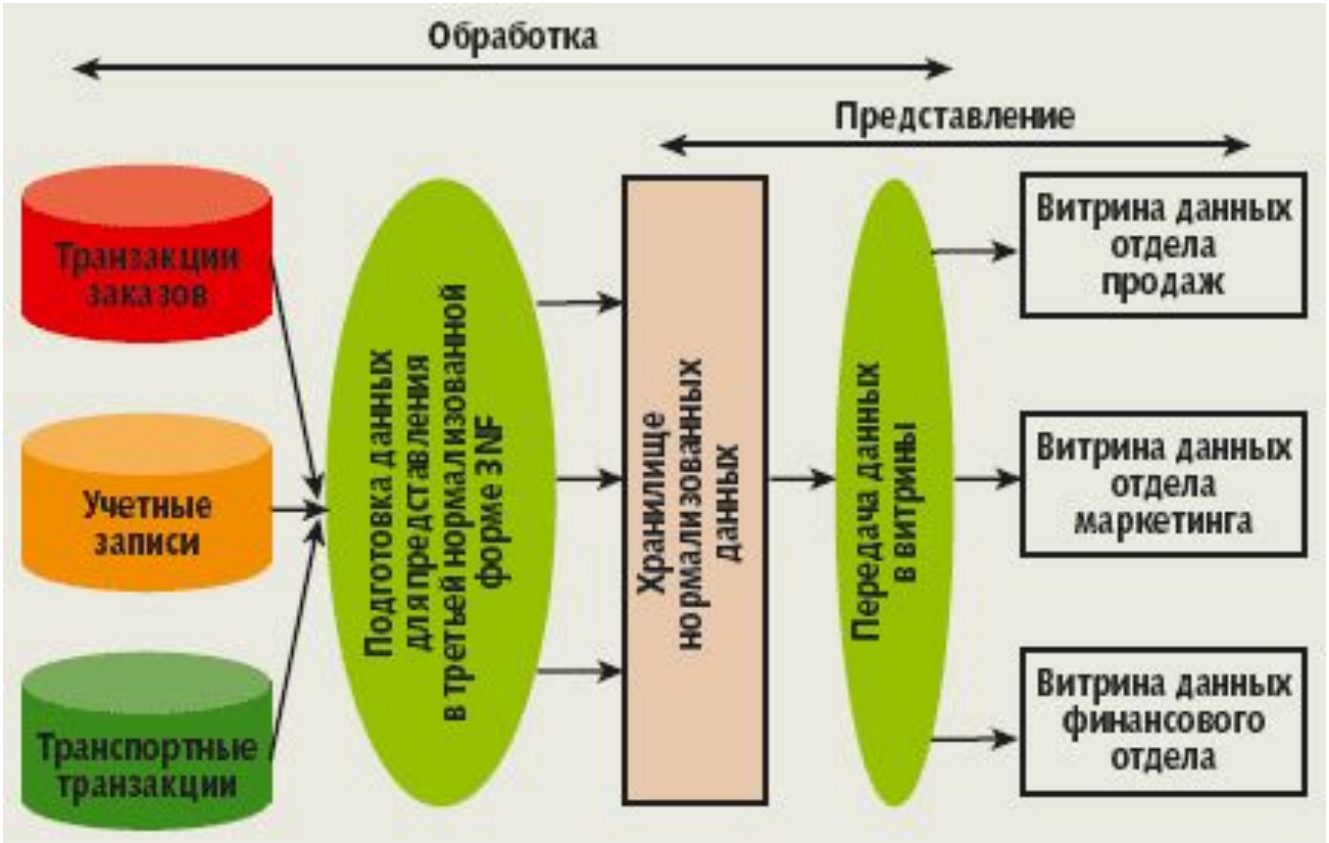
Два основных подхода к архитектуре Хранилищ данных:

- корпоративная информационная фабрика (Corporate Information Factory, CIF) Билла Инмона
- хранилище данных с архитектурой шины или звезды (Data Warehouse Bus, BUS) Ральфа Кимболла (Ralph Kimball).

Принципы создания и использования Хранилища данных

Термин "OLAP" неразрывно связан с термином "хранилище данных". Хотя OLAP и не представляет собой необходимый атрибут хранилища данных, он все чаще применяется для анализа накопленных в этом хранилище сведений. Задача хранилища - предоставить "сырье" для анализа в одном месте и в простой, понятной структуре, т.к. аналитиков в большинстве случаев интересуют не детальные, а обобщенные показатели.

CIF - нормализованные хранилище данных. Данные из источников конвертируются в РБД, которая преобразуется в 3-ю нормальную форму, содержащую атомарные данные. Это Хранилище используется для того, чтобы наполнить информацией дополнительные репозитории данных, подготовленные для анализа. Они включают специализированные Хранилища для изучения и "добычи" данных (Data Mining), а также отдельные витрины данных. Конечные витрины данных создаются для обслуживания бизнес-отделов или для реализации бизнес-функций и используют пространственную модель для структурирования суммарных данных. Атомарные данные остаются доступными через нормализованное хранилище данных.

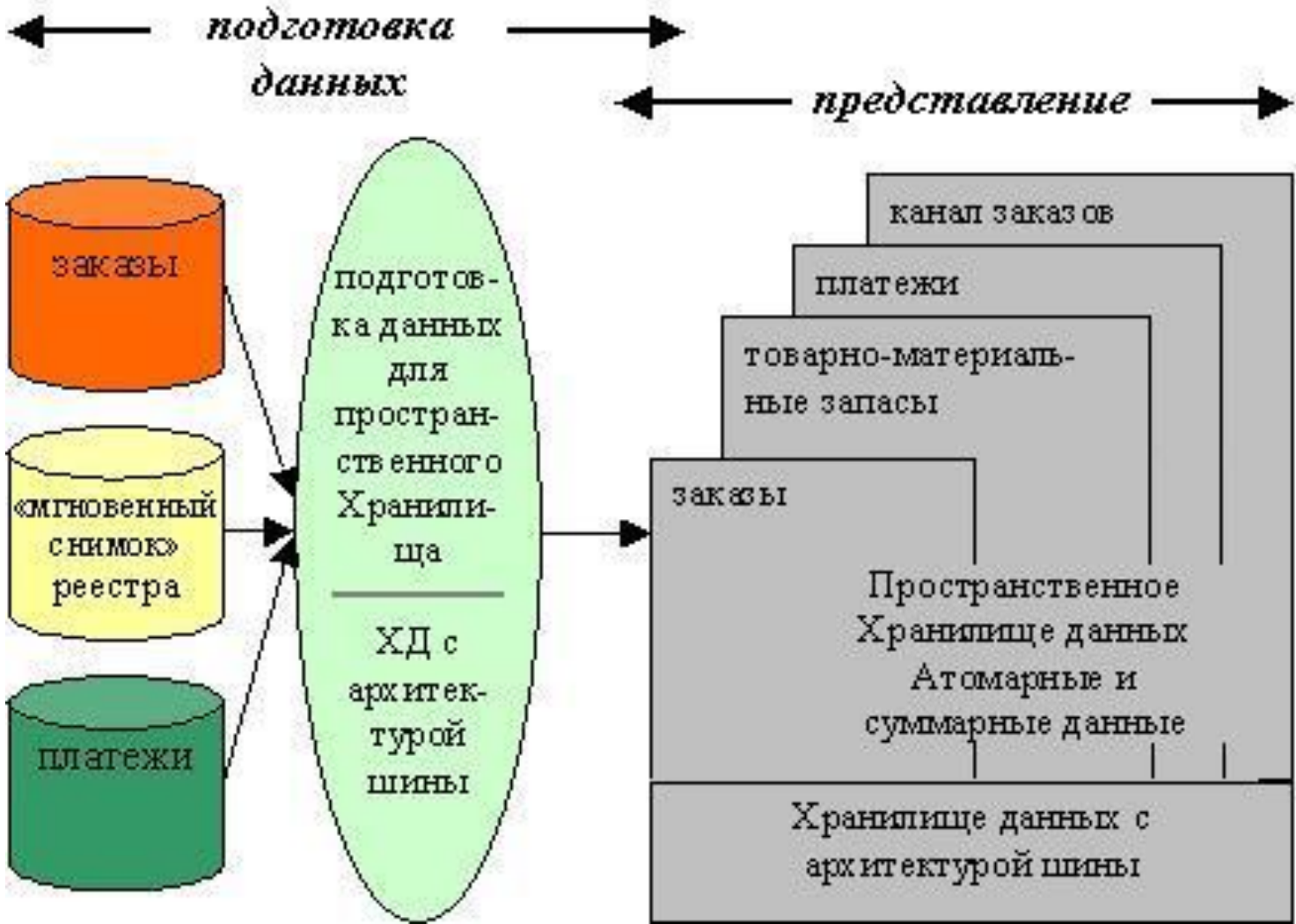


Таким образом, хранилище данных CIF - это физически целостный объект, проект корпоративного масштаба, охватывающий все отделы и обслуживающий нужды всех пользователей корпорации (пример – сервер Crystal Enterprise).

В качестве отличительных характеристик подхода Билла Инмона к архитектуре хранилищ данных можно назвать следующие:

- использование реляционной модели организации **атомарных** данных;
- использование пространственной модели для организации **суммарных** данных;
- использование итеративного или "спирального" подхода при создании больших хранилищ данных. Это позволяет вносить изменения в блоки данных или программных кодов и избавляет от необходимости перепрограммировать значительные объемы данных в хранилище. Потенциальные ошибки также будут локализованы в пределах небольшого массива;
- использование третьей нормальной формы для организации атомарных данных, что обеспечивает высокую степень детальности интегрированных данных и, соответственно, предоставляет корпорациям широкие возможности для манипулирования ими и изменения формата и способа представления данных по мере необходимости.

Data Warehouse Bus. Пространственная модель (dimensional model) - данные организованы не по 3-ей нормальной форме, а в виде тематических таблиц, каждая из которых содержит характеристику отдельных категорий информации (dimensions) - справочники. Центральная таблица связана со всеми описательными таблицами, но последние напрямую не связаны между собой (так называемая архитектура "звезда" или star scheme).



Основная цель пространственной модели - минимизировать время выполнения запроса, поэтому допускается денормализация данных.

С этой же целью данные группируются вокруг центральной задачи (или вопроса), которую придется выполнять наиболее часто. В этой модели еще на этапе подготовки данных они преобразуются в требуемую информацию.

Пространственная модель Хранилища данных содержит ту же атомарную информацию, что и нормализованная модель Билла Инмона, но информация структурирована по-другому, чтобы облегчить ее использование и выполнение запросов. Эта модель включает как атомарные данные, так и обобщающую информацию (агрегаты в связанных таблицах или многомерных кубах) в соответствии с требованиями производительности или пространственного распределения данных. Запросы в процессе выполнения обращаются ко все более низкому уровню детализации без дополнительного перепрограммирования со стороны пользователей или разработчиков приложения. Матрица корпоративного Хранилища данных с архитектурой шины выявляет и усиливает связи между показателями бизнес-

В отличие от подхода Билла Инмона, пространственные модели строятся для обслуживания бизнес-процессов (которые, в свою очередь, связаны с бизнес-показателями или бизнес-событиями), а не бизнес-отделов. Например, данные о заказах, которые должны быть доступны для корпоративного использования, вносятся в пространственное Хранилище данных только один раз, в отличие от CIF-подхода, в котором их пришлось бы трижды копировать в витрины данных отделов маркетинга, продаж и финансов. Итак, типичные черты подхода Ральфа Кимболла:

- использование пространственной модели организации данных с архитектурой "звезда" (star scheme).
- использование двухуровневой архитектуры, которая включает стадию подготовки данных, недоступную для конечных пользователей, и хранилище данных с архитектурой шины. В состав последнего входят несколько витрин атомарных данных, несколько витрин агрегированных данных и персональная витрина данных, но оно не содержит одного физически целостного или централизованного хранилища данных.

Хранилище данных с архитектурой шины обладает следующими характеристиками:

  - оно пространственное, включает как данные о транзакциях, так и суммарные данные и витрины данных, посвященные только одной предметной области или имеющие только одну таблицу фактов (fact table);
  - оно может содержать множество витрин данных в пределах одной базы данных.

- хранилище данных не является единым физическим репозиторием (в отличие от подхода Билла Инмона). Это "виртуальное" хранилище. Это коллекция витрин данных, каждая из которых имеет архитектуру типа "звезда".

Различия двух типов Хранилищ данных

- Различные подходы к построению БД, составляющих основу Хранилища. Если Ральф Кимболл использует пространственную организацию БД с так называемой архитектурой "звезда" как на стадии подготовки, так и презентации данных, то Билл Инмон комбинирует два подхода. В его модели атомарные данные организованы в РБД и находятся в нормализованном Хранилище данных, причем суммарные данные доступны для использования через специализированные Хранилища, средства data mining и OLAP; что же касается зависимых витрин данных, то только они организованы с помощью пространственных моделей, как и у Ральфа Кимболла. Таким образом, по сути дела архитектуры отличаются только способами обращения с атомарными данными: их пространственной организацией у Кимболла и нормализованной - у Инмона. Хранилище данных у Инмона – аналог РБД, в которой для выбора данных необходимо перебрать связанные таблицы, у Кимболла – куб, в котором все данные собраны в одной таблице.
- Вопрос физической организации Хранилища. Если у Инмона это физически целостный реально существующий объект, то Хранилище Кимболла - скорее "виртуальный" объект. Это коллекция витрин данных, которые могут быть пространственно разобщенными. Кимбалл считает возможным объединить с помощью шины отдельные витрины данных в информационную инфраструктуру, имеющую топологию *звезды*, а Инмон считает необходимым загружать все данные в *единое хранилище*.

Data Mining

Data Mining переводится как "добыча" или "раскопка данных" ("обнаружение знаний в базах данных" (knowledge discovery in databases) и "интеллектуальный анализ данных"). **Data Mining** - это процесс обнаружения в сырых данных ранее неизвестных, нетривиальных, практически полезных и доступных интерпретации знаний, необходимых для принятия решений в различных сферах человеческой деятельности .

Без продуктивной переработки потоки сырых данных образуют никому не нужную свалку. Требования к переработке данных:

- Данные имеют неограниченный объем
- Данные являются разнородными (количественными, качественными, текстовыми)
- Результаты должны быть конкретны и понятны
- Инструменты для обработки сырых данных должны быть просты в использовании

Математическая статистика – основной инструмент анализа данных использовала **концепцию усреднения по выборке**, приводящую к операциям над фиктивными величинами (**средняя температура пациентов по больнице**, **средняя высота дома на улице**, **состоящей из дворцов и лачуг и т.п**.). Методы математической статистики оказались полезными, главным образом, для проверки заранее сформулированных гипотез (verification-driven data mining) и для "грубого" разведочного анализа, составляющего основу оперативной аналитической обработки данных (online analytical processing, OLAP).

В основу современной технологии Data Mining (discovery-driven data mining) положена концепция шаблонов (паттернов), отражающих фрагменты многоаспектных взаимоотношений в данных. Эти шаблоны представляют собой закономерности, свойственные подвыборкам данных, которые могут быть компактно выражены в понятной человеку форме.

Выделяют пять стандартных типов закономерностей, которые позволяют выявлять методы Data Mining:

- ассоциация
- последовательность
- классификация
- кластеризация
- прогнозирование

Методы Data Mining используются на основе информационных хранилищ данных.

Особенность областей применения Data Mining заключается в их сложной системной организации. Они относятся к надкибернетическому уровню организации систем, закономерности которого не могут быть достаточно точно описаны на языке статистических или иных аналитических математических моделей. Данные в указанных областях неоднородны, нестационарны и часто отличаются высокой

Розничная торговля

Предприятия розничной торговли сегодня собирают подробную информацию о каждой отдельной покупке, используя кредитные карточки с маркой магазина и компьютеризованные системы контроля:

- анализ покупательской корзины предназначен для выявления товаров, которые покупатели стремятся приобретать вместе. Знание покупательской корзины необходимо для улучшения рекламы, выработки стратегии создания запасов товаров и способов их раскладки в торговых залах.
- исследование временных шаблонов для решения о создании товарных запасов. Оно дает ответы на вопросы типа "Если сегодня покупатель приобрел видеокамеру, то через какое время он вероятнее всего купит новые батарейки и пленку?"
- создание прогнозирующих моделей дает возможность торговым предприятиям узнавать характер потребностей различных категорий клиентов с определенным поведением, например, покупающих товары известных дизайнеров или посещающих распродажи. Эти знания нужны для разработки точно направленных, экономичных мероприятий по продвижению товаров.

Банковское дело

- выявление мошенничества с кредитными карточками. Путем анализа прошлых транзакций, которые впоследствии оказались мошенническими, банк выявляет некоторые стереотипы такого мошенничества.
- сегментация клиентов. Разбивая клиентов на различные категории, банки делают свою маркетинговую политику более целенаправленной и результативной, предлагая различные виды услуг разным группам клиентов.
- прогнозирование изменений клиентуры. Data Mining помогает банкам строить прогнозные модели ценности своих клиентов, и соответствующим образом обслуживать каждую категорию.

Телекоммуникации

В области телекоммуникаций методы Data Mining помогают компаниям более энергично продвигать свои программы маркетинга и ценообразования, чтобы удерживать существующих клиентов и привлекать новых. Среди типичных мероприятий отметим следующие:

- анализ записей о подробных характеристиках вызовов. Назначение такого анализа - выявление категорий клиентов с похожими стереотипами пользования их услугами и разработка привлекательных наборов цен и услуг;
- выявление лояльности клиентов. Data Mining можно использовать для определения характеристик клиентов, которые, один раз воспользовавшись услугами данной компании, с большой долей вероятности останутся ей верными. В итоге средства, выделяемые на маркетинг, можно тратить там, где отдача больше всего.

Страхование и медицина

Страховые компании в течение ряда лет накапливают большие объемы данных. Здесь обширное поле деятельности для методов Data Mining:

- выявление мошенничества. Страховые компании могут снизить уровень мошенничества, отыскивая определенные стереотипы в заявлениях о выплате страхового возмещения, характеризующих взаимоотношения между юристами, врачами и заявителями.
- анализ риска. Путем выявления сочетаний факторов, связанных с оплаченными заявлениями, страховщики могут уменьшить свои потери по обязательствам. Известен случай, когда в США крупная страховая компания обнаружила, что суммы, выплаченные по заявлениям людей, состоящих в браке, вдвое превышает суммы по заявлениям одиноких людей. Компания отреагировала на это новое знание пересмотром своей общей политики предоставления скидок семейным клиентам.

Известно много экспертных систем для постановки медицинских диагнозов. Они построены главным образом на основе правил, описывающих сочетания различных симптомов различных заболеваний. С помощью таких правил узнают не только, чем болен пациент, но и как нужно его лечить. Правила помогают выбирать средства медикаментозного воздействия, определять показания - противопоказания, ориентироваться в лечебных процедурах, создавать условия наиболее эффективного лечения, предсказывать исходы назначенного курса лечения и т. п. Технологии Data Mining позволяют обнаруживать в медицинских данных шаблоны, составляющие основу указанных правил

Другие приложения в бизнесе

- развитие автомобильной промышленности. При сборке автомобилей производители должны учитывать требования каждого отдельного клиента, поэтому им нужны возможность прогнозирования популярности определенных характеристик и знание того, какие характеристики обычно заказываются вместе;
- политика гарантий. Производителям нужно предсказывать число клиентов, которые подадут гарантийные заявки, и среднюю стоимость заявок;
- поощрение часто летающих клиентов. Авиакомпании могут обнаружить группу клиентов, которых данными поощрительными мерами можно побудить летать больше. Например, одна авиакомпания обнаружила категорию клиентов, которые совершали много полетов на короткие расстояния, не накапливая достаточно миль для вступления в их клубы, поэтому она таким образом изменила правила приема в клуб, чтобы поощрять число полетов так же, как и мили.

Молекулярная генетика и генная инженерия

Наиболее остро и вместе с тем четко задача обнаружения закономерностей в экспериментальных данных стоит в молекулярной генетике и генной инженерии. Здесь она формулируется как определение так называемых маркеров, под которыми понимают генетические коды, контролирующие те или иные фенотипические признаки живого организма. Такие коды могут содержать сотни, тысячи и более связанных элементов.

На развитие генетических исследований выделяются большие средства. В последнее время в данной области возник особый интерес к применению методов Data Mining. Известно несколько крупных фирм, специализирующихся на применении этих методов для расшифровки генома человека и растений.

Прикладная химия

Методы Data Mining находят широкое применение в прикладной химии (органической и неорганической). Здесь нередко возникает вопрос о выяснении особенностей химического строения тех или иных соединений, определяющих их свойства. Особенно актуальна такая задача при анализе сложных химических соединений, описание которых включает сотни и тысячи структурных элементов и их связей.

Классы систем Data Mining

- Предметно-ориентированные аналитические системы. Совокупность нескольких десятков методов прогноза динамики цен и выбора оптимальной структуры инвестиционного портфеля, основанных на различных эмпирических моделях динамики рынка. Используют несложный статистический аппарат, но максимально учитывают сложившуюся своей области специфику - профессиональный язык, системы различных индексов и пр. ($300–$1000).
- Статистические пакеты. Последние версии почти всех известных статистических пакетов включают наряду с традиционными статистическими методами также элементы Data Mining. Основное внимание в них уделяется все же классическим методикам — корреляционному, регрессионному, факторному анализу и другим. Недостатком систем этого класса считают требование к специальной подготовке пользователя. Большинство методов, входящих в состав пакетов опираются на статистическую парадигму, в которой главными фигурантами служат усредненные характеристики выборки. А эти характеристики, как указывалось выше, при исследовании реальных сложных жизненных феноменов часто являются фиктивными величинами. Мощные современные статистические пакеты являются слишком "тяжеловесными" для массового применения в финансах и бизнесе ($1000 - $15000) - SAS (компания SAS Institute), SPSS (SPSS), STATGRAPICS (Manugistics), STATISTICA, STADIA и другие.

Нейронные сети

Большой класс систем, архитектура которых имеет аналогию (как теперь известно, довольно слабую) с построением нервной ткани из нейронов.

В одной из архитектур имитируется работа нейронов в составе иерархической сети, где каждый нейрон более высокого уровня соединен своими входами с выходами нейронов нижележащего слоя. На нейроны самого нижнего слоя подаются значения входных параметров, на основе которых нужно принимать какие-то решения, прогнозировать развитие ситуации и т. д. Эти значения рассматриваются как сигналы, передающиеся в следующий слой, ослабляясь или усиливаясь в зависимости от числовых значений (весов), приписываемых межнейронным связям.

Для практического применения сеть надо "натренировать" на полученных ранее данных, для которых известны значения входных параметров и правильные ответы на них. Тренировка состоит в подборе весов межнейронных связей, обеспечивающих наибольшую близость ответов сети к известным правильным ответам.

Основным недостатком нейросетевой парадигмы является необходимость иметь большой объем обучающей выборки. Другой недостаток заключается в том, что даже натренированная нейронная сеть представляет собой черный ящик. Знания, зафиксированные как веса нескольких сотен межнейронных связей, совершенно не поддаются анализу и интерпретации человеком.

Примеры нейросетевых систем — BrainMaker (CSS), NeuroShell (Ward Systems Group), OWL (HyperLogic). Стоимость их довольно значительна: $1500–8000.

Системы рассуждений на основе аналогичных случаев

Идея систем case based reasoning — CBR — на первый взгляд крайне проста. Для того чтобы сделать прогноз на будущее или выбрать правильное решение, эти системы находят в прошлом близкие аналоги наличной ситуации и выбирают тот же ответ, который был для них правильным. Поэтому этот метод еще называют методом "ближайшего соседа" (nearest neighbour). В последнее время распространение получил также термин memory based reasoning, который акцентирует внимание, что решение принимается на основании всей информации, накопленной в памяти.

Системы CBR показывают неплохие результаты. Главным их минусом считают то, что они вообще не создают каких-либо моделей или правил, обобщающих предыдущий опыт — в выборе решения они основываются на всем массиве доступных исторических данных, поэтому невозможно сказать, на основе каких конкретно факторов CBR-системы строят свои ответы.

Другой минус заключается в произволе, который допускают системы CBR при выборе меры "близости". От этой меры самым решительным образом зависит объем множества прецедентов, которые нужно хранить в памяти для достижения удовлетворительной классификации или прогноза.

Примеры систем, использующих CBR — KATE tools (Acknosoft, Франция), Pattern Recognition Workbench (Unica, США).

Деревья решений (decision trees)

Деревья решения являются одним из наиболее популярных подходов к решению задач Data Mining. Они создают иерархическую структуру классифицирующих правил типа "ЕСЛИ... ТО..." (if-then), имеющую вид дерева. Для принятия решения, к какому классу отнести некоторый объект или ситуацию, требуется ответить на вопросы, стоящие в узлах этого дерева, начиная с его корня. Вопросы имеют вид "значение параметра А больше х?". Если ответ положительный, осуществляется переход к правому узлу следующего уровня, если отрицательный — то к левому узлу; затем снова следует вопрос, связанный с соответствующим узлом.

Популярность подхода связана как бы с наглядностью и понятностью. Но деревья решений принципиально не способны находить «лучшие» (наиболее полные и точные) правила в данных. Они реализуют наивный принцип последовательного просмотра признаков и «цепляют» фактически осколки настоящих закономерностей, создавая лишь иллюзию логического вывода.

Вместе с тем, большинство систем используют именно этот метод. Самыми известными являются See5/C5.0 (RuleQuest, Австралия), Clementine (Integral Solutions, Великобритания), SIPINA (University of Lyon, Франция), IDIS (Information Discovery, США), KnowledgeSeeker (ANGOSS, Канада). Стоимость этих систем варьируется от 1 до 10 тыс. долл.

Эволюционное программирование

Проиллюстрируем современное состояние данного подхода на примере системы PolyAnalyst — отечественной разработке, получившей сегодня общее признание на рынке Data Mining. В данной системе гипотезы о виде зависимости целевой переменной от других переменных формулируются в виде программ на внутреннем языке программирования. Процесс построения программ строится как эволюция в мире программ (этим подход немного похож на генетические алгоритмы). Когда система находит программу, более или менее удовлетворительно выражающую искомую зависимость, она начинает вносить в нее небольшие модификации и отбирает среди построенных дочерних программ те, которые повышают точность расчета целевой переменной. Таким образом система "выращивает" несколько генетических линий программ, которые конкурируют между собой в точности выражения искомой зависимости. Специальный модуль системы PolyAnalyst переводит найденные зависимости с внутреннего языка системы на понятный пользователю язык (математические формулы, таблицы и пр.).

Другое направление эволюционного программирования связано с поиском зависимости целевых переменных от остальных в форме функций какого-то определенного вида. Например, в одном из наиболее удачных алгоритмов этого типа — методе группового учета аргументов (МГУА) зависимость ищут в форме полиномов. В настоящее время из продающихся в России систем МГУА реализован в системе NeuroShell компании Ward Systems Group. Стоимость систем до $ 5000.

Генетические алгоритмы

Первый шаг при построении генетических алгоритмов — это кодировка в БД исходных логических закономерностей, которые именуют хромосомами, а весь набор таких закономерностей называют популяцией хромосом. Далее для реализации концепции отбора вводится способ сопоставления различных хромосом. Популяция обрабатывается с помощью процедур репродукции, изменчивости (мутаций), генетической композиции. Эти процедуры имитируют биологические процессы. В ходе работы процедур на каждой стадии эволюции получаются популяции со все более совершенными индивидуумами.

Генетические алгоритмы удобны тем, что их легко распараллеливать. Например, можно разбить поколение на несколько групп и работать с каждой группой независимо, обмениваясь время от времени несколькими хромосомами. Генетические алгоритмы имеют ряд недостатков. Критерий отбора хромосом и используемые процедуры являются эвристическими и далеко не гарантируют нахождения «лучшего» решения. Как и в реальной жизни, эволюцию может «заклинить» на какой-либо непродуктивной ветви. И, наоборот, можно привести примеры, как два неперспективных родителя, которые будут исключены из эволюции генетическим алгоритмом, оказываются способными произвести высокоэффективного потомка. Это особенно становится заметно при решении высокоразмерных задач со сложными внутренними связями.

Примером может служить система GeneHunter фирмы Ward Systems Group. Его стоимость — около $1000.

Алгоритмы ограниченного перебора

Алгоритмы ограниченного перебора были предложены в середине 60-х годов М. М. Бонгардом для поиска логических закономерностей в данных. Эти алгоритмы вычисляют частоты комбинаций простых логических событий в подгруппах данных. Примеры простых логических событий: X = a; X < b; X > a; a < X < b и др., где X — какой либо параметр, «a» и «b» — константы. Ограничением служит длина комбинации простых логических событий (у М. Бонгарда она была равна 3). На основании анализа вычисленных частот делается заключение о полезности той или иной комбинации для установления ассоциации в данных, для классификации, прогнозирования и пр.

Наиболее ярким современным представителем этого подхода является система WizWhy предприятия WizSoft. Пример использования системы WizWhy - обнаружение правила, объясняющего низкую урожайность некоторых сельскохозяйственных участков. Автор WizWhy утверждает, что его система обнаруживает ***BCE*** логические if-then правила в данных. На самом деле максимальная длина комбинации в if-then правиле в системе WizWhy равна 6, поэтому система выдает решение за приемлемое время только для сравнительно небольшой размерности данных.

Тем не менее, система WizWhy является на сегодняшний день одним из лидеров на рынке продуктов Data Mining. Это не лишено оснований. Система постоянно демонстрирует более высокие показатели при решении практических задач, чем все остальные алгоритмы. Стоимость системы около $ 4000, количество продаж — 30000.

Системы для визуализации многомерных данных

В той или иной мере средства для графического отображения данных поддерживаются всеми системами Data Mining. Вместе с тем, весьма внушительную долю рынка занимают системы, специализирующиеся исключительно на этой функции. Примером здесь может служить программа DataMiner 3D словацкой фирмы Dimension5.

В подобных системах основное внимание сконцентрировано на дружелюбности пользовательского интерфейса, позволяющего ассоциировать с анализируемыми показателями различные параметры диаграммы рассеивания объектов (записей) базы данных. К таким параметрам относятся цвет, форма, ориентация относительно собственной оси, размеры и другие свойства графических элементов изображения. Кроме того, системы визуализации данных снабжены удобными средствами для масштабирования и вращения изображений. Стоимость систем визуализации может достигать нескольких сотен долларов.