**Алексеева Т.В.,**

**Америди Ю.В.,**

**Лужецкий М.Г.**

**<http://www.e-biblio.ru/book/bib/01_informatika/IAS/Book.html>**

**Информационно-аналитические системы**

**Москва, 2005**

**Содержание**

[Тема 1. Информация в контексте современной экономики](http://www.e-biblio.ru/book/bib/01_informatika/IAS/Book.html#_Toc99807759)

[Тема 2. Анализ информации на предприятии.](http://www.e-biblio.ru/book/bib/01_informatika/IAS/Book.html#_Toc99807760)

[Тема 3. Информационно – аналитическая система как инструмент проведения экономического анализа.](http://www.e-biblio.ru/book/bib/01_informatika/IAS/Book.html#_Toc99807761)

[Тема 4. Хранилища данных](http://www.e-biblio.ru/book/bib/01_informatika/IAS/Book.html#_Toc99807762)

[Тема 5. Оперативный анализ данных](http://www.e-biblio.ru/book/bib/01_informatika/IAS/Book.html#_Toc99807763)

[Тема 6. Интеллектуальный анализ данных](http://www.e-biblio.ru/book/bib/01_informatika/IAS/Book.html#_Toc99807764)

[Тема 7. Инструментальные средства автоматизации аналитической работы и планирования.](http://www.e-biblio.ru/book/bib/01_informatika/IAS/Book.html#_Toc99807765)

[Тема 8. Программные инструментальные средства информационно – аналитических систем](http://www.e-biblio.ru/book/bib/01_informatika/IAS/Book.html#_Toc99807766)

**Тема 1.**  **Информация в контексте современной экономики**

**Содержание**

1.     Информационное пространство как среда анализа.

2.     Понятие информационного пространства.

3.     Структура информационного пространства

4.     Элементы структуры информационного пространства. Понятие показателя.

5.     Пространственная интерпретация понятия показатель.

6.     Единое информационное пространство предприятия.

**1. Информационное пространство как среда анализа**

Деятельность предприятия, любого объекта или даже самой маленькой системы неразрывно связана с информацией, которая окружает их повсюду. Она может быть представлена в различных формах – от информации, хранящейся в виде последовательных соединений различных атомов в простейшей молекуле ДНК, до более привычных нам в повседневной жизни бумажных документов, радио и телевидения и, конечно же, сети Интернет, с её необозримыми информационными просторами.

На каждом из уровней функционирования какого-либо объекта можно выделить источники информации различной природы, непосредственно влияющие на выполнение того или иного действия. Простым примером может быть процесс принятия решения о выдаче кредита в банке. Здесь задействована информация, представленная:

  в явном виде - в виде документов, которые клиент банка передаёт вместе с заявкой на выдачу кредита;

  в виде описания аналогичных ситуаций - накопленная информация в записях в базе данных кредитного бюро, описывающая кредитную историю данного конкретного клиента. Из неё сотрудник кредитного отдела формирует представление о качестве клиента и качестве выполнения бравшихся им до этого обязательств;

  информация, для получения которой необходимо выполнить набор операций по приданию фактической информации качественной характеристики - информация, скрытая в отчётных документах клиента, характеризующая его финансовое положение;

  внешняя по отношению к клиенту и сотруднику кредитного отдела информация, в виде законов, инструкций, ограничений, связанных с выдачей кредитов вообще;

  информация, хранящаяся в голове сотрудника кредитного отдела и представляющая его опыт и интуицию.

Так, даже в такой небольшой по меркам банка операции по принятию первичного решения о выдаче или не выдаче кредита сотруднику банка приходится обрабатывать гигантские объёмы информации, к которым до появления электронно-вычислительных систем просто невозможно было получить оперативного доступа и, как следствие, аналогичный процесс мог достаточно сильно растягиваться во времени.

         Но что же такое информация, о которой все постоянно говорят, употребляя такие привычные слова как средства массовой информации, информационные технологии, информационные системы?

Информация  (по законодательству РФ) – это сведения о лицах, предметах, фактах, событиях, явлениях и процессах независимо от формы их представления. Получается, что под информацией можно понимать представленные в любой форме прошлые или будущие значения конкретных характеристик того или иного объекта. Тогда что такое, например, интуиция – что она характеризует. Или другой пример – обычный бухгалтерский баланс – в нём явно представлена информация о фактической структуре активов и пассивов, но в явном виде нет информации об их качестве, но которую можно получить, выполнив ряд операций.

В качестве альтернатив существуют и другие трактовки термина информация.

Шеннон отражает **количественно-информационный** подход, определяет информацию **как меру неопределённости**(энтропию) события. Количество информации в том или ином сообщении зависит от вероятности его получения: чем более вероятным является сообщение, тем меньшее количество информации содержится в нём. Этот подход учитывает такую сторону информации, как новизна, неожиданность сообщения.

Глушков рассматривает информацию **как свойство материи**. Его теория основана на утверждении, что информацию содержат любые сообщения, воспринимаемые человеком или приборами. Иными словами, информация как свойство материи создаёт представление о её природе и структуре, упорядоченности, разнообразии и т.д.

Афанасьев основывает свою концепцию на логико-семантическом подходе, при котором информация трактуется как **часть знания**, **которая используется для активного действия** (полезная, "работающая" часть знания).

         А в советском энциклопедическом словаре даётся следующее определение информации: «Информация — это общенаучное понятие, включающее обмен сведениями между людьми, человеком и автоматом, автоматом и автоматом, обмен сигналами в животном и растительном мире; передачу признаков от клетки к клетке, от организма к организму».

         Если рассматривать современное предприятие как целостную сложную и открытую систему, то между его элементами идёт постоянный обмен информацией, которая на самом верхнем уровне представлена в виде человеческой речи, бумажных документов, данных в информационных системах.

Эта информация чаще всего используется для принятия решений на различных уровнях управления в организационной структуре от самого низшего (например, обычного корпоративного секретаря) – до высшего руководителя (председателя правления банка). При этом, ввиду достаточно больших потоков информации, которые необходимо обработать для принятия решения, требуется сначала преобразовать этот поток в его качественную или количественную характеристику, которую может адекватно воспринять человек, а уже затем принимать то или иное решение.

Такие потоки информации, этапы обработки этой информации и те, кто участвует в обработке информации входят в общее понятие информационного пространства предприятия, а деятельность по преобразованию потока информации в его сжатую характеристику представляет собой аналитическую деятельность и является неотъемлемым элементом совокупного информационного пространства.

Сначала более детально остановимся на самом понятии информационного пространства.

**2. Понятие информационного пространства.**

Предприятие, корпорация, любой хозяйственный комплекс является системой. Систему можно представить с одной стороны объектом как единым целым, с другой стороны как совокупность (множество) связанных между собой и взаимодействующих составных частей - объектов, но меньшего масштаба. Информационное отображение  физических объектов  или  процессов  называют информационным объектом (или информационным ресурсом).

         Помимо информационных объектов также стоит остановиться и на таком понятии, как информационная культура, которое непосредственно связано с работой над информационными  объектами. **Информационная культура** - это область **корпоративной культуры,**определяющая качество обмена информацией и знаниями в организации. Формирование информационной культуры как культуры работы с информацией, охватывающей все процессы: сбор, обработку, хранение, использование, является основой для построения системы управления знаниями в компании.

Информационная культура является **cистемообразующим фактором,** акселератором процессов управления знаниями, поскольку:

  позволяет осуществлять обмен знаниями;

  **обмен возможен только при наличии единого информационного языка;**

  благодаря единому языку формируется коммуникативная среда, ориентированная на совместную деятельность, совместный результат, сотрудничество;

  осуществляется личностный рост, в частности, в аспекте **информационных компетенций** сотрудников.

  Для того, чтобы создавать новые знания, работники должны быть способны использовать информацию и знания, полученные из разных источников.

**Единый информационный язык** является необходимым условием успешных корпоративных коммуникаций (информационного обмена в компании), поскольку эффективные коммуникации и обмен знаниями невозможны без общего языка.

Формирование информационной культуры - это процесс, в контексте которого рождаются также **информационные компетенции** специалистов, то есть, знания, умения и навыки, позволяющие ориентироваться в информационной среде, участвовать в её формировании и способствовать информационным коммуникациям.

Информационная культура, понимаемая нами как совокупность вышеперечисленных компонентов, формирует **единую информационную систему компании,** которая является:

  **открытой,** т.е. информация рассматривается как свободный, открытый для всех ресурс, не ограниченный никакими барьерами внутри корпорации;

  **самодостаточной,** т.е. система способна как формировать информационный ресурс, так и сама быть ресурсом;

  **саморегулирующейся,** т.е. осуществляется взаимное влияние всех ее компонентов;

  **саморазвивающейся,** т.е. внешние информационные ресурсы система способна сделать своим достоянием, так же и внутренние информационные ресурсы системы доступны внешним пользователям.

Информационная система, обладающая всеми этими свойствами, а также единым информационным языком, формирует единое информационное пространство, реализуя высокий уровень информационной культуры.

Иными словами можно сказать, что информационное пространство - совокупность  банков и баз данных, технологий их сопровождения и использования, информационных телекоммуникационных систем, функционирующих на основе общих принципов и обеспечивающих информационное взаимодействие организаций и граждан и удовлетворение их информационных потребностей.

Оно состоит из различных квантов или массивов информации в виде разного рода письменных (знаковых) и фиксированных на носителях информации кодограмм, буквенно-цифровых на естественном языке, устных и визуальных   сообщений.   Все  виды   сообщений   передаются непосредственно потребителям информации или по каналам связи, могут быть сохранены в различном виде с помощью современных технических средств и по мере необходимости могут неоднократно воспроизводиться.

Подавляющая часть сообщений, как правило, "загружается" в информационную систему и становится основой её информационного обеспечения.

Сообщения могут содержать оперативную информацию о технологических процессах, хозяйственных операциях; экономическую, техническую, организационно-распорядительную, отчётную документацию и т.д. Рассмотрим более детально структуру типового информационного пространства.

**3. Структура информационного пространства**

Основными компонентами информационного пространства являются:

1) информационные ресурсы;

2) средства и технологии информационного взаимодействия;

3) информационная инфраструктура.

Под информационным ресурсом в широком смысле понимается - совокупность данных, организованных для эффективного получения достоверной информации. Закон же устанавливает следующее определение «информационные ресурсы - это отдельные документы и отдельные массивы документов, документы и массивы документов в информационных системах: библиотеках, архивах, фондах, банках данных, других видах информационных систем».

Все информационные ресурсы можно классифицировать по различным признакам. Так различают:

по отношению к предприятию

  Внутренние;

  Внешние;

по способу хранения (передачи)

  Бездокументарные;

  Документарные;

  Электронные;

по типу источника (пользователя)

  Персональные;

  Коллективные (группа, структурное подразделение);

  Обще-корпоративные;

по доступности

  Общедоступные;

  Частично закрытые (только для группы лиц);

  Закрытые (для нескольких лиц).

Если же рассматривать информационные ресурсы, в рамках информационных систем, то можно выделить 2 типа ИР по содержанию:

  Фактуальные:

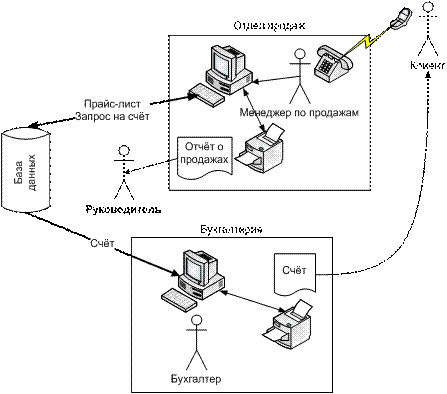
o       «Учётная информация» в информационных системах (в базах или хранилищах данных, файлах);

o       Электронные документы («текстовые» документы, хранящиеся в электронном виде);

  Субъективные:

o       Знания - вид информации, отражающей опыт специалиста (эксперта) в определенной предметной области, его понимание множества текущих ситуаций и способы перехода от одного описания объекта к другому.

Все информационные ресурсы имеют различную степень своей структурированности, т.е. возможности разделения на информационные элементы и фиксации их позиции в конкретном информационном ресурсе. Для примера рассмотрим часть информационного пространства, компании занимающейся продажами какого-либо товара (рис. 1):



**Рисунок 1.**   Пример части информационного пространства предприятия

         Здесь информационные ресурсы присутствуют в различных формах:

  Звонок клиента менеджеру по продажам по телефону;

  База данных, в которой хранится прайс-лист;

  Счёт, который бухгалтер выставляет в своей бухгалтерской ИС, печатает и отправляет клиенту;

  Отчёт о продажах.

Все эти информационные ресурсы имеют различную структуру и степень, с которой можно охарактеризовать наличие этой структуры. Соответственно, в зависимости от того, с какой степенью структурированности обрабатываются в рамках информационного пространства информационные ресурсы и различают степень структурированности самого пространства.

Свойство структурированности обеспечивает возможность представления информации в виде документов и манипулирования данными с помощью программно-технических средств информационных систем.

Различает пять степеней структурированности ИП:

  неструктурированное ИП (НИЦ); (пример - человеческая речь)

  слабо структурированное (ССИП); (пример - естественный письменный язык)

  структурированное (СИП); (пример - информация отличается преобладанием структурированных компонентов, подготовленная к загрузке в информационную систему)

  формализованно - структурированное (ФСИП); (пример - имеются алгоритмы получения любых значений элементов данных; обеспечиваются операции по управлению данными)

  машинно - структурированное (МСИП), (пример - представлены в формализованном виде все информационные объекты и их взаимосвязи, процессы преобразования информации описаны на языках программирования, обеспечивается взаимодействие пользователя и ЭИС на естественном языке.)

**4. Элементы структуры информационного пространства. Понятие показателя.**

В теории экономических ИС рассматривается понятие единицы информации, под которой понимают набор символов, которому придаётся определенный смысл. Выделяют несколько уровней единиц информации в зависимости от смыслового (семантического) значения, и его наполненности.

По  возрастанию  содержательности  понятия  определены следующие единицы информации:

  Реквизит;

  Составная единица информации (СЕИ);

  Показатель;

  База данных;

  Хранилище данных.

**Реквизит** - это информационное отображение свойства объекта, какого-либо процесса или явления. Сообщения состоят из определений свойств объектов, предметов, явлений, складывающихся некоторым образом из соответствующих реквизитов. Следует заметить, что синонимом понятия реквизит является атрибут, термин, широко используемый в литературе по базам данных.

**Составная единица информации** собирается из набора соответствующих определению данного объекта реквизитов и представляет собой информационное отображение объекта или его части.

Разновидностью составной единицы информации является **показатель**. Это сложное понятие. Имеются его различные определения. Одни авторы подчёркивают сущностный смысл или характер, привязанный к предметной области, в частности экономической. Другие исходят из формально-структурного подхода, ориентированного на структуризацию содержащейся в показателе информации в целях приспособления его структуры для эффективного использования в информационной   системе.   Результаты   такой   структуризации используются в информационно – аналитических  системах.

**Показатель** представляет высказывание с законченным смыслом, включающее как название переменной величины, так и её конкретное количественное значение со всеми качественными признаками,  необходимыми  для   идентификации   последнего.

Показатель образуется из набора реквизитов или терминов.

Реквизиты составляют две группы:

  **реквизиты-признаки**,   выражающие   качественные   отличия показателя, его смысловое содержание, в частности экономическое;

  **реквизиты - основания**, содержащие количественные значения показателя.

Показатель теряет смысл без какого-либо из названных реквизитов. В совокупности они образуют высказывание (сообщение), имеющее законченный предметный смысл, что позволяет утверждать, что показатель является наименьшей составной единицей информации, которая достаточна для документообразования, передачи, хранения и восприятия сообщений,

При структуризации   информационного  пространства разрабатывается система показателей, анализируется их собственная структура, В ходе этой работы необходимо исследовать общие закономерности, выявить категории показателей - члены общей структурной формулы описания показателей.

В общем виде структура показателя выглядит следующим образом:

**Р(И, Х),**где

Р - показатель (может быть экономическим);

И - набор реквизитов (терминов), идентифицирующих смысловое значение показателя;

Х - количественное значение показателя.

Например, для фирмы, торгующей автомобилями, этот показатель может иметь следующий вид:

**Объем продаж** ** (автомобиль, значение объема продаж)**

Идентификатор в свою очередь можно представить в виде двух частей:

**И(S, Q),**    где:

S - составленное из реквизитов наименование показателя, выявляющее его предметный смысл;

Q — дополнительные признаки показателя, составленные также из реквизитов и уточняющие его количественное значение.

Например:

**И (модель автомобиля, время)**

Выделенные реквизиты могут быть в свою очередь составными. Для уточнения связей между ними строятся схемы, детализирующие объект до такой степени, что дальнейшая детализация невозможна.

Реквизиты самого нижнего уровня называются единичными. Другие, расположенные на более верхних уровнях - множественными.

Дальнейший анализ проведем, начав с дополнительных признаков. Они могут состоять из единиц измерения Е, уровня показателя У, времени В, субъекта С. Представим эту структуру в виде соотношения:

**Q  (Е, С, В, У)**

Таким образом,   **И  (S (Е, С, В, У))**

Например:

**И** ** (модель автомобиля (менеджер, регион, год, покупатель))**

Наименование показателя может быть слитным (определенным одним реквизитом) или иметь свою структуру и, в свою очередь, состоять из реквизитов, таких как формальная (вычисляемая) характеристика показателя - Ф; обозначение отображаемого технологического или бизнес - процесса - П; объект измерения, подсчета – О. Тогда

**S  (Ф, П, О)**

Например:

**S  (модель автомобиля, фирма изготовитель, год выпуска)**

Таким образом, общая структурная формула показателя примет вид:

**И(S, Q)**

**S  (Ф, П, 0)**

**Q  (Е, С, В,У)**

**И  ((Ф, П, 0) (Е, С, В, У))**

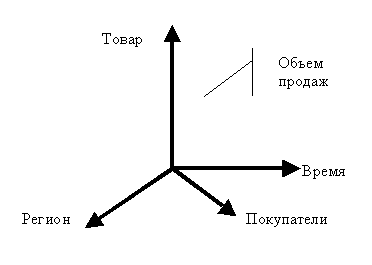
**Р  ((Ф, П, 0) (Е, С, В, У)Х)**

Данная структура может отображать документ вида:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Объем продаж | | | | | | |
| Автомобиль | | | Время | | | |
| Модель автомобиля | Фирма изготовитель | Год выпуска | Менеджер | Регион | Год | Покупатель |
|  |  |  |  |  |  |  |

**5. Пространственная интерпретация понятия показатель.**

Для образного восприятия сложных понятий удобна их пространственная или геометрическая интерпретация. Представим пространство признаков показателя на системе координат (рис. 2). Число измерений признаков этого пространства равно числу осей его системы координат, на которых откладывают отрезки, соответствующие реквизитам-признакам.

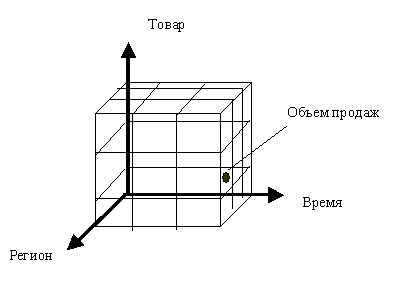


**Рисунок 2.** Система координат многомерного информационного пространства показателя.

Множественность значений реквизита выражается в длине отрезка. Тогда составляющие его единичные реквизиты укладываются на этом отрезке. Для трёхмерного пространства это будет куб. В свою очередь можно построить пространство системы взаимоувязанных показателей в виде набора соприкасающихся трёхмерных кубов.

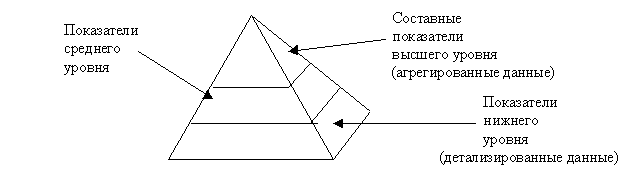
Как правило, на осях откладываются реквизиты-признаки в виде отрезков равной длины, а их количественное наполнение содержится в реквизитах-основаниях, которые не отображаются геометрически.

Пример такого пространства показателей приведён на рисунке 3.



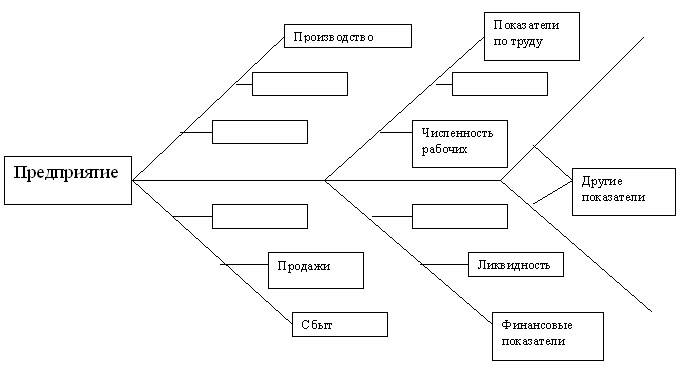
**Рисунок 3.** Трёхмерный гиперкуб информационного пространства показателя "Объем продаж"

Широко распространена также геометрическая интерпретация системы показателей в виде пирамиды, в основании которой лежат единичные признаки. Пирамида делится по высоте на "этажи", которые соответствуют иерархии признаков или системы показателей (рис. 4).



**Рисунок 4.** Пирамида показателей

Соответствующие структуры представляют также в виде перевёрнутого дерева или рыбьего скелета" - диаграммы Ишикава (рис. 5).



**Рисунок 5.** Диаграмма Ишикава

**6. Единое информационное пространство предприятия.**

Комплексная автоматизация функции управления требует создания единого информационного пространства на любом современном предприятии, в котором обычные сотрудники и руководство смогут осуществлять свою деятельность, руководствуясь едиными правилами доступа, представления и обработки информации.

Начальным этапом создания такой системы является построение модели предметной области или другими словами модели информационного пространства для конкретного бизнеса и позиционирование в ней своего предприятия (рис. 6).



**Рисунок 6.**  Информационное пространство предприятия

Исходя из составных элементов ИП, выделенных ранее можно выделить 3 измерения любого полноценного информационного пространства.

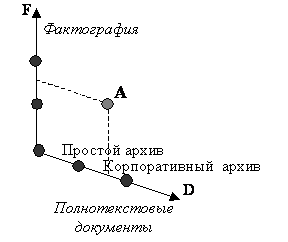
Первая ось (F) характеризует уровень организации хранения фактографической информации, которая привязана к специфике конкретного рода деятельности компании или организации, а также баз знаний, которые требуются информационным системам и сотрудникам для работы с информацией.

Вторая ось (D) - полнотекстовые документы, отражает необходимость организации взаимодействия как внутри корпорации, так и вне ее. В этих документах наряду с фактографической информацией содержится слабо и неструктурированная информация, не подлежащая автоматизированной аналитической обработке. Все взаимоотношения между субъектами бизнеса сопровождаются документами, которые становятся осязаемым отражением результата взаимодействия.

Третья ось (R) вносит в информационное пространство третье измерение - регламент процессов прохождения информации, а именно: описание того какие процедуры, когда и как должны выполняться.

Рассмотренная модель информационного пространства не является застывшим образованием, данным нам в ощущениях - прежде чем сформировалось современное представление о контурах этой модели, она претерпела три основные фазы своей эволюции.

**Фаза первая -** фактографическая. Начало любой деятельности знаменуется обычно периодом накопления первичной информации, имеющей жесткую структуру и атрибутику. Условно эту фазу можно представить в виде одной единственной оси (рис. 7).



**Рисунок 7.** Эволюция модели информационного пространства

Точка на этой оси - это текущее состояние информационной системы организации. Движение по оси вверх характеризует накопление фактографической информации и, начиная с определенного момента  можно отметить второй этап первой фазы - возникновение понятия "операция". Информация теперь представляется как некоторый привязанный к бизнес - процессам предприятия агрегат из имеющихся характеристик (атрибутов). После возникновения привязки к конкретным бизнес - процессам дальнейшая эволюция в одномерном пространстве уже невозможна - необходим новый качественный скачок к новой фазе.

**Фаза вторая -** полнотекстовая. Расширение организации и увеличение круга решаемых задач требуют использования полнотекстовых документов, включающих уже не только тексты, но и любые другие способы представления: графики, таблицы, видео и т.п. виды конструкторско-технологической документации. Возникает новая ось - полнотекстовые или, лучше, мультимедийные документы, а точка в новом, уже двумерном, пространстве характеризует информационную систему предприятия, где кроме фактографической базы документов имеются уже хранилища и архивы информации.

Хранилища позволяют накапливать информацию в различных форматах, предполагают наличие ее структуризации и возможностей поиска. Если на предприятии уже используется автоматизация, то хранилище - это не что иное, как электронный архив.

На данной оси также имеются свои этапы - с определенного момента развития хранилища можно уже говорить не об индивидуальном, а о корпоративном архиве, обслуживающем деятельность рабочих групп. Точка на плоскости эволюции, достигнутой во второй фазе, характеризует информационную систему, позволяющую отображать фактографическую информацию в виде полнотекстовых документов, имеющих необходимое количество атрибутов. Однако, как только речь пошла о корпорации, двумерного пространства для соответствующей ей системы опять становится недостаточно - необходим новый скачок к очередной фазе.

**Фаза третья** - регламентирующая. Информационная система в масштабах корпорации невозможна без решения вопросов обмена информацией, согласования или соблюдения регламента работы с информацией.

Третья ось в информационном пространстве предприятия, как и две другие имеет свое деление на этапы. Первоначальный этап движения по оси характеризуется наличием упрощенного регламента работы с информацией. Количественное накопление атрибутов и расширение возможностей различной работы с информацией сопровождается постепенным переходом ко второму этапу, отличительная черта которого - появление системы аналитической обработки информации. При дальнейшем движении вдоль этой оси можно говорить о появлении единой информационной системы предприятия.

Оси "F" и "D" определяют специфику деятельности организации, регламентируемую положением третьей координаты (R) пространства модели. При этом модель не зависит от технологии обработки информации, принятой на предприятии - все решает только цель деятельности, будь то государственная организация, торговая компания или промышленная фирма.

         В результате такой эволюции информационное пространство современного предприятия часто состоит из интегрированных  в различной степени компонентов:

  телефонная связь и телефонные станции от простейших АТС, до программных АТС, функционирующих на базе сетевого протокола IP, принимающие и передающие звонки через обычную компьютерную сеть. Такие станции позволяют программировать логику обработки вызовов и распределять поступающие информационные потоки на соответствующих сотрудников и (или) аудио хранилища, хранить и передавать детальную информацию о звонках (CDR – Call Detail Record) и содержание разговора во внешние информационные системы;

  факс сервера;

  SMS сервера;

  корпоративная электронная почта, сайт, хранилище файлов;

  набор баз данных и (или) одно или несколько хранилищ данных;

  корпоративные информационные системы, такие как ERP, CRM, SCM, PLM, MRP, MRPII и другие, обеспечивающие сбор и первичную обработку информации;

  аналитические системы, обеспечивающие всесторонний анализ хранимой информации;

  специализированные учётные системы, такие как бухгалтерские, складские, банковские и др.

  электронные базы знаний;

  системы поддержки принятия решений, которые оказывают поддержку в принятии решений, но его не принимают;

  другие информационные системы;

  и, конечно же – все традиционные формы и каналы поступления, обработки  информационных ресурсов.

**Тема 2.  Анализ информации на предприятии.**

**Содержание**

1.     Виды и задачи анализа на предприятии.

2.     Содержание экономических показателей.

3.     Классификация методов анализа

4.     Источники данных для проведения анализа.

**1. Виды и задачи анализа на предприятии.**

В современных условиях, характеризующихся высокой динамичностью рынка, возрастанием значимости фактора конкуренции особое значение имеет совершенствование деятельности предприятий. Оно включает широкий круг вопросов,направленных на эффективную организацию процесса деятельности предприятий, приспособленных к условиям, в которых эта деятельность осуществляется. Однако такое совершенствование подразумевает детальное изучение и анализ процесса.

Управление бизнес-процессами требует комплексного решения многочисленных проблем, обусловленных как внешними, так и внутренними факторами. Многообразие информации, характеризующей и описывающей деятельность предприятия или фирмы, приводит к необходимости применения новых методов и инструментальных средств ее обработки.

С переходом к рыночной экономике управление российскими предприятиями претерпело существенные изменения. Основные из них связаны с необходимостью ориентации деятельности фирм на удовлетворение потребностей рынка. Так вот, планы деятельности предприятий теперь берут начало от целей, стратегий и программ маркетинга, т.е. обращены вовне. Вместе с тем, для динамичных, развивающихся, структурно и функционально адаптирующихся к быстро изменяющимся условиям организаций очень важно и управление их внутренним развитием. Такое управление сплошь уникально. Жизненная важность грамотности представления постоянно меняющихся и трудно предсказуемых данных о рыночных потребностях и деятельности фирмы требует, в первую очередь, проведения грамотного разностороннего анализа при управлении этой областью деятельности фирм.

На данный момент решения на предприятиях все же нередко принимаются на интуитивной основе. Ошибочный прогноз в отношении целесообразности того или иного решения может вызвать неприятные последствия, связанные как с недополучением дохода, так и с неполной загрузкой мощностей.

Кроме того, эффективность деятельности предприятия зависит и от своевременности выполнения операций в процессе управления. Более полно представлять информацию для принятия решения и анализировать её помогает использование разного рода аналитических методов и информационных средств.

Когда мы говорим о применении их в экономке, то имеем в виду не просто проведение различного рода экономических расчетов, а использование для изучения экономических закономерностей, получения новых теоретических выводов, нахождение наилучших экономических решений. Математические методы и модели экономики, отражающие с помощью математических соотношений основные свойства экономических процессов и явлений, также представляют собой эффективный инструмент исследования сложных экономических проблем.

В общем виде процесс управления можно свести к ответу на 3 вопросы:

     Где мы находимся (оценка текущего состояния),

     Куда мы хотим прийти (моделирование идеального или желаемого состояния),

     Как мы туда попадём (анализ и прогнозирование процесса перехода из текущего состояния в желаемое).

Для сложных систем характерно, что управлять ими приходится, как правило, в условиях неполноты информации, незнания основных закономерностей функционирования и постоянного изменения множества внешних и внутренних факторов. Поэтому процесс управления имеет итерационный характер. После принятия решения и применения управляющего воздействия необходимо вновь оценить состояние, в котором находится система, и решить вопрос о том, правильно ли мы движемся по намеченному пути. Если отклонения нас не удовлетворяют, то необходимо переопределить процесс управления. Идеальное состояние при этом также не статично, следовательно, процесс анализа и отслеживания изменений должен стать непрерывным. Причём желательно иметь возможность «дрейфа» текущего состояния для просчёта движения на несколько шагов вперед. Кроме того, анализ прошлого опыта даёт зачастую много полезной информации, поэтому необходимо иметь как можно более полную информацию о прошлых текущих и идеальных состояниях.

Современные информационные технологии при поиске ответов на поставленные вопросы позволяют аналитику формулировать и решать следующие классы задач:

1.     Аналитико-расчётные – вычисление заданных показателей и статистических характеристик деятельности на основе исторической информации из БД.

2.     Визуализация данных – наглядное графическое и табличное представление имеющейся информации, в т.ч. аналитической.

3.     Добыча знаний (извлечение) – определение взаимосвязей и взаимозависимостей бизнес-процессов на основе имеющейся информации.

4.     Имитационные – проведение компьютерых экспериментов с математическими моделями, описывающими поведение сложных систем в течение заданного или формируемого интервала времени. Задачи этого класса применяются для анализа возможных последствий принятия того или иного управленческого решения (анализ «что - если»).

5.     Синтез-управление – используются для определения допустимых управляющих воздействий, обеспечивающих достижение заданной цели. Задачи этого типа применяются для оценки достижимости намеченных целей, определения множества возможных управляющих воздействий, приводящих к заданной цели.

6.     Оптимизационные – основаны на интеграции имитационных, управленческих и статистических методов моделирования и прогнозирования. Задачи данного класса позволяют выбрать из множества возможных управлений те из них, которые обеспечивают наиболее эффективное (с точки зрения определённого критерия) продвижение к поставленной цели.

Бизнес-процесс создания адекватных моделей осложняется двумя объективно существующими моментами.

Первый из них состоит в том, что накоплению у бизнесменов личного опыта в ходе повседневной деятельности препятствует динамичное изменение экономической ситуации, которое особенно характерно для современной России.

Второй момент заключается в том, что в предпринимательской деятельности, да ещё в условиях свободного рынка, отсутствует возможность проведения целенаправленных экспериментов, которые предшествуют открытию гипотезы и позволяют проверять её на практике.

Остановимся на понятии экономического анализа. Все определения предмета экономического анализа, которые содержатся в работах российских экономистов, принято разграничивать на несколько характерных групп.

  К первой из них относятся формулировки М.З. Рубинова, П.И. Савичева, М.Ф. Дъячкого, М.И. Баканова, в соответствии с которыми «под предметом экономического анализа понимаются ***хозяйственные процессы*, происходящие на предприятиях**».

  Вторая группа определений выделяется трактовкой предмета экономического анализа как непосредственно ***хозяйственной деятельности*предприятий, объединений и вышестоящих звеньев**. Такой подход характерен для работ И.И. Поклада, Н.В. Дембинского, С.Б. Барнгольц, которые   подчеркивают   необходимость   отражения   хозяйственной деятельности в плановых, отчетных, учетных и других источниках информации. Другими словами «**предметом анализа могут быть только *документированные события***».

  У Н.Г. Чумаченко «**предмет экономического анализа трансформируется в *информационный поток*о планомерной хозяйственной деятельности предприятий, который исследуется в целях достижения *обоснованности* принимаемых   решений**». Специфичность   такого   определения обуславливается рассмотрением экономического анализа в системе управления производством в качестве одного из его инструментов.

  Отличается от рассмотренных выше определение предмета анализа, данное В.П.Копняевым, который отнес к нему «***использование всех видов ресурсов* (материальных, трудовых и денежных), которыми располагает предприятие для выполнения плана**»

  И,   наконец,   в   последней группе определений, данных Е.В. Долгополовым, Г.В. Савицкой, Н.П. Любушиным, под предметом экономического анализа понимается ***экономика предприятий*, рассматриваемая как совокупность производственных отношений в процессе производства и распределения созданного продукта, использования производственных ресурсов**.

А.Д. Шеремет *связывает* предмет анализа через *производственные отношения* с хозяйственным расчетом, технической стороной производства, природными условиями и социальным развитием коллективов. В его учебнике дано определение: «**Предмет экономического анализа** **- *хозяйственные процессы* предприятий, их *социально-экономическая эффективность* и конечные финансовые *результаты* деятельности, складывающиеся под воздействием *объективных и субъективных*факторов и отражающиеся через систему *экономической информации*».**

Как и во всех остальных видах процессов в экономическом анализе выделяют и субъектов экономического анализа. Субъектом в общенаучном понимании «является сам человек или сообщество людей, все человечество, т.е. живые личности, творчески относящиеся к объекту познания». Субъекты экономического анализа, прежде всего, разделяются на так называемых:

  производителей и

  потребителей

результатов исследования.

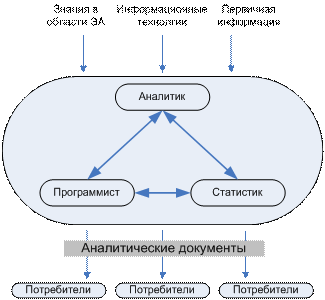
Развитие науки экономического анализа в этом случае зависит от взаимодействия с одной стороны потребностей и требований пользователей, с другой вооруженности труда разработчиков. Пользователи результатов науки определяют значимость науки.

Принципиально можно выделить следующих субъектов-производителей ЭА (рис. 1):

  Аналитик – ему непосредственно адресуются научные знания в области ЭА; связан с использованием ИТ;

  Программист и математик (или системный аналитик); связан с разработкой ИТ;

  Статистик (или бухгалтер); связан с оказанием информационных услуг.



**Рисунок 1.** Субъекты экономического анализа

Результаты экономического анализа для практики могут стать:

  информационным продуктом

  информационной услугой

  информационной технологией.

Услуга и продукт, однако, есть результат технологии. Поэтому объектом экономического анализа для различных субъектов будет именно технология обработки информации (ТОИ). ТОИ, являясь результатом отношений субъектов, должна принимать соответствующий вид в результате решения противоречия этих отношений. Значит, с одной стороны ТОИ классифицируется  с точки зрения потребителей, а с другой – с точки зрения производителей.

Процессы анализа и планирования могут быть  как контролируемыми, т.е. управляемыми субъектом, так и неконтролируемыми, на которые он, субъект, не в состоянии оказать достаточно заметно влияющее воздействие.

При анализе имеющейся информации наблюдаемые процессы можно разделить по этому признаку (влияем или не влияем на данный процесс) с дальнейшим выявлением существенных факторов, степени их влияния на рассматриваемый процесс и т.д. Такого рода анализ, сочетающийся с прогнозированием, является непременным условием эффективного планирования деятельности предприятия, обоснования принимаемых управленческих решений.

Наблюдаемые или исследуемые процессы - это в первую очередь, протекающие на предприятии (в корпорации) бизнес - процессы. Под ними подразумевают совокупность работ по выполнению какой-либо задачи предприятия. На основе анализа хода этих процессов внешних условий, внутреннего состояние предприятия, в том числе финансового, делаются определенные выводы, вырабатываются или корректируются цели предприятия. В соответствие с выработанными целевыми установками ставятся задачи, осуществляется планирование мероприятий и деятельности предприятия в целом. При этом учитываются принятые критерии оценки, вырабатываются управленческие решения по реализации планов. Эту часть анализа назовем **стратегическим анализом**.

В процессе реализации планов должен осуществляться контроль и анализ хода их реализации, который назовём **текущим анализом**. Результаты его являются материалом для выработки решений по корректировке, с одной стороны уже состоявшихся управленческих решении, а с другой - по доработке самих планов или даже целевых установок в случае значительных отклонений полученных показателей от запланированных, отсутствия ресурсов или в связи с какими-либо другими обстоятельствами.

В целом содержание анализа состоит в систематизации, оценке полученных параметров в соответствии с принятой системой показателей, изучении и оценке факторов влияющих на деятельность предприятия, выявлении его сильных и слабых сторон, определении возможностей и рисков.

Традиционно «поставщиком» информации для проведения анализа являются отчеты или справки, составленные на основе сведений, хранящихся в системах обработки данных (OLTP-системах). Однако по мере развития методов анализа и переполнения этих систем данными пришло понимание того, что подобный способ получения аналитической информации стал неэффективным. В качестве основных причин можно назвать следующие:

1.     Получить требуемые для принятия решения данные за приемлемое время зачастую не представляется возможным, несмотря на обилие средств автоматизации на предприятии и даже на наличие нужных данных в системах обработки. Такой «информационно-аналитический голод» при кажущемся изобилии информации обусловливают несколько факторов. В том числе:

a.      Необходимые для анализа данные часто рассредоточены по разным OLTP-системам, функционирующим в организации в силу исторических причин. Примером проблем такого рода, например, в банке является оперативное управление позициями корреспондентских счетов и планирование этих позиций, для которых необходимы данные платежного календаря от большого количества подразделений банка.

b.     Организация данных в OLTP-системах не ориентирована на решение задач анализа. Данные организованы в целях оптимальной поддержки конкретной совокупности технологических операций, которые хорошо структурированы, формализованы и повторяются изо дня в день. Наиболее важным требованием для продуктов такого класса является производительность при выполнении процессов актуализации данных, что, несомненно, отражается в структурах хранения данных.

c.     Для задач анализа характерно использование внешних данных (из внешних источников), которые в БД OLTP-систем отсутствуют вообще. Ими могут быть, например, рыночные процентные ставки, индексы, данные по конкурентам, корреляция между различными финансовыми инструментами и т.п. Ввод информации такого рода в базы OLTP ничем не оправдан, поскольку они не используются в процессах обработки.

2.     Реализация аналитических приложений на базах данных OLTP-систем мешает оперативной обработке данных, снижает производительность системы и повышает стоимость ее сопровождения. Ведь для задач анализа и принятия решений зачастуютребуются «исторические сведения» за достаточно длительный период. В БД OLTP-систем со временем накапливается большой объем архивной информации, которая практически не используется в процессах оперативной обработки данных. Она лишь усложняет сопровождение и эксплуатацию этих систем (в результате чего возрастают финансовые затраты), что в конечном итоге негативно сказывается на производительности труда работников. Кроме того, реализация многих задач анализа требует обработки значительных объемов информации, применения сложных алгоритмов обработки – это приводит к существенным затратам вычислительных ресурсов. Решение таких задач в рамках баз данных OLTP-системы заметно снижает производительность основных процессов обработки.

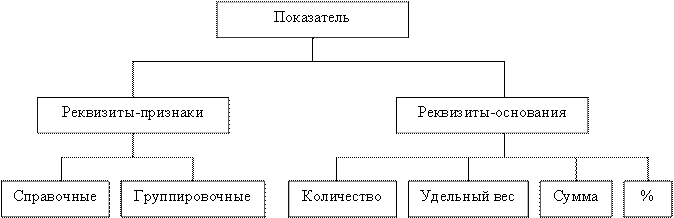
**2. Содержание экономических показателей.**

Для оценки ситуации на предприятии или другом объекте при подготовке и принятии решений по управлению объектом должна быть выработана или принята система оценок, которая ложится в основу аналитической работы с имеющейся и требуемой информацией. Систему оценок можно рассматривать как совокупность показателей деятельности объекта с их критериальными значениями.

Для экономической сферы деятельности это будут экономические показатели. Системы экономических показателей основываются на элементах структуры информационного пространства.

С экономической точки зрения под показателем деятельности предприятия понимают "конкретное проявление экономической категории в характеристике объекта".  Это наиболее строгое определение  экономической  сущности  показателя  из встречающихся в литературе.

Экономический показатель является составной единицей информации, отражающей количественную характеристику некоторого процесса предметной области – реквизит-основание вместе с однозначно определяющими его качество реквизитами-признаками (рис. 2).



**Рисунок 2.**  Схема структуры экономического показателя.

Реквизиты основания подразделяются по типу алгоритмов их получения на количественные, стоимостные, процентные, удельные веса и др. Множество реквизитов-признаков по степени формализации делятся на два подмножества:

  справочные реквизиты-признаки – как правило, наименования, предназначенные для понимания показателя экономистом-пользователем;

  группировочные реквизиты-признаки – это закодированные аналоги справочных признаков, предназначенные для логической обработки информации на ПК.

Отдельным показателем невозможно достаточно полно отобразить и оценить состояние экономического объекта, поэтому применяются системы показателей.

Система экономических показателей отображает финансово-хозяйственную деятельность как на уровне предприятия, то есть на микроуровне, так и на макроуровне. Два этих раздела показателей касаются как каждого предприятия, корпорации (объединения), так и государственных   органов   и   негосударственных   организаций регионального, государственного и межгосударственного масштаба (уровня). Показатели отражаются в документации, которая ведется на предприятии, используются в оценках его состояния и динамики процессов, происходящих на предприятии, в информационных системах. Естественно, что при анализе их используют непосредственно и интерпретируют различным образом в целях извлечения знаний, формирования выводов и т.д.

В связи с этим централизованно разработаны системы реквизитов соответствующих  обязательных  к  использованию  документов: государственные стандарты документооборота, формы документов, системы кодирования статистической, учётной, финансовой отчётности и другой документации. На предприятии разрабатывается внутренняя система показателей и соответствующих реквизитов.

Предприятие регулярно обменивается информацией с внешней средой в виде представления разного рода отчётности и других материалов, получения различных руководящих и установочных документов, информационных материалов и т. д.

Предприятие представляет в государственные органы ежеквартально и нарастающим итогом за полугодие, год:

  внешнюю финансовую отчётность в территориальные государственные органы – налоговую инспекцию и финансовое управление;

  статистическую отчетность о различных сторонах деятельности предприятия в территориальные органы Госкомстата РФ;

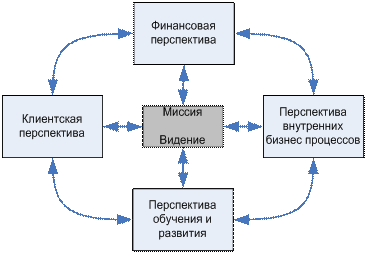
  отчётность о финансово-хозяйственной деятельности в государственные фонды (предстоит их реформирование в плане объединения и упрощения отчетности).

В государственных органах проведены структуризация, формализация, классификация и кодирование многих видов документов на основе разработки соответствующей системы реквизитов и показателей, которая постоянно видоизменяется и далека от совершенства, а самое главное – от общепринятых в мировой практике международных стандартов, систем отчётности, классификации, кодирования. Особенно страдает от такого положения дел финансовая сфера управленческой деятельности.

В качестве одной из последних концепций, отражающей идеологию построения и использования систем экономических показателей является концепция BSC (Balanced Scorecard)  или система сбалансированных показателей. Идеологами данной концепции являются два американских экономиста П. Нортон и Р. Каплан, которые и описали данную концепцию в 1992 году. Основная цель внедрения данного подхода на предприятии – получение гибкого инструмента по управлению эффективностью бизнеса на базе анализа значений различных наборов показателей на различных уровнях организационной структуры.

Базовой идеей данной концепции является утверждение о том, что «менеджерам, так же как и пилотам самолётов требуется инструментарий для управления различными аспектами окружающей среды и производительностью направленный на то, чтобы полёт был успешным».

В рамках концепции выделяется базовое окружение набора показателей. Так в центре располагаются миссия и видение бизнеса компании, на котором основываются показатели из четырёх перспектив: финансовая перспектива, взаимоотношения с клиентами, внутренние бизнес процессы, обучение и развитие (рис. 3).



**Рисунок 3.** Среда системы сбалансированных показателей

В качестве примеров показателей различных перспектив можно выделить следующие:

       Финансовая перспектива

o      Рост и структура выручки;

o      Уменьшение стоимости/ увеличение продуктивности;

o      Использование активов/ инвестиционная стратегия;

       Клиентская перспектива

o      Доля рынка;

o      Приобретение клиентов;

o      Удовлетворённость клиентов;

o      Прибыльность клиентов;

       Перспектива внутренних бизнес - процессов

o      Инновации;

o      Операции;

o      После продажное обслуживание;

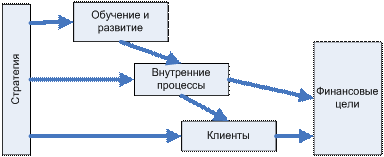
       Перспектива обучения и развития

o      Удовлетворённость сотрудников;

o      Способность удержания сотрудников;

o      Продуктивность работы сотрудников.

После определения набора показателей происходит их связывание с общей стратегией компании и создание целостной системы экономических показателей (рис.4).



**Рисунок 4.** Влияние стратегии на создание системы экономических показателей

Таким образом, для каждой перспективы необходимо определить цели и установить между ними причинно-следственные связи.

**3. Классификация методов анализа**

Что же такое метод вообще? Слово **“метод”** происходит от греческого “methodos”, что буквально означает “путь к чему-либо”. В толковом словаре можно встретить следующие понимания:

Метод –*способ теоретического исследования или практического осуществления чего-нибудь.*

Методика –*совокупность методов практического выполнения. Прием - способ осуществления чего-нибудь.*

Способ -*действие или система действий, применяемых при осуществлении чего-нибудь.*

Отсюда можно сказать, что метод - есть совокупность общих способов теоретического исследования предмета и выработка приемов  практического применения с использованием категориально-понятийного аппарата. В результате выработки приемов формируется методика практического выполнения.

К *общим способам* теоретического исследования относятся:

  наблюдение,

  сравнение,

  моделирование,

  системный подход,

  анализ и синтез,

  индукция и дедукция,

  абстрагирование,

  формализация,

  исторический и логический методы

  и др.

**Техническими приемами анализа** являются - способы четкого и упрощенного выражения различных ресурсов и явлений в их взаимосвязи и взаимозависимости, а также из­мерение влияния тех или иных факторов или причин на измене­ние уровня соответствующих показателей для сравнения анало­гичных показателей, приведения их к сопоставимому виду, упрощения цифровых величин и др.

Так, Каракоз выделяет следующие **группы технических приёмов**:

  *простые, элементарные* (сравнение показателей работы, упрощение цифровых величин показателей, группировка и детализация показателей работы, отвечающая требованиям логических приемов — индукции и дедукции);

  *сложные* (элими­нирование показателей — цепные подстановки, приемы разниц и относительных величин показателей);

  *математические* (корреля­ционной связи, линейного и динамического программирования, сетевых графиков в планировании, теории массового обслужи­вания).

Характерными особенностями метода экономического анализа являются:

  использование си­стемы показателей, всесторонне характеризующих хозяйствен­ную деятельность;

  изучение причин изменения этих показа­телей;

  выявление и измерение взаимосвязи между ними в целях повышения социально-экономической эффективности.

Существует большое количество методов анализа, которые делятся на группы по различным признакам. Рассмотрим систему признаков, характеризующих методы анализа. Их можно сгруппировать:

***По целям - это:***

        оценка состояния и результатов деятельности предприятия;

        постоянный контроль рациональности ведения хозяйственной деятельности, выявление резервов для обеспечения выполнения поставленных задач;

        прогнозирование хода внутренних процессов на предприятии и внешних факторов, влияющих на его деятельность.

***По временному фактору анализ разделяют на:***

        использующий прошлую информацию, отражённую в документации и на различных носителях и содержащуюся в информационной системе - анализ фактов;

        на базе как прошлой, так и обращённой в будущее, то есть прогнозной информации - анализ событий и отклонений,

        анализ будущей информации - по существу оценка бюджетов и планов, их альтернатив.

***По масштабности решаемых или обслуживаемых задач:***

        стратегический, сюда можно отнести оценку эффективности целей, долгосрочные прогнозы, исторические оценки процессов и явлений и т.д.

        оперативный - это оценка текущего состояния, выявление узких мест и отклонений,

        система раннего предупреждения.

***По предметным областям***, в рассматриваемом случае – экономики различают анализ:

        в маркетинге;

        в производственной или основной деятельности;

        в логистике;

        в обеспечении ресурсами;

        в финансовой;

        в сфере инвестиций и инноваций.

***По методам различают***:

        сравнительный по подразделениям, предприятиям, регионам, временным периодам и т.д.;

        анализ отклонений;

        функционально-стоимостный;

        анализ цепочки создания стоимости и конкурентный анализ по Портеру;

        анализ полей бизнеса

        бенчмаркинг;

        интеллектуальный анализ

В процессе анализа используются различные математические методы, в том числе:

        математической статистики;

        многомерного статистического анализа,

        эконометрики;

        алгебры - линейной, логики, предикатов, нечёткой логики;

        численные методы анализа.

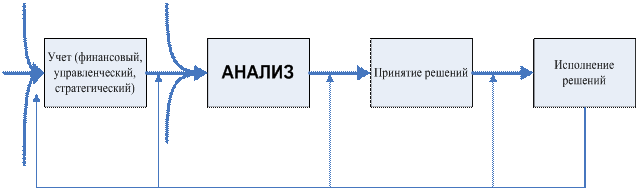
Необходимо заметить, что какой-либо конкретный аналитический процесс или  аналитическая работа могут характеризоваться одновременно несколькими из перечисленных выше признаков.

**4. Источники данных для проведения анализа.**

Аналитическая   работа   на   предприятии   осуществляется специальной группой. Она может быть автономной или включённой в какое-либо подразделение. В последнее время создаются подразделения контроллинга, в чьи функции в качестве основной включается эта деятельность. В отдельных, особо сложных ситуациях пользуются услугами консультантов. На малых предприятиях эта работа может быть возложена на одного из заместителей руководителя или эксперта.

Дня   уяснения   функций   ИАС   необходимо   изучить информационный обмен, связанный с аналитической работой. В общей постановке анализ основан на переработке информации, которую аналитики должны где-то получить, и выдаче информации заинтересованным лицам или организационным единицам. Место аналитического процесса в цепочке других, связанных с управлением процессов показано на рисунке 5.

|  |  |
| --- | --- |
|  |  |
|  | |  | | --- | | Источники внешней информации | |

Подпись: Источники внутренней информации

**Рисунок 5.**  Место анализа в цепочке принятия управленческих решений

Все источники информации для анализа можно разделить на внутренние и внешние.

***К внутренним источникам относятся:***

  бухгалтерский учёт, включая аналитический и складской;

  статистический учёт,

  управленческий учёт;

  деловая переписка;

  материалы различных исследований и обследований, выполненных на предприятии

  текущая документация, в том числе материалы ревизий и аудиторских проверок и т.д.;

  зафиксированные данные опросов;

  устная информация;

  информация из баз данных, эксплуатирующихся на предприятии ЭИС и автономных автоматизированных рабочих мест (АРМ);

Из перечисленных видов учёта бухгалтерский и статистический относятся к обязательным видам учёта.

***К внешним источникам информации относятся:***

  установочная информация из  государственных  органов и вышестоящих организаций (для зависимых предприятий) это правовые и руководящие документы, инструкции и т.д., определяющие условия функционирования,

  информация из специализированных информационных организаций и их информационных хранилищ, к ним относятся различные фонды, финансовые и биржевые и т.д.;

  библиотечные фонды и информационные хранилища;

  средства массовой и специализированной информации;

  глобальные информационные ресурсы, например сеть Интернет и другие;

  данные деловой разведки и прочие возможные источники информации.

С другой стороны служба анализа выдает информацию заинтересованным потребителям. Основные её потребители - лица, принимающие решения (ЛПР).

На предприятии должен быть установлен порядок доступа к такой информации  по  причине  её  особой  ценности  и  подчас конфиденциальности.

С точки зрения аналитических систем информацию можно разделить на:

  Агрегированные данные

  Исторические данные

  Прогнозируемые данные

**Агрегированные данные**. Пользователя, занимающегося анализом, редко интересуют детализированные данные. Более того, чем выше уровень пользователя (руководителя, управляющего, аналитика), тем выше уровень агрегации данных, используемых им для принятия решения. Рассмотрим в качестве примера фирму по продаже автомобилей. Коммерческого директора такой фирмы мало интересует вопрос: "Какого цвета "Жигули" успешнее всего продает один из ее менеджеров - Петров: белого или красного?" Для него важно, какие модели и какие цвета предпочитают в данном регионе.

**Исторические данные**. Важнейшим свойством данных в аналитических задачах является их исторический характер. После того как зафиксировано, что Петров в июне 2004г. продал 2 автомобиля "Волга" и 12 автомобилей "Жигули", данные об этом событии становятся историческим (свершившимся) фактом. И после того, как информация об этом факте получена, верифицирована и заведена в БД, она может быть сколько угодно раз считана оттуда, но уже не может и не должна быть изменена.

Другим неотъемлемым свойством исторических данных является обязательная спецификация времени, которому эти данные соответствуют. Причем время является не только наиболее часто используемым критерием выборки, но и одним из основных критериев, по которому данные упорядочиваются в процессе обработки и представления пользователю.

Во многих организациях используются как общепринятые, так и собственные календарные циклы (финансовый год может начинаться не в январе как календарный, а, например, в июне); время является стандартным параметром практически любой аналитической, статистической или финансовой функции (прогноз, нарастающий итог, переходящий запас, скользящее среднее и т.д.).

**Прогнозируемые данные**. Когда говорится о неизменности и статичности данных в аналитических системах, имеется в виду неизменность исключительно исторических данных (данных, описывающих уже произошедшие события). Такое предположение ни в коем случае не распространяется на прогнозируемые данные (данные о событии, которое еще не происходило). И этот момент является весьма существенным.

Например, если мы строим прогноз об объеме продаж на июнь 2005г. для менеджера Петрова, то, по мере поступления фактических (исторических) данных за 2004г., эта цифра может и будет многократно изменяться и уточняться. Более того, достаточно часто прогнозирование и моделирование затрагивает не только будущие, еще не произошедшие, но и прошлые, уже свершившиеся события. Например, анализ: "а, что будет (было бы)..., если (бы)..?", строится на предположении о том, что значения некоторых данных, в том числе и из прошлого, отличны от реальных.

На первый взгляд, мы сами противоречим себе, говоря о неизменности данных, как основополагающем свойстве аналитической системы. Но это не так. Это кажущееся противоречие наоборот подчеркивает и усиливает значимость требований к неизменности исторических данных. Сколько бы мы не упражнялись (например, при анализе: "а что... если..?") со значением объема продаж за июнь 2004г., значения исторических (реальных) данных должны оставаться неизменными.

С технической точки зрения источники данных для проведения анализа могут быть представлены в следующих основных формах:

  в виде файлов определённого формата (ранее самым распространённым форматом были DBF файлы, сейчас же всё большую долю начинает занимать формат XML);

  в виде привычных большинству специалистов реляционных баз данных, в которых хранится первичная или агрегированная информация;

  в виде хранилищ данных, которые собирают внутри себя информацию из различных предметных транзакционных баз данных и производят её агрегацию и систематизацию;

  в виде информации, получаемой из отчётов, когда в процессе выполнения аналитической работы результаты (агрегаты) одного отчёта становятся источником данных для других отчётов, тем самым, являясь элементом-источником общей технологической цепочки при осуществлении анализа;

  информация, получаемая напрямую при вызове удалённых процедур во внешних по отношению к аналитической подсистем или даже информационных систем. Данная технология реализована во многих языках программирования, на большинстве технологических платформ и основывается на комплексной идеологии удалённого вызова процедур (RPC – Remote Procedure Call);

  традиционные Интернет сайты различных СМИ, информационных агентств и обычных компаний в зависимости от типа и содержания требуемой информации;

  «ВЕБ сервисы», взаимодействие с которыми основывается на протоколе SOAP и XML и которые по основной концепции должны быть зарегистрированы в каталогах и доступны по протоколу UDDI – основная цель данной концепции – создать единую гибкую инфраструктуру автоматизированного взаимодействия информационных систем различных компаний с целью оказания каких-либо информационных услуг (сервисов) через традиционную сеть Интернет. Отсюда пошло и само название – WEB Services (ВЕБ сервисы).

**Тема 3.   Информационно – аналитическая система как инструмент проведения экономического анализа.**

**Содержание**

1.     Общее понятие информационно – аналитической системы.

2.     Функции и сферы применения ИАС

3.     Классификация аналитических систем

4.     Концепции построения ИАС

5.     Общая структура информационной аналитической системы

**1. Общее понятие информационно – аналитической системы.**

Современный этап развития рыночных отношений в российской экономике характеризуется началом экономического подъема. Сегодня все большее число организаций приходит к пониманию того, что без наличия своевременной и объективной информации о состоянии рынка, прогнозирования его перспектив, постоянной оценки эффективности функционирования собственных структур и анализа взаимоотношений с бизнес - партнерами и конкурентами их дальнейшее развитие становится практически невозможным. Начинают  приобретать определяющее значение знания о протекающих хозяйственных процессах. На успех ведения дела влияют как объективные, так и субъективные факторы. К объективным факторам можно отнести:

        закономерности протекания хозяйственных процессов,

        правовую среду,

        неписаные правила и традиции ведения дел,

        экономическую конъюнктуру и т.д.

Большое значение имеет субъективный фактор, под которым понимается влияние на ход бизнес - процессов работников предприятия и в особенности лиц, принимающих решения (ЛПР).

Для выработки и принятия соответствующих складывающейся обстановке решений необходима информация, которая должна удовлетворять требованиям полноты, достоверности, своевременности (актуальности), полезности.

Основополагающую роль в подготовке принятия решений играет его обоснование по имеющейся у ЛПР информации. Её, как правило, получают из различных внутренних и внешних источников. В интересах выработки   адекватного   решения   используются   внутренние информационные ресурсы, которые   складываются из отражения деятельности (функционирования) объекта в документах, других видах и способах сбора, обработки, хранения информации. А также внешние по отношению к объекту информационные ресурсы, например (если это предприятие) - корпорации, отрасли, региона, а также глобальные – из средств массовой информации, специальной литературы, всемирной информационной сети Интернет и т.д.

Таким образом, границы информационного пространства как отображения деятельности предприятия и его взаимодействия с внешней средой, в рамках которого принимаются решения, выходят далеко за пределы предприятия.

Эти обстоятельства вынуждают использовать имеющиеся в настоящее время весьма развитые программно-технические средства. Широкое и эффективное применение этих средств стало одним из факторов выживаемости и успеха предприятия в условиях острой конкурентной   борьбы.   Получили   широкое   распространение автоматизированные информационные системы.

Проблема анализа исходной информации для принятия решений оказалась настолько серьезной, что появилось отдельное направление или вид информационных систем – **информационно – аналитические системы (ИАС).**

Информационно-аналитические системы (ИАС) призваны на основе данных, получаемых в режиме реального времени, помогать в принятии управленческих решений. ИАС - это современный высокоэффективный инструмент поддержки принятия стратегических, тактических и оперативных управленческих решений на основе наглядного и оперативного предоставления всей необходимой совокупности данных пользователям, ответственным за анализ состояния дел и принятие управленческих решений. Комплекс информационно-аналитических систем затрагивает всю управленческую вертикаль: корпоративную отчетность, финансово-экономическое планирование и стратегическое планирование.

**2. Функции и сферы применения ИАС**

Основное назначение ИАС — динамическое представление и многомерный анализ исторических и текущих данных, анализ тенденций, моделирование и прогнозирование результатов различных управленческих решений.

Основными функциями информационно-аналитической системы являются:

        Извлечение данных из различных источников, их преобразование и загрузка в хранилище;

        Хранение данных;

        Анализ данных, в том числе оперативный и интеллектуальный;

        Подготовка результатов оперативного  и интеллектуального анализа для эффективного их восприятия потребителями.

Результатом применения средств ИАС являются с одной стороны — регламентные аналитические отчеты, ориентированные на нужды пользователей различных категорий, с другой — средства интерактивного анализа информации и быстрого построения отчетов пользователями-непрограммистами с использованием привычных понятий предметной области.

Функцию сбора и хранения информации с сопутствующей доработкой в информационно-аналитических системах, выполняют  информационные хранилища (Data Warehouse).

В связи с большим объемом и сложностью анализ данных имеет два направления - оперативный анализ данных   (информации),   широко  распространена  англоязычная аббревиатура названия – On-line Analytical Processing (OLAP). Основной задачей оперативного или OLAP-анализа является быстрое (в пределах секунд) извлечение необходимой аналитику для обоснования или принятия решения информации. Эту функцию выполняют всевозможные OLAP – средства.

Интеллектуальный анализ информации - имеет также широко распространенное в русской специальной литературе англоязычное название  Data Mining.  Он предназначен для фундаментального исследования проблем в той или иной предметной области. Требования по времени менее жёстки, но используются более сложные методики. Ставятся, как правило, задачи и получают результаты стратегического значения. Эту функцию выполняют всевозможные средства Data Mining.

Жестких границ между OLAP и интеллектуальным анализом нет, но при решении сложных задач приходится использовать весьма мощные специальные программные средства.

С технической точки зрения ИАС – это набор процедур, методов и регламентов, приводящих к регулярному плановому сбору, хранению, анализу и предоставлению информации, используемой для принятия управленческих решений.

Информационно-аналитические системы являются надстройкой над уже функционирующими на предприятии информационными приложениями и не требуют их замены; эти системы аккумулируют данные по всем видам деятельности компании - от состояния складов до финансовой и бухгалтерской отчетности.

Информационно - аналитические системы верхнего уровня служат для принятия стратегических решений. Они позволяют руководителю решать следующие задачи:

  составление консолидированной отчетности и предоставление сводной информации о деятельности предприятия (финансовые, производственные и другие показатели, динамика их изменений и тенденции),

  анализ деятельности дочерних предприятий, филиалов и подразделений компании (анализ доходности, затрат, выполнения плана),

  анализ финансовой деятельности (основные финансовые показатели, тенденции, взаиморасчеты), оптимизация финансовых потоков, реальная оценка себестоимости продукции,

  проведение комплексной оценки деятельности предприятия, основанной на постоянном контроле четырех наиболее существенных ее аспектов (финансы, отношения с внешним миром, внутреннее состояние компании, инновации),

  анализ сбытовых процессов (составление плана, контроль исполнения распоряжений, расчеты за отгруженную продукцию, прогноз поступления средств, прогноз спроса).

Информационно-аналитические системы подразделений предполагают большую детализацию и более сложную аналитическую обработку. Эти системы помогают подготовить информацию для принятия решений в области сбыта, продуктового предложения, финансового планирования.

Различают два вида информационно-аналитических систем по режиму и темпу анализа:

  статические - имеют заранее разработанный сценарий обработки данных при весьма ограниченных возможностях вариаций запросов;

  динамические - обеспечивают обработку нерегламентированных запросов и гибкую систему подготовки отчётов;

Можно выделить следующие принципы построения ИАС на предприятии:

  объединение всех информационных процессов предприятия;

  встраивание системы в уже сложившуюся организационную структуру предприятия;

  координация усилий всех подразделений предприятия при выполнении поставленных задач;

  открытость системы для дальнейшего развития;

  комплексное использование всех доступных методов анализа;

  информационная этика - "от каждого - в общую копилку, и из неё - каждому".

**3. Классификация аналитических систем**

Для обозначения аналитических технологий и средств в целом принято использовать термин "Business Intelligence" или, сокращенно, - BI. Понятие BI объединяет различные средства и технологии анализа и обработки данных масштаба предприятия. На их основе создаются BI-системы. Их цель – повысить качество информации для принятия управленческих решений. BI-системы ранее были известны под названием Систем Поддержки Принятия Решений (СППР, DSS- Decision Support System). В качестве синонимов понятия "СППР" оперируют также понятиями "аналитическая система" или "управленческая система". Сейчас же класс систем BI является независимым классом систем, в который входят системы класса СППР.

По оценкам IDC рынок BI состоит из 5 сегментов:

1.     OLAP-продукты,

2.     инструменты добычи данных,

3.     средства построения Хранилищ и Витрин данных,

4.     управленческие информационные системы и приложения,

5.     инструменты конечного пользователя для выполнения запросов и построения отчетов,

6.     системы СППР.

**Полный классификатор аналитических систем**

|  |  |  |
| --- | --- | --- |
| OLAP-продукты | Способ хранения данных | MOLAP |
| ROLAP |
| HOLAP |
| Место размещения OLAP-машины | OLAP-серверы |
| OLAP-клиенты |
| Степень готовности к применению | OLAP-компоненты |
| Инструментальные OLAP-системы |
| OLAP-приложения |
| Инструменты добычи данных | Метод Data Mining | Фильтрация |
| Деревья решений |
| Генетические алгоритмы |
| Ассоциативные правила |
| Нейронные сети |
| Способ предоставления | В составе OLAP-систем |
| В виде самостоятельных систем Data Mining |
| Средства построения Хранилищ и Витрин данных | Средства проектирования Хранилищ данных | В составе СУБД |
| Универсальные средства |
| Студии |
| Средства извлечения, преобразования и загрузки данных | В составе СУБД |
| Универсальные средства |
| Готовые предметно-ориентированные ХД |  |
| Управленческие информационные системы и приложения | Вид решаемой задачи | Анализ финансового состояния |
| Инвестиционный анализ |
| Подготовка бизнес-планов |
| Маркетинговый анализ |
| Управление проектами |
| Бюджетирование |
| Финансовое управление |
| Масштаб решаемой задачи | Автоматизация труда одного специалиста |
| Для коллективной работы группы сотрудников |
| Для применения в территориально распределенной корпорации |
| Технологическое построение | Монолитные |
| Настраиваемые |
| Инструменты конечного пользователя для выполнения запросов и построения отчетов | В составе OLAP-систем |  |
| В виде систем Query & Reporting |  |
| Системы СППР | В рамках данного материала системы СППР подробно не рассматриваются, так как это является отдельной специфической областью интеллектуальных информационных систем. | |

Рассмотрим более подробно каждый сегмент.

**OLAP-продукты**

На сегодняшний день в мире разработано множество продуктов, реализующих OLAP-технологии. Чтобы легче было ориентироваться среди них, существует несколько классификаций OLAP-продуктов:

        по способу хранения данных,

        по месту нахождения OLAP-машины,

        по степени готовности к применению.

***Рассмотрим классификацию систем по способу хранения данных.*** Основная идея OLAP заключается в построении многомерных таблиц, которые будут доступны для запросов пользователей. Многомерные таблицы (многомерные кубы) строятся на основе исходных и агрегатных данных. И исходные и агрегатные данные для многомерных таблиц могут храниться как в реляционных, так и многомерных базах данных. Поэтому в настоящее время применяются три способа хранения данных: MOLAP(Multidimensional OLAP), ROLAP (Relational OLAP) и HOLAP (Hybrid OLAP).

Соответственно, OLAP-продукты по способу хранения данных делятся на три аналогичные категории:

  В случае MOLAP, исходные и агрегатные данные хранятся в многомерной БД или в многомерном локальном кубе. Такой способ хранения обеспечивает высокую скорость выполнения OLAP-операций. Но многомерная база в этом случае чаще всего будет избыточной. Куб, построенный на ее основе, будет сильно зависеть от числа измерений. При увеличении количества измерений объем куба будет экспоненциально расти. Иногда это может привести к "взрывному росту" объема данных, парализующему в результате запросы пользователей.

  В ROLAP-продуктах исходные данные хранятся в реляционных БД или в плоских локальных таблицах на файл-сервере. Агрегатные данные могут помещаться в служебные таблицы в той же БД. Преобразование данных из реляционной БД в многомерные кубы происходит по запросу OLAP-средства. При этом скорость построения куба будет сильно зависеть от типа источника данных и порой приводит к неприемлемому времени отклика системы.

  В случае использования Гибридной архитектурыисходные данные остаются в реляционной базе, а агрегаты размещаются в многомерной. Построение OLAP-куба выполняется по запросу OLAP-средства на основе реляционных и многомерных данных. Такой подход позволяет избежать взрывного роста данных. При этом можно достичь оптимального времени исполнения клиентских запросов.

***Следующая классификация - по месту размещения OLAP-машины.*** По этому признаку OLAP-продукты делятся на OLAP-серверы и OLAP-клиенты.

  В серверных OLAP-средствах вычисления и хранение агрегатных данных выполняются сервером. Клиентское приложение получает только результаты запросов к многомерным кубам, которые хранятся на сервере. Некоторые OLAP-серверы поддерживают хранение данных только в реляционных базах, другие - только в многомерных. Многие современные OLAP-серверы поддерживают все три способа хранения данных: MOLAP, ROLAP и HOLAP. Одним из самых распространенным в настоящее время серверным решением является OLAP-сервер корпорации Microsoft.

  OLAP-клиентустроен по-другому. Построение многомерного куба и OLAP-вычисления выполняются в памяти клиентского компьютера. OLAP-клиенты также делятся на ROLAP и MOLAP. А некоторые могут поддерживать оба варианта доступа к данным. Среди одних из первых клиентских OLAP-средств можно назвать Oracle Discoverer. Те же возможности обеспечивает и отечественная разработка – продукты Аналитической платформы Контур от компании Intersoft Lab.

У каждого из этих подходов есть свои "плюсы" и "минусы". Нельзя однозначно говорить о преимуществах серверных средств перед клиентскими и наоборот. На практике такой выбор является результатом компромисса "эксплуатационных показателей", стоимости программного обеспечения и затрат на разработку, внедрение и сопровождение аналитической системы.

***Следующая классификация OLAP-продуктов - по степени готовности к применению.*** Различают: OLAP-компоненты, инструментальные OLAP – системы и конечные OLAP-приложения.

  OLAP-компонента– это инструмент разработчика. С ее помощью разрабатываются клиентские OLAP-программы. Различают MOLAP и ROLAP-компоненты: MOLAP-компоненты являются инструментами генерации запросов к OLAP-серверу. Они также обеспечивают визуализацию полученных данных. ROLAP-компоненты содержат собственную OLAP-машину. OLAP-машина обеспечивает построение OLAP-кубов в оперативной памяти и отображает их на экране. Одна из наиболее доступных, но в то же время и одна из самых слабых OLAP-компонент – Decision Cube в составе Borland Delphi.

  Инструментальные OLAP-системы– это программные продукты, предназначенные для создания аналитических приложений. Различают две категории инструментальных OLAP-систем: системы для программирования и системы для быстрой настройки. Системы для программирования – это среда разработчика аналитических систем. В ней, путем программирования запросов к данным, алгоритмов расчета и OLAP-интерфейсов можно создать OLAP-приложение для конечного пользователя. Представителем этого класса программного обеспечения является аналитическая платформа Knosys Pro Clarity. С другой стороны, OLAP-системы для быстрой настройки – это средства, которые предоставляют визуальный интерфейс для создания OLAP-приложений без программирования. Такие системы включают визуальный генератор запросов, встроенные алгоритмы агрегации и инструменты настройки пользовательских OLAP-интерфейсов. В такой технологии реализована большая часть инструментов пакета BusinessObjects и Аналитической платформы Контур.

  Наконец, к третьей категории OLAP-продуктов по степени готовности к применению относятся конечные OLAP-приложения. Это готовые прикладные решения для конечного пользователя. Они требуют только установки, и, не всегда, настройки под специфику пользователя. Пример такого решения – OLAP-приложения системы "Контур Стандарт", подготовленные для анализа данных в различных отраслях и для решения различных аналитических задач.

**Инструменты добычи данных**

Knowledge Discovery in Databases (KDD)– это процесс поиска полезных знаний в "сырых" данных. KDD включает в себя вопросы подготовки данных, выбора информативных признаков, очистки данных, применения методов "раскапывания данных" (Data Mining), а также обработки и интерпретации полученных результатов.

Центральным элементом этой технологии являются методы Data Mining, позволяющие обнаруживать знания при помощи математических правил:

 Фильтрация. Необходимость в фильтрации возникает, когда нужно отделить полезную информацию от искажающего его шума за счет сглаживания, очистки, редактирования аномальных значений, устранения незначащих факторов, понижения размерности информации и т.д. Применение фильтрации в системах анализа данных относится к первичной обработке данных и позволяет повысить качество исходных данных, а, следовательно, и точность результата анализа.

 Деревья решений. Они позволяют представлять правила в иерархической, последовательной структуре, где каждому объекту соответствует единственный узел, дающий решение. Под правилом понимается логическая конструкция, представленная в виде «если..., то...». Деревья решений применяются при решении задач поиска оптимальных решений на основе описанной модели поведения.

 Ассоциативные правила. Они позволяют находить закономерности между связанными событиями. Примером такого правила служит утверждение, что в том случае, если произошло событие А, то произойдет и событие В с вероятностью C. Впервые это задача была предложена для нахождения типичных шаблонов покупок, совершаемых в супермаркетах, поэтому иногда ее еще называют анализом рыночной корзины (market basket analysis).

 Генетические алгоритмы. Они применяются при решении задач оптимизации. Эти методы были открыты при изучении эволюции и происхождения видов. Генетические алгоритмы нужны для настройки нейронных сетей, а также решения различных задач, когда можно составить описание возможных вариантов решения в виде вектора параметров, и известен критерий, определяющий эффективность каждого варианта. Генетические алгоритмы применяются для составления расписаний, портфелей ценных бумаг, заполнения контейнеров при перевозке (пересылке) грузов, выбор маршрутов движения, конфигурации оборудования и т.д.

 Нейронные сети. Они реализуют алгоритмы на основе сетей обратного распространения ошибки, самоорганизующихся карт Кохонена, RBF-сетей, сетей Хэмминга и других подобных алгоритмов анализа данных. Нейронные сети применяются для решения самых различных задач - восстановление пропусков в данных, поиск закономерностей, классификация и кластеризация данных, прогнозирование и моделирование.

Инструменты добычи данных поставляются заказчикам двумя способами:

  в составе OLAP-систем,

  в виде самостоятельных систем Data Mining.

Функциональность Data Mining в той или иной степени полноты реализации включена в аналитические системы различных производителей – Oracle, Hyperion, SAS и т.д. Однако, наиболее «продвинутыми» в этом плане являются специализированные системы математического анализа данных. В России авторитетным разработчиком систем в технологии KDD является компания "Лаборатория BaseGroup".

**Средства построения Хранилищ и Витрин данных**

Хранилища и Витрины данных создаются с применением специализированных средств построения Хранилищ\витрин данных. К этим средствам относятся:

  средства проектирования Хранилищ данных,

  средства извлечения, преобразования и загрузки данных,

  готовые предметно-ориентированные ХД.

Средства проектирования Хранилищ данных входят в состав реляционных и многомерных СУБД от таких производителей как Microsoft, Oracle, IBM, Sybase и других. Также часто применяются универсальные CASE-инструменты, такие как BPWin и ErWin. После описания структур хранения данных специальными системными утилитами выполняется их генерация. Такой подход к созданию Хранилища данных позволяет построить индивидуальное Хранилище или Витрину данных в сжатые сроки. В тоже время такой подход затрудняет перенос наработок от одного заказчика к другому и обмен практическим опытом в решении аналитических задач.

Альтернативным способом построения Хранилищ данных является применение других специализированных средств – Студий для построения Хранилищ данных. Такие продукты предлагают набор шаблонов и заготовок для быстрого создания Хранилища. В составе Студии может предоставляться базовая модель Хранилища данных, ориентированная на определенную бизнес-сферу. С помощью таких инструментов можно значительно быстрее создать Хранилище данных, воспользовавшись опытом предыдущих решений и начать его эксплуатацию. Продукты этого класса, в частности, предлагает компания Sybase – это продукт Industry Warehouse Studio.

ETL-средства (extraction, transformation, loading) - средства извлечения, преобразования и загрузки данных) обеспечивают три основных процесса, используемые при переносе данных из одного приложения или системы в другие. ETL-средства извлекают информацию из исходной базы данных, преобразуют ее в формат, поддерживаемый базой данных назначения, а затем загружают в нее преобразованную информацию. Эти средства обычно входят в состав функциональности реляционных и многомерных СУБД или Студий для построения Хранилищ данных. Однако существуют и специализированные системы, реализующие только ETL-функции. Классической ETL-системой является, например, продукт Ascential DataStage компании Ascential Software.

И, наконец, существует еще один способ построения Хранилищ и Витрин данных – это применение готовых предметно-ориентированных Хранилищ данных. Это самый надежный способ построить Хранилище данных в сжатые сроки. Готовые к эксплуатации Хранилища данных характеризуются наличием в них механизмов средств построения Хранилищ/Витрин данных, взаимосвязанных посредством единого словаря метаданных. К ним относятся - процедуры извлечения, преобразования, очистки и загрузки данных, функции генерации баз данных и процедур обработки, механизмы построения выборок данных, интерфейсы просмотра и анализа данных. Ограничением в применении готовых Хранилищ данных является их предметная ориентация. Например, финансовое Хранилище данных невозможно применить для решения задач оптимизации химического производства. Примером готового предметно-ориентированного Хранилища данных является система Контур Корпорация от компании Intersoft Lab. Применение предметно-ориентированных Хранилищ данных отражает общемировую тенденцию развития рынка BI, наметившуюся в последнее время – предоставления платформ для "быстрой" разработки аналитических приложений.

**Управленческие информационные системы и приложения**

Существует еще один очень разносторонний класс аналитических систем. Это – конечные решения для управленцев и аналитиков. Исторически сложилось так, что технологическая основа реализации таких систем существенно различается. Одни из них построены на современных аналитических инструментах, другие – с применением базовых информационных технологий. Чтобы легче ориентироваться в этих системах вводится 3 классификации:

  по виду решаемой задачи,

  по масштабу решаемой задачи,

  по технологическому построению.

Аналитические системы классифицируются по виду задач, решаемых с их помощью. Среди видов задач можно выделить:

 Анализ финансового состояния банка или предприятия, выполняемый по внешним публичным данным, таким как баланс, отчет о финансовых результатах, иногда - приложение к балансу и отчет о движении денежных средств. Системы - Audit Expert (Про-Инвест), Альт-финансы (Альт), АБФИ (Вестона),Аналитик, АФСП, АДП (ИНЭК) и другие.

 Инвестиционный анализ – для комплексной оценки эффективности инвестиционных проектов и принятия решения об их финансировании, Project Expert (Про-Инвест), Альт-Инвест (Альт) и другие.

 Подготовка бизнес-планов, учитывающих вариации схем производства, сбыта и финансирования, комплексного анализа маркетинговой ситуации, чувствительности проекта по основным параметрам. Системы - Project Expert (Про-Инвест), Альт-Инвест (Альт) и другие.

 Маркетинговый анализ, позволяющий оценить положение компании на рынке, провести сравнительный анализ ее сбытовой деятельности с конкурентами, сформировать оптимальную структуру сбыта, определить доходность различных сегментов рынка и товаров, долю рынка компании, темпы роста и другое. Системы - Marketing Expert (Про-Инвест), Касатка и другие.

 Управление проектами, применяемое для разработки расписания исполнения проекта, определения критического пути и резервов времени исполнения операций проекта; потребности проекта в финансировании, материалах и оборудовании, анализ рисков и планирование расписания с учетом рисков и так далее. Системы - MS Project (Microsoft), Open Plan (Welcom Software Technology) и другие.

 Бюджетирование, обеспечивающее планирование, учет и анализ по центрам финансовой ответственности, бизнесам, продуктам в разрезе активов и пассивов, доходов и расходов, выполнение аллокаций и расчет финансового результата. Системы - Hyperion Pillar, Comshare MPC, Контур Корпорация. Бюджет (Intersoft Lab) и другие.

 Финансовое управление, включающее помимо задач бюджетирования задачи финансового планирования, управленческого учета, трансфертного управления ресурсами, оценки бизнесов по методу ABC, анализа активов, пассивов, рисков. Системы -Oracle Financial Services Applications (Oracle),Контур Корпорация. Финансовое управление (Intersoft Lab) и другие.

На практике встречается гораздо больше видов задач, но здесь был перечислен ряд только тех задач, которые нашли свое воплощение в тиражных аналитических системах. Некоторые из этих систем могут решать только одну задачу, другие являются комплексом, включающим в себя широкий перечень взаимосвязанных задач.

Аналитические системы также классифицируются по масштабу решаемой задачи:

  Системы автоматизации труда одного специалиста. Это так называемые DeskTop-системы, предназначенные для автоматизации труда узкого специалиста. Как правило, для эксплуатации таких систем не требуется помощь службы автоматизации.

  Системы для коллективной работы группы сотрудников. Такие системы содержат средства, обеспечивающие коллективную работу пользователей в режиме реального времени с единой базой данных в рамках прав доступа. Такие системы уже требуют выполнения функций администрирования и сопровождения.

  Системы для применения в территориально распределенной корпорации. Эти системы включают в себя свойства аналитических систем для групп пользователей, а также средства взаимодействия с удаленными подразделениями корпорации (филиалами) в виде технологий сбора данных, дистрибуции НСИ и отчетов. Системы данного класса сложны в эксплуатации, но при этом обеспечивают наиболее полное решение аналитических и управленческих задач.

По технологическому построению аналитические системы можно условно разделить на монолитные и настраиваемые:

  Монолитные аналитические системы характеризуются тем, что аналитическая методика в них реализуется в коде самой программы, а не в виде настройки универсального аналитического инструмента. В этом случае аналитическая система не требует, как правило, выполнения существенных работ по своей настройке (за исключением систем класса Data Mining). Она может использоваться практически сразу после установки. В то же время такие системы не "гибкие" и плохо поддаются изменениям в соответствии с требованиями пользователя. Монолитные системы разрабатываются с применением базовых средств программирования и СУБД.

  Настраиваемые аналитические системы характеризуются тем, что при их создании применяются универсальные аналитические инструменты и специализированные средства, такие как OLAP, Студии, ETL, Data Mining. Их применение позволяет повысить качество аналитической системы, дает перспективы развития, но в то же время приводит к удорожанию конечного решения. Прикладные аналитические системы, выполненные в виде настроек универсальных аналитических инструментов, требуют большего объема работ при внедрении, однако позволяют реализовывать уникальные аналитические методики, принятые в организации.

Компромиссом между этими двумя классами систем является реализация аналитической методики в виде Приложения универсальной аналитической системы. Такой подход позволяет выполнять их тиражирование независимо друг от друга. Но подобных систем в настоящее время на рынке представлено крайне мало.

**Инструменты конечного пользователя для выполнения запросов и построения отчетов**

Системы данного класса (Query & Reporting) предназначены для формирования запросов к информационным системам в пользовательских терминах, а также их исполнение, интеграцию данных из разных источников, просмотр данных с возможностями детализации и обобщения и построение полноценных отчетов, как экранных, так и печатных. Предполагается, что уровень подготовки специалиста, создающего отчеты, может быть приравнен к опыту среднего пользователя Excel. Поэтому пользователь составляет запрос к источнику данных, используя заранее подготовленный программистом каталог терминов (семантический слой). Визуализация результатов запроса может быть представлена пользователю в различном виде – плоские или многомерные таблицы, графики, диаграммы, различные специализированные интерфейсы.

Инструменты конечного пользователя для выполнения запросов и построения отчетов поставляются двумя способами:

  в составе OLAP-систем,

  в виде специализированных систем Query & Reporting.

Практически каждая система класса OLAP снабжена средствами Query & Reporting. Эти средства могут быть как встроенными в основной продукт (примеры – Business Objects, "Контур Стандарт", Oracle Discoverer), так и выделенными в отдельный продукт (например, система Impromptu в составе продуктов Cognos).

Также существуют и специализированные системы генерации и дистрибуции отчетов. Наиболее распространенные из них – это продукты компаний Crystal Decisions и Actuate. В то же время эти системы имеют в своем составе собственные OLAP-средства. Поэтому провести четкую грань между OLAP-системами и системами класса Query & Reporting практически невозможно. Пример – продукты компании MicroStrategy, которые различные аналитики и издания с равной регулярностью относят к продуктам обоих классов.

**4. Концепции построения ИАС**

Современный уровень развития аппаратных и программных средств с некоторых пор сделал возможным повсеместное ведение баз данных оперативной информации на разных уровнях управления. В процессе своей деятельности промышленные предприятия, корпорации, ведомственные структуры, органы государственной власти и управления накопили большие объемы данных. Они хранят в себе большие потенциальные возможности по извлечению полезной аналитической информации, на основе которой можно выявлять скрытые тенденции, строить стратегию развития, находить новые решения.

В последние годы в мире оформился ряд новых концепций хранения и анализа корпоративных данных:

1)     Хранилища данных, или Склады данных (Data Warehouse);

2)     Оперативная аналитическая обработка (On-Line Analytical Processing, OLAP);

3)     Интеллектуальный анализ данных - ИАД (Data Mining).

Технологии OLAP тесно связаны с технологиями построения Data Warehouse и методами интеллектуальной обработки - Data Mining. Поэтому наилучшим вариантом является комплексный подход к их внедрению.

Для того чтобы существующие хранилища данных способствовали принятию управленческих решений, информация должна быть представлена аналитику в нужной форме, то есть он должен иметь развитые инструменты доступа к данным хранилища и их обработки.

Очень часто информационно-аналитические системы, создаваемые в расчете на непосредственное использование лицами, принимающими решения, оказываются чрезвычайно просты в применении, но жестко ограничены в функциональности. Такие статические системы называются в литературе Информационными системами руководителя (ИСР), или Executive Information Systems (EIS). Они содержат в себе предопределенные множества запросов и, будучи достаточными для повседневного обзора, неспособны ответить на все вопросы к имеющимся данным, которые могут возникнуть при принятии решений. Результатом работы такой системы, как правило, являются многостраничные отчеты, после тщательного изучения которых у аналитика появляется новая серия вопросов. Однако каждый новый запрос, непредусмотренный при проектировании такой системы, должен быть сначала формально описан, закодирован программистом и только затем выполнен. Время ожидания в таком случае может составлять часы и дни, что не всегда приемлемо. Таким образом, внешняя простота статических ИСР, за которую активно борется большинство заказчиков информационно-аналитических систем, оборачивается катастрофической потерей гибкости.

Динамические ИАС, напротив, ориентированы на обработку нерегламентированных запросов аналитиков к данным. Работа аналитиков с этими системами заключается в интерактивной последовательности формирования запросов и изучения их результатов.

Но динамические ИАС могут действовать не только в области оперативной аналитической обработки (OLAP); поддержка принятия управленческих решений на основе накопленных данных может выполняться в трех базовых сферах:

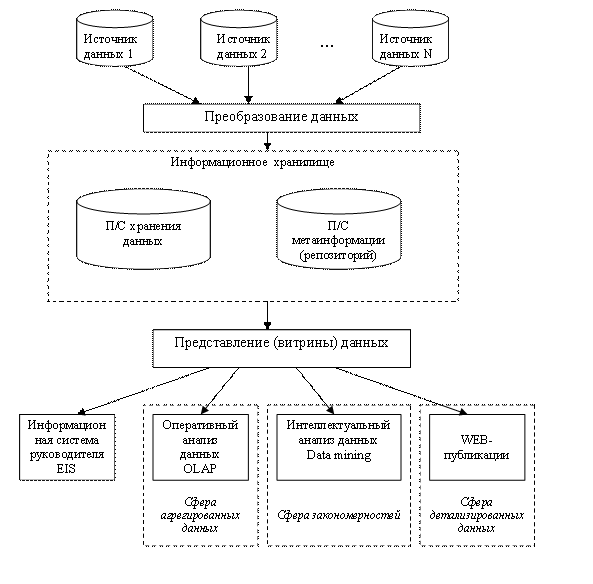
1. Сфера детализированных данных. Это область действия большинства систем, нацеленных на поиск информации. В большинстве случаев реляционные СУБД отлично справляются с возникающими здесь задачами. Общепризнанным стандартом языка манипулирования реляционными данными является SQL. Информационно-поисковые системы, обеспечивающие интерфейс конечного пользователя в задачах поиска детализированной информации, могут использоваться в качестве надстроек как над отдельными базами данных транзакционных систем, так и над общим хранилищем данных.

2. Сфера агрегированных показателей. Комплексный взгляд на собранную в хранилище данных информацию, ее обобщение и агрегация, гиперкубическое представление и многомерный анализ являются задачами систем оперативной аналитической обработки данных (OLAP). Здесь можно или ориентироваться на специальные многомерные СУБД, или оставаться в рамках реляционных технологий. Во втором случае заранее агрегированные данные могут собираться в БД звездообразного вида, либо агрегация информации может производиться на лету в процессе сканирования детализированных таблиц реляционной БД.

3. Сфера закономерностей. Интеллектуальная обработка производится методами интеллектуального анализа данных (ИАД, Data Mining), главными задачами которых являются поиск функциональных и логических закономерностей в накопленной информации, построение моделей и правил, которые объясняют найденные аномалии и/или прогнозируют развитие некоторых процессов.

**5. Общая структура информационной аналитической системы**

Полная структура информационно-аналитической системы, построенной на основе хранилища данных, показана на рисунке 1. В конкретных реализациях отдельные компоненты этой схемы часто отсутствуют.

****

**Рисунок 1.** Структура информационной аналитической  системы (ИАС)

Рассмотрим состав основных подсистем.

**Подсистема хранения данных**

Многомерное хранилище данных может быть организовано в виде одной из следующих структур:

1.     физической структуры, называемой MOLAP, в которую с определенной периодичностью загружаются данные из файлов – источников, принадлежащих базам оперативных данных

2.     виртуальной структуры, называемой ROLAP, которая динамически используется при запросах. ROLAP – система рассматривается просто как надстройка над реляционными базами данных, обеспечивающая удобный интерфейс пользователя. Типичными инструментальными средствами, поддерживающими ROLAP, является Business Objects.

3.     гибридной структуры, называемой НOLAP, которая используется при построении многоуровневых информационных хранилищ, применяемых на разных уровнях управления больших корпораций.

Анализ параметров использования ROLAP и MOLAP информационных хранилищ показывает, что внедрение и эксплуатация ROLAP - систем является более простым и дешевым по сравнению с MOLAP – системами, но уступают последним в эффективности оперативного анализа данных.

**Подсистема метаинформации**

Репозиторий представляет собой описание структуры информационного хранилища: состава показателей, иерархии агрегаций измерений, форматов данных, используемых функций, физического размещения на сервере, прав доступа пользователей, частоты обновления.

В репозитории задается  схема отображения структуры файлов-источников данных на структуре ИХ, а также схема отображения структуры ИХ на витринах данных. Через репозиторий осуществляется интерпретация запросов к ИХ на проведение оперативного анализа данных.

**Подсистема преобразования данных (загрузки хранилища)**

Подсистема загрузки ИХ создается только для MOLAP – систем. Для ROLAP – систем в процессе выполнения запросов осуществляется преобразование данных из файлов – источников. В том и другом случаях требуется выполнение следующих основных функций:

  сбор данных

  очистка данных

  агрегирование данных

Сбор данных предполагает передачу данных из источников в ИХ в соответствии со схемой отображения, представленной в репозитории.

В процессе очистки данных осуществляется проверка целостности, исключение дублирования данных, отбраковка случайных данных, восстановление отсутствующих  данных, приведение данных к единому формату.

В случае необходимости агрегирования данных осуществляется суммирование итогов по заданным в репозитории признакам.

**Подсистема представления данных (организация витрин данных)**

Под витриной данных понимается предметно-ориентированное хранилище данных, как правило, агрегированной информации, предназначенное для использования группой пользователей в рамках конкретного вида деятельности предприятия, например маркетинга и т.д.

Как правило, витрины данных являются подмножествами общего хранилища данных компании, которое служит для них источником. Обычно общее ИХ и витрины данных разрабатываются параллельно.

**Подсистема оперативного анализа данных**

Подсистема оперативного анализа данных, как правило, используется лицами, подготавливающими информацию для принятия решений, путем выполнения различных статистических группировок исходных данных.

В рамках пользовательского интерфейса для оперативного анализа данных используются следующие базовые операции:

  **Поворот.** Добавление нового признака анализа.

  **Проекция**. Выборка подмножества по задаваемой совокупности измерений. При этом значения, лежащие на оси проекции, суммируются.

  **Раскрытие.** Осуществляется декомпозиция признака агрегации на компоненты, например, признак года разбивается на кварталы. При этом автоматически детализуются числовые показатели.

  **Свертка**. Операция обратная раскрытию. При этом значения детальных показателей суммируются в агрегируемый показатель.

  **Сечение или срез.** Выделение подмножества данных по конкретным значениям одного или нескольких измерений.

**Подсистема интеллектуального анализа данных**

Подсистема интеллектуального анализа данных используется специальной категорией пользователей – аналитиков, которые на основе ИХ обнаруживают закономерности в деятельности предприятия и на рынке, используемые в дальнейшем для обоснования  стратегических и тактических решений. Интеллектуальный анализ требует более сложных методов анализа по сравнению со  статическими группировками и выполняется путем проведения множества сеансов.

Типичными задачами интеллектуального анализа данных являются:

  Установление корреляций, причинно-следственных связей и временных связей событий, например определение местоположения прибыльных предприятий.

  Классификация ситуаций,  позволяющая обобщать конкретные события в классы, например определение типичного профиля покупателя конкретных видов продукции.

  Прогнозирование развития ситуаций, например прогнозирование цен, объемов продаж, производства.

К основным методам интеллектуального анализа данных относятся:

       Методы многомерного статистического анализа,

       Индуктивные методы построения деревьев решений,

       Нейронные сети.

**Подсистема «Информационная система руководителя»**

Информационная система руководителя предназначена для лиц, непосредственно принимающих решения. Поэтому интерфейс таких систем должен быть в наибольшей степени упрощенным. Обычно в качестве интерфейса руководителям предприятия предлагается набор стандартных отчетов и графиков, настраиваемых на потребности руководителя через систему меню. Часто в качестве интерфейса предлагаются  диаграммы Ишикава, представляющие  собой саморазворачивающееся  дерево показателей, в котором листья ветвей раскрашиваются в разные цвета, символизирующие характер состояния показателя (нормальный, тревожный, кризисный). Лист любой ветви дерева может быть развернут а таблицу значений показателя или график.

**Подсистема WEB – публикации**

Подсистема WEB – публикации предполагает преобразование полученной из ИХ информации в HTML – вид, доступный для ее просмотра удаленными клиентами с помощью браузеров Интернета.

**Тема 4. Хранилища данных**

**Содержание**

1.     Пространственная интерпретация данных

2.     Понятие хранилища данных

3.     Структура хранилищ данных

4.     Вопросы реализации Хранилищ Данных

5.     Хранилище данных предприятия

**1. Пространственная интерпретация данных**

Программные инструментальные средства, обеспечивающие автоматизацию аналитических работ в целях поддержки принятия решений, в литературе получили два распространенных названия: ОLАР - системы и информационные Хранилища.

Как правило, все инструментальные средства, предназначенные для автоматизации аналитических работ, приспособлены для обработки многомерных   массивов   информации, для хранения которых используются многомерные базы данных.

Информационное пространство, отображающее функционирование объекта, многомерно. Естественно стремление аналитика и ЛПР к тому, чтобы иметь дело с моделью данных в наиболее естественном виде. Это обстоятельство привело к тому, что с помощью современных программно-технических средств, имеющих широкие возможности  интерпретации данных, были созданы соответствующие многомерные модели.

В последнее десятилетие XX века основной моделью данных использованной в многочисленных инструментальных средствах создания и поддержки баз данных - СУБД была реляционная модель. Данные в ней представлены в виде множества связанных ключевыми полями двумерных таблиц – отношений (табл. 1).

Таблица 1

**Реляционная модель представления данных**

|  |  |  |
| --- | --- | --- |
| **Модель** | **Месяц** | **Объем** |
| "Жигули" | Июнь | 12 |
| "Жигули" | Июль | 24 |
| "Жигули" | Август | 5 |
| "Москвич" | Июнь | 2 |
| "Москвич" | Июль | 18 |
| «Волга» | Июль | 19 |

А теперь представим, что у нас не три модели, а 30 и не три, а 12 различных месяцев. В случае построчного (реляционного) представления мы получим отчет в 360 строк (30х12), который займет не менее 5-6 страниц. В случае же многомерного (в нашем случае двухмерного) представления мы получим достаточно компактную таблицу 12 на 30, которая вполне уместится на одной странице и которую, даже при таком объеме данных, можно реально оценивать и анализировать (табл. 2).

Таблица 2

**Многомерная модель представления данных**

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Июнь** | **Июль** | **Август** |
| **"Жигули"** | 12 | 24 | 5 |
| **"Москвич"** | 2 | 18 | No |
| **"Волга"** | No | 19 | No |

И когда говорится о многомерной организации данных, вовсе не подразумевается то, что данные представляются конечному пользователю (визуализируются) в виде четырех или пятимерных гиперкубов. Это невозможно, да и пользователю более привычно и комфортно иметь дело с двухмерным табличным представлением и двухмерной бизнес - графикой.

Многомерная модель данных представляет исследуемый объект в виде многомерного куба, чаще используют трехмерную модель.

По осям или граням куба откладываются измерения или реквизиты - признаки. Реквизиты - основания являются наполнением ячеек куба.

**Многомерное представление при описании структур данных**

Основными понятиями, с которыми оперирует пользователь и проектировщик в многомерной модели данных, являются:

  **измерение**(Dimension);

  **ячейка**(Cell). Иногда вместо термина "Ячейка" используется термин "Показатель" (Measure).

**Измерение** - это множество однотипных данных, образующих одну из граней гиперкуба. Например - Дни, Месяцы, Кварталы, Годы - это наиболее часто используемые в анализе временные Измерения. Примерами географических измерений являются: Города, Районы, Регионы, Страны и т.д.

В многомерной модели данных Измерения играют роль индексов, используемых для идентификации конкретных значений (Показателей), находящихся в Ячейках гиперкуба.

В свою очередь, **Показатель -** это поле (обычно цифровое), значения которого однозначно определяются фиксированным набором Измерений. В  зависимости от того, как формируются его значения, Показатель может быть определен, как:

  Переменная (Variable) - значения таких Показателей один раз вводятся из какого-либо внешнего источника или формируются программно и затем в явном виде хранятся в многомерной базе данных (МБД);

  Формула (Formula) - значения таких Показателей вычисляются по некоторой заранее специфицированной формуле.

То есть для Показателя, имеющего тип Формула, в БД хранится не его значения, а формула, по которой эти значения могут быть вычислены.

Заметим, что это различие существует только на этапе проектирования и полностью скрыто от конечных пользователей.

В примере каждое значение поля Объем продаж однозначно определяется комбинацией полей:

*Модель автомобиля;*

*Месяц продаж.*

Но в реальной ситуации для однозначной идентификации значения Показателя, скорее всего, потребуется большее число измерений, например:

*Модель автомобиля;*

*Менеджер;*

*Время (например, Год).*

**Измерения:**

*Модель автомобиля – Жигули, Москвич, Волга*

*Время (Год) - 1994, 1995, 1995*

*Менеджер - Петров, Смирнов, Яковлев*

**Показатель:**

*Объем Продаж*

И в терминах многомерной модели речь будет идти уже не о двухмерной таблице, а о трехмерном гиперкубе:

o первое Измерение - Модель автомобиля;

o второе Измерение - Менеджер, продавший автомобиль;

o третье Измерение - Время (Год);

на пересечении граней которого находятся значения Показателя Объем продаж.

Заметим, что, в отличие от Измерений, не все значения Показателей должны иметь и имеют реальные значения. Например, Менеджер Петров в 1994 г. мог еще не работать в фирме, и в этом случае все значения Показателя Объем продаж за этот год будут иметь неопределенные значения.

**Гиперкубические и поликубические модели данных**

В различных МСУБД используются два основных варианта организации данных:

        Гиперкубическая модель;

        Поликубическая модель.

В чем состоит разница? Системы, поддерживающие Поликубическую модель предполагают, что в МБД может быть определено несколько гиперкубов с различной размерностью и с различными Измерениями в качестве их граней. Например, значение Показателя Рабочее Время Менеджера, скорее всего, не зависит от Измерения Модель Автомобиля и однозначно определяется двумя Измерениями: День и Менеджер. В Поликубической модели в этом случае может быть объявлено два различных гиперкуба:

Двухмерный - для Показателя Рабочее Время Менеджера;

Трехмерный - для Показателя Объем Продаж.

В случае же Гиперкубической модели предполагается, что все Показатели должны определяться одним и тем же набором Измерений. То есть только из-за того, что Объем Продаж определяется тремя Измерениями, при описании Показателя Рабочее Время Менеджера придется также использовать три Измерения и вводить избыточное для этого Показателя Измерение Модель Автомобиля.

**Методы извлечения информации из кубов данных**

Для извлечения информации из кубов данных используются различные операции манипулирования Измерениями:

1)     **Формирование "Среза".**

Пользователя редко интересуют все потенциально возможные комбинации значений Измерений. Более того, он практически никогда не работает одновременно сразу со всем гиперкубом данных. Подмножество гиперкуба, получившееся в результате фиксации значения одного или более Измерений, называется Срезом (Slice). Например, если мы ограничим значение Измерения Модель Автомобиля = "ВАЗ2108", то получим подмножество гиперкуба (в нашем случае - двухмерную таблицу), содержащее информацию об истории продаж этой модели различными менеджерами в различные годы.

**2)    Операция "Вращение".**

Изменение порядка представления (визуализации) Измерений (обычно применяется при двухмерном представлении данных) называется Вращением (Rotate). Эта операция обеспечивает возможность визуализации данных в форме, наиболее комфортной для их восприятия. Например, если менеджер первоначально вывел отчет, в котором Модели автомобилей были перечислены по оси X, а Менеджеры по оси Y, он может решить, что такое представление мало наглядно, и поменять местами координаты (выполнить Вращение на 90 градусов).

**3)    Отношения и Иерархические Отношения.**

В нашем примере значения Показателей определяются только тремя измерениями. На самом деле их может быть гораздо больше и между их значениями обычно существуют множество различных Отношений (Relation) типа "один ко многим".

Например, каждый Менеджер может работать только в одном подразделении, а каждой модели автомобиля однозначно соответствует фирма, которая ее выпускает:

*Менеджер ->Подразделение;*

*Модель Автомобиля ->Фирма-Производитель.*

Заметим, что для Измерений, имеющих тип Время (таких как День, Месяц, Квартал, Год), все Отношения устанавливаются автоматически, и их не требуется описывать.

В свою очередь, множество Отношений может иметь иерархическую структуру - Иерархические Отношения (Hierarchical Relationships). Вот только несколько примеров таких Иерархических Отношений:

*День -> Месяц -> Квартал -> Год;*

*Менеджер -> Подразделение -> Регион -> Фирма -> Страна;*

*Модель Автомобиля -> Завод-Производитель -> Страна.*

И часто более удобно не объявлять новые Измерения и затем устанавливать между ними множество Отношений, а использовать механизм Иерархических Отношений. В этом случае все потенциально возможные значения из различных Измерений объединяются в одно множество. Например, мы можем добавить к множеству значений Измерения Менеджер ("Петров", "Сидоров", "Иванов", "Смирнов"), значения Измерения Подразделение ("Филиал 1", "Филиал 2", "Филиал 3") и Измерения Регион ("Восток", "Запад") и затем определить между этими значениями Отношение Иерархии.

4)     **Операция Агрегации**.

С точки зрения пользователя, Подразделение, Регион, Фирма, Страна являются точно такими же Измерениями, как и Менеджер. Но каждое из них соответствует новому, более высокому уровню агрегации значений Показателя Объем продаж. В процессе анализа пользователь не только работает с различными Срезами данных и выполняет их Вращение, но и переходит от детализированных данных к агрегированным, т.е. производит операцию Агрегации (Drill Up). Например, посмотрев, насколько успешно в 2004 г. Петров продавал модели "Жигули" и "Волга", управляющий может захотеть узнать, как выглядит соотношение продаж этих моделей на уровне Подразделения, где Петров работает. А затем получить аналогичную справку по Региону или Фирме.

**5)    Операция Детализации.**

Переход от более агрегированных к более детализированным данным называется операцией Детализации (Drill Down). Например, начав анализ на уровне Региона, пользователь может захотеть получить более точную информацию о работе конкретного Подразделения или Менеджера.

**2. Понятие хранилища данных**

Термин "OLAP" неразрывно связан с термином "хранилище данных" (Data Warehouse).

Приведем определение, сформулированное "отцом-основателем" хранилищ данных Биллом Инмоном: "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.

Зачем строить хранилища данных - ведь они содержат заведомо избыточную информацию, которая и так "живет" в базах или файлах оперативных систем? Ответить можно кратко: анализировать данные оперативных систем напрямую невозможно или очень затруднительно. Это объясняется различными причинами, в том числе разрозненностью данных, хранением их в форматах различных СУБД и в разных "уголках" корпоративной сети. Но даже если на предприятии все данные хранятся на центральном сервере БД (что бывает крайне редко), аналитик почти наверняка не разберется в их сложных, подчас запутанных структурах.

Таким образом, задача хранилища - предоставить "сырье" для анализа в одном месте и в простой, понятной структуре.

Есть и еще одна причина, оправдывающая появление отдельного хранилища - сложные аналитические запросы к оперативной информации тормозят текущую работу компании, надолго блокируя таблицы и захватывая ресурсы сервера.

Под хранилищем можно понимать не обязательно гигантское скопление данных - главное, чтобы оно было удобно для анализа. Вообще говоря, для маленьких хранилищ предназначается отдельный термин - Data Marts (киоски или витрины данных).

В основе концепции Хранилищ Данных лежат две основополагающие идеи:

   Интеграция ранее разъединенных **детализированных** данных:

      исторические архивы,

      данные из традиционных СОД,

      данные из внешних источников

в едином Хранилище Данных, их согласование и возможно агрегация.

   Разделение наборов данных используемых для операционной обработки и наборов данных используемых для решения задач анализа.

Предметом концепции Хранилищ Данных являются сами данные. То есть, её предметом являются не способы описания и отображения объектов предметной области, а собственно данные, как самостоятельный объект предметной области, порожденной в результате функционирования ранее созданных информационных систем. Основные требования к данным в хранилищах  приведены в таблице 2.

***Таблица 2.***

**Основные требования к данным в Хранилище Данных**

|  |  |
| --- | --- |
| Предметная ориентированность | Все данные о некотором предмете (бизнес объекте) собираются (обычно из множества различных источников), очищаются, согласовываются, дополняются, агрегируются и представляются в единой, удобной для их использования в бизнес анализе форме. |
| Интегрированность | Все данные о разных бизнес объектах, взаимно согласованы и хранятся в едином общекорпоративном Хранилище. |
| Неизменчивость | Исходные (исторические) данные, после того как они были согласованы, верифицированы и внесены в общекорпоративное Хранилище, остаются неизменными и используются исключительно в режиме чтения. |
| Поддержка хронологии | Данные хронологически структурированы и отражают историю, за достаточный, для выполнения задач бизнес анализа и прогнозирования, период времени. |

Для правильного понимания данной концепции необходимо понимание следующих принципиальных моментов: Концепция Хранилищ Данных - это концепция подготовки данных для анализа.

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

   Концепция Хранилищ Данных предполагает не просто единый логический взгляд на данные организации. Она предполагает реализацию единого интегрированного источника данных.

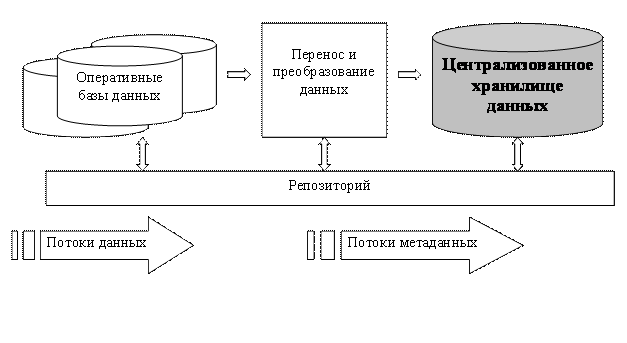
Без поддержки хронологии (наличия исторических данных) нельзя говорить о решении задач прогнозирования и анализа тенденций. Но наиболее критичными и болезненными, оказываются вопросы, связанные с согласованием данных.

Основным требованием аналитика, является даже не столько оперативность, сколько достоверность ответа. Но достоверность, в конечном счете, и определяется согласованностью. Пока не проведена работа по взаимному согласованию значений данных из различных источников, сложно говорить об их достоверности.

Реализация ИХ может быть осуществлена несколькими способами:

**Централизованное хранилище данных**

Такой подход означает, что при нескольких источниках информации - операционных базах данных создаётся единое централизованное  хранилище (рис. 1).

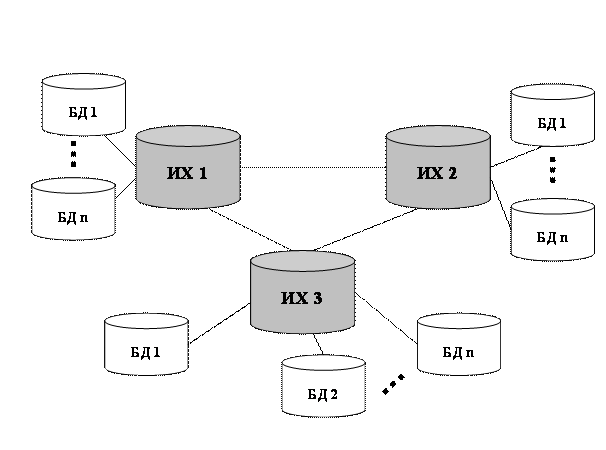


**Рисунок 1.** Единое централизованное  хранилище

Вся поступающая в ИХ информация должна быть преобразована в принятую в данном**ИХ** структуру. Передача данных из операционных БД в ИХ, которая сопровождается доработкой, может быть организована по заданному временному графику и правилам доработки

**Распределенное хранилище данных**

Возможен и имеет место противоположный подход к хранению данных на основе распределения функций ИХ по местам их возникновения или группировки нескольких операционных БД вокруг локального или регионального информационного хранилища. Эти хранилища могут быть ориентированы на определённую предметную область или на регион в корпоративных структурах. Система локальных хранилищ действует в качестве распределённого хранилища (рис. 2).



**Рисунок 2.** Распределенное хранилище данных

Не исключается и наличие центрального хранилища, но в такой структуре требования к его размерности значительно облегчаются.

**Автономные витрины данных**

Одним из вариантов организации централизованного хранения и представления информации является концепция витрин данных. При таком подходе информация, относящаяся к крупной предметной области – например, информационному пространству крупной корпоративной системы, имеющей несколько достаточно самостоятельных направлений деятельности, группируется по этим направлениям в специально организованных базах данных, которые называют витринами данных. Этот подход является развитием концепции распределенного ИХ в части придания функций предметной  ориентированности некоторым локальным ИХ.

Такой подход позволяет обойтись сравнительно менее ресурсоемкими аппаратными и программными средствами, обеспечивает повышение адаптируемости системы к изменяющимся условиям, расширяет доступность для внедрения. Пользователь предприятия или другого подразделения корпорации получает своё ИХ, обслуживающее местные потребности.

**Единое интегрированное хранилище и много витрин данных**

Эта структура ИХ объединяет две концепции: единого интегрированного хранилища и связанных с ним и получающих из него информацию витрин данных. В таком варианте имеется крупное информационное  хранилище агрегированной  и  подработанной информации, которое может удовлетворить потенциальные запросы по отдельным направлениям деятельности.

Здесь очевидны преимущества: данные заранее агрегируются, обеспечивается единая хронология, согласованы различные форматы, устраняются противоречивость и неоднозначность данных - информация приобретает необходимую кондицию для быстрого и достаточно полного удовлетворения необходимого множества запросов,

Недостатком     является     необходимость     применения высокопроизводительных аппаратных средств и специализированных многомерных или гибридных программных инструментальныхсредств.

**3. Структура хранилищ данных**

ИХ представляет собой базу  обобщенной информации, формируемую из множества внешних и внутренних источников, на основе которой выполняются статистические группировки и интеллектуальный анализ данных.

В основе ИХ лежит понятие многомерного информационного пространства или гиперкуба, в ячейках которого хранятся анализируемые числовые показатели (например: объемы продаж, инвестиций, оборота и др.) Измерениями (осями) гиперкуба являются признаки анализа (например: время, группа продукции, регион и др.) При хранении признаки анализа отделяются от фактических данных.

Основными составляющими структуры хранилищ данных являются таблица фактов (fact table) и таблицы измерений (dimension tables).

**Таблица фактов**

Таблица фактов является основной таблицей хранилища данных. Как правило, она содержит сведения об объектах или событиях, совокупность которых будет в дальнейшем анализироваться. Обычно говорят о четырех наиболее часто встречающихся типах фактов. К ним относятся:

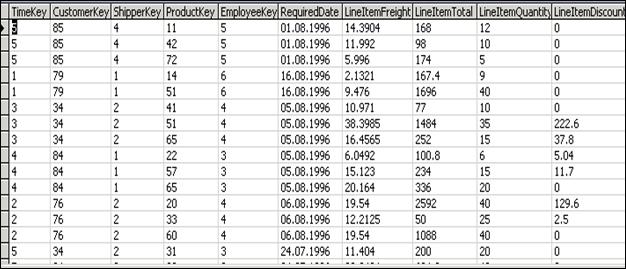
   факты, связанные с транзакциями (Transaction facts). Они основаны на отдельных событиях (типичными примерами которых являются телефонный звонок или снятие денег со счета с помощью банкомата);

   факты, связанные с «моментальными снимками» (Snapshot facts). Основаны на состоянии объекта (например, банковского счета) в определенные моменты времени, например на конец дня или месяца. Типичными примерами таких фактов являются объем продаж за день или дневная выручка;

   факты, связанные с элементами документа (Line-item facts). Основаны на том или ином документе (например, счете за товар или услуги) и содержат подробную информацию об элементах этого документа (например, количестве, цене, проценте скидки);

   факты, связанные с событиями или состоянием объекта (Event or state facts). Представляют возникновение события без подробностей о нем (например, просто факт продажи или факт отсутствия таковой без иных подробностей).

Таблица фактов, как правило, содержит уникальный составной ключ, объединяющий первичные ключи таблиц измерений. Чаще всего это целочисленные значения либо значения типа «дата/время» — ведь таблица фактов может содержать сотни тысяч или даже миллионы записей, и хранить в ней повторяющиеся текстовые описания, как правило, невыгодно — лучше поместить их в меньшие по объему таблицы измерений. При этом как ключевые, так и некоторые неключевые поля должны соответствовать будущим измерениям OLAP-куба. Помимо этого таблица фактов содержит одно или несколько числовых полей, на основании которых в дальнейшем будут получены агрегатные данные. Пример таблицы фактов, которая может быть построена на основе базы данных Northwind, приведен на рис. 3.



**Рисунок 3**. Пример таблицы фактов

  В данном примере измерениям будущего куба соответствуют первые шесть полей, а агрегатным данным — последние четыре.

Отметим, что для многомерного анализа пригодны таблицы фактов, содержащие как можно более подробные данные (то есть соответствующие членам нижних уровней иерархии соответствующих измерений). В данном случае предпочтительнее взять за основу факты продажи товаров отдельным заказчикам, а не суммы продаж для разных стран — последние все равно будут вычислены OLAP-средством. Исключение можно сделать, пожалуй, только для клиентских OLAP-средств, поскольку в силу ряда ограничений они не могут манипулировать большими объемами данных.

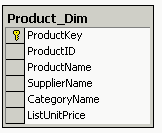
Отметим, что в таблице фактов нет никаких сведений о том, как группировать записи при вычислении агрегатных данных. Например, в ней есть идентификаторы продуктов или клиентов, но отсутствует информация о том, к какой категории относится данный продукт или в каком городе находится данный клиент. Эти сведения, в дальнейшем используемые для построения иерархий в измерениях куба, содержатся в таблицах измерений.

**Таблицы измерений**

Таблицы измерений содержат неизменяемые либо редко изменяемые данные. В подавляющем большинстве случаев эти данные представляют собой по одной записи для каждого члена нижнего уровня иерархии в измерении. Таблицы измерений также содержат как минимум одно описательное поле (обычно с именем члена измерения) и, как правило, целочисленное ключевое поле для однозначной идентификации члена измерения. Если будущее измерение, основанное на данной таблице измерений, содержит иерархию, то таблица измерений также может содержать поля, указывающие на «родителя» данного члена в этой иерархии. Нередко (но не всегда) таблица измерений может содержать и поля, указывающие на «прародителей», и иных «предков» в данной иерархии (это обычно характерно для сбалансированных иерархий), а также дополнительные атрибуты членов измерений, содержавшиеся в исходной оперативной базе данных (например, адреса и телефоны клиентов).

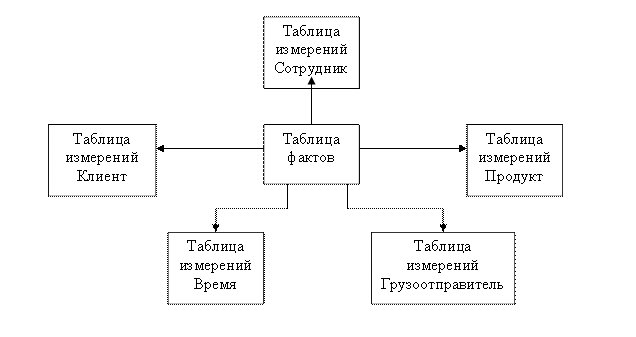
Каждая таблица измерений должна находиться в отношении «один ко многим» с таблицей фактов.

Отметим, что скорость роста таблиц измерений должна быть незначительной по сравнению со скоростью роста таблицы фактов; например, добавление новой записи в таблицу измерений, характеризующую товары, производится только при появлении нового товара, не продававшегося ранее. Структура таблицы измерений приведена на рисунке 4.



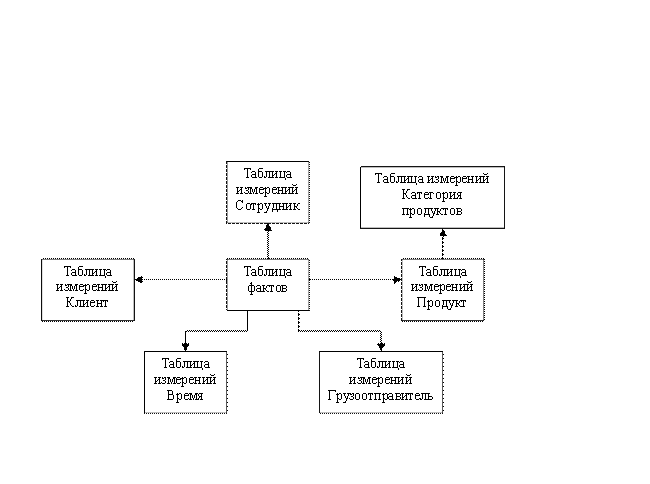
**Рисунок 4.** Таблица измерений

Одно измерение куба может содержаться как в одной таблице (в том числе и при наличии нескольких уровней иерархии), так и в нескольких связанных таблицах, соответствующих различным уровням иерархии в измерении. Если каждое измерение содержится в одной таблице, такая схема хранилища данных носит название «звезда» (star schema). Пример такой схемы приведен на рис. 5.



**Рисунок 5.**Пример схемы «звезда»

Если же хотя бы одно измерение содержится в нескольких связанных таблицах, такая схема хранилища данных носит название «снежинка» (snowflake schema). Дополнительные таблицы измерений в такой схеме, обычно соответствующие верхним уровням иерархии измерения,  находятся в соотношении «один ко многим» с главной таблицей измерений, соответствующей нижнему уровню иерархии. Пример схемы «снежинка» приведен на рис.6.

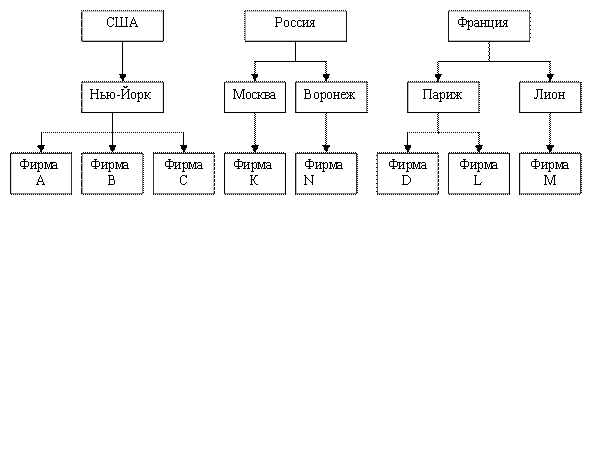
****

**Рисунок 6.**  Пример схемы «снежинка»

Отметим, что даже при наличии иерархических измерений с целью повышения скорости выполнения запросов к хранилищу данных нередко предпочтение отдается схеме «звезда».

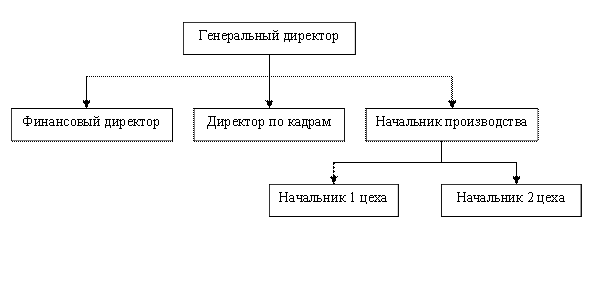
Говоря об измерениях, следует упомянуть о том, что значения, могут иметь различные уровни детализации. Например, нас может интересовать суммарная стоимость заказов, сделанных клиентами в разных странах, либо суммарная стоимость заказов, сделанных иногородними клиентами или даже отдельными клиентами. Естественно, результирующий набор агрегатных данных во втором и третьем случаях будет более детальным, чем в первом. Заметим, что возможность получения агрегатных данных с различной степенью детализации соответствует одному из требований, предъявляемых к хранилищам данных, — требованию доступности различных срезов данных для сравнения и анализа.

Поскольку в рассмотренном примере в общем случае в каждой стране может быть несколько городов, а в городе — несколько клиентов, можно говорить об иерархиях значений в измерениях. В этом случае на первом уровне иерархии располагаются страны, на втором — города, а на третьем — клиенты рис. 7.



**Рисунок 7.** Иерархия в измерении, связанная с географическим положением клиентов

Отметим, что иерархии могут быть сбалансированными (balanced), как, например, иерархия, представленная на рис. 3, а также иерархии, основанные на данных типа "дата—время", и несбалансированными (unbalanced). Типичный пример несбалансированной иерархии — иерархия типа "начальник—подчиненный" (рис. 8).



**Рисунок 8.**  Несбалансированная иерархия

Существуют также иерархии, занимающие промежуточное положение между сбалансированными и несбалансированными (они обозначаются термином ragged — "неровный"). Обычно они содержат такие члены, логические "родители" которых находятся не на непосредственно вышестоящем уровне. Например, в географической иерархии есть уровни Страна, Города и Штаты, но при этом в наборе данных имеются страны, не имеющие штатов или регионов между уровнями Страна и Города.

**4. Вопросы реализации Хранилищ Данных**

Аналитические системы предъявляют высокие требования к аппаратному и программному обеспечению. И, приступая к построению аналитической системы, следует понимать, что её реализация практически невозможна без разрешения таких вопросов как:

   Неоднородность программной среды.

   Распределенность.

   Защиты данных от несанкционированного доступа.

   Построения и ведения многоуровневых справочников метаданных.

   Эффективное хранение и обработка очень больших объемов данных.

**Неоднородность программной среды.**

Основой Хранилищ Данных являются не внутренние, а внешние источники данных: различного рода информационные системы, электронные архивы, общедоступные и коммерческие электронные каталоги, справочники, статистические сборники. Как правило, сегодня в любой организации реально функционирует множество несвязанных или слабо связанных систем обработки данных. В большинстве случаев, они создавались в различное время, различными коллективами разработчиков и реализованы на основе различных программных и аппаратных средств. Таким образом, сама основа, на которой будет строиться Хранилище Данных, чаще всего уже является крайне неоднородной. Добавьте сюда средства выгрузки, транспортировки, реализации целевой БД Хранилища Данных.

Очевидно, что в таких условиях, даже говорить об однородности программных средств чрезвычайно сложно. И практически всегда, задача построения Хранилища Данных, это задача построения единой согласовано функционирующей информационной системы, на основе неоднородных программных средств и решений. И уже сам выбор средств реализации Хранилища Данных становится чрезвычайно сложной задачей. Здесь должно учитываться множество факторов, включая, взаимную совместимость различных программных компонент, легкость их освоения и использования, эффективность функционирования, стабильность и даже формы, уровень и потенциальную перспективность взаимоотношений различных фирм производителей.

**Распределенность.**

Хранилища Данных уже по своей природе являются распределенным решением.

В основе концепции Хранилищ Данных, лежит физическое разделение узлов, в которых выполняется операционная обработка, от узлов в которых выполняется анализ данных. И хотя, при реализации такой системы, нет необходимости в строгой синхронизации данных в различных узлах, средства асинхронной асимметричной репликации данных являются неотъемлемой частью практически любого решения.

**Метаданные**

Наличие метаданных и средств их представления конечным пользователям является одним из основополагающих факторов успешной реализации Хранилища Данных. Более того, без наличия актуальных, максимально полных и легко понимаемых пользователем описаний данных, Хранилище Данных превращается в обычный, но очень дорогостоящий электронный архив.

Первой же задачей, с которой сталкиваешься при проектировании и реализации системы Хранилищ Данных, является необходимость одновременной работы с самыми разнородными внешними источниками данных, несогласованностью их структур и форматов, масштабами и количеством архивов, которые должны быть переработаны и загружены. И при построении такой системы, разработчику сложно обойтись без высокоуровневых средств описания информационной модели системы. Причем, эта модель должна содержать описания не только целевых структур данных в БД Хранилища, но и структур данных в источниках их получения (различных информационных системах, архивах, электронных справочниках и т.д.), правила, процедуры и периодичность их выборки и выгрузки, процедуры и места согласования и агрегации.

Здесь следует сделать несколько замечаний относительно выбора конкретных средств проектирования. Как уже было сказано выше, характерными свойствами аналитической системы, является:

o Разнородность компонент.

o Ориентированность на нерегламентированную работу с данными.

Рассмотрим, как это влияет на выбор и требования к средствам проектирования. С одной стороны, из-за разнородности программных и системных компонент образующих Хранилища и малой доли регламентированных пользовательских приложений, чаще всего результатом проектирования системы будет не готовый к исполнению программный продукт, а база метаданных, содержащая всестороннее многоуровневое описание целевой информационной системы. С другой стороны в аналитических системах, именно вопросы полноты, актуальности, простоты использования и понимания метаданных приобретают особую актуальность.

**Роль метаданных в системах Хранилищ Данных**

В случае информационных систем ориентированных на аналитическую работу с данными (таблица 4)  наличие метаданных и средств их представления конечным пользователям является одним из основополагающих факторов успешной реализации системы. Для конечного пользователя, база метаданных является тем же самым, что и путеводитель для туриста, попавшего в незнакомый город. Прежде чем сформулировать свой вопрос к системе, менеджер должен понять, какая информация в ней есть, её актуальность, насколько ей можно доверять и даже, сколько времени может занять формирование ответа. Поэтому, для конечного пользователя крайне важно и желательно, чтобы в системе содержались не только описания собственно структур данных, их взаимосвязей, предвычисленных уровней агрегации, но, и:

   ***Источников получения данных***. Аналитику желательно не просто знать о том, какие данные есть в системе, но и источники их получения, и степень их достоверности. Например, одна и та же информация может попасть в Хранилище Данных из различных источников. В этом случае, пользователь должен иметь возможность узнать какой источник выбран в качестве основного и каким образом выполнялась согласование и очистка исходных данных.

   ***Периодичности обновления***. Пользователю желательно не просто знать, какому моменту времени соответствуют те или иные данные, но и когда они будут обновлены.

   ***Собственников данных***. Пользователю будет полезно знать, какие еще данные есть в системе, кто является их собственником и какие шаги он должен предпринять, чтобы получить к ним доступ.

   ***Статистические оценки запросов***. Еще до выполнения запроса пользователю желательно иметь хотя бы приблизительную оценку времени, которое потребуется для получения ответа, и представлять, каков будет объем этого ответа.

Таблица 4.

**Уровни метаданных в Хранилище Данных**

|  |  |
| --- | --- |
| Уровень приложения (внешних источников данных) | Описывает структуру данных в операционных БД и других источниках данных. Обычно, этот уровень достаточно сложен для понимания неподготовленного пользователя и является приложение ориентированным |
| Уровень ядра Хранилища Данных | Описывает логическую и физическую структуру и взаимосвязи данных в Хранилище Данных. |
| Уровень конечного пользователя | Описывает структуры данных в Хранилище Данных в терминах предметной области конечного пользователя. |

**Вопросы защиты данных**

Собрав в одном месте всю информацию об истории развития организации, ее успехах и неудачах, о взаимоотношениях с поставщиками и заказчиками, об истории развития и состоянии рынка, менеджеры получают уникальную возможность для анализа прошлой деятельности, сегодняшнего дня и построения обоснованных прогнозов на будущее. Однако не следует забывать и о том, что если не обеспечены надлежащие средства защиты и ограничения прав доступа, вы можете снабдить этой информацией и ваших конкурентов.

Одним из первых же вопросов, встающих при обсуждении проекта Хранилища Данных, является вопрос защиты данных. Чисто психологически, многих пугают не столько затраты на реализацию системы Хранилищ Данных, а то, что доступ к критически значимой информации может получить кто-либо, не имеющий на это права.

В таких системах, часто оказывается недостаточно защиты обеспечиваемой в стандартных конфигурациях коммерческих СУБД. Региональный менеджер должен видеть только те данные, которые относятся к его региону, а менеджер подразделения не должен видеть данные, относящиеся ко всей фирме. Но, для повышения эффективности доступа к данным, в целевой БД Хранилища Данных, все эти данные, как правило, хранятся в виде единой фактологической таблицы. Следствием этого, является то, что средства реализации должны поддерживать ограничения доступа не только на уровне отдельных таблиц и их колонок, но и отдельных строк в таблице.

Не менее остро стоят и вопросы авторизации и идентификации пользователей, защиты данных в местах их преобразования и согласования, в процессе их передачи по сети (шифрование паролей, текстов запросов, данных).

**5. Хранилище данных предприятия**

Тема Хранилищ данных имеет необычайную актуальность для современных российских предприятий. Причин этому две. Во-первых, большинство средних и крупных предприятий уже прошли стадию первичной автоматизации, то есть автоматизации бухгалтеров. Во-вторых, происходит быстрое укрупнение предприятий за счет их слияний, а также развития региональной сети. Поэтому настало время автоматизации менеджеров среднего и высшего звена.

**Задачи Хранилища данных**

В классическом представлении под целью создания Хранилища данных  понимается  поддержка принятия решений, другим словами обеспечение всех менеджеров предприятия полной, достоверной, согласованной и своевременной информацией из единого источника. Для реализации этой цели Хранилище данных должно выполнять ряд задач:

1.     Консолидация данных

2.     Интеграция данных

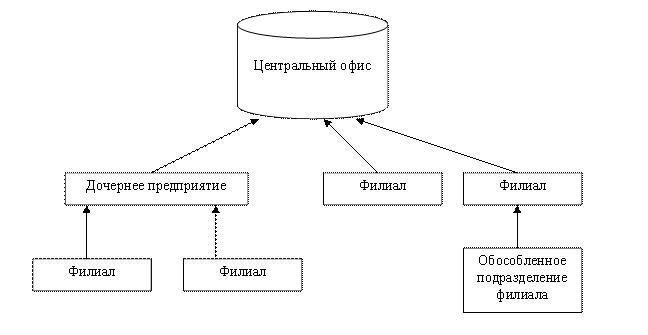
3.     Агрегация данных

4.     Расчеты производных показателей

5.     Предоставление данных для поддержки принятия решений (DSS)

**Консолидация данных**

Консолидация данных – это сбор в единую базу данных из удаленных филиалов многофилиального предприятия, или предприятий, входящих в холдинг (рис. 9). Консолидированные данные необходимы центральному руководству, чтобы осуществлять глобальное управление бизнесом, внедрять единую политику в филиалах и осуществлять  контроль над их деятельностью.

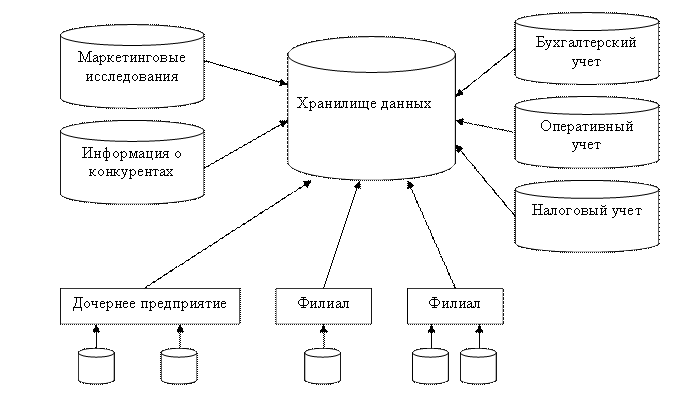
****

**Рисунок 9**.  Консолидация данных многоуровневого предприятия

Задача консолидации осложняется тем, что часто распределенные структуры создаются путем слияния предприятий, уже имеющих некоторый уровень автоматизации, обученный определенным системам персонал. Поэтому во многих случаях в филиалах работают различные системы автоматизации. Единственным способом консолидации данных в этих условиях является применение  разрозненных программ сбора показателей отчетности или единого Хранилища данных.

**Интеграция данных**

Интеграция данных – это объединение данных, которые изначально вводятся в разные  системы. Сами эти системы могут располагаться в одной локальной сети, но иметь различные платформы и внутреннюю архитектуру (рис. 10). Такая ситуация практически неизбежна во всех предприятиях занимающихся сложным бизнесом. Как правило, один единственный поставщик не может создать систему, в которой одинаково хорошо решены вопросы бухгалтерского учета и автоматизации производственного цикла,  управления кадрами и документооборота и так далее.

****

**Рисунок 10.**Интеграция данных из разнородных источников

Кроме того, существуют задачи, например, маркетингового анализа, привлечения клиентов, анализа конкурентной среды, которые по своей природе требуют получения (покупки) информации от разных поставщиков. Эта информация поставляется в виде разнообразных баз данных или электронных таблиц и требует загрузки в общую базу данных для совместного анализа.

**Агрегация данных**

Агрегация данных – это вычисление обобщенных показателей для поддержки стратегического или тактического управления из детальных данных. Например, все записи о продажах двухсот тысяч наименований товаров тысяче оптовых покупателям за каждый день года преобразуются в данные о продажах десяти категорий товаров пяти категориям покупателей в разрезе месяцев и кварталов года и регионов (рис. 11). Эти данные используются впоследствии менеджерами для принятия решений об изменениях направлений бизнеса, расширении рынка,  анализа сезонных колебаний спроса на товары разных категорий.

****

**Рисунок 11.**Предварительная агрегация данных по разным разрезам

Предварительный расчет агрегированных показателей применяется для того, чтобы руководитель получал ответы на подобные запросы предельно быстро. В то же время в хранилище собираются максимально детальные данные, что позволяет строить отчеты в произвольных аналитических разрезах, вычисляя агрегаты по мере возникновения в них потребности.

**Расчеты производных показателей**

В управленческой практике собранные из подразделений первичные оперативные данные используются для расчета сложных финансовых и оперативных показателей, таких как прибыль на капитал, средневзвешенные цены, ликвидность, доходность клиента и т.д. Хранилище данных предоставляет формульный язык для настройки алгоритмов расчета показателей и специальные механизмы быстрого выполнения расчетов над огромными массивами первичной информации.

**Предоставление данных для поддержки принятия решений (DSS)**

Как уже указывалось выше, изначально концепция Хранилища данных была разработана с единственной целью – для информационной поддержки принятия решений. Поэтому предполагалось, что данные Хранилища должны быть неизменяемы. Пользовательский интерфейс  обеспечивает всего две основные функции – выпуск отчетов для печати и интерактивный анализ данных. В связи с этим в качестве front-end можно применять универсальные системы выполнения запросов, анализа данных и выпуска отчетов. Эти инструменты позволяют свести к минимуму затраты на разработку отчетов, во многих случаях сводя создание новых форм отчетов к настройке, выполняемой самим пользователем.

**Тема 5. Оперативный анализ данных**

**Содержание**

1.     Место OLAP в информационной структуре предприятия

2.     Оперативная аналитическая обработка данных

3.     Требования к средствам оперативной аналитической обработки

4.     Классификация OLAP-продуктов

5.     Принципы работы OLAP-клиентов

6.     Выбор архитектуры OLAP-приложения

7.     Сферы применения OLAP-технологий

8.     Пример использования  OLAP-технологий для анализа в сфере продаж

**1. Место OLAP в информационной структуре предприятия**

Термин "OLAP" неразрывно связан с термином "хранилище данных" (Data Warehouse).

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.

Задача хранилища - предоставить "сырье" для анализа в одном месте и в простой, понятной структуре.

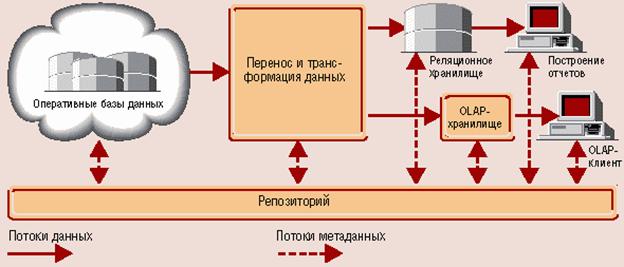
Есть и еще одна причина, оправдывающая появление отдельного хранилища - сложные аналитические запросы к оперативной информации тормозят текущую работу компании, надолго блокируя таблицы и захватывая ресурсы сервера.

Под хранилищем можно понимать не обязательно гигантское скопление данных - главное, чтобы оно было удобно для анализа.

Централизация и удобное структурирование - это далеко не все, что нужно аналитику. Ему ведь еще требуется инструмент для просмотра, визуализации информации. Традиционные отчеты, даже построенные на основе единого хранилища, лишены одного - гибкости. Их нельзя "покрутить", "развернуть" или "свернуть", чтобы получить желаемое представление данных. Вот бы ему такой инструмент, который позволил бы разворачивать и сворачивать данные просто и удобно! В качестве такого инструмента и выступает OLAP.

Хотя OLAP и не представляет собой необходимый атрибут хранилища данных, он все чаще и чаще применяется для анализа накопленных в этом хранилище сведений.

Место OLAP  в информационной структуре предприятия (рис. 1).



**Рисунок 1**. Место OLAP  в информационной структуре предприятия

Оперативные данные собираются из различных источников, очищаются, интегрируются и складываются в реляционное хранилище. При этом они уже доступны для анализа при помощи различных средств построения отчетов. Затем данные (полностью или частично) подготавливаются для OLAP-анализа. Они могут быть загружены в специальную БД OLAP или оставлены в реляционном хранилище. Важнейшим его элементом являются метаданные, т. е. информация о структуре, размещении и трансформации данных. Благодаря им обеспечивается эффективное взаимодействие различных компонентов хранилища.

Подытоживая, можно определить OLAP как совокупность средств многомерного анализа данных, накопленных в хранилище.

**2. Оперативная аналитическая обработка данных**

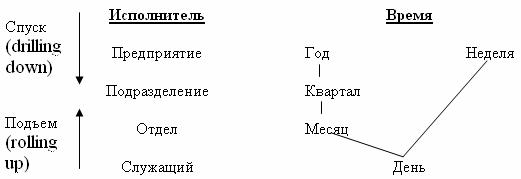
В основе концепции OLAP лежит принцип многомерного представления данных. В 1993 году  E. F. Codd рассмотрел недостатки реляционной модели, в первую очередь, указав на невозможность "объединять, просматривать и анализировать данные с точки зрения множественности измерений, то есть самым понятным для корпоративных аналитиков способом", и определил общие требования к системам OLAP, расширяющим функциональность реляционных СУБД и включающим многомерный анализ как одну из своих характеристик.

По Кодду, многомерное концептуальное представление данных (multi-dimensional conceptual view) представляет собой множественную перспективу, состоящую из нескольких независимых измерений, вдоль которых могут быть проанализированы определенные совокупности данных.

Одновременный анализ по нескольким измерениям определяется как многомерный анализ. Каждое измерение включает направления консолидации данных, состоящие из серии последовательных уровней обобщения, где каждый вышестоящий уровень соответствует большей степени агрегации данных по соответствующему измерению.

Так, измерение Исполнитель может определяться направлением консолидации, состоящим из уровней обобщения "предприятие - подразделение - отдел - служащий". Измерение Время может даже включать два направления консолидации - "год - квартал - месяц - день" и "неделя - день", поскольку счет времени по месяцам и по неделям несовместим. В этом случае становится возможным произвольный выбор желаемого уровня детализации информации по каждому из измерений.

Операция спуска (drilling down) соответствует движению от высших ступеней консолидации к низшим; напротив, операция подъема (rolling up) означает движение от низших уровней к высшим (рис. 2).



**Рисунок 2.**   Измерения и направления консолидации данных

**3. Требования к средствам оперативной аналитической обработки**

Многомерный подход возник практически одновременно и параллельно с реляционным. Однако, только начиная с середины девяностых годов, а точнее с 1993 г., интерес к **МСУБД** начал приобретать всеобщий характер. Именно в этом году появилась новая программная статья одного из основоположников реляционного подхода **Э. Кодда**, в которой он сформулировал 12 основных требований к средствам реализации **OLAP** (табл. 1).

*Таблица 1.*

|  |  |  |
| --- | --- | --- |
| 1 | Многомерное представление данных | Средства должны поддерживать многомерный на концептуальном уровне взгляд на данные. |
| 2 | Прозрачность | Пользователь не должен знать о том, какие конкретные средства используются для хранения и обработки данных, как данные организованы и откуда они берутся. |
| 3 | Доступность | Средства должны сами выбирать и связываться с наилучшим для формирования ответа на данный запрос источником данных. Средства должны обеспечивать автоматическое отображение их собственной логической схемы в различные гетерогенные источники данных. |
| 4 | Согласованная производительность | Производительность практически не должна зависеть от количества Измерений в запросе. |
| 5 | Поддержка архитектуры клиент-сервер | Средства должны работать в архитектуре клиент-сервер. |
| 6 | Равноправность всех измерений | Ни одно из измерений не должно быть базовым, все они должны быть равноправными (симметричными). |
| 7 | Динамическая обработка разреженных матриц | Неопределенные значения должны храниться и обрабатываться наиболее эффективным способом. |
| 8 | Поддержка многопользовательского режима работы с данными | Средства должны обеспечивать возможность работать более чем одному пользователю. |
| 9 | Поддержка операций на основе различных измерений | Все многомерные операции (например Агрегация) должны единообразно и согласованно применяться к любому числу любых измерений. |
| 10 | Простота манипулирования данными | Средства должны иметь максимально удобный, естественный и комфортный пользовательский интерфейс. |
| 11 | Развитые средства представления данных | Средства должны поддерживать различные способы визуализации (представления) данных. |
| 12 | Неограниченное число измерений и уровней агрегации данных | Не должно быть ограничений на число поддерживаемых Измерений. |

**Правила оценки программных продуктов класса OLAP**

Набор этих требований, послуживших фактическим определением OLAP, следует рассматривать как рекомендательный, а конкретные продукты оценивать по степени приближения к идеально полному соответствию всем требованиям.

Позже определение Кодда было переработано в так называемый тест FASMI, требующий, чтобы OLAP-приложение предоставляло возможности быстрого анализа разделяемой многомерной информации.

**Тест FASMI**

Помнить 12 правил Кодда слишком обременительно для большинства людей. Оказались, что можно  резюмировать OLAP-определение только пятью ключевыми словами: Быстрый Анализ Разделяемой Многомерной Информации - или, кратко - FASMI (в переводе с английского: ***F***ast ***A***nalysis of ***S***hared ***M***ultidimensional ***I***nformation).

Это определение впервые было сформулировано в начале 1995 года и с тех пор не нуждалось в пересмотре.

**FAST (Быстрый)** - означает, что система должна обеспечивать выдачу большинства ответов пользователям в пределах приблизительно пяти секунд. При этом самые простые запросы обрабатываются в течение одной секунды и очень немногие - более 20-ти секунд. Исследования показали, что конечные пользователи воспринимают процесс неудачным, если результаты не получены по истечении 30 секунд.

На первый взгляд может казаться удивительным, что при получении отчета за минуту, на который не так давно требовались дни, пользователь очень быстро начинает скучать во время ожиданий, и проект оказывается намного менее успешным, чем в случае мгновенного ответа, даже ценой менее детального анализа.

**ANALYSIS (Анализ)** означает, что система может справляться с любым логическим и статистическим анализом, характерным для данного приложения, и обеспечивает его сохранение в виде, доступном для конечного пользователя.

Не так важно, выполнен ли этот анализ в собственных инструментальных средствах поставщика или в связанном внешнем программном продукте типа электронной таблицы, просто все требуемые функциональные возможности анализа должны обеспечиваться интуитивным способом для конечных пользователей. Средства анализа могли бы включать определенные процедуры, типа анализа временных рядов, распределения затрат, валютных переводов, поиска целей, изменения многомерных структур, непроцедурного моделирования, выявления исключительных ситуаций, извлечения данных и другие операции зависимые от приложения. Такие возможности широко отличаются среди продуктов, в зависимости от целевой ориентации.

**SHARED (Разделяемой)** означает, что система осуществляет все требования защиты конфиденциальности (возможно до уровня ячейки) и, если множественный доступ для записи необходим, обеспечивает блокировку модификаций на соответствующем уровне. Не во всех приложениях есть необходимость обратной записи данных. Однако количество таких приложений растет, и система должна быть способна обработать множественные модификации своевременным, безопасным способом.

**MULTIDIMENSIONAL (Многомерной) -**это ключевое требование. Если бы нужно было определить OLAP одним словом, то выбрали бы его. Система должна обеспечить многомерное концептуальное представление данных, включая полную поддержку для иерархий и множественных иерархий, поскольку это определенно наиболее логичный способ анализировать бизнес и организации. Не установлено минимальное число измерений, которые должны быть обработаны, поскольку оно также зависит от приложения, и большинство продуктов OLAP имеет достаточное количество измерений для тех рынков, на которые они нацелены.

**INFORMATION (Информации)** - это все. Необходимая информация должна быть получена там, где она необходима. Однако многое зависит от приложения. Мощность различных продуктов измеряется в терминах того, сколько входных данных они могут обрабатывать, но не сколько гигабайт они могут хранить. Мощность продуктов весьма различна - самые большие OLAP продукты могут оперировать, по крайней мере, в тысячу раз большим количеством данных по сравнению с самыми маленькими. По этому поводу следует учитывать много факторов, включая дублирование данных, требуемую оперативная память, использование дискового пространства, эксплуатационные показатели, интеграцию с информационными хранилищами и т.п.

Тест FASMI - разумное и понятное определение целей, на достижение которых ориентированы OLAP.

**4. Классификация OLAP-продуктов**

Итак, суть OLAP заключается в том, что исходная для анализа информация представляется в виде многомерного куба, и обеспечивается возможность произвольно манипулировать ею и получать нужные информационные разрезы - отчеты. При этом конечный пользователь видит куб как многомерную динамическую таблицу, которая автоматически суммирует данные (факты) в различных разрезах (измерениях), и позволяет интерактивно управлять вычислениями и формой отчета. Выполнение этих операций обеспечивается OLAP-машиной (или машиной OLAP-вычислений).

На сегодняшний день в мире разработано множество продуктов, реализующих OLAP-технологии. Чтобы легче было ориентироваться среди них, используют классификации OLAP-продуктов: по способу хранения данных для анализа и по месту нахождения OLAP-машины. Рассмотрим подробнее каждую категорию OLAP-продуктов.

**Классификация по способу хранения данных**.

Многомерные кубы строятся на основе исходных и агрегатных данных. И исходные и агрегатные данные для кубов могут храниться как в реляционных, так и многомерных базах данных. Поэтому в настоящее время применяются три способа хранения данных: MOLAP (Multidimensional OLAP), ROLAP (Relational OLAP) и HOLAP (Hybrid OLAP). Соответственно, OLAP-продукты по способу хранения данных делятся на три аналогичные категории:

1.  В случае MOLAP, исходные и агрегатные данные хранятся в многомерной БД или в многомерном  локальном кубе.

2.  В ROLAP-продуктах исходные данные хранятся в реляционных БД или в плоских локальных таблицах на файл-сервере. Агрегатные данные могут помещаться в служебные таблицы в той же БД. Преобразование данных из реляционной БД в многомерные кубы происходит по запросу OLAP-средства.

3.  В случае использования HOLAP архитектуры исходные данные остаются в реляционной базе, а агрегаты размещаются в многомерной. Построение OLAP-куба выполняется по запросу OLAP-средства на основе реляционных и многомерных данных.

**Классификация по месту размещения OLAP-машины.**

По этому признаку OLAP-продукты делятся на OLAP-серверы и OLAP-клиенты:

  В серверных OLAP-средствах вычисления и хранение агрегатных данных выполняются отдельным процессом - сервером. Клиентское приложение получает только результаты запросов к многомерным кубам, которые хранятся на сервере. Некоторые OLAP-серверы поддерживают хранение данных только в реляционных базах, некоторые - только в многомерных. Многие современные OLAP-серверы поддерживают все три способа хранения данных: MOLAP, ROLAP и HOLAP.

**MOLAP**

MOLAP - это Multidimensional On-Line Analytical Processing, то есть Многомерный OLAP. Это означает, что сервер для хранения данных использует  многомерную базу данных (МБД). Смысл использования МБД очевиден. Она может эффективно хранить многомерные по своей природе данные, обеспечивая средства быстрого обслуживания запросов к базе данных. Данные передаются от источника данных в многомерную базу данных, а затем база данных подвергается агрегации. Предварительный расчет - это то, что ускоряет OLAP-запросы, поскольку расчет сводных данных уже произведен. Время запроса становится функцией исключительно времени, необходимого для доступа к отдельному фрагменту данных и выполнения расчета. Этот метод поддерживает концепцию, согласно которой работа производится единожды, а результаты затем используются снова и снова. Многомерные базы данных являются относительно новой технологией. Использование МБД имеет те же недостатки, что и большинство новых технологий. А именно - они не так устойчивы, как реляционные базы данных (РБД), и в той же мере не оптимизированы. Другое слабое место МБД заключается в невозможности использовать большинство многомерных баз в процессе агрегации данных, поэтому требуется время для того, чтобы новая информация стала доступна для анализа.

**ROLAP**

ROLAP - это Relational On-Line Analytical Processing, то есть Реляционный OLAP. Термин ROLAP обозначает, что OLAP-сервер базируется на реляционной базе данных. Исходные данные вводятся в реляционную базу данных, обычно по схеме "звезда" или схеме "снежинка", что способствует сокращению времени извлечения. Сервер обеспечивает многомерную модель данных с помощью оптимизированных SQL-запросов.

Существует ряд причин для выбора именно реляционной, а не многомерной базы данных. РБД - это хорошо отработанная технология, имеющая множество возможностей для оптимизации. Использование в реальных условиях дало в результате более проработанный продукт. К тому же, РБД поддерживают более крупные объемы данных, чем МБД. Они как раз и спроектированы для таких объемов. Основным аргументом против РБД является сложность запросов, необходимых для получения информации из большой базы данных с помощью SQL. Неопытный SQL-программист мог бы с легкостью обременить ценные системные ресурсы попытками выполнить какой-нибудь подобный запрос, который в МБД выполняется гораздо проще.

**Агрегированные/Предварительно агрегированные данные**

Быстрая реализация запросов является императивом для OLAP. Это один из базовых принципов OLAP - способность интуитивно манипулировать данными требует быстрого извлечения информации. В целом, чем больше вычислений необходимо произвести, чтобы получить фрагмент информации, тем медленнее происходит отклик. Поэтому, чтобы сохранить маленькое время реализации запросов, фрагменты информации, обращение к которым обычно происходит наиболее часто, но которые при этом требуют вычисления, подвергаются предварительной агрегации. То есть они подсчитываются и затем хранятся в базе данных в качестве новых данных. В качестве примера типа данных, который допустимо рассчитать заранее, можно привести сводные данные - например, показатели продаж по месяцам, кварталам или годам, для которых действительно введенными данными являются ежедневные показатели.

Различные поставщики придерживаются различных методов отбора параметров, требующих предварительной агрегации и числа предварительно вычисляемых величин. Подход к агрегации влияет одновременно и на базу данных и на время реализации запросов. Если вычисляется больше величин, вероятность того, что пользователь запросит уже вычисленную величину, возрастает, и поэтому время отклика сократиться, так как не придется запрашивать изначальную величину для вычисления. Однако, если вычислить все возможные величины - это не лучшее решение - в таком случае существенно возрастает размер базы данных, что сделает ее неуправляемой, да и время агрегации будет слишком большим. К тому же, когда в базу данных добавляются числовые значения, или если они изменяются, информация эта должна отражаться на предварительно вычисленных величинах, зависящих от новых данных. Таким образом, и обновление базы может также занять много времени в случае большого числа предварительно вычисляемых величин. Поскольку обычно во время агрегации база данных работает автономно, желательно, чтобы время агрегации было не слишком длительным.

  OLAP-клиент устроен по-другому. Построение многомерного куба и OLAP-вычисления выполняются в памяти клиентского компьютера. OLAP-клиенты также делятся на ROLAP и MOLAP. А некоторые могут поддерживать оба варианта доступа к данным.

У каждого из этих подходов, есть свои "плюсы" и "минусы". Вопреки распространенному мнению о преимуществах серверных средств перед клиентскими, в целом ряде случаев применение OLAP-клиента для пользователей может оказаться эффективнее и выгоднее использования OLAP-сервера.

Разработка аналитических приложений с помощью клиентских OLAP-средств – процесс быстрый и не требующий специальной подготовки исполнителя. Пользователь, знающий физическую реализацию базы данных, может разработать аналитическое приложение самостоятельно, без привлечения ИТ-специалиста.

При использовании OLAP-сервера необходимо изучить 2 разные системы, иногда от различных поставщиков, – для создания кубов на сервере, и для разработки клиентского приложения.

OLAP-клиент предоставляет единый визуальный интерфейс для описания кубов и настройки к ним пользовательских интерфейсов.

Итак, в каких случаях применение OLAP-клиента для пользователей может оказаться эффективнее и выгоднее использования OLAP-сервера?

 Экономическая целесообразность применения OLAP-сервера возникает, когда объемы данных очень велики и непосильны для OLAP-клиента, иначе более оправдано применение последнего. В этом случае OLAP-клиент сочетает в себе высокие характеристики производительности и низкую стоимость.

 Мощные ПК аналитиков – еще один довод в пользу OLAP-клиентов. При применении OLAP-сервера эти мощности не используются.

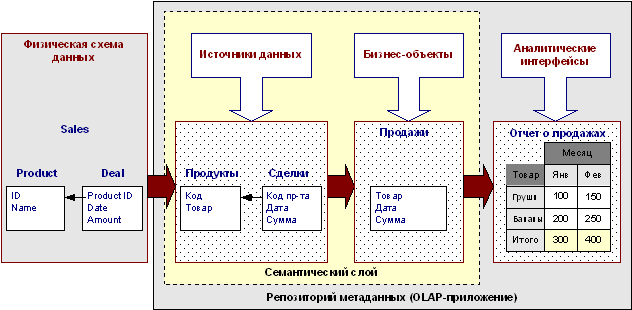
Среди преимуществ OLAP-клиентов можно также назвать следующее:

 Затраты на внедрение и сопровождение OLAP-клиента существенно ниже, чем затраты на OLAP-сервер.

 При использовании OLAP-клиента со встроенной машиной передача данных по сети производится один раз. При выполнении OLAP-операций новых потоков данных не порождается.

**5. Принципы работы OLAP-клиентов**

Рассмотрим процесс создания OLAP-приложения с помощью клиентского инструментального средства (рис. 1).



**Рисунок 1.**   Создание OLAP-приложения с помощью клиентского ROLAP-средства

Принцип работы ROLAP-клиентов – предварительное описание семантического слоя, за которым скрывается физическая структура исходных данных. При этом источниками данных могут быть: локальные таблицы, РСУБД. Список поддерживаемых источников данных определяется конкретным программным продуктом. После этого пользователь может самостоятельно манипулировать понятными ему объектами в терминах предметной области для создания кубов и аналитических интерфейсов.

Принцип работы клиента OLAP-сервера иной. В OLAP-сервере при создании кубов пользователь манипулирует физическими описаниями БД. При этом в самом кубе создаются пользовательские описания. Клиент OLAP-сервера настраивается только на куб.

При создании семантического слоя источники данных – таблицы Sales и Deal – описываются понятными конечному пользователю терминами и превращаются в «Продукты» и «Сделки». Поле «ID» из таблицы «Продукты» переименовывается в «Код», а «Name» - в «Товар» и т.д.

Затем создается бизнес-объект «Продажи». Бизнес-объект – это плоская таблица, на основе которой формируется многомерный куб. При создании бизнес-объекта таблицы «Продукты» и «Сделки» объединяются по полю «Код» товара. Поскольку для отображения в отчете не потребуются все поля таблиц – бизнес-объект использует только поля «Товар», «Дата» и «Сумма».

Далее на базе бизнес-объекта создается OLAP-отчет. Пользователь выбирает бизнес-объект и перетаскивает его атрибуты в области колонок или строк таблицы отчета.

В нашем примере на базе бизнес-объекта «Продажи» создан отчет по продажам товаров по месяцам.

При работе с интерактивным отчетом пользователь может задавать условия фильтрации и группировки такими же простыми движениями «мышью». В этот момент ROLAP-клиент обращается к данным в кэше. Клиент же OLAP-сервера генерирует новый запрос к многомерной базе данных. Например, применив в отчете о продажах фильтр по товарам, можно получить отчет о продажах интересующих нас товаров.

Все настройки OLAP-приложения могут храниться в выделенном репозитории метаданных, в приложении или в системном репозитории многомерной базы данных. Реализация зависит от конкретного программного продукта.

Все, что включается в состав этих приложений, представляет собой стандартный взгляд на интерфейс, заранее определенные функции и структуру, а также быстрые решения для более или менее стандартных ситуаций. Например, популярны финансовые пакеты. Заранее созданные финансовые приложения позволят специалистам использовать привычные финансовые инструменты без необходимости проектировать структуру базы данных или общепринятые формы и отчеты.

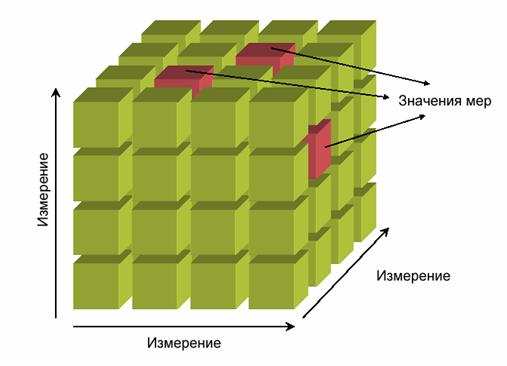
Интернет является новой формой клиента. Кроме того, он несет на себе печать новых технологий; множество интернет-решений существенно отличаются по своим возможностям в целом и в качестве OLAP-решения - в частности. Существует масса преимуществ в формировании OLAP-отчетов через Интернет. Наиболее существенным представляется отсутствие необходимости в специализированном программном обеспечении для доступа к информации. Это экономит предприятию кучу времени и денег.

**6. Выбор архитектуры OLAP-приложения**

При реализации информационно-аналитической системы важно не ошибиться в выборе архитектуры OLAP-приложения. Дословный перевод термина On-Line Analytical Process — «оперативная аналитическая обработка» — часто воспринимается буквально в том смысле, что поступающие в систему данные оперативно анализируются. Это заблуждение — оперативность анализа никак не связана с реальным временем обновления данных в системе. Эта характеристика относится к времени реакции OLAP-системы на запросы пользователя. При этом зачастую анализируемые данные представляют собой снимок информации «на вчерашний день», если, например, данные в хранилищах обновляются раз в сутки.

В этом контексте более точен перевод OLAP как «интерактивная аналитическая обработка». Именно возможность анализа данных в интерактивном режиме отличает OLAP-системы от систем подготовки регламентированных отчетов.

Другой особенностью интерактивной обработки в формулировке родоначальника OLAP Э. Кодда является возможность «объединять, просматривать и анализировать данные с точки зрения множественности измерений, т. е. самым понятным для корпоративных аналитиков способом». У самого Кодда термин OLAP обозначает исключительно конкретный способ представления данных на концептуальном уровне — многомерный. На физическом уровне данные могут храниться в реляционных базах данных, однако на деле OLAP-инструменты, как правило, работают с многомерными базами данных, в которых данные упорядочены в виде гиперкуба (рис. 1).



**Рисунок 1.**  OLAP – куб (гиперкуб, метакуб)

При этом актуальность этих данных определяется моментом наполнения гиперкуба новыми данными.

Очевидно, что время формирования многомерной базы данных существенно зависит от объема загружаемых в нее данных, поэтому разумно ограничить этот объем. Но как при этом не сузить возможности анализа и не лишить пользователя доступа ко всей интересующей информации? Существует два альтернативных пути: Analyze then query («Сначала проанализируй — затем запроси дополнительную информацию») и Query then analyze («Сначала запроси данные — затем анализируй»).

Последователи первого пути предлагают загружать в многомерную базу данных обобщенную информацию, например, месячные, квартальные, годовые итоги по подразделениям. А при необходимости детализации данных пользователю предлагается сформировать отчет по реляционной базе, содержащей требуемую выборку, например, по дням для данного подразделения или по месяцам и сотрудникам выбранного подразделения.

Сторонники второго пути, напротив, предлагают пользователю, прежде всего, определиться с данными, которые он собирается анализировать и именно их загружать в микрокуб — небольшую многомерную базу данных. Оба подхода отличаются на концептуальном уровне и имеют свои достоинства и недостатки.

К достоинствам второго подхода следует отнести «свежесть» информации, которую пользователь получает в виде многомерного отчета — «микрокуба». Микрокуб формируется на основе только что запрошенной информации из актуальной реляционной базы данных. Работа с микрокубом осуществляется в интерактивном режиме — получение срезов информации и ее детализация в рамках микрокуба осуществляется моментально. Другим положительным моментом является то, что проектирование структуры и наполнение микрокуба осуществляется пользователем «на лету», без участия администратора баз данных. Однако подход страдает и серьезными недостатками. Пользователь, не видит общей картины и должен заранее определяться с направлением своего исследования. В противном случае запрошенный микрокуб может оказаться слишком мал и не содержать всех интересующих данных, а пользователю придется запрашивать новый микрокуб, затем новый, затем еще и еще. Подход Query then analyze реализует инструментальное средство BusinessObjects одноименной компании и инструментальные средства платформы Контур компании Intersoft Lab.

При подходе Analyze then query объем данных, загружаемых в многомерную базу данных, может быть достаточно велик, наполнение должно выполняться по регламенту и может занимать достаточно много времени. Однако все эти недостатки окупаются впоследствии, когда пользователь имеет доступ практически ко всем необходимым данным в любой комбинации. Обращение к исходным данным в реляционной базе данных осуществляется лишь в крайнем случае, когда необходима детальная информация, например, по конкретной накладной.

На работе единой многомерной базы данных практически не сказывается количество обращающихся к ней пользователей. Они лишь читают имеющиеся там данные в отличие от подхода Query then analyze, при котором количество микрокубов в предельном случае может расти с той же скоростью, что и количество пользователей.

При данном подходе увеличивается нагрузка на ИТ-службы, которые кроме реляционных вынуждены обслуживать еще и многомерные базы данных. Именно эти службы несут ответственность за своевременное автоматическое обновление данных в многомерных базах данных.

Наиболее яркими представителями подхода «Analyze then query» являются инструментальные средства PowerPlay и Impromptu компании Cognos.

Выбор и подхода, и инструмента его реализующего, зависит в первую очередь от преследуемой цели: всегда приходится балансировать между экономией бюджета и повышением качества обслуживания конечных пользователей. При этом надо учитывать, что в стратегическом плане создание информационно-аналитических систем преследует цели достижения конкурентного преимущества, а не избежания расходов на автоматизацию. Например, корпоративная информационно-аналитическая система может предоставлять необходимую, своевременную и достоверную информацию о компании, публикация которой для потенциальных инвесторов обеспечит прозрачность и предсказуемость данной компании, что неизбежно станет условием ее инвестиционной привлекательности.

**7. Сферы применения OLAP-технологий**

OLAP применим везде, где есть задача анализа многофакторных данных. Вообще, при наличии некоторой таблицы с данными, в которой есть хотя бы одна описательная колонка (измерение) и одна колонка с цифрами (меры или факты) OLAP-инструмент, как правило, будет эффективным средством анализа и генерации отчетов.

Рассмотрим некоторые сферы применения OLAP-технологий, взятые из реальной жизни.

***1. Продажи***

На основе анализа структуры продаж решаются вопросы необходимые для принятия управленческих решений: об изменении ассортимента товаров, цен, закрытии и открытии магазинов, филиалов, расторжении и подписании договоров с дилерами, проведения или прекращения рекламных кампаний и т.д.

***2. Закупки***

Задача обратно противоположная анализу продаж. Многие предприятия закупают комплектующие и материалы у поставщиков. Торговые предприятия закупают товары для перепродажи. Возможных задач при анализе закупок множество, от планирования денежных средств на основе прошлого опыта, до контроля за менеджерами, выбирающими поставщиков.

***3. Цены***

С анализом закупок смыкается анализ рыночных цен. Целью этого анализа является оптимизация расходов, выбор наиболее выгодных предложений.

***4. Маркетинг***

Под маркетинговым анализом будем иметь ввиду только область анализа покупателей или клиентов-потребителей услуг. Задачей этого анализа является правильное позиционирование товара, выявление групп покупателей для целевой рекламы, оптимизация ассортимента. Задача OLAP в данном случае - дать пользователю инструмент быстрого, со скоростью мысли, получения ответов на вопросы, интуитивно возникающие по ходу анализа данных.

***5. Склад***

Анализ структуры остатков на складе в разрезе видов товаров, складов, анализ сроков хранения товаров, анализ отгрузки по получателям и многие другие важные для предприятия виды анализа возможны при наличии в организации складского учета.

***6. Движение денежных средств***

Это целая область анализа, имеющая множество школ и методик. OLAP-технология может служить инструментом реализации или усовершенствования этих методик, но никак не их заменой. Анализируются денежные обороты безналичных и наличных средств в разрезе бизнес-операций, контрагентов, валют и времени с целью оптимизации потоков, обеспечения ликвидности, и т.д. Состав измерений сильно зависит от особенностей бизнеса, отрасли, методики.

***7. Бюджет***

Одна из самых благодатных областей применения OLAP-технологий. Не даром ни одна современная система бюджетирования не считается завершенной без наличия в ее составе OLAP-инструментария для анализа бюджета. Большинство бюджетных отчетов легко строятся на основе OLAP-систем. При этом отчеты отвечают на очень широкую гамму вопросов: анализ структуры расходов и доходов, сравнение расходов по определенным статьям у разных подразделений, анализ динамики и тенденций расходов на определенные статьи, анализ себестоимости и прибыли.

***8. Бухгалтерские счета***

Классический балансовый отчет, состоящий из номера счета и содержащий входящие остатки, обороты и исходящие остатки может быть отлично проанализирован в OLAP-системе. Кроме того, OLAP-система может автоматически и очень быстро вычислять консолидированные балансы многофилиальной организации, балансы за месяц, квартал и год, агрегированные балансы по иерархии счетов, аналитические балансы на основании аналитических признаков.

***9. Финансовая отчетность***

Технологично построенная система отчетности  есть ни что иное, как набор именованных показателей со значениями на дату, которые требуется сгруппировать и просуммировать в различных разрезах для получения конкретных отчетов. Когда это так, то отображение и печать отчетов наиболее просто и дешево реализуются в OLAP-системах. В любом случае система внутренней отчетности предприятия не так консервативна и может быть перестроена в целях экономии средств на технические работы по созданию отчетов и получения возможностей многомерного оперативного анализа.

***10. Посещаемость сайта***

Лог-файл Интернет-сервера многомерен по природе, а значит подходит для OLAP-анализа. Фактами являются: количество посещений, количество хитов, время проведенное на странице и другая информация, имеющаяся в логе.

***11. Объемы производства***

Это еще один пример статистического анализа. Таким образом, можно анализировать объемы выращенного картофеля, выплавленной стали, произведенного товара.

***12. Потребление расходных материалов***

Представьте себе завод, состоящий из десятков цехов, в которых расходуются охлаждающие, промывочные жидкости, масла, ветошь, наждачная бумага - сотни наименований расходных материалов. Для точного планирования, оптимизации издержек требуется тщательный анализ фактического потребления расходных материалов.

***13. Использование помещений***

Еще один вид статистического анализа. Примеры: анализ загруженности учебных аудиторий, сдаваемых в аренду зданий и помещений, использования залов для конференций и пр.

***14. Текучесть кадров на предприятии***

Анализ текучести кадров на предприятии в разрезе филиалов, отделов, профессий, уровня образования, пола, возраста, времени.

**15. Пассажирские перевозки**

Анализ количества проданных билетов и сумм в разрезе сезонов, направлений, видов вагонов (классов), типов поездов (самолетов).

Этим списком не ограничиваются сферы применения OLAP - технологий. Для примера рассмотрим технологию OLAP-анализа в сфере продаж.

**8. Пример использования  OLAP-технологий для анализа в сфере продаж**

Проектирование многомерного представления данных для OLAP-анализа начинается с формирования карты измерений. Например, при анализе продаж может быть целесообразно, выделить отдельные части рынка (развивающиеся, стабильные, крупные и мелкие потребители, вероятность появления новых потребителей и т.п.) и оценить объемы продаж по продуктам, территориям, покупателям, сегментам рынка, каналам сбыта и размерам заказов. Эти направления образуют координатную сетку многомерного представления продаж — структуру его измерений.

Поскольку деятельность любого предприятия протекает во времени, первый вопрос, который возникает при анализе, это вопрос о динамике развития бизнеса. Правильная организация оси времени позволит качественно ответить на этот вопрос. Обычно ось времени делится на годы, кварталы и месяцы. Возможно еще большее дробление на недели и дни. Структура временного измерения формируется с учетом периодичности поступления данных; может обуславливаться также периодичностью востребования информации.

Измерение «группы товаров» разрабатывается так, чтобы в максимальной степени отразить структуру продаваемой продукции. При этом важно соблюсти определенный баланс, чтобы, с одной стороны, избежать излишней детализации (количество групп должно быть обозримым), а с другой — не упустить существенный сегмент рынка.

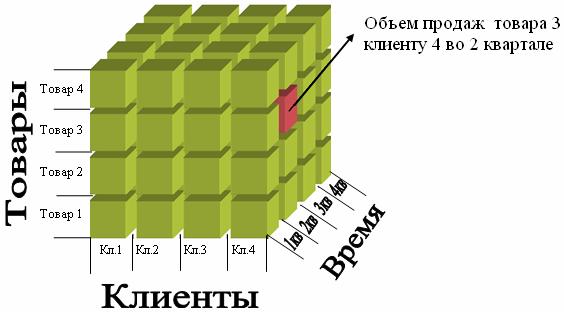
Измерение «Клиенты» отражает структуру продаж по территориально-географическому признаку. В каждом измерении могут существовать свои иерархии, например, в данном измерении это может быть структура: Страны – Регионы – Города – Клиенты.

Для анализа эффективности деятельности подразделений следует создать свое измерение. Например, можно выделить два уровня иерархии: департаменты и входящие в них отделы, что и должно найти отражение в измерении «Подразделения».

По сути, измерения «Время», «Товары», «Заказчики» достаточно полно определяют пространство предметной области.

Дополнительно, полезно разбить это пространство на условные области, взяв за основу вычисляемые характеристики, например, диапазоны объема сделок в стоимостном выражении. Тогда весь бизнес можно разделить на ряд стоимостных диапазонов, в котором он осуществляется. В данном примере можно ограничиться следующими показателями: суммами продаж товаров, количеством проданных товаров, величиной дохода, количеством сделок, количеством клиентов, объемом закупок у производителей.

OLAP – куб для анализа будет иметь вид (рис. 2):

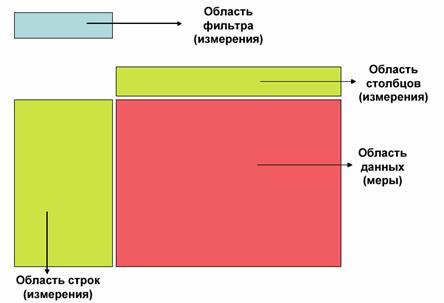


**Рисунок 2.**OLAP – куб для анализа объема продаж

Вот именно такой трехмерный массив в терминах OLAP и называется кубом. На самом деле, с точки зрения строгой математики кубом такой массив будет далеко не всегда: у настоящего куба количество элементов во всех измерениях должно быть одинаковым, а у кубов OLAP такого ограничения нет. Куб OLAP совсем не обязательно должен быть трехмерным. Он может быть и двух-, и многомерным - в зависимости от решаемой задачи. Серьезные OLAP-продукты рассчитаны на количество измерений порядка 20. Более простые настольные приложения поддерживают где-то 6 измерений.

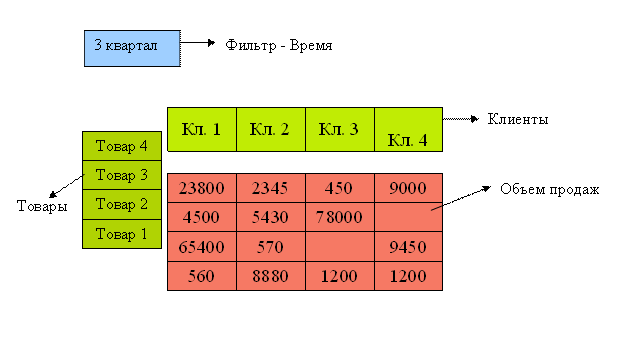
Должны быть заполнены далеко не все элементы куба: если нет информации о продажах Товара 2 Клиенту 3 в третьем квартале, значение в соответствующей ячейке просто не будет определено.

Однако куб сам по себе для анализа не пригоден. Если еще можно адекватно представить или изобразить трехмерный куб, то с шести- или девятнадцатимерным дело обстоит значительно хуже. Поэтому перед употреблением из многомерного куба извлекают обычные двумерные таблицы. Эта операция называется "разрезанием" куба. Аналитик как бы берет и "разрезает" измерения куба по интересующим его меткам. Этим способом аналитик получает двумерный срез куба  (отчет) и с ним работает. Структура отчета представлена на рисунке 3.



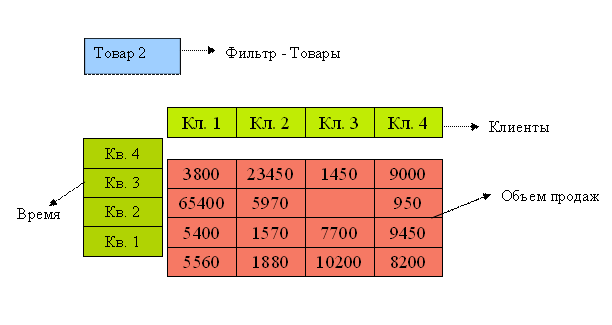
**Рисунок 3.**Структура аналитического отчета

Разрежем наш OLAP – куб и получим отчет о продажах за третий квартал, он будет иметь следующий вид (рис.4).

****

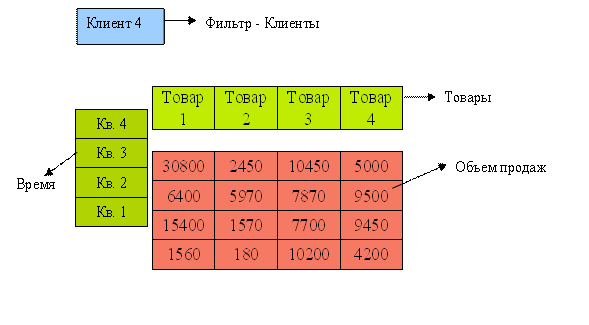
**Рисунок 4.** Отчет о продажах за третий квартал

Можно разрезать куб вдоль другой оси и получить отчет о продажах группы товаров 2 в течение года (рис. 5).

****

**Рисунок 5.**Поквартальный отчет о продажах товара 2.

Аналогично можно проанализировать отношения с клиентом 4, разрезав куб по метке Клиенты (рис. 6)



**Рисунок 6.** Отчет о поставках товаров клиенту 4.

Можно детализировать отчет по месяцам или говорить о поставках товаров в определенный филиал клиента.

**Тема 6. Интеллектуальный анализ данных**

**Содержание**

1.     Интеллектуальный анализ данных

2.     Стадии ИАД

3.     Методы АИД

4.     Типы закономерностей

5.     Типовые задачи для методов ИАД

6.     Области применения Data mining

7.     Классы систем Data Mining

8.     Интеграция OLAP и ИАД

**1. Интеллектуальный анализ данных**

Мы живем в веке информации. Трудно переоценить значение данных, которые мы непрерывно собираем в процессе нашей деятельности, в управлении бизнесом или производством, в банковском деле, в решении научных, инженерных и медицинских задач.

Мощные компьютерные системы, хранящие и управляющие огромными базами данных, стали неотъемлемым атрибутом жизнедеятельности, как крупных корпораций, так и даже небольших компаний.

Тем не менее, наличие данных само по себе еще недостаточно для улучшения показателей работы. Нужно уметь трансформировать "сырые" данные в полезную для принятия важных бизнес решений информацию.

В этом и состоит основное предназначение технологий Data mining.

Data Mining переводится как "добыча" или "раскопка данных". Нередко рядом с Data Mining встречаются слова "обнаружение знаний в базах данных" (knowledge discovery in databases) и "интеллектуальный анализ данных". Их можно считать синонимами Data Mining. Возникновение всех указанных терминов связано с новым витком в развитии средств и методов обработки данных.

 В связи с совершенствованием технологий записи и хранения данных на людей обрушились колоссальные потоки информационной руды в самых различных областях. Деятельность любого предприятия (коммерческого, производственного, медицинского, научного и т.д.) теперь сопровождается регистрацией и записью всех подробностей его деятельности. Что делать с этой информацией? Стало ясно, что без продуктивной переработки потоки сырых данных образуют никому не нужную свалку.

Специфика современных требований к такой переработке следующая:

        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.

Традиционная математическая статистика, долгое время претендовавшая на роль основного инструмента анализа данных, откровенно спасовала перед лицом возникших проблем. 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 (on-line 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. Поиск шаблонов производится методами, не ограниченными рамками априорных предположений о структуре выборки и виде распределений значений анализируемых показателей.

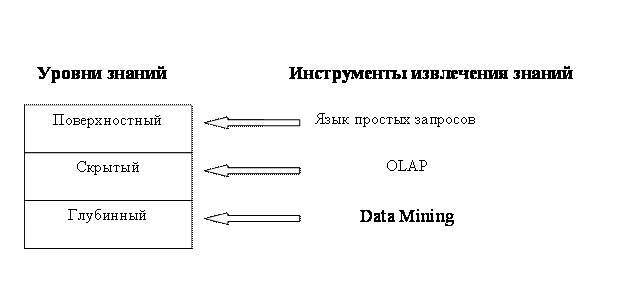
Примерами заданий на такой поиск при использовании Data Mining могут служить следующие вопросы:

1.     Встречаются ли точные шаблоны в описаниях людей, подверженных повышенному травматизму?

2.     Имеются ли характерные портреты клиентов, которые, по всей вероятности, собираются отказаться от услуг телефонной компании?

3.     Существуют ли стереотипные схемы покупок для случаев мошенничества с кредитными карточками?

Важное положение Data Mining — нетривиальность разыскиваемых шаблонов. Это означает, что найденные шаблоны должны отражать неочевидные, неожиданные регулярности в данных, составляющие так называемые скрытые знания (hidden knowledge). К обществу пришло понимание, что сырые данные содержат глубинный пласт знаний, при грамотной раскопке которого могут быть обнаружены настоящие самородки.На рисунке 1 показаны уровни знаний и инструменты для их извлечения.



**Рисунок 1.**  Уровни знаний, извлекаемых из 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.

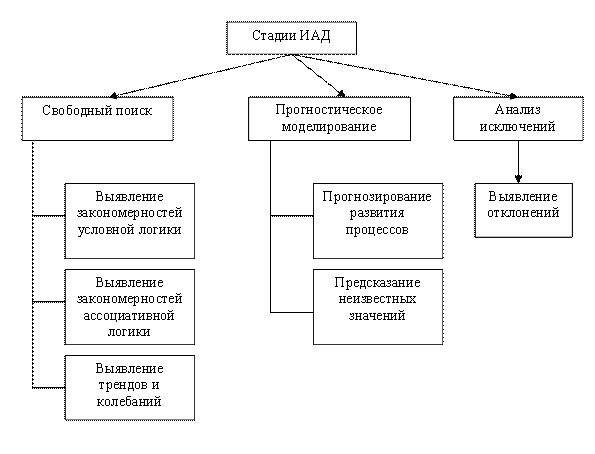
**2. Стадии ИАД**

В общем случае процесс интеллектуального анализа данных (ИАД) состоит из трёх стадий (рис. 2):

1)     выявление закономерностей (свободный поиск);

2)     использование выявленных закономерностей для предсказания неизвестных значений (прогностическое моделирование);

3)     анализ исключений, предназначенный для выявления и толкования аномалий в найденных закономерностях.



**Рисунок 2**.  Стадии процесса интеллектуального анализа данных

**1. Свободный поиск (Discovery)**

Свободный поиск определяется как процесс исследования исходной БД на предмет поиска скрытых закономерностей без предварительного определения гипотез относительно вида этих закономерностей. Другими словами, сама программа берет на себя инициативу в деле поиска интересных аномалий, или шаблонов, в данных, освобождая аналитика от необходимости обдумывания и задания соответствующих запросов. Этот подход особенно ценен при исследовании больших баз данных, имеющих значительное количество скрытых закономерностей, большинство из которых было бы упущено при непосредственном поиске путем прямых запросов пользователя к исходным данным.

В качестве примера свободного поиска по инициативе системы рассмотрим исследование реестра физических лиц. Если инициатива принадлежит пользователю, он может построить запрос типа "Каков средний возраст директоров предприятий отрасли промышленности строительных материалов, расположенных в Иванове и находящихся в собственности субъекта Федерации?" и получить ответ - 48. В системе, обеспечивающей стадию свободного поиска, пользователь может поступить иначе и запросить у системы найти что-нибудь интересное относительно того, что влияет на атрибут Возраст. Система начнет действовать так же, как и аналитик-человек, т. е. искать аномалии в распределении значений атрибутов, в результате чего будет произведен список логических правил типа "ЕСЛИ ..., ТО ...", в том числе, например:

   ЕСЛИ Профессия="Программист", ТО Возраст<=30 в 61% случаев;

   ЕСЛИ Профессия="Программист", ТО Возраст<=60 в 98% случаев.

Аналогично, при исследовании реестра юридических лиц аналитика может заинтересовать атрибут Форма\_собственности. В результате свободного поиска могут быть получены правила:

   ЕСЛИ Основной\_вид\_деятельности="Общеобразовательные детские школы",   
ТО Форма\_собственности="Муниципальная собственность" в 84% случаев;

   ЕСЛИ Вид\_деятельности="Наука и научное обслуживание",   
ТО Форма\_собственности="Частная собственность" в 73% случаев.

Стадия свободного поиска может выполняться посредством:

   индукции правил условной логики (как в приведенных примерах) - с их помощью, в частности, могут быть компактно описаны группы похожих обучающих примеров в задачах классификации и кластеризации;

   индукции правил ассоциативной логики - то есть того, что было определено в рамках классификации задач ИАД по типам извлекаемой информации как выявление ассоциаций и последовательностей;

   определения трендов и колебаний в динамических процессах, то есть исходного этапа задачи прогнозирования.

Стадия свободного поиска, как правило, должна включать в себя не только генерацию закономерностей, но и проверку их достоверности на множестве данных, не принимавшихся в расчет при их формулировании.

**2. Прогностическое моделирование (Predictive Modeling)**

Здесь, на второй стадии ИАД, используются плоды работы первой, то есть найденные в БД закономерности применяются для предсказания неизвестных значений:

   при классификации нового объекта мы можем с известной уверенностью отнести его к определенной группе результатов рассмотрения известных значений его атрибутов;

   при прогнозировании динамического процесса результаты определения тренда и периодических колебаний могут быть использованы для вынесения предположений о вероятном развитии некоторого динамического процесса в будущем.

Возвращаясь к рассмотренным примерам, продолжим их на данную стадию. Зная, что некто Иванов - программист, можно быть на 61% уверенным, что его возраст <=30 годам, и на 98% - что он <=60 годам. Аналогично, можно сделать заключение о 84% вероятности того, что некоторое новое юридическое лицо будет находиться в муниципальной собственности, если его основной вид деятельности - "Общеобразовательные детские школы".

Следует отметить, что свободный поиск раскрывает общие закономерности, т. е. индуктивен, тогда как любой прогноз выполняет догадки о значениях конкретных неизвестных величин, следовательно, дедуктивен. Кроме того, результирующие конструкции могут быть как прозрачными, т. е. допускающими разумное толкование (как в примере с произведенными логическими правилами), так и нетрактуемыми - "черными ящиками" (например, про построенную и обученную нейронную сеть никто точно не знает, как именно она работает).

**3. Анализ исключений (Forensic Analysis)**

Предметом данного анализа являются аномалии в раскрытых закономерностях, то есть необъясненные исключения. Чтобы найти их, следует сначала определить норму (стадия свободного поиска), вслед за чем выделить ее нарушения. Так, определив, что 84% общеобразовательных школ отнесены к муниципальной форме собственности, можно задаться вопросом - что же входит в 16%, составляющих исключение из этого правила? Возможно, им найдется логическое объяснение, которое также может быть оформлено в виде закономерности. Но может также статься, что мы имеем дело с ошибками в исходных данных, и тогда анализ исключений может использоваться в качестве инструмента очистки сведений в хранилище данных.

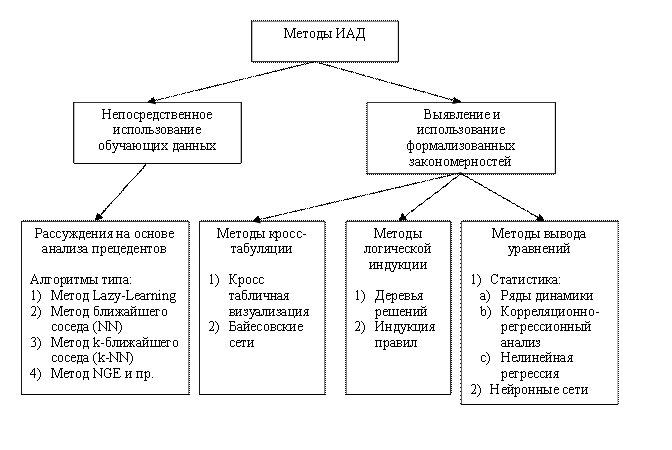
**3. Методы ИАД**

Все методы ИАД подразделяются на две большие группы по принципу работы с исходными обучающими данными.

В первом случае исходные данные могут храниться в явном детализированном виде и непосредственно использоваться для прогностического моделирования и/или анализа исключений; это так называемые методы рассуждений на основе анализа прецедентов. Главной проблемой этой группы методов является затрудненность их использования на больших объемах данных, хотя именно при анализе больших хранилищ данных методы ИАД приносят наибольшую пользу.

Во втором случае информация вначале извлекается из первичных данных и преобразуется в некоторые формальные конструкции (их вид зависит от конкретного метода). Согласно предыдущей классификации, этот этап выполняется на стадии свободного поиска, которая у методов первой группы в принципе отсутствует. Таким образом, для прогностического моделирования и анализа исключений используются результаты этой стадии, которые гораздо более компактны, чем сами массивы исходных данных. При этом полученные конструкции могут быть либо "прозрачными" (интерпретируемыми), либо "черными ящиками" (нетрактуемыми).

Две эти группы и примеры входящих в них методов представлены на рисунке 3.



**Рисунок 3.**   Классификация технологических методов ИАД

**1. Непосредственное использование обучающих данных**

Обобщенный алгоритм Lazy-Learning, относящийся к рассматриваемой группе, выглядит так: на вход классификатора подается пример, на выходе ожидается предсказание включающего его класса. Каждый пример представляется точкой в многомерном пространстве свойств (атрибутов), принадлежащей некоторому классу. Каждый атрибут принимает непрерывные значения либо дискретные значения из фиксированного набора. Для примера возвращается его наиболее вероятный класс.

Индивидуальной особенностью алгоритма k-ближайшего соседа является метод определения в нем апостериорной вероятности принадлежности примера классу.

Функция близости, определяемая как  множество k ближайших соседей примера во множестве известных обучающих примеров, близость которых к классифицируемому примеру определяется функцией расстояния.

Метод ближайшего соседа является частным случаем метода k-ближайшего соседа при k=1.

Более сложные алгоритмы типа Lazy-Learning основываются на том же обобщенном алгоритме, но или иначе определяют апостериорные вероятности принадлежности примеров классам, или (как, например, Nested Generalized Exemplars Algoritm(NGE)) усложняют расчет функции.

Особенность этой группы методов состоит в том, что предсказание неизвестных значений выполняется на основе явного сравнения нового объекта (примера) с известными примерами. В случае большого количества обучающих примеров, чтобы не сканировать последовательно все обучающее множество для классификации каждого нового примера, иногда используется прием выборки относительно небольшого подмножества "типичных представителей" обучающих примеров, на основе сравнения с которыми и выполняется классификация. Однако, этим приемом следует пользоваться с известной осторожностью, так как в выделенном подмножестве могут не быть отражены некоторые существенные закономерности.

Что касается самого известного представителя этой группы - метода k-ближайшего соседа, - он более приспособлен к тем предметным областям, где атрибуты объектов имеют преимущественно численный формат, так как определение расстояния между примерами в этом случае является более естественным, чем для дискретных атрибутов.

**2. Выявление и использование формализованных закономерностей**

Методы этой группы извлекают общие зависимости из множества данных и позволяют затем применять их на практике. Они отличаются друг от друга:

   по типам извлекаемой информации (которые определяются решаемой задачей);

   по способу представления найденных закономерностей.

Формализм, выбранный для выражения закономерностей, позволяет выделить три различных подхода, каждый из которых уходит своими корнями в соответствующие разделы математики:

   методы кросс-табуляции;

   методы логической индукции;

   методы вывода уравнений.

Логические методы наиболее универсальны в том смысле, что могут работать как с численными, так и с другими типами атрибутов. Построение уравнений требует приведения всех атрибутов к численному виду, тогда как кросс-табуляция, напротив, требует преобразования каждого численного атрибута в дискретное множество интервалов.

**Методы кросс-табуляции**

Кросс-табуляция является простой формой анализа, широко используемой в генерации отчетов средствами систем оперативной аналитической обработки (OLAP). Двумерная кросс-таблица представляет собой матрицу значений, каждая ячейка которой лежит на пересечении значений атрибутов. Расширение идеи кросс-табличного представления на случай гиперкубической информационной модели является, как уже говорилось, основой многомерного анализа данных, поэтому эта группа методов может рассматриваться как симбиоз многомерного оперативного анализа и интеллектуального анализа данных.

Кросс-табличная визуализация является наиболее простым воплощением идеи поиска информации в данных методом кросс-табуляции. Строго говоря, этот метод не совсем подходит под отмеченное свойство ИАД - переход инициативы к системе в стадии свободного поиска. На самом деле кросс-табличная визуализация является частью функциональности OLAP. Здесь система только предоставляет матрицу показателей, в которой аналитик может увидеть закономерность. Но само предоставление такой кросс-таблицы имеет целью поиск "шаблонов информации" в данных для поддержки принятия решений, то есть удовлетворяет приведенному определению ИАД. Поэтому неслучайно, что множество авторов все же относит кросс-табличную визуализацию к методам ИАД.

К методам ИАД группы кросс-табуляции относится также использование байесовских сетей (Bayesian Networks),  в основе которых лежит теорема Байеса теории вероятностей для определения апостериорных вероятностей составляющих полную группу попарно несовместных событий по их априорным вероятностям. Байесовские сети активно использовались для формализации знаний экспертов в экспертных системах, но с недавних пор стали применяться в ИАД для извлечения знаний из данных.

Можно отметить четыре достоинства байесовских сетей как средства ИАД:

  поскольку в модели определяются зависимости между всеми переменными, легко обрабатываются ситуации, когда значения некоторых переменных неизвестны;

  построенные байесовские сети просто интерпретируются и позволяют на этапе прогностического моделирования легко производить анализ по сценарию "что - если";

  подход позволяет естественным образом совмещать закономерности, выведенные из данных, и фоновые знания, полученные в явном виде (например, от экспертов);

  использование байесовских сетей позволяет избежать проблемы переподгонки (overfitting), то есть избыточного усложнения модели, чем страдают многие методы (например, деревья решений и индукция правил) при слишком буквальном следовании распределению зашумленных данных.

  Байесовские сети предлагают простой наглядный подход ИАД и широко используются на практике.

**Методы логической индукции**

Методы данной группы являются, пожалуй, наиболее выразительными, в большинстве случаев оформляя найденные закономерности в максимально "прозрачном" виде. Кроме того, производимые правила, в общем случае, могут включать как непрерывные, так и дискретные атрибуты. Результатами применения логической индукции могут быть построенные деревья решений или произведенные наборы символьных правил.

***Деревья решений***

Деревья решений являются упрощенной формой индукции логических правил. Основная идея их использования заключается в последовательном разделении обучающего множества на основе значений выбранного атрибута, в результате чего строится дерево, содержащее:

  терминальные узлы (узлы ответа), задающие имена классов;

  нетерминальные узлы (узлы решения), включающие тест для определенного атрибута с ответвлением к поддереву решений для каждого значения этого атрибута.

В таком виде дерево решений определяет классификационную процедуру естественным образом: любой объект связывается с единственным терминальным узлом. Эта связь начинается с корня, проходит путь по дугам, которым соответствуют значения атрибутов, и доходит до узла ответа с именем класса.

***Индукция правил***

Популярность деревьев решений проистекает из быстроты их построения и легкости использования при классификации. Более того, деревья решений могут быть легко преобразованы в наборы символьных правил - генерацией одного правила из каждого пути от корня к терминальной вершине. Однако, правила в таком наборе будут неперекрывающимися, потому что в дереве решений каждый пример может быть отнесен к одному и только к одному терминальному узлу. Более общим (и более реальным) является случай существования теории, состоящей из набора неиерархических перекрывающихся символьных правил. Значительная часть алгоритмов, выполняющих индукцию таких наборов правил, объединяются стратегией отделения и захвата (separate-and-conquer), или покрывания (covering). Эта стратегия индукции характеризуется следующим образом:

  произвести правило, покрывающее часть обучающего множества;

  удалить покрытые правилом примеры из обучающего множества (отделение);

  последовательно обучиться другим правилам, покрывающим группы оставшихся примеров (захват), пока все примеры не будут объяснены.

*Сравнение возможностей деревьев решений и индукции правил*

Индукция правил и деревья решений, будучи способами решения одной задачи, значительно отличаются по своим возможностям. Несмотря на широкую распространенность деревьев решений, индукция правил по ряду причин представляется более предпочтительным подходом.

1.     Деревья решений часто довольно сложны и тяжелы для понимания.

2.     Непременное требование неперекрываемости правил в алгоритмах обучения деревьев решений навязывает жесткое ограничение на возможность выражения существующих закономерностей. Одна из проблем, вытекающих из этого ограничения - проблема дублированного поддерева. Часто случается, что идентичные поддеревья оказываются в процессе обучения в разных местах дерева решений вследствие фрагментации пространства исходных примеров, обязательной по ограничению на неперекрываемость правил. Индукция отделения и захвата не ставит такого ограничения и, следовательно, менее чувствительна к этой проблеме.

3.     Построение деревьев решений затруднено при большом количестве исходной информации (что чаще всего имеет место при интеллектуальном анализе хранилищ данных). Для решения этой проблемы часто выделяют относительно небольшое подмножество имеющихся обучающих примеров и на его основе сооружают дерево решений. Такой подход во многих случаях приводит к потере информации, скрытой в проигнорированных при индукции примерах.

С другой стороны, индукция правил осуществляется значительно более сложными (и медленными) алгоритмами, чем индукция деревьев решений. Особенно большие трудности возникают с поступрощением построенной теории, в отличие от простоты подрезания деревьев решений: отсечение ветвей в дереве решений никогда не затронет соседние ветви, тогда как отсечение условий правила оказывает влияние на все перекрывающиеся с ним правила.

С другой стороны, отсечение условий от правила означает его обобщение, то есть в новом виде оно будет покрывать больше положительных и больше отрицательных примеров. Следовательно, эти дополнительные положительные и отрицательные примеры должны быть исключены из обучающего множества, дабы не воздействовать на индукцию последующих правил

Следовательно, исходя из проведенного сравнения, можно заключить, что построение деревьев решений оправдано в несложных задачах при небольшом количестве исходной информации благодаря простоте и быстроте их индукции. Однако при анализе больших объемов данных, накопленных в хранилищах, использование методов индукции правил предпочтительнее, несмотря на их относительную сложность.

**3. Методы вывода уравнений**

Методы вывода уравнений пытаются выразить закономерности, скрытые в данных, в форме математических выражений. Поэтому они способны работать только с атрибутами численного типа, тогда как другие атрибуты должны быть искусственно закодированы численными значениями. Отсюда вытекает несколько проблем, ограничивающих использование этих методов на практике. Тем не менее, они широко применяются во многих приложениях.

*Статистика*

Классические методы статистического анализа применяются в средствах ИАД чаще всего для решения задачи прогнозирования.

1.     Выявление тенденций динамических рядов. Тенденцию среднего уровня можно представить в виде графика или аналитической функции, вокруг значения которой варьируют фактические значения уровней исследуемого процесса. Часто тенденции среднего уровня называют детерминированной компонентой процесса. Детерминированная компонента обычно представляется достаточно простой аналитической функцией - линейной, параболической, гиперболической, экспоненциальной, - параметры которой подбираются согласно историческим данным для лучшей аппроксимации исторических данных.

2.     Гармонический анализ. Во многих случаях сглаживание рядов динамики с помощью определения тренда не дает удовлетворительных результатов, так как в остатках наблюдается автокоppеляция. Причиной автокоppелиpованности остатков могут быть нередко встречающиеся в pядах динамики заметные периодические колебания относительно выделенной тенденции. В таких случаях следует прибегать к гармоническому анализу, то есть к выделению из динамического ряда периодической составляющей. По результатам выделения из динамического ряда тренда и периодической составляющей может выполняться статистический прогноз процесса по принципу экстраполяции, по предположению, что параметры тренда и колебаний сохранятся для прогнозируемого периода.

3.     Корреляционно-регрессионный анализ. В отличие от функциональной (жестко детерминированной) связи, статистическая (стохастически детерминированная) связь между переменными имеет место тогда, когда с изменением значения одной из них вторая может в определенных пределах принимать любые значения с некоторыми вероятностями, но ее среднее значение или иные статистические характеристики изменяются по определенному закону. Частным случаем статистической связи, когда различным значениям одной переменной соответствуют различные средние значения другой, является корреляционная связь. Метод корреляционно-регрессионного анализа хорошо изучен и широко применяется на практике. Получаемые в результате применения анализа корреляционно-регрессионные модели (КРМ) обычно достаточно хорошо интерпретируемы и могут использоваться в прогностическом моделировании. Но невозможно применять этот вид анализа, не имея глубоких знаний в области статистики. Теоретическая подготовка аналитика играет здесь особенно важную роль, поэтому немногие существующие средства ИАД предлагают метод корреляционно-регрессионного анализа в качестве одного из инструментов обработки данных.

4.     Корреляция рядов динамики. Проблема изучения причинных связей во времени очень сложна, и полное решение всех задач такого изучения до сих пор не разработано. Основная сложность состоит в том, что при наличии тренда за достаточно длительный промежуток времени большая часть суммы квадратов отклонений связана с трендом; при этом, если два признака имеют тренды с одинаковым направлением изменения уровней, то это вовсе не будет означать причинной зависимости. Следовательно, чтобы получить реальные показатели корреляции, необходимо абстрагироваться от искажающего влияния трендов - вычислить отклонения от трендов и измерить корреляцию колебаний. Однако, не всегда допустимо переносить выводы о тесноте связи между колебаниями на связь рядов динамики в целом.

*Нейронные сети*

Искусственные нейронные сети как средство обработки информации моделировались по аналогии с известными принципами функционирования биологических нейронных сетей. Их структура базируется на следующих допущениях:

  обработка информации осуществляется во множестве простых элементов - нейронов;

  сигналы между нейронами передаются по связям от выходов ко входам;

  каждая связь характеризуется весом, на который умножается передаваемый по ней сигнал;

  каждый нейрон имеет активационную функцию (как правило, нелинейную), аргумент которой рассчитывается как сумма взвешенных входных сигналов, а результат считается выходным сигналом.

Таким образом, нейронные сети представляют собой наборы соединенных узлов, каждый из которых имеет вход, выход и активационную функцию (как правило, нелинейную). Они обладают способностью обучаться на известном наборе примеров обучающего множества. Обученная нейронная сеть представляет собой "черный ящик" (нетрактуемую или очень сложно трактуемую прогностическую модель), которая может быть применена в задачах классификации, кластеризации и прогнозирования.

  Обучение нейронной сети заключается в подстройке весовых коэффициентов, связывающих выходы одних нейронов со входами других. Обучение сети может производиться по одному из двух базовых сценариев:

  обучение с учителем (supervised training) - наиболее типичный случай, когда для каждого вектора значений входных переменных примера обучающего множества известен желаемый вектор значений выходных переменных; такой способ обучения применяется в задачах классификации и прогнозирования;

  обучение без учителя (unsupervised learning) - механизм настройки весов сети в случае, когда известны только значения входных переменных примеров обучающего множества; обученные таким способом нейронные сети выполняют задачу кластеризации.

Имеется ряд недостатков, ограничивающих использование нейронных сетей в качестве инструмента ИАД.

1.     Обученные нейронные сети являются нетрактуемыми моделями - "черными ящиками", поэтому логическая интерпретация описанных ими закономерностей практически невозможна (за исключением простейших случаев).

2.     Будучи методом группы вывода уравнений, нейронные сети могут обрабатывать только численные переменные. Следовательно, переменные других типов, как входные, так и выходные, должны быть закодированы числами. При этом недостаточно заменить переменную, принимающую значения из некоторой области определения, одной численной переменной, так как в этом случае могут быть получены некорректные результаты. Таким образом, при большом количестве нечисловых переменных с большим количеством возможных значений использование нейронных сетей становится совершенно невозможным.

Главной проблемой обучения нейронных сетей является синтез структуры сети, способной обучиться на заданном обучающем множестве. Нет никакой гарантии, что процесс обучения сети определенной структуры не остановится, не достигнув допустимого порога ошибки, или не попадет в локальный минимум. Хотя многослойные сети широко применяются для классификации и аппроксимации функций, их структурные параметры до сих пор должны определяться путем проб и ошибок.

Таким образом, нейронные сети - довольно мощный и гибкий инструмент ИАД - должны применяться с известной осторожностью и подходят не для всех проблем, требующих интеллектуального анализа корпоративных данных.

**4. Типы закономерностей**

Выделяют пять стандартных типов закономерностей, которые позволяют выявлять  методы Data Mining: ассоциация, последовательность, классификация, кластеризация и прогнозирование.

***Ассоциация*** имеет место в том случае, если несколько событий связаны друг с другом. Например, исследование, проведенное в супермаркете, может показать, что 65% купивших кукурузные чипсы берут также и "кока-колу", а при наличии скидки за такой комплект "колу" приобретают в 85% случаев. Располагая сведениями о подобной ассоциации, менеджерам легко оценить, насколько действенна предоставляемая скидка.

Если существует цепочка связанных во времени событий, то говорят о ***последовательности***. Так, например, после покупки дома в 45% случаев в течение месяца приобретается и новая кухонная плита, а в пределах двух недель 60% новоселов обзаводятся холодильником.

С помощью ***классификации*** выявляются признаки, характеризующие группу, к которой принадлежит тот или иной объект. Это делается посредством анализа уже классифицированных объектов и формулирования некоторого набора правил.

***Кластеризация*** отличается от классификации тем, что сами группы заранее не заданы. С помощью кластеризации средства Data Mining самостоятельно выделяют различные однородные группы данных.

Основой для всевозможных систем ***прогнозирования***служит историческая информация, хранящаяся в БД в виде временных рядов. Если удается построить, найти шаблоны, адекватно отражающие динамику поведения целевых показателей, есть вероятность, что с их помощью можно предсказать и поведение системы в будущем.

Особенно широко методы ИАД применяются в бизнес-приложениях аналитиками и руководителями компаний. Для этих категорий пользователей разрабатываются инструментальные средства высокого уровня, позволяющие решать достаточно сложные практические задачи без специальной математической подготовки. Актуальность использования ИАД в бизнесе связана с жесткой конкуренцией, возникшей вследствие перехода от «рынка производителя» к «рынку потребителя». В этих условиях особенно важно качество и обоснованность принимаемых решений, что требует строгого количественного анализа имеющихся данных. При работе с большими объемами накапливаемой информации необходимо постоянно оперативно отслеживать динамику рынка, а это практически невозможно без автоматизации аналитической деятельности.

5. **Типовые задачи для методов ИАД**

***Прогнозирование*** – одна из самых распространенных задач ИАД. В частности, при планировании и составлении бюджета необходимо прогнозировать объемы продаж и другие параметры с учетом многочисленных взаимосвязанных факторов – сезонных, региональных, общеэкономических и т.д. Можно также выявлять корреляции в продажах, например «покупке компьютера, как правило, сопутствует покупка блока бесперебойного питания».

***Маркетинговый анализ***. Чтобы разработать эффективный маркетинговый план, нужно знать, каким образом на уровень продаж влияют такие факторы как стоимость товара, затраты на продвижение продукции и рекламу. Нейросетевые модели позволяют менеджерам и аналитикам прогнозировать подобное влияние.

***Анализ работы персонала***. Производительность труда служащих зависит от уровня подготовки, от оплаты труда, опыта работы, взаимоотношений с руководством и т.д. Проанализировав влияние этих факторов, можно выработать методику повышения производительности труда, а также предложить оптимальную стратегию подбора кадров в будущем.

***Анализ эффективности продажи товаров по почте***. Если компания занимается рассылкой рекламы и образцов продукции по почте, то имеет смысл оценить эффективность подобной деятельности. При этом можно выявить круг потенциальных покупателей, и оценить вероятность совершения ими покупки. Кроме того, можно опробовать различные формы переписки и выбрать наиболее удачные.

***Профилирование клиентов***. С помощью нейросетевых моделей можно среди многочисленных клиентов фирмы выбрать тех, сотрудничество с которыми наиболее выгодно – получить портрет «типичного клиента компании». Кроме того, можно выяснить, почему работа с некоторыми из заказчиков стала неэффективной, и выработать стратегию поиска подходящих клиентов в будущем. Эта задача чаще всего решается менеджерами по продажам, а для банков, например, особый интерес представляет возможность оценки кредитоспособности клиентов.

***Оценка потенциальных клиентов***. Планируя предварительные переговоры имеет смысл определить, с какой долей вероятности они закончатся заключением договора (или продажей продукции). Анализ опыта работы с клиентами позволяет выявить характерные особенности тех заявок, которые закончились реальными продажами. Используя результаты данного анализа, менеджеры могут остановиться на более перспективных заявках клиентов.

***Анализ результатов маркетинговых исследований***. Чтобы оценить реакцию покупателей на политику компании в области распространения продукции, ценообразования, а также на характеристики самой продукции, необходимо, наряду с анализом продаж, проводить опросы покупателей. Это позволяет усовершенствовать процесс принятия решений по ценам и характеристикам выпускаемой продукции (дизайн, функциональность, упаковка).

***Анализ работы региональных отделений компании.*** С помощью нейросетевых моделей можно сравнивать результаты деятельности региональных отделений или филиалов компании и определять, от чего зависит эффективность их работы (географическое положение, численность персонала, ассортимент продукции/услуг и т.д.). Результаты используются для оптимизации работы «отстающих» отделений, а также при планировании создания новых филиалов.

***Сравнительный анализ конкурирующих фирм***. Почему некоторые компании процветают и удерживают прочные позиции на рынке, а другие нет? Какие сферы бизнеса самые выгодные? Чтобы ответить на эти вопросы можно сравнить деятельность конкурирующих компаний и выяснить, какие факторы определяют прибыльность их бизнеса.

Очевидно, что перечисленные виды задач актуальны практически для всех отраслей бизнеса: банковского дела и страхования (выявление злоупотреблений с кредитными карточками, оценка кредитных рисков, оценка закладных, выявление профилей пользователей, оценка эффективности региональных отделений, вероятность подачи заявки на выплату страховки и др.), финансовых рынков (прогнозирование, анализ портфелей, моделирование индексов), производства (прогнозирование спроса, контроль качества, оценка дизайна продукции), торговли и т. д.

**Почему растет популярность Data mining?**

Необходимость автоматизированного интеллектуального анализа данных стала очевидной в

1.     ***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. Человеческий ум, даже такой тренированный, как ум профессионального аналитика, просто не в состоянии своевременно анализировать столь огромные информационные потоки.

2.     ***Другой причиной роста популярности 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.

3.     ***И, наконец, data mining дешевле***. Оказывается, что выгоднее инвестировать деньги в решения data mining, чем постоянно содержать целую армию высокоподготовленных и дорогих профессиональных статистиков.

Data mining 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.

**6. Области применения Data mining**

|  |  |  |
| --- | --- | --- |
| 1. | Маркетинг | Рыночная сегментация, идентификация целевых групп, построение профиля клиента |
| 2. | Банковское дело: | Анализ кредитных рисков, привлечение и удержание клиентов, управление ресурсами |
| 3. | Кредитные компании | Детекция подлогов, формирование "типичного поведения" обладателя кредитки, анализ достоверности клиентских счетов, cross-selling программы |
| 4. | Страховые компании | Привлечение и удержание клиентов, прогнозирование финансовых показателей |
| 5. | Розничная торговля | Анализ деятельности торговых точек, построение профиля покупателя, управление ресурсами |
| 6. | Биржевые трейдеры | Выработка оптимальной торговой стратегии, контроль рисков |
| 7. | Телекоммуникация и энергетика | Привлечение клиентов, ценовая политика, анализ отказов, предсказание пиковых нагрузок, прогнозирование поступления средств |
| 8. | Налоговые службы и аудиторы | Детекция подлогов, прогнозирование поступлений в бюджет |
| 9. | Фармацевтические компании | Предсказание результатов будущего тестирования препаратов, программы испытания |
| 10. | Медицина | Диагностика, выбор лечебных воздействий, прогнозирование исхода хирургического вмешательства |
| 11. | Управление производством | Контроль качества, материально-техническое обеспечение, оптимизация технологического процесса |
| 12. | Ученые и инженеры | Построение эмпирических моделей, основанных на анализе данных, решение научно-технических задач |

Рассмотрим некоторые бизнес-приложения Data Mining.

***Розничная торговля***

Предприятия розничной торговли сегодня собирают подробную информацию о каждой отдельной покупке, используя кредитные карточки с маркой магазина и компьютеризованные системы контроля. Вот типичные задачи, которые можно решать с помощью Data Mining в сфере розничной торговли:

  *анализ покупательской корзины* (анализ сходства) 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* помогает торговым предприятиям принимать 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***

Достижения технологии Data Mining используются в банковском деле для решения следующих распространенных задач:

  *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***

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.

***Другие приложения в бизнесе***

Data Mining может применяться во множестве других областей:

  *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.

**7. Классы систем Data Mining**

Data Mining является мультидисциплинарной областью, возникшей и развивающейся на базе достижений прикладной статистики, распознавания образов, методов искусственного интеллекта, теории баз данных и др. Отсюда обилие методов и алгоритмов, реализованных в различных действующих системах Data Mining. Многие из таких систем интегрируют в себе сразу несколько подходов. Тем не менее, как правило, в каждой системе имеется какая-то ключевая компонента, на которую делается главная ставка. Приведем классификацию указанных ключевых компонент с краткой характеристикой для каждого  класса.

**Индустриальные системы Data mining**

В настоящее время большинство ведущих в мире производителей программного обеспечения предлагает свои продукты и решения в области Data mining. Как правило - это масштабируемые системы, в которых реализованы различные математические алгоритмы анализа данных. Они имеют развитый графический интерфейс, богатые возможности в визуализации и манипулирования с данными, предоставляют доступ к различным источникам данных, функционирующих в архитектуре клиент/сервер на Intel или UNIX платформах. Вот несколько примеров таких систем:

  [PolyAnalyst](http://www.megaputer.ru/doc.php?production/dm/products_dm.html) (Мегапьютер Интеллидженс)

  Intelligent Miner (IBM)

  Interprise Miner (SAS)

  Clementine (Integral Solutions)

  MineSet (Silicon Graphics)

  Knowledge Studio (Angoss Software)

**Предметно-ориентированные аналитические системы**

Предметно-ориентированные аналитические системы очень разнообразны. Эти системы решают узкий класс специализированных задач. Наиболее широкий подкласс таких систем, получивший распространение в области исследования финансовых рынков, носит название "технический анализ". Он представляет собой совокупность нескольких десятков методов прогноза динамики цен и выбора оптимальной структуры инвестиционного портфеля, основанных на различных эмпирических моделях динамики рынка. Эти методы часто используют несложный статистический аппарат, но максимально учитывают сложившуюся в своей области специфику (профессиональный язык, системы различных индексов и пр.).

На рынке имеется множество программ этого класса. Как правило, они довольно дешевы (обычно $300–1000).

Хорошим примером являются программы технического анализа финансовых рынков:

  MetaStock (Equis International, USA)

  SuperCharts (Omega Research, USA)

  Candlestick Forecaster (IPTC, USA)

  Wall Street Money (Market Arts, USA)

**Статистические пакеты**

Это мощные математические системы, предназначенные для статистической обработки данных любой природы. Они включают многочисленные инструменты статистического анализа, имеют развитые графические средства. Главный недостаток систем этого класса - их невозможно эффективно применять для анализа данных, не имея глубоких знаний в области статистики. Неподготовленный пользователь должен пройти специальный курс обучения.

Примеры систем:

  SAS (SAS Institute, USA)

  SPSS (SPSS, USA)

  Statgraphics (Statistical Graphics, USA)

**Нейроннoсетевые пакеты**

Это широкий класс разнообразных систем, представляющих собой иерархические сетевые структуры, в узлах которых находятся так называемые нейроны. Сети тренируются на примерах, и во многих случаях дают хорошие результаты предсказаний. Основными недостатками нейронных сетей являются необходимость иметь очень большой объем обучающей выборки, а также  трудности в интерпретации результатов. Тренированная нейронная сеть представляет собой "умный черный ящик", работу которого невозможно понять и контролировать.

Примеры нейронно-сетевых пакетов:

  BrainMaker (CSS, USA)

  NeuroShell (Ward Systems Group, USA)

  OWL (Hyperlogic, USA)

**Пакеты, реализующие алгоритмы "Decision trees"**

Деревья решения являются одним из наиболее популярных подходов к решению задач Data Mining. Этот метод используется только для решения задач классификации. Это является его серьезным ограничением. Результатом работы метода является иерархическая древовидная структура классификационных правил типа "IF...THEN...". 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 A больше x?". Если ответ положительный, осуществляется переход к правому узлу следующего уровня, если отрицательный — то к левому узлу; затем снова следует вопрос, связанный с соответствующим узлом. Достоинством метода является естественная способность классификации на множество классов.

Примеры систем:

  C5.0 (Rule Quest, Australia)

  SIPINA (University of Lyon, France)

  IDIS (Information Discovery, USA)

**Системы рассуждений на основе аналогичных случаев (case based reasoning - CBR)**

Идея систем case based reasoning - CBR - 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 называется методом "ближайшего соседа" (nearest neighbour). Системы CBR показывают очень хорошие результаты в самых разнообразных задачах.

Главный их минус заключается в том, 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:

  KATE tools (Acknosoft, Франция),

  Pattern Recognition Workbench (Unica, США).

**Генетические алгоритмы**

Строго говоря, интеллектуальный анализ данных - далеко не основная область применения генетических алгоритмов, которые, скорее, нужно рассматривать как мощное средство решения разнообразных комбинаторных задач и задач оптимизации. Тем не менее, генетические алгоритмы вошли сейчас в стандартный инструментарий методов data mining. Этот метод назван так потому, что в какой-то степени имитирует процесс естественного отбора в природе.

Первый шаг при построении генетических алгоритмов — это кодировка исходных логических закономерностей в базе данных, 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.

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.

Пример системы:

GeneHunter фирмы Ward Systems Group.

**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 некотором 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. 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.

**8. Интеграция OLAP и ИАД**

Оперативная аналитическая обработка и интеллектуальный анализ данных - две составные части процесса поддержки принятия решений. Но сегодня большинство систем OLAP заостряет внимание только на обеспечении доступа к многомерным данным, а большинство средств ИАД, работающих в сфере закономерностей, имеют дело с одномерными перспективами данных. Эти два вида анализа должны быть тесно объединены, то есть системы OLAP должны фокусироваться не только на доступе, но и на поиске закономерностей. Как заметил N. Raden, "многие компании создали ... прекрасные хранилища данных, идеально разложив по полочкам горы неиспользуемой информации, которая сама по себе не обеспечивает ни быстрой, ни достаточно грамотной реакции на рыночные события".

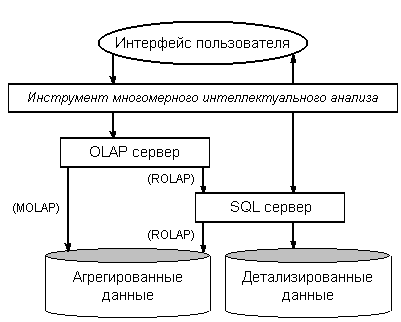
K. Parsaye вводит составной термин "OLAP Data Mining" (многомерный интеллектуальный анализ) для обозначения такого объединения (рис. 4). J. Han предлагает еще более простое название - "OLAP Mining", и предлагает несколько вариантов интеграции двух технологий.

1.     "Cubing then mining". Возможность выполнения интеллектуального анализа должна обеспечиваться над любым результатом запроса к многомерному концептуальному представлению, то есть над любым фрагментом любой проекции гиперкуба показателей.

2.     "Mining then cubing". Подобно данным, извлечённым из хранилища, результаты интеллектуального анализа должны представляться в гиперкубической форме для последующего многомерного анализа.

3.     "Cubing while mining". Этот гибкий способ интеграции позволяет автоматически активизировать однотипные механизмы интеллектуальной обработки над результатом каждого шага многомерного анализа (перехода между уровнями обобщения, извлечения нового фрагмента гиперкуба и т. д.).

К сожалению, очень немногие производители предоставляют сегодня достаточно мощные средства интеллектуального анализа многомерных данных в рамках систем OLAP. Проблема также заключается в том, что некоторые методы ИАД (байесовские сети, метод k-ближайшего соседа) неприменимы для задач многомерного интеллектуального анализа, так как основаны на определении сходства детализированных примеров и не способны работать с агрегированными данными.



**Рисунок 4.** Архитектура системы многомерного интеллектуального анализа данных

Рынок систем Data Mining экспоненциально развивается. В этом развитии принимают участие практически все крупнейшие корпорации.

Системы Data Mining применяются по двум основным направлениям:

1)     как массовый продукт для бизнес-приложений;

2)     как инструменты для проведения уникальных исследований (генетика, химия, медицина и пр.).

Лидеры Data Mining связывают будущее этих систем с использованием их в качестве интеллектуальных приложений, встроенных в корпоративные хранилища данных.

Несмотря на обилие методов Data Mining, приоритет постепенно все более смещается в сторону логических алгоритмов поиска в данных if-then правил. С их помощью решаются задачи прогнозирования, классификации, распознавания образов, сегментации БД, извлечения из данных “скрытых” знаний, интерпретации данных, установления ассоциаций в БД и др. Результаты таких алгоритмов эффективны и легко интерпретируются.

Вместе с тем, главной проблемой логических методов обнаружения закономерностей является проблема перебора вариантов за приемлемое время. Известные методы либо искусственно ограничивают такой перебор,  либо строят деревья решений, имеющих принципиальные ограничения эффективности поиска if-then правил. Другие проблемы связаны с тем, что известные методы поиска логических правил не поддерживают функцию обобщения найденных правил и функцию поиска оптимальной композиции таких правил. Удачное решение указанных проблем может составить предмет новых конкурентоспособных разработок.

Компьютерные технологии автоматического интеллектуального анализа данных переживают бурный расцвет. На российском рынке эта технология пока делает лишь первые шаги. Отчасти это можно объяснить высокой стоимостью систем data mining, но, как показывает история развития других сегментов компьютерного рынка России, сам по себе этот фактор вряд ли является определяющим. Скорее здесь проявляется действие некоторых специфичных для России негативных факторов, резко уменьшающих эффективность применения технологии data mining.

Начнем с характеристики российской специфики. Компьютерные системы поддержки принятия решений, в принципе, могут основываться на двух подходах.

Первый, более традиционный, заключается в том, что в системе фиксируется опыт эксперта, который и используется для выработки оптимального в данной ситуации решения.

Второй, системы пытаются найти решения на основе анализа исторических данных, описывающих поведение изучаемого объекта, принятые в прошлом решения, их результаты и т.д.

С этим очевидным обстоятельством связана главная трудность продвижения технологии data mining в России: отличительная черта большинства российских предприятий - сравнительно небольшой срок их существования. Характерный "возраст" накопленных ими баз данных составляет 2-3 года, и, как показывает опыт, информации, содержащейся в этих базах данных, часто оказывается недостаточно для выработки на ее основе эффективной стратегии принятия решений с помощью систем data mining.

Другой отличительной чертой российской экономики, как на макро-уровне, так и на уровне отдельных предприятий является ее нестабильность; кроме того, она подвержена и действию многочисленных неожиданно возникающих факторов. В то время как на Западе предприятия в основном работают в рамках уже устоявшейся законодательной базы, в сложившихся структурах товарных, финансовых и информационных потоков, российские предприятия вынуждены подстраиваться под постоянно меняющиеся правила игры.

Наконец, еще одно обстоятельство влияет на применение систем добычи знаний в российских условиях. Оно связано с тем, что люди, ответственные за принятие решений в бизнесе и финансах, обычно не являются специалистами по статистике и искусственному интеллекту и поэтому не могут непосредственно использовать системы интеллектуального анализа данных, требующие сложной настройки или специальной подготовки данных. Поэтому важными факторами, определяющими коммерческий успех систем интеллектуального анализа данных в России, являются простота в использовании и высокая степень автоматизма.

**Тема 7. Инструментальные средства автоматизации аналитической работы и планирования.**

**Содержание**

1.     Инструментальные средства поддержки аналитической работы и их классификация.

2.     Аналитические инструментальные средства пакетов прикладных программ широкого применения.

3.     Специализированные информационно-аналитические системы.

**1.   Инструментальные средства поддержки аналитической работы и их классификация**

         В реальности, аналитическая работа на современных предприятиях пока ещё носит частично автоматизированный характер ввиду ряда причин:

        отсутствия опыта у IT специалистов и знания рынка средств автоматизации поддержки аналитической деятельности и возможностей этих средств;

        дороговизна профессиональных средств поддержки аналитической деятельности;

        отсутствие формализации бизнес процессов, протекающих на предприятии, что делает невозможным регламентный сбор информации;

        сильная децентрализация собираемой первичной информации по отдельным системам;

        отсутствие интеграции между информационными системами, работающими в рамках разных предметных технологий внутри одного предприятия;

        и др.

Поэтому до сих пор разрабатываются и пользуются большой популярностью отдельные информационные системы, автоматизирующие части общей аналитической технологии. Исходя из понятия и содержания аналитической работы можно выделить следующие направления инструментов поддержки аналитической работы, автоматизирующих её в различных пределах:

        универсальные программные продукты, которые можно использовать для сбора и просмотра данных (excel, Access и др.);

        универсальные программные продукты, которые могут производить групповые действия с произвольными наборами данных (Excel, Statistica, Mathcad) – отсутствие мета-моделей предметной области и заранее подготовленного набора отчётов;

        отдельные программные продукты, которые могут автоматизировать не профильные для аналитической работы операции (например, копирование данных, рассылка подготовленных отчётов и т.д.);

        специализированные аналитические системы, поддерживающие ведение мета-моделей предметной области и содержащие репозиторий заранее подготовленных к использованию отчётов, но автоматизирующие лишь часть из общей концепции BI;

        полноценные профессиональные ИАС, автоматизирующие большинство аспектов и задач аналитической деятельности на предприятии.

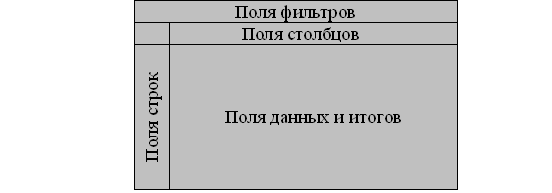
Остановимся более детально на некоторых системах.

**2 Аналитические инструментальные средства пакетов прикладных программ широкого применения**

         Скорее всего, никто не будет возражать против того, что пакет программ Microsoft Office является самым распространённым среди большинства пользователей ПК – все пользуются текстовым редактором Word для составления бланков, рефератов, дипломов, договоров, отчётов и других документов. Все также пользуются электронной таблице Excel для создания таблиц, отчётов, диаграмм и проведения расчётов. Но, умея пользовать базовой функциональностью Excel, лишь немногие знают, что в их руках находится гибкий аналитический инструментарий. Остановимся более подробно на его аналитических возможностях.

         Основным инструментом анализа внутри Excel являются сводные таблицы. Сводная таблица Excel (PivotTable) представляет собой интерактивную таблицу, применяемую для суммирования или статистического анализа большого количества исходных данных, обычно содержащихся в одном из диапазонов ячеек Excel, либо являющихся результатом запроса к какой–либо базе данных. Строки и столбцы подобной интерактивной таблицы базируются на данных из нескольких столбцов исходной таблицы. Ячейки сводной таблицы представляют собой суммы значений одного из числовых полей исходной таблицы. При этом суммирование производиться для тех строк исходной таблицы, в которых значения столбцов, образовавших столбцы и строки сводной таблицы, совпадают  с именами строк и столбцов сводной таблицы, на пересечении которых находится данная ячейка. В некоторых источниках сводные таблицы называют перекрестными таблицами или кросс-таблицами.

В Excel поля сводной таблицы можно менять местами, добавлять или удалять, а сами суммируемые данные можно фильтровать по значению других наборов столбцов исходной таблицы с целью более подробного рассмотрения какой-либо конкретной области этих значений.  Это позволяет создавать различные способы отображения суммарных значений для одной и той же исходной таблицы. Структура сводной таблицы изображена на рисунке 1.



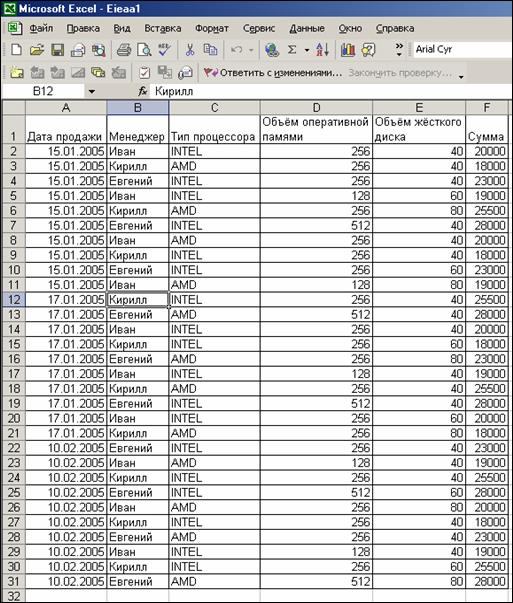
**Рисунок 1.** Структура сводной таблицы

Пользователь может изменить вид сводной таблицы, перемещая имена полей с панели инструментов сводной таблицы (PivotTable) на сводную таблицу или возвращая их обратно на инструментальную панель. Список полей исходной таблицы появляется на панели инструментов PivotTable в момент начала редактирования сводной таблицы и исчезает после прекращения ее редактирования.

Следует заметить, что сводная таблица Excel может быть построена на основании данных любого ODBC- или OLE DB-источника данных, а том числе данных OLAP – кубов, доступных с помощью  соответствующих OLE DB-провайдеров. Помимо того Excel допускает использование более одного поля исходной таблицы для формирования строк, столбцов и фильтров, что позволяет создавать сводные таблицы с иерархическими метаданными.

При необходимости в Excel можно построить сводную диаграмму, синхронизированную со сводной таблицей. Для этого достаточно нажать соответствующую кнопку на панели инструментов PivotTable, и, если нужно отредактировать внешний вид диаграммы. Отметим, что с помощью панелей инструментов, а также выпадающих списков на осях диаграммы и легенде, можно управлять отображением данных на сводной диаграмме, например, выполнять операцию drill-down, при этом сводная таблица будет меняться синхронно с диаграммой.

Для того, чтобы начать пользоваться аналитическими возможностями необходимо, чтобы в распоряжении Excel стали доступны какие-либо исходные данные. Источником данных может быть как собственный лист Excel, так и любой внешний источник (файлы с данными или СУБД, доступ к которым можно получить с использованием стандартных драйверов ODBC). Для упрощения знакомства будем использовать данные из собственного листа Excel (рис.2). На рисунке приведены данные о продаже компьютеров в салоне.



**Рисунок 2.** Данные на листе Excel

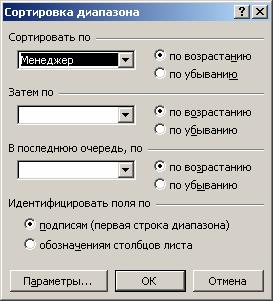
         Первое, что требуется рассчитать – это количество проданных компьютеров каждым из менеджеров. Вы наверняка скажите: «что тут считать». Всего лишь 30 строк – легко вручную пересчитать. А если строк будет не 30, а 30000 – что тогда делать? Для решения этой простейшей задачи можно воспользоваться функцией расчёта итогов (пункт меню Данные-> Итоги):

 выделяем все столбцы с данными;

  выполняем сортировку данных по полю менеджер:

o пункт меню Данные-> Сортировка;

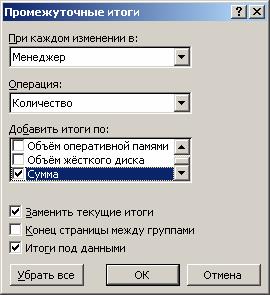
o в выпадающем списке «сортировать по» выбираем «менеджер»



o       нажимаем кнопку «ОК»

 снимаем выделение, установив курсор на любую ячейку с данными;

 выбираем пункт Данные -> Итоги;

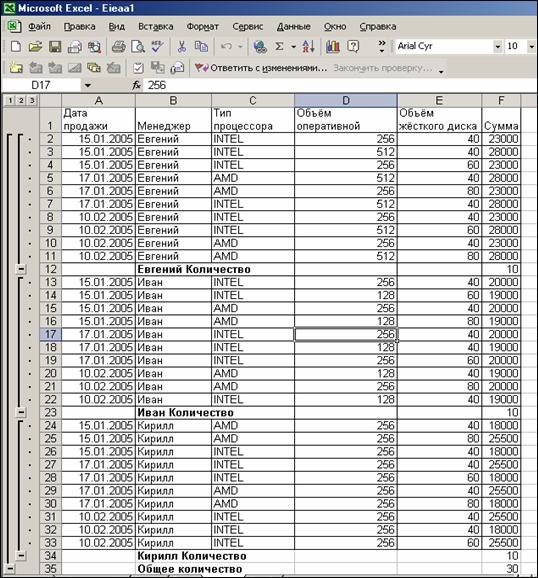


O выбираем в списке «При каждом изменении в:» поле «менеджер». (данная опция необходима для того, чтобы при построчном анализе исходных данных счётчик обнулялся при каждом изменении значения выбранного поля);

O в поле «операция» выбираем «количество», т.к. цель – посчитать количество проданных каждым менеджером компьютеров. Помимо количества возможно также использование операций сумма, среднее, максимум, минимум, произведение, количество чисел, смещённое отклонение, несмещённое отклонение, смещённая дисперсия, несмещённая дисперсия;

O нажимаем кнопку «ОК»

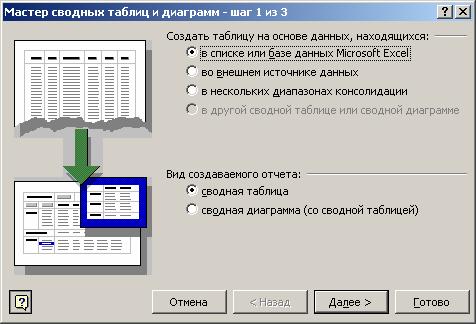
В результате выполнения описанных действий на листе посчитаются итоги (рис. 3).



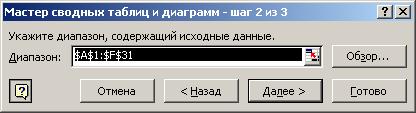
**Рисунок 3.** Количественные итоги продаж компьютеров по менеджерам

         Так можно решить простейшую задачу подсчёта итогов. Немного усложним задачу и попытаемся рассчитать на какую сумму каждый из менеджеров продал компьютеров различных производителей. Её решение использованием базовой функциональности итогов крайне сложно и будет требовать от пользователя помимо расчёта итогов выполнения ручных операций. Здесь на помощь приходит такой инструмент как «сводная таблица». Итак, возвращаем данные в исходное состояние и выполняем следующую последовательность действий:

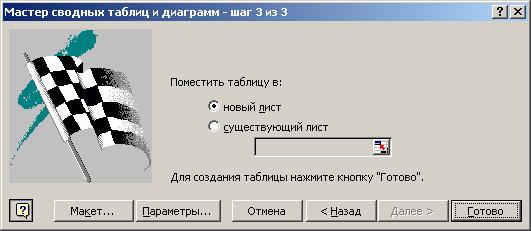
        выбираем пункт меню Данные -> Сводная таблица;



        из первого списка выбираем «в списке или базе данных Microsoft Excel», а из второго «сводная таблица» и нажимаем кнопку «Далее>»;

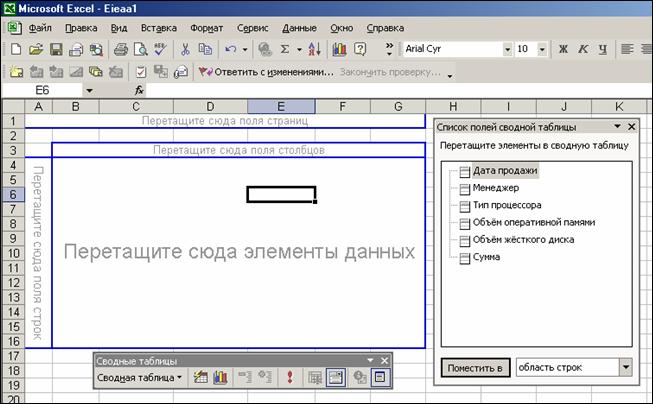


        выделяем диапазон, в котором находятся данные для анализа. Диапазон должен включать и стоку заголовка с названиями колонок. Нажимаем кнопку «Далее>»;



        выбираем поместить таблицу в «новый лист» и нажимаем кнопку «Готово».

Результат представлен на рисунке 4.



**Рисунок 4.** Результат работы мастера сводных таблиц.

          Теперь можно заняться построением отчётов. Для этого необходимо перенести требуемые поля из списка полей сводной таблицы (правая часть) в соответствующие зоны сводной таблицы (левая часть). Узнаем, на какую сумму каждый из менеджеров продал ПК с процессором Intel в зависимости от объёма жёсткого диска. Для этого выполним следующие действия:

        перенесём поле «менеджер» в зону поля столбцов;

        перенесём поле «объём жёсткого диска» в зону поля строк;

        перенесём поле «сумма» в зону элементов данных.

В результате выполнения данных действий получится общая информация по продажам в разрезе менеджеров и объёма жёсткого диска ПК (рис. 5).



**Рисунок 5.** Продажи ПК

Теперь добавим условие на процессоры Intel:

        перенесём поле «тип процессора» в поля страниц;

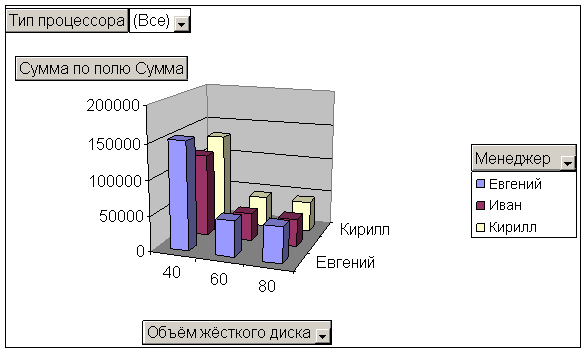
        выберем тип процессоров Intel, нажав на выпадающий список рядом с полем и выбрав соответствующее значение

Результат представлен на рисунке 6.



**Рисунок 6.** Результат построения отчёта (сводная таблица)

Помимо сводной таблицы в Excel имеется аналогичная возможность построения сводных диаграмм (рис. 7).



**Рисунок 7.** Результат построения отчёта (сводная диаграмма)

Таким образом, при помощи средства MS Excel подготовленный пользователь сможет без труда сформировать достаточно сложные аналитические таблицы и диаграммы и использовать их в процессе принятия решений.

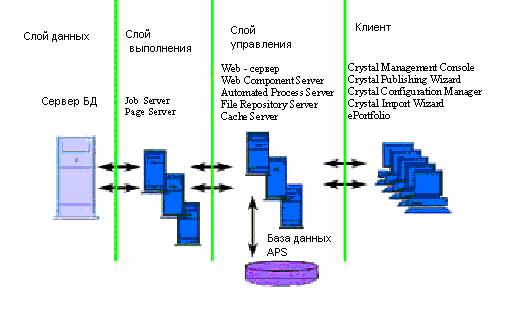
Помимо Excel существуют также более специализированные инструменты, которые предназначены именно для анализа, но не имеют привязки к конкретной предметной области.

**Crystal Enterprise**(по материалам www.interface.ru)



В настоящее время Crystal Enterprise занимает лидирующую позицию на рынке генераторов отчетов. Помимо трех редакций этого продукта (Standard, Developer и Professional), поставляемых непосредственно Seagate Software, Crystal Reports входит в состав некоторых средств разработки BI-средств (рис. 8).

Отличительными особенностями последней, восьмой версии продукта являются: поддержка самых разнообразных механизмов доступа к данным; обширный спектр средств автоматической генерации макетов стандартных отчетов; высокоточный графический дизайнер макетов отчетов; средства интеграции со средствами разработки приложений и с приложениями Microsoft Office в виде модулей расширений; поддержка публикации отчетов в Internet с точным воспроизведением их в браузере; наличие в комплекте поставки ASP-серверов для генерации отчетов, а также возможность создания простейших OLAP-отчетов.



**Рисунок 8.**  Структура продуктов Crystal Enterprise

***Клиентский слой*** Crystal Enterprise позволяет выполнять задачи администрирования всей системы, публиковать и просматривать отчеты. Он включает следующие модули:

  **Crystal Management Console** (CMC) - основанное на web-технологии средство администрирования, позволяющее создавать пользователей и группы пользователей Crystal Enterprise, изменять их права, устанавливать опции публикации отчетов, изменять настройки серверов и вносить или удалять лицензионные ключи. Доступ к CMC может быть осуществлен через web-браузер.

  **Crystal Publishing Wizard**  - приложение, позволяющее администратору и пользователю включить отчет в Crystal Enterprise. Администратор может управлять правами пользователя на включение отчета в Crystal Enterprise и на публикацию отчета.

  **Crystal Configuration Manager** - приложение, позволяющее администратору запускать, останавливать и удалять сервера.

  **Crystal Import Wizard** позволяет импортировать пользователей, папки и отчеты из Seagate Info или предыдущих версий Crystal Enterprise.

  **ePortfolio** - основанное на web-технологии приложение, позволяющее пользователю просматривать отчеты и устанавливать расписание выполнения отчетов. Crystal Enterprise поддерживает работу с отчетами формата Crystal Reports версий 6, 7 и 8. В качестве альтернативы ePortfolio может быть использован интерфейс Software Development Kit (SDK) для создания собственных приложений, реализующих все функции ePortfolio.

***Слой управления*** решает задачи безопасности, посылает запросы к соответствующим серверам и хранит версии отчетов с сохраненными данными (report instances).

  **Web - сервер**. Crystal Enterprise поддерживает работу с шестью типами серверов - Microsoft IIS, Netscape, Enterprise Server, Netscape Fast Track Server, Lotus Domino, Apache и CGI.

  **Web Component Server** отвечает за связь между Web - сервером и прочими компонентами Crystal Enterprise. Web Component Server обрабатывает запросы от браузера и сервера Crystal Server Pages (CSP).

  **Automated Process Server** (APS ) выполняет три основные функции - обеспечение задач безопасности (обеспечение привилегий пользователей), управление объектами (папками и отчетами) и управление серверами. APS создает базу данных пользователей Crystal Enterprise и определяет права каждого пользователя на определенные операции. База данных APS может быть создана на следующих СУБД: Oracle 7.3 и 8.1, Sybase Adaptive Server 11.5, Informix 7.3, Microsoft Access, Microsoft SQL Server 6.5, 7 и 2000. APS взаимодействуя с сервером Job Server отслеживает выполнение отчетов в заданное время, сохраняет пути к отчетам и поддерживает иерархию папок. Кроме того, APS постоянно отслеживает состояние серверов, формируя таблицу состояния серверов, которая используется для обработки запросов к отчетам.

  **File Repository Server**. Для каждого экземпляра Crystal Enterprise создается один входящий (Input) и один исходящий (Output) сервер File Repository. Входящий сервер содержит все отчеты, которые были добавлены в Crystal Enterprise пользователями с помощью модулей Crystal Publishing Wizard, Crystal Management Console или Crystal Import Wizard. Исходящий сервер содержит отчеты, выполненные сервером Job Server.

  **Cache Server** сохраняет страницы отчета, сгенерированные сервером Page Server. По запросу Cache Server пересылает страницы отчета серверу Web Component Server. Хранение страниц на сервере Cache Server позволяет минимизировать обращение к серверу баз данных при обращении к отчету.

***Слой выполнения*** обращается непосредственно к данным и генерирует отчеты. Он включает следующие компоненты:

   **Job Server** обрабатывает файлы отчетов (.rpt) в соответствии с расписанием APS и генерирует версии отчета, содержащие данные. Для генерации версий отчета Job Server связывается с базами данных.

   **Page Server** отвечает за получение страниц отчета от сервера Cache Server и генерацию страниц в формате Encapsulated Page Format (EPF). Страницы EPF содержат информацию о форматировании отчета. Данные для отчета могут быть сохранены вместе с отчетом или получены из базы данных.

Дополнительная информация доступна на Web-сайте фирмы по адресу: [http://www.seagatesoftware.com/products/crystalreports](http://www.seagatesoftware.com/products/crystalreports" \t "_blank).

**Hyperion Essbase (Hyperion Solutions)**



OLAP-сервер Hyperion Essbase представляет собой серверное OLAP-средство, основанное на многозвенной архитектуре и предназначенное для многопользовательского доступа к OLAP-данным и проведения аналитических расчетов. OLAP-сервер Hyperion Essbase поддерживает Windows NT, UNIX и AS/400 и может быть использован совместно с различными источниками данных и аналитическими приложениями. Отметим, что на базе Hyperion Essbase созданы OLAP-решения IBM.

Открытая архитектура Hyperion Essbase поддерживает доступ к OLAP-данным этого сервера с помощью самых разнообразных средств — электронных таблиц, построителей запросов, генераторов отчетов, — позволяя этим средствам выступать в роли клиентских приложений. Essbase поддерживает более 50 клиентских продуктов, в частности Seagate Software Crystal Reports и Crystal Info, средства просмотра OLAP-данных, такие как Business Objects и Cognos PowerPlay, электронные таблицы ведущих производителей. Кроме того, Hyperion Solutions поставляет собственные BI-приложения на базе Essbase.

Дополнительная информация доступна на Web-сайте фирмы по адресу: [http://www.essbase.com/](http://www.essbase.com/" \t "_blank).

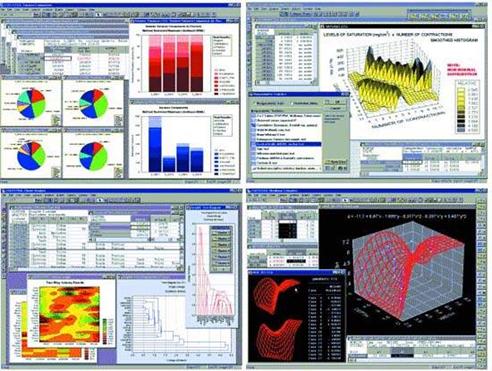
**STATISTICA (StatSoft)**

*STATISTICA Base* предлагает широкий набор методов для анализа (рис. 9):

**Описательные статистики и графики**. Программа вычисляет практически все используемые описательные статистики общего характера: медиану, моду, квартили, заданные пользователем процентили, среднее значение и стандартное отклонение, квартильный размах, доверительные интервалы для среднего, асимметрию и эксцесс (и их стандартные ошибки), гармоническое и геометрическое среднее, а также многие специальные описательные статистики.

**Группировка**. Практически все описательные статистики и графики могут быть построены для данных, категоризованных (сгруппированных) по значениям одной или нескольких группирующих переменных. Например, с помощью нескольких щелчков мыши можно сгруппировать имеющиеся данные о людях по *полу* и *возрасту*, а затем просмотреть категоризованные гистограммы, диаграммы размаха, нормальные вероятностные графики, диаграммы рассеяния и т.д.

**Корелляция.**В системе имеется большой набор методов для исследования корреляций между переменными. Прежде всего, могут быть вычислены все основные характеристики связи между переменными, в том числе: коэффициент корреляции Пирсона *r*, коэффициент ранговой корреляции Спирмена *R*,*тау*(*b*, *c*) Кендалла, *Гамма*, тетрахорический *r*, *Фи*, *V* Крамера, коэффициент сопряженности *C*, *D* Соммера, коэффициенты неопределенности, частные и получастные корреляции, автокорреляции, различные меры расхождения и т.д.



**Рисунок 9.**  Окно STATISTICA

**Диаграмма рассеивания, матричная диаграмма рассеивания, анализ по группам.**Как и во всех других диалоговых окнах вывода, здесь доступны различные общие параметры графического вывода, позволяющие проводить дальнейшее изучение закономерностей и взаимосвязей между переменными; например, двух- и трехмерные диаграммы рассеяния (с метками наблюдений или без них) служат для выявления зависимостей по подмножествам наблюдений или последовательностям переменных. Корреляционные матрицы могут быть категоризованы группирующими переменными и представляться графически в виде категоризованных диаграмм рассеяния.

**Вероятностный калькулятор.** Из любой панели инструментов системы *STATISTICA* доступен удобный интерактивный *Калькулятор вероятностных распределений*. Он поддерживает множество типов стандартных распределений (*бета*, *Коши*, *хи-квадрат*, *экспоненциальное*, *экстремальное (Гумбеля)*, *F*, *гамма*, *Лапласа*, *логнормальное*, *логистическое*, *Парето*, *Релея*, *t (Стьюдента)*, *Вейбулла* и *Z (нормальное)*).

Модуль **Непараметрическая статистика**содержит полный набор непараметрических статистик, включая все стандартные тесты и некоторые специальные прикладные статистики, в частности, критерий Вальда-Вольфовица, *U* тест Манна-Уитни (с точными вероятностями вместо нормальных аппроксимаций для малых выборок), критерии Колмогорова-Смирнова, критерий Вилкоксона парных сравнений, ранговый дисперсионный анализ Краскела-Уоллиса, медианный тест, критерий знаков, ранговый дисперсионный анализ Фридмана, *Q-* критерий Кохрена, критерий МакНемара, коэффициент конкордации Кендалла, *тау* (*b*, *c*) Кендалла, ранговая корреляция Спирмена *R*, точный критерий Фишера, критерии *хи-квадрат*, статистики *V-квадрат*, *Фи*, *Гамма*, *d*Соммера, коэффициенты сопряженности и другие

**Mathcad 11 (MathSoft)**



Программа Mathcad сочетает в себе:

  набор мощных инструментов для технических расчетов с полиграфическим качеством написания формул

  гибкий, полнофункциональный текстовый редактор.

С помощью эффективной среды решения задач программы Mathcad можно выполнять работу и демонстрировать результаты в одном и том же документе – на рабочей странице Mathcad. Прекрасное взаимодействие с другими инженерными, графическими и бизнес приложениями делает Mathcad необходимым элементом любого многогранного решения. Мощные средства Интернет-опубликования ускоряет процесс ознакомления с документами коллег и других Mathcad пользователей.

В отличие от другого технического программного обеспечения Mathcad осуществляет математические расчеты в той же последовательности, в которой Вы их записываете. Вводятся уравнения, данные для построения графика функции и текстовые примечания в любом месте страницы, при этом математические выражения в Mathcad записываются в полиграфическом формате.

Единственная разница с обычным текстом, включающим математические формулы и графики состоит в том, что Mathcad уравнения и графики – «живые». Изменение значений переменных, данных графика или уравнений приведет к немедленному перевычислению рабочей страницы.

Набор математических функций и методов вычислений, входящих в Mathcad настолько велик, что его можно сравнить с математической энциклопедией с живыми формулами.

Например, Mathcad содержит все элементарные математические функции и большое количество специальных функций; обрабатывает данные, в том числе статистическими методами, находит подгоночные функции; строит двух- и трехмерные графики; решает численно и аналитически системы дифференциальных уравнений, как обыкновенных, так и с частными производными, а также решает множество других задач.

**KXEN (Knowledge Extraction Engines) Analytic Framework** – это инструмент для построения описательных и предсказательных моделей. Применение KXEN позволяет организации перейти к анализу на высшем уровне: от «портрета» ситуации – к пониманию причин происходящих процессов и прогнозированию.

Технологию KXEN называют эволюцией подхода Data Mining по следующим причинам:

 рекордно высокая скорость работы без ущерба для точности и надежности результатов;

 способность включать в анализ сотни и даже тысячи переменных (в т.ч. транзакционные данные);

 удобство в применении: ориентация на бизнес-пользователей, автоматизация наиболее трудоемких этапов моделирования, интуитивно понятный интерфейс;

 легкость интеграции в существующие системы и приложения (открытые API).

**3 Специализированные информационно-аналитические системы**

**Программные продукты Business Objects**

****

Компания Business Objects с момента своего основания специализируется на создании аналитического программного обеспечения. Линейка продуктов компании Business Objects включает в свой состав средства бизнес-анализа (business intelligence), интеграции данных (data integration) и управления эффективностью деятельности (enterprise performance management).

***Категории продуктов***

 **Отчетность**  
К данной категории относятся генераторы отчетности Crystal Rports и средства создания систем корпоративной отчетности Crystal Enterprise.

  **Доступ к данным и анализ информации**

В данную категорию входят средства доступа к данным, создания отчетности и анализа информации BusinessObjetcs, WebIntelligence, BusinessQuery, Crystal Analysis и средства создания информационной инфраструктуры, входящие в пакет BusinessObjects Enterprise.

  **Управление эффективностью деятельности**

К данной категории продуктов относятся средства визуализации данных и создания информационных панелей Dashboard Manager и средства построения систем управления эффективностью деятельности на основе анализа ключевых показателей Performance Manager.

 **Корпоративные аналитические приложения**

В данную категорию входят пакетные аналитики для основных предметных областей деятельности BusinessObjects Analytics, средства настройки и создания собственных аналитических приложений Application Foundation и модель корпоративного хранилища данных BusinessObjects Warehouse, обеспечивающая быстрое внедрение аналитических приложений.

 **Интеграция данных**

К данной категории относятся средства извлечения, преобразования и загрузки данных BusinessObjects Data Integrator и «готовые» витрины данных Data Marts, предназначенные для быстрого извлечения данных из корпоративных приложений SAP, PeopleSoft, Oracle, J.D. Edwards и Siebel.

**Аналитическая платформа Контур (Intersoft Lab www.iso.ru)**

Программные продукты АПК относятся к классу систем Business Intelligence и предоставляют технологию оперативной аналитической обработки данных (OLAP). В составе АПК для создания системы корпоративных отчетов предлагаются программы:

  **"Контур Дизайнер кубов"** для проектирования мобильных отчетов из различных информационных систем и настройки сценариев пакетного создания отчетов.

  **"Контур Генератор кубов"** для пакетного создания мобильных отчетов по заданным сценариям из различных информационных систем.

  **"Контур OLAPBrowser"** для выпуска мобильных отчетов через Интранет/Интернет, в локальной сети и на ПК.

  **"Контур Стандарт"** для выпуска отчетов в режиме on-line из информационных систем.

Программы АПК позволяют выпускать мобильные отчеты по данным любых корпоративных источников информации: учетных систем, Хранилища данных, локальных таблиц и др. С помощью мобильных отчетов можно передавать корпоративную отчетность удаленным пользователям по e-mail и через Интернет/Интранет:

**ФИНАНСОВОЕ ХРАНИЛИЩЕ ДАННЫХ "КОНТУР КОРПОРАЦИЯ" (Intersoft Lab www.iso.ru)**

Финансовое Хранилище данных "Контур Корпорация" предлагается холдингам, промышленным группам, организациям и банкам с многофилиальной структурой для создания единой информационной среды и обеспечения эффективного управления бизнесом.

Финансовое Хранилище данных (ФХД) позволяет консолидировать финансовую информацию из различных корпоративных систем и предоставляет быстрый доступ ко всем деловым данным организации. ФХД "Контур Корпорация" обеспечит высшее руководство, менеджеров среднего звена, аналитиков и бухгалтеров информацией для эффективного управления и контроля деятельности многофилиальной организации.

 ФХД "Контур Корпорация" решает следующие задачи:

  **Предоставление акционерам и топ-менеджерам полной и достоверной информации о бизнесе** ФХД "Контур Корпорация" станет источником достоверной информации для управления крупным бизнесом, в котором участвует множество организаций. В ФХД, установленном в Управляющей компании, можно объединить деловую информацию из всех учетных систем дочерних предприятий и филиалов и любых других баз данных. Интегрированная бизнес-информация станет основой для решения задач финансового управления организацией, будет использована для расчета финансовых показателей и подготовки отчетности для акционеров и топ-менеджеров. Руководство в любой момент может получить адекватную информацию о деятельности каждой дочерней организации и обеспечить контроль исполнения стратегических целей компании.

  **Автоматизация различных технологий управления** "Контур Корпорация" предоставляет инструментарий для автоматизации комплексных технологий управления многофилиальной организацией: бюджетирования, ведения сводной главной книги бухгалтерского и управленческого учета и др.

  **Автоматизация работ по выпуску консолидированной отчетности** ФХД "Контур Корпорация" позволяет автоматизировать деятельность сотрудников бухгалтерии филиалов и головной конторы по подготовке обязательной и внутрифирменной консолидированной отчетности.

  **Аудит филиалов и дочерних предприятий**Хранилище данных обеспечивает интеграцию первичной бухгалтерской информации из всех филиалов и дочерних предприятий. Сотрудники бухгалтерии Головной конторы или Управляющей компании смогут проводить аудит на месте, не выезжая в филиалы. Видеть бухгалтерские данные можно вплоть до отдельных проводок.

  **Обеспечение архивными данными для оценки развития бизнеса**В Хранилище данных можно интегрировать и накапливать архивные данные за любой период. Эта информация может быть использована аналитиками для оценки динамики развития бизнеса, построения прогнозов.

  **Предоставление современного инструментария для бизнес-анализа** Инструменты OLAP-анализа и встроенный генератор отчетов предоставят информацию в удобном виде для каждого пользователя: руководителя, аналитика, бухгалтера.

Возможны различные масштабы построения ФХД "Контур Корпорация" - от Единого Хранилища данных, функционирующего в Управляющей компании или Головной конторе, до Распределенного Хранилища, когда в организации используется несколько Хранилищ данных.

**Тема 8.  Программные инструментальные средства информационно – аналитических систем**

**Содержание**

1.     Назначение и сферы применения продуктов компании Intersoft Lab.

2.     Состав и назначение систем Аналитической Платформы Контур.

3.     Понятие микрокуба.

4.     Назначение и принципы работы системы «Контур Стандарт».

5.     Работа с системой «Контур Стандарт».

6.     Пример анализа продаж по данным БД “Northwind” с помощью «Контур Стандарт».

**1. Назначение и сферы применения продуктов компании Intersoft Lab.**

Системы «Контур Дизайнер Кубов», «Контур Генератор Кубов», «Контур OLAPBrowser» и «Контур Стандарт» входят в состав Аналитической платформы Контур (АПК), которую разрабатывает компания Intersoft Lab. Аналитическая платформа Контур предназначена для анализа финансовой, статической и другой бизнес-информации и выпуска отчетов.

Системы Контур используют современную технологию оперативной аналитической обработки бизнес-данных OLAP (On-line Analytical Processing). На сегодняшний день OLAP –это самый мощный вид табличного генератора отчетов. Системы Контур позволяют получать произвольные экранные отчеты для интерактивного анализа и распечатывать отчеты на бумаге.

С помощью программ могут быть реализованы различные решения для подготовки отчетности организации:

***Система доставки отчетов удаленным пользователям***

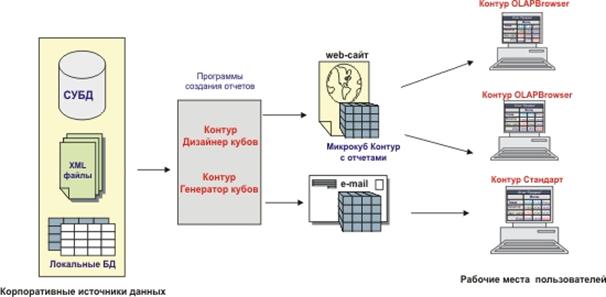
Для организаций с многофилиальной сетью, подразделения которых территориально распределены, можно создать систему корпоративной отчетности в сети Интернет. По данным корпоративных источников будут создаваться мобильные отчеты для анализа различных аспектов деятельности организации и публиковаться на web-сайте компании, размещаться на ftp-сервере или рассылаться сотрудникам удаленных подразделений по e-mail. Мобильные отчеты ориентированы на передачу больших объемов данных через Интернет и позволяют значительно уменьшить трафик и время передачи отчетности.

С помощью "Контур Дизайнера кубов" по данным корпоративных систем проектируются отчеты и настраиваются сценарии их пакетного создания. "Контур Генератор кубов" позволяет создавать отчеты по настроенным сценариям с заданной периодичностью (раз в день, раз в час или иначе) так, чтобы пользователи всегда получали актуальную информацию. Чтобы получить готовые отчеты руководителям и сотрудникам филиалов достаточно зайти на web-сайт компании и кликнуть по ссылке, либо открыть "Контур OLAPBrowser" для просмотра отчетов.

Возможен и другой способ подготовки отчетов через web-сайт. Пользователи могут самостоятельно на web-сайте формировать запрос на создание отчета из корпоративного источника данных (определять его параметры), а "Контур Генератор кубов" обеспечит быстрое создание и размещение отчета на web-сайте. В результате пользователь получит готовый отчет с актуальной информацией из первоисточника.

Доставлять мобильные отчеты удаленным пользователям можно также по электронной почте или, размещая их на ftp-сервере для скачивания. Пользователи смогут просматривать полученные отчеты с помощью программ "Контур OLAPBrowser" и "Контур Стандарт".

Система доставки отчетов удаленным пользователям представлена на рисунке 1.

  
  
**Рисунок 1.** Система доставки отчетов удаленным пользователям.

Технология подготовки отчетов через Интернет может использоваться не только для сотрудников организации, но и для ее акционеров, партнеров и клиентов. В виде микрокубов могут публиковаться отчеты о финансовых результатах компании для акционеров и клиентов. В интерактивных отчетах микрокуба можно предоставлять объемные прайс-листы компании, в них клиентам будет удобно просматривать и находить нужные позиции.

***Система подготовки отчетности в локальной сети***

Программы АПК могут использоваться для создания системы подготовки отчетности компании внутри офиса. Можно организовать выпуск отчетности по данным автоматизированных систем, в которых ведется учет текущих операций финансово-хозяйственной деятельности. IT-специалисты с помощью программ проектирования и генерации мобильных отчетов смогут быстро создавать любую отчетность по данным учетных систем, необходимую руководству и менеджерам.

Использование технологии мобильных отчетов, позволит уменьшить количество нерегламентированных запросов к учетной системе, которые замедляют выполнение транзакций, снизить нагрузку на сервер и оптимизировать выпуск аналитических отчетов. Создание отчетов может выполняться автоматически по расписанию в нерабочее время, когда операционисты не вводят данные в учетную систему.

Мобильные отчеты могут быть размещены на сервере или на клиентских машинах, где пользователи смогут просматривать их программами "Контур OLAPBrowser" и "Контур Стандарт". Отчеты будут предоставлены в виде интерактивной таблицы, пользователи смогут управлять представлением информации в ней и получать на ее основе десятки разнообразных отчетов и графиков. Таким образом, IT-специалисты будут освобождены от рутинной работы по реализации однотипных жестких форм отчетов.

Если руководству необходимо получать оперативные отчеты, то обеспечит выпуск отчетности из учетной системы или другого источника данных в режиме on-line программа "Контур Стандарт".

Система подготовки отчетности по локальной сети представлена на рисунке 2.

  
  
**Рисунок 2.** Система подготовки отчетности в локальной сети.

***Персональная система отчетности***

Программы "Контур OLAPBrowser" и "Контур Стандарт" можно использовать для создания персональной системы подготовки отчетности.

С помощью "Контур OLAPBrowser" руководители, менеджеры и аналитики смогут самостоятельно создавать единичные мобильные отчеты из источника, просматривать и анализировать готовые отчеты.

"Контур Стандарт" позволит напрямую обращаться к источнику данных и выпускать интерактивные отчеты в режиме on-line.

**2. Состав и назначение систем Аналитической Платформы Контур**

В состав АПК входят программы:

   **"Контур Дизайнер кубов"** для проектирования мобильных отчетов из различных информационных систем и настройки сценариев пакетного создания отчетов.

   **"Контур Генератор кубов"** для пакетного создания мобильных отчетов по заданным сценариям из различных информационных систем.

   **"Контур OLAPBrowser"** для выпуска мобильных отчетов через Интранет/Интернет, в локальной сети и на ПК.

   **"Контур Стандарт"** для выпуска отчетов в режиме on-line из информационных систем.

Системы «Контур Дизайнер Кубов», «Контур Генератор Кубов» и «Контур OLAPBrowser» используются в едином промышленном цикле работы с микрокубами.

«Контур Дизайнер кубов» – система для проектирования мирокубов и сценариев их генерации. Она предоставляет визуальную среду для описания корпоративных источников данных, запросов к источникам, структуры микрокубов и сценариев генерации. Все выполненные настройки сохраняются в репозитории метаданных или приложении системы.

Репозиторий (приложение) представляет собой файл в формате XML. Из репозитория сценарии генерации микрокубов можно сохранять в виде отдельных файлов XML-формата. Сценарий используется системой «Контур Генератор кубов» для генерации микрокубов.

«Контур Генератор кубов» – программа для массовой генерации микрокубов по заданным сценариям. Читая сценарий, «Контур Генератор кубов» выполняет запрос к базе данных, создает микрокуб и наполняет его данными. На основе одного сценария может быть создан один куб или пакет кубов. Чтобы выполнять массовую генерацию кубов, в сценарии определяются мастер-поля. Например, в сценарии «Продажа товаров» можно задать мастер-поле «Наименование покупателя». Тогда «Контур Генератор» кубов создаст столько микрокубов, сколько указано покупателей в этом поле. В каждом кубе будет храниться информация по продажам товара одному покупателю.

«Контур Генератор кубов» является утилитой командной строки и может принимать внешние параметры. Они позволяют передавать сценарию дополнительные условия фильтрации для выборки данных, которая помещается в кубы. Например, передавая сценарию параметр «Дата», можно генерировать кубы с данными за разные периоды.

«Контур Генератор Кубов» может запускаться системным планировщиком по заданному расписанию или внешними системами по событию. Например, Web-сервер предоставляет пользователю страницу запросов на генерацию микрокубов. Введенные параметры сценария передаются программе, она запускается и создает кубы.

Просмотр и OLAP-анализ информации микрокубов в корпоративной сети и в сети Интернет обеспечивает программа «Контур OLAPBrowser».

«Контур OLAPBrowser» – специализированный броузер для просмотра и OLAP-анализа микрокубов в Интернет, локальной сети, на пользовательском компьютере и обычной работы в Интернет. Программа предлагается в трех редакциях, ориентированных на различных пользователей:

1.     «Контур OLAPBrowser. Аналитик» – редакция системы с расширенными возможностями для опытных пользователей, поддерживает: просмотр микрокубов, OLAP-анализ, создание и редактирование микрокубов.

2.     «Контур OLAPBrowser. Обозреватель» – система для конечных пользователей информации: руководителей и менеджеров. Обеспечивает просмотр и OLAP-анализ готовых микрокубов.

Технология работы с микрокубами представлена на рисунке 3.



**Рисунок 3.**  Технология работы с микрокубами.

***Стандартный алгоритм работы с системами АПК Контур***

1. Проектирование структуры микрокуба и сценариев генерации в системе «Контур Дизайнер Кубов»:

o       Создание приложения.

o       Описание баз данных.

o       Описание взаимосвязи источников данных.

o       Построение запросов к модели данных.

o       Проектирование структуры микрокуба и настройка OLAP-интерфейсов для просмотра данных микрокуба.

o       Создание сценария генерации микрокубов.

2. Генерация микрокубов по сценарию с помощью программы «Контур Генератор Кубов».

3. Просмотр и OLAP-анализ микрокубов в системе «Контур OLAPBrowser» и «Контур Стандарт».

**3.     Понятие микрокуба.**

Микрокуб Контур – это мобильный контейнер аналитических отчетов и данных для анализа. В микрокубе хранятся данные, выгруженные из корпоративных информационных систем, алгоритмы расчета вычисляемых показателей и формы OLAP-отчетов. Данные в микрокубе подготовлены для многомерного анализа и оперативного получения показателей в различных аналитических разрезах. При помещении в микрокуб объем исходной информации сжимается в десятки раз. Компактный размер позволяет передавать микрокубы по Интернет-протоколам и пересылать по электронной почте.

Микрокубы – это эффективный способ информационно-аналитического обеспечения различных специалистов и подразделений организации. В микрокубы можно поместить информацию из бухгалтерских и ERP-систем, Хранилищ данных, специализированных баз данных подразделений, Web-сайта и других корпоративных источников. Руководители, менеджеры и аналитики, работая с подготовленными к анализу данными, смогут быстро получать интересующие цифры и показатели.

Микрокубы – это удобное средство информационного взаимодействия с удаленными подразделениями компании и внешними организациями. В виде микрокубов филиалы могут получать из Центрального офиса управленческую и аналитическую отчетность, информацию о корпоративных клиентах и пр. Размещая микрокубы на Web-сайте или корпоративном портале, можно обеспечить доступ к данным удаленным участникам бизнеса. В микрокубах можно предоставлять информацию клиентам организации, например многоуровневые электронные прайс-листы, выписки по счетам клиентов и др.

Микрокубы – это новое качество информационного сервиса. Информационные агентства и электронные биржи могут публиковать на своих сайтах микрокубы с аналитическими обзорами, проспектами биржевых индексов, котировками ценных бумаг различных эмитентов, акций, облигаций и векселей, биржевые курсы валют и др. В результате подписчики будут обеспечены информацией и современными инструментами ее анализа.

Применяя микрокубы, можно наладить эффективный информационных сервис в корпоративной сети и в Интернет.

Основные принципы организации микрокуба:

1. OLAP-машина расположена на стороне клиента. Это необходимо для использования мощности персональных компьютеров и для исключения центрального сервера, требующего постоянного обслуживания и монополизирующего информацию.

2. Данные и OLAP-машина автономны. Многомерная база данных не является неотъемлемой и неотделяемой частью OLAP-системы и может свободно перемещаться и обрабатываться произвольной системой, оснащенной стандартной OLAP-машиной.

3. Данные и метаданные расположены в одном файле. Это позволяет отображать данные универсальными программами и исключает необходимость инсталляции и настройки конкретного приложения пользователем.

4. В одном файле сохраняется одна база данных, неограниченное количество алгоритмов расчета вычисляемых полей и неограниченное количество форм отчетов. В результате появляется универсальный контейнер аналитического приложения.

5. Не существует ограничений на способы отображения и манипуляций данными. Кроме динамической таблицы данные могут отображаться в виде неограниченного количества диаграмм, настройки которых также сохраняются в едином файле-контейнере.

6. Объем файла должен быть минимальным. Это необходимо для передачи его по Интернет и электронной почте и исключает хранение в файле предварительно рассчитанных агрегатов, поскольку приводит к эффекту "взрывного роста данных".

7. Использование всех средств доставки данных. OLAP-машина может получать данные в виде потока по распространенным протоколам - file, html, ftp.

**4. Назначение и принципы работы системы «Контур Стандарт».**

Система «Контур  Стандарт» использует современную технологию оперативной аналитической обработки бизнес-данных OLAP. На сегодняшний день OLAP – это самый мощный вид табличного генератора отчетов. Система позволяют получать произвольные экранные отчеты для интерактивного анализа и распечатывать отчеты на бумаге.

Систему можно использовать для анализа в различных предметных областях: управление снабжением по данным ERP-системы, анализ продаж по данным CRM-системы, финансовый анализ по показателям обязательной отчетности и т.д.:

O руководители могут оперативно получать интересующие показатели.

O менеджеры отделов продаж и закупок – контролировать динамику продаж, текущее состояние счетов клиента и складских запасов.

O специалисты службы маркетинга – решать такие задачи как сегментирование клиентской базы, анализ спроса, оценка потребительского интереса по данным о посещаемости сайта компании (click-stream анализ).

O аналитики – реализовать индивидуальную модель оценки различных аспектов деятельности организации, и пр.

«Контур Стандарт» можно эксплуатировать практически в любой сфере, где требуется анализировать числовые данные и получать множество представлений одного набора данных в разных отчетах. Инструменты OLAP-анализа позволяют огромные массивы деловой информации показать в аналитических отчетах в виде наглядной картины состояния бизнеса.

**Принципы работы системы «Контур Стандарт»**

Система «Контур Стандарт» – инструмент анализа данных различных информационных систем и генерации отчетов. Обеспечивает доступ к локальным таблицам и популярным СУБД, позволяет выполнять запросы к БД и выпускать отчеты с интересующими показателями.

«Контур Стандарт» относится к классу OLAP-клиентов со встроенной OLAP-машиной. То есть система не хранит данные, а обращается с запросами к имеющимся источникам данных, выполняет промежуточные вычисления в памяти клиентского компьютера и отображает полученные результаты в OLAP-интерфейсах.

С помощью «Контур Стандарт» можно анализировать данные информационных систем, как в режиме реального времени, так и в off-line режиме. Для этого OLAP-система «Контур Стандарт» поддерживает две модели доступа к данным: ROLAP (Relational OLAP) и MOLAP (Multidimensional OLAP).

1. **ROLAP**. Программа обеспечивает прямой доступ к реляционным БД. «Контур Стандарт» в режиме on-line обращается к БД, получает плоские выборки, преобразует их в многомерное представление и отображает данные в аналитических интерфейсах.

Описание источников данных, запросов и пользовательских интерфейсов сохраняются в файле OLAP-приложения.

ROLAP-архитектура используется для анализа корпоративных данных в режиме реального времени. Например, on-line анализ требуется для управления риском ликвидности, когда есть необходимость постоянно получать актуальную информацию.

2. **MOLAP**. «Контур Стандарт» позволяет сохранять данные, полученные из реляционной БД, в многомерную локальную базу данных и предоставляет интерфейс для их просмотра. В терминах АПК файл, который содержит локальную базу данных и OLAP- интерфейс для работы с ней, называется микрокубом. MOLAP-технология значительно ускоряет процесс получения аналитических показателей. При этом работа ведется в режиме off-line от корпоративных БД.

MOLAP-архитектура используется для информационной поддержки удаленных пользователей, например, для обмена данными с филиалами или для размещения информации на Web-сайтах, а также для организации работы с популярными запросами к корпоративным БД.

**Работа в ROLAP-архитектуре**

Создание OLAP-приложения в «Контур Стандарт» заключается в последовательном описании трех взаимосвязанных слоев: слоя источников данных, слоя бизнес-объектов и слоя интерфейса:

1. **Описание источников данных**– это создание «словаря», в котором в терминах предметной области описывается физическая структура хранения данных для анализа: таблицы и их поля.

Например, описание физической таблицы *Product*с полями *ID*и *Name*хранится в словаре как источник *Продукты*с полями *Код*и *Товар*, а таблицы *Deal*с полями *Product ID, Date*и *Amount*– как источник данных *Сделки*с полями *Код продукта, Дата*и *Сумма*.

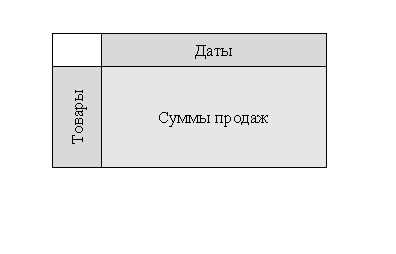
2. **Создание бизнес-объектов (запросов)**– это описание произвольных SQL-запросов к источникам данных. Оно выполняется без программирования путем выбора полей для отображения, связывания таблиц по ключевым полям, установки условий фильтрации данных и т.д.

Так, например, при описании бизнес-объекта *Продажи*из списка полей источников данных *Продукты*и *Сделки*для отображения выбираются поля *Товар, Дата*и *Сумма*, а сами таблицы связываются по ключу *Код продукта*.

Слой источников данных и слой бизнес-объектов составляют словарь объектов предметной области, который скрывает от пользователя реальную структуру хранения данных в базе.

3. **Описание пользовательских интерфейсов (отчетов)**– это первоначальная настройка аналитических отчетов, которая выполняется на основании запросов. Она заключается в определении состава запрашиваемых из источника данных полей для анализа, их расположения в отчете и алгоритмов работы с данными.

Например, при создании OLAP-таблицы *«Отчет о продажах»*строками назначаются *Товары*, колонками *Даты*, а значениями на пересечениях строк и колонок – *Суммы продаж*(рис. 2).



**Рисунок 2.** Структура OLAP - таблицы

Все настройки сохраняются в файле OLAP-приложения, который можно использовать в других копиях и редакциях программы.

**Работа в MOLAP-архитектуре**

OLAP-отчет можно сохранить в микрокуб для работы в off-line режиме и передачи удаленным пользователям. «Контур Стандарт» позволяет сохранять в микрокубах единичные отчеты.

Для массового производства микрокубов и работы с ними предназначены специализированные системы Аналитической платформы Контур: «Контур Дизайнер кубов», «Контур Генератор кубов» и «Контур OLAPBrowser».

**Редакции системы «Контур Стандарт»**

Система «Контур Стандарт» имеет несколько редакций, ориентированных на различных пользователей:

**«Контур Стандарт. Дизайнер»**– полнофункциональная редакция системы, предназначенная для специалистов, которые будут выполнять настройку на источники данных, и конечных пользователей аналитической информации. Включает следующие функции:

• создание OLAP-приложений (настройка источников данных, построение запросов и отчетов).

•  доступ к базам данных в режиме on-line.

•  выпуск динамических отчетов для выполнения OLAP-анализа.

• сохранение полученных данных в локальных микрокубах. Локальные микрокубы можно смотреть как в системе «Контур Стандарт», так и в других системах АПК.

**«Контур Стандарт. Аналитик»**– система для анализа данных с использованием готовых приложений. Обеспечивает аналитика следующей функциональностью:

• доступ к базам данных в режиме on-line.

• выпуск динамических отчетов для выполнения OLAP-анализа.

• настройка новых отчетов на основе имеющихся в приложении запросов.

• сохранение данных в локальных микрокубах.

**«Контур Стандарт. Обозреватель»**– система для просмотра и анализа готовых отчетов.

Обеспечивает пользователя следующей функциональностью:

• доступ к базам данных в режиме on-line.

• выпуск динамических отчетов для выполнения OLAP-анализа.

**5. Работа с системой «Контур Стандарт»**

Работа с системой «Контур Стандарт» организуется в два этапа:

1. Создание OLAP-приложения – пакета аналитических отчетов для конечного пользователя: руководителя, аналитика, маркетолога и т.д. Для этого используется система «Контур Стандарт» в редакции «Дизайнер».

2. Анализ и выпуск отчетов с помощью готового OLAP-приложения. Для этого может применяться система «Контур Стандарт» в редакциях «Дизайнер», «Аналитик», «Обозреватель» или «Инспектор».

**OLAP-приложение**(файл формата \*.caf) – отдельный программный ресурс (репозиторий метаданных), в котором хранятся настройки источников данных, запросов к источникам данных и аналитические отчеты. Создание OLAP-приложения состоит в последовательном описании этих настроек.

**Источники данных**– локальные таблицы и реляционные БД автоматизированных систем, в которых хранятся корпоративные данные для анализа (таблицы MS Excel, БД бухгалтерских или ERP-систем, Хранилищ данных и т.д.). «Контур Стандарт» поддерживает прямой доступ к локальным таблицам (Dbase и Paradox) и к таблицам и хранимым процедурам реляционных СУБД (MS SQL, Oracle, Sybase, MS Access и т.д.).

**Запросы**– SQL-запросы к источникам данных, описываемые в системе на уровне правил объединения таблиц источников данных по ключевым полям, условия фильтрации данных и набора возвращаемых полей. Результатом выполнения запроса является плоская выборка данных, отображаемая в отчетах.

**Отчеты**– пользовательские интерфейсы для анализа данных. В «Контур Стандарт» можно создать 4 вида отчетов:

• Таблица.

• Тренд.

• Кластерный анализ.

• OLAP-отчет.

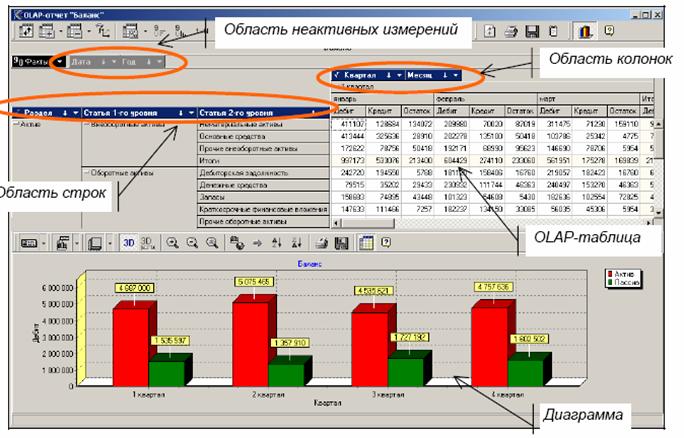
Все пользовательские отчеты могут быть проиллюстрированы графиками, распечатаны, выгружены в офисные приложения (MS Excel, MS Word, html-формат) или сохранены в csv-формате.

**Таблицы**– плоские списочные отчеты, сопровождаемые графиками. Пользователь может быстро перерисовывать графики, фильтруя значения измерений и управляя фактами и измерениями по осям. Это позволяет оперативно менять «точку зрения» на анализируемые данные в поисках зависимостей и тенденций.

**Тренды**– отчеты, отражающие изменение показателей во времени и динамику этих изменений. Особенно эффективно использование трендов для изучения сезонных колебаний и прогнозирования тенденций.

**Кластерный анализ**– интерфейс, позволяющий объединять объекты в группы (кластеры) по заданным признакам, сравнивать группы, выявлять среди них крупные (наиболее влиятельные) и мелкие (наименее влиятельные). Этот интерфейс полезен, например, при создании приложений для маркетингового анализа.

**OLAP-отчеты**– управляемые динамические OLAP-таблицы, которые сопровождаются синхронной диаграммой (графиком). OLAP-отчет может содержать данные, полученные по запросу как из БД, так и из локального микрокуба (рис. 4).

**

**Рисунок 4.**  OLAP - отчет

**OLAP-таблица**– таблица, автоматически суммирующая данные (факты) в различных разрезах (измерениях) и позволяющая интерактивно управлять вычислениями и формой отчета. Измерения отображаются в названиях строк и столбцов таблицы, соответствующие им факты и итоги (агрегированные факты) – в ячейках таблицы. Колонки и строки являются инструментами управления таблицей. Пользователь может перемещать их, фильтровать, сортировать, детализировать/обобщать и выполнять другие OLAP-операции. При этом таблица автоматически вычисляет новые итоги (агрегаты). Управляя OLAP-таблицей, можно из одного набора данных сформировать множество отчетов.

В OLAP-таблице можно условно выделить несколько рабочих областей (рис. 4):

**Область активных измерений**– строки и столбцы OLAP-таблицы.

**Область неактивных измерений**(фильтры) – область, содержащая измерения, не отображенные в OLAP-таблице, но влияющие на представленные в ней данные.

**Область фактов**– таблица с числовыми данными, над которыми выполняются вычисления.

**Диаграмма**– синхронное с OLAP-таблицей графическое представление данных. Диаграмма строится по расположенным в крайнем левом положении элементам динамической таблицы, то есть по крайне левому измерению-строке, измерению-колонке и фактам (крайне левому или по всем). Чтобы изменить измерение (строку или колонку), по которому строится диаграмма, надо переместить ее в крайне левое положение в таблице. Изменить порядок отображения фактов можно с помощью специального фильтра.

**Стандартный алгоритм работы с системой «Контур Стандарт»**

1. Создание нового OLAP-приложения:

• Создание файла OLAP-приложения.

• Описание источников данных.

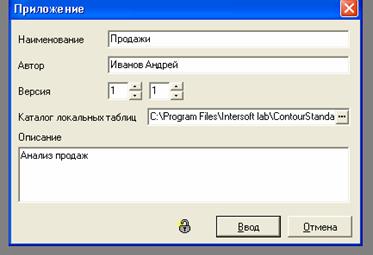
• Построение запросов к данным.

• Настройка пользовательских отчетов.

2. Анализ данных и выпуск отчетов.

**Создание приложения**

Работа в системе начинается с создания файла нового приложения **Продажи** (рис. 5).



**Рисунок 5**. Создание приложения.

Для создания приложения в поле **Каталог локальных таблиц** необходимо указать название папки, в которой хранятся локальные таблицы.

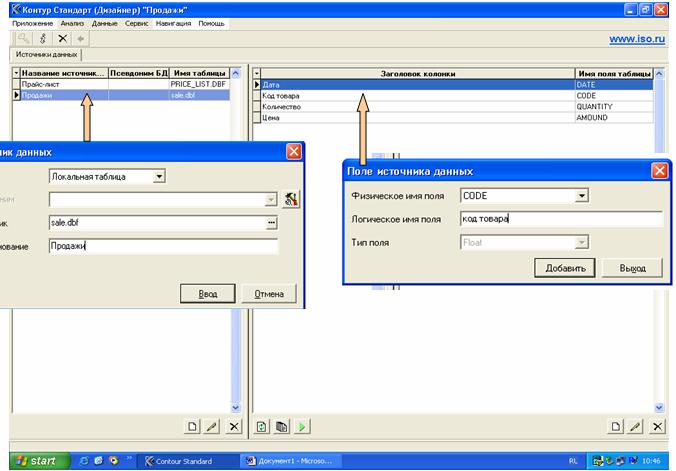
Далее должны быть описаны все источники данных, которые будут использоваться в системе. Для описания источника данных нужно указать доступ к источнику данных: для локальных таблиц это имя и путь к таблице.

**Описание источников данных**

Описание источников выполняется в закладке ***«Источники данных***», которая вызывается из меню **Сервис**по команде **Источники данных**.

Закладка «Источники данных » (рис. 6) разделена на две части. Левая часть закладки предназначена для описания источников данных, здесь отображается таблица, содержащая перечень источников данных.

Правая часть закладки предназначена для описания структуры источника данных, и содержит перечень полей текущего источника данных.

****

**Рисунок 6**. Описание источников данных приложения

При необходимости можно переименовать таблицы источников данных и поля и присвоить им названия в терминах предметной области.

**Построение запроса к источникам данных**

Запрос – это SQL-запрос к источнику данных. В «Контур Стандарт» запрос описывается на уровне правил объединения таблиц источников данных по ключевым полям, условий фильтрации данных и набора возвращаемых полей.

В одном приложении может быть описано неограниченное количество запросов для получения данных, которые в виде плоских выборок отображаются в аналитических отчетах.

Готовые запросы удобно хранить в тематических папках. Для хранения запросов приложения нужно создать папку и дать ей название.

Создание запросов выполняется в закладке «Запросы», которая вызывается из меню **Сервис** по команде **Запросы**.

Закладка «Запросы» состоит из двух частей (рис. 7). В левой части отображается заданная пользователем иерархическая структура тематических папок. Причем, папка любого уровня может содержать произвольное количество SQL-запросов к источнику данных.

В правой части закладки «Запросы » отображается перечень запросов, хранящихся  в текущей папке.



**Рисунок 7**. Создание запросов.

Для построения запроса к данным необходимо указать:

1. Название запроса.

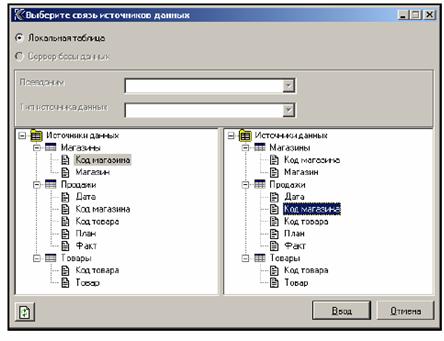
2. Список полей запроса.

3. Условия запроса:

  Связи между таблицами - источниками данных, если запрос выполняется к нескольким источникам данных (рис. 8).

  Условия фильтрации данных и правила сочетания условий фильтрации, которые можно задать с помощью логических операторов.

4. Краткое описание запроса.



**Рисунок 8.**  Создание связи между таблицами

**Создание пользовательских отчетов**

Пользовательские отчеты или аналитические интерфейсы системы «Контур Стандарт» предназначены для интерактивного анализа данных. В системе для анализа можно создать 4 вида отчетов:

1.     OLAP-отчет – для получения различных информационных срезов на базе одного отчета.

2.     Кластерный анализ – для оценки групп показателей.

3.     Тренд – для исследования динамики показателей.

4.     Таблица – для быстрого построения различных графиков и диаграмм.

Создание всех отчетов Приложения выполняется в закладке «Отчеты», которая вызывается командой меню **Отчеты**. Готовые отчеты удобно хранить в тематических папках.

Закладка «Отчеты» состоит из двух частей (рис. 9). В левой части отображается заданная пользователем иерархическая структура тематических папок. Причем, папка любого уровня может содержать произвольное количество отчетов любого типа.

В правой части закладки «Отчеты » отображается перечень отчетов, хранящихся в текущей папке. Работа пользователя с отчетами осуществляется с использованием команд локального меню, вызываемого по правой кнопке мыши, или с помощью следующих кнопок:

 Создание OLAP-отчета из базы данных.

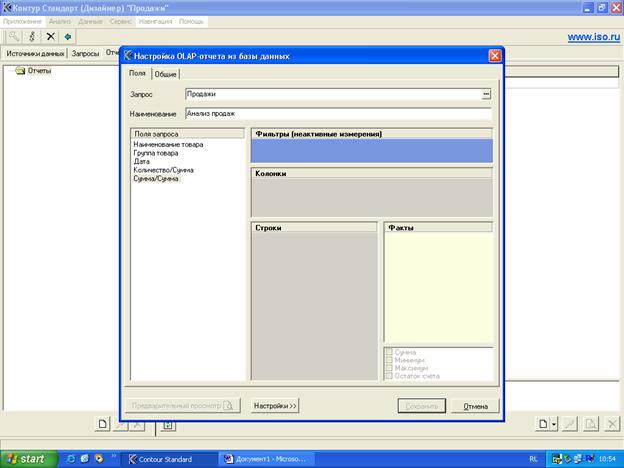
 Создание OLAP-отчета из микрокуба.

 Создание тренда.

 Создание таблицы.

 Создание интерфейса кластерного анализа.

**OLAP-отчеты** – это интерфейсы для OLAP-анализа данных из БД или микрокубов. OLAP-отчет включает динамическую таблицу и синхронную диаграмму. В таблице автоматически суммируются числовые данные в различных разрезах и рассчитываются итоговые показатели. С помощью OLAP-отчетов можно быстро посмотреть на одни и те же данные с разных «точек зрения».



**Рисунок 9.** Создание OLAP-отчета

Для создания OLAP-отчета необходимо указать:

1. Запрос, на основании которого будут поступать данные из источников.

2.  Название отчета.

3. Поля, которые будут являться строками и столбцами (активными измерениями), фильтрами (неактивными измерениями) и фактами (анализируемыми числовыми значениями) OLAP-таблицы (рис. 10).

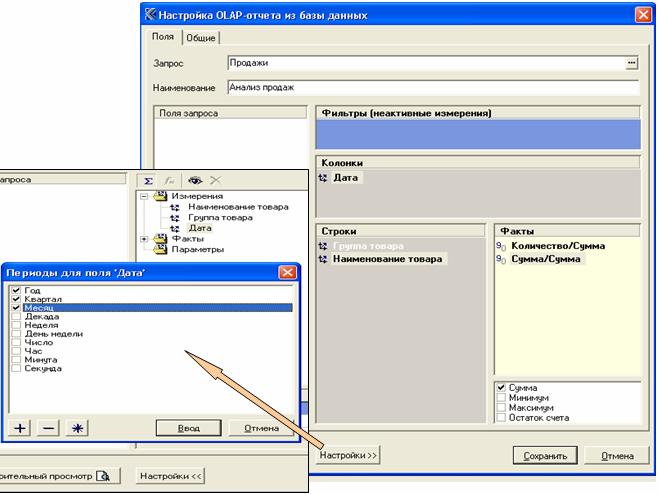
4. Факты, вычисляемые по встроенным алгоритмам или рассчитываемые по формулам, старшие временные периоды и др.

5. Заголовок диаграммы.

6. Действия, которые допустимо производить над данными таблицы.

7. Поля детализации данных.

8. Краткое описание OLAP-отчета.

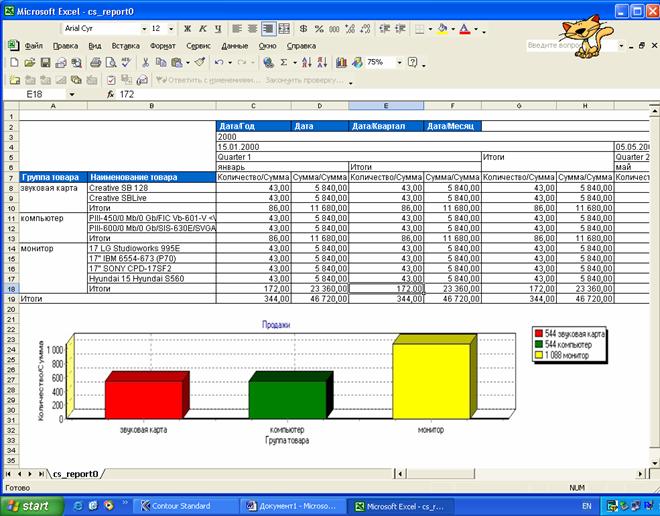
****

**Рисунок 10.** Определение измерений таблицы и настройка OLAP-отчета

OLAP – таблица позволяет интерактивно управлять вычислениями и формой отчета. Инструментами управления динамической таблицей являются ее элементы – колонки и строки. Пользователь может перемещать их, удалять, фильтроватьи выполнять другие OLAP-операции. При этом OLAP-машина автоматически вычисляет новые промежуточные и окончательные итоги. Таблица иллюстрируется графиком, который синхронно отображает изменения в таблице.

OLAP-отчеты можно применять для прикладного анализа практической любой направленности, если требуется обобщение, детализация или "другой" взгляд на одни и те же данные.

После создания и настройки OLAP – отчета можно произвести выгрузку данных отчета в форматы MS Word, MS Excel, html (рис. 11).



**Рисунок 11.** OLAP – отчет в Microsoft Excel.

**Таблицы** – это плоские списочные отчеты, сопровождаемые графиками. В таблице можно быстро строить различные графики и диаграммы (рис. 12). Это позволяет оперативно менять «точку зрения» на анализируемые данные в поисках зависимостей и тенденций.

Для создания таблицы необходимо указать:

1. Запрос, на основании которого будут поступать данные из источников.

2. Список полей таблицы из полей запроса. Для таблицы не задается явное разделение элементов на измерения и факты. При работе с такой таблицей можно быстро отображать отношения фактов и измерений в виде отдельных графиков.

3. Поля таблицы, для которых нужно рассчитать итоговые значения.

4. Название отчета.

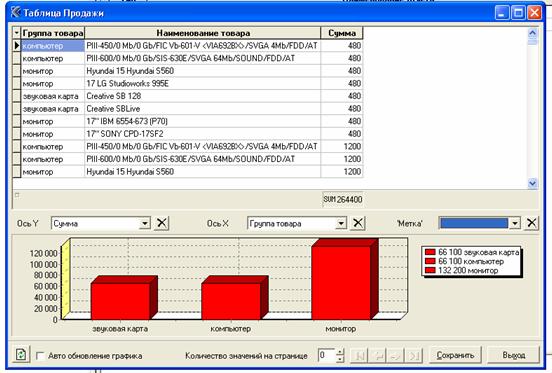
5. Краткое описание интерфейса «Таблица».

Для управления данными таблицы используются следующие операции:

1. Сортировка данных таблицы – кликните левой кнопкой мыши на заголовке измерения таблицы. В результате данные будут отсортированы в порядке возрастания. Заголовок того столбца, по которому проведена сортировка, выделяется подчеркиванием.

2. Установка фильтра на значения измерений таблицы – кликните правой кнопкой мыши на заголовке измерения таблицы и из локального меню выберите команду **Фильтр**. В окне «**Фильтрация**» задаются условие фильтра.

Для снятия фильтра необходимо выбрать из локального меню команду **Удалить фильтр**.

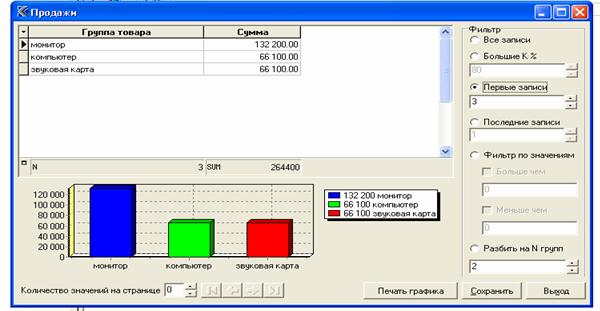


**Рисунок 12.** Таблица

Основное полезное свойство интерфейса «Таблица» – это возможность перестраивать график, меняя измерения, которые откладываются по его осям. Таким образом, можно анализировать зависимость факта от различных измерений и других фактов.

**Кластерный анализ.**

Интерфейс кластерного анализа позволяет объединять объекты в группы (кластеры) по заданным признакам для сравнительного анализа (рис. 13).



**Рисунок 13**. Интерфейс кластерного анализа

Для создания интерфейса кластерного анализа необходимо указать:

1.Запрос, на основании которого будут поступать данные из источников.

2. Поле запроса, по которому будут группироваться (агрегироваться) факты.

3. Поле запроса, которое будет являться анализируемым фактом.

4. Алгоритм агрегации фактических значений.

5. Название отчета.

6. Краткое описание интерфейса «Кластерный анализ».

При необходимости можно откорректировать названия элементов интерфейса.

Интерфейс кластерного анализа предназначен для проведения сравнительного анализа групп объектов, объединенных по заданным признакам.

В интерфейсе кластерного анализа используйте следующие механизмы фильтрации:

1. «Все записи » – отображение всех записей таблицы кластерного анализа.

2. «Большие К %» – отображение первых записей с наибольшим значением факта, сумма которых составляет К % от общей суммы.

3. «Первые записи » – отображение первых N-записей с наибольшим значением факта.

4. «Последние записи » – отображение последних N-записей с наименьшим значением факта.

5. «Фильтр по значениям » – отображение записей, отсортированных по введенному значению факта (больше или меньше данного значения).

6. «Разбить на N групп » – разделение итоговой суммы факта на N-значений.

После применения фильтров графическая иллюстрация синхронно перестраивается в соответствии с данными таблицы.

С помощью кластерного анализа  можно выявить наиболее или наименее влияющие на какой-либо показатель группы. Этот интерфейс полезен при создании приложений для маркетингового анализа. Например, с его помощью можно выполнять сегментирование товарной продукции, то есть определять наиболее и наименее выгодные группы товаров в разрезе потребителей и регионов.

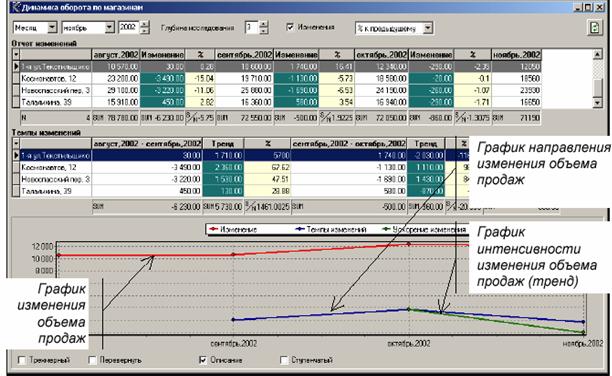
**Тренд (отчет изменений)**

Тренд – это отчет, отражающий изменение показателей во времени и динамику этих изменений.

Трендовая модель, изучающая явление во времени, полезна при оценке сезонных колебаний, прогнозирования тенденций и др.

Период и глубина исследований выбирается пользователем. В зависимости от данных можно анализировать динамику показателя по дням, месяцам, кварталам и годам. При необходимости изменения показателя можно показать в процентах.

Отчет изменений строится на основе выборки, критериями которой являются: одно измерение, один факт и дата. Интерфейс отчета состоит из следующих блоков (рис. 14):



**Рисунок 14.**  Тренд.

1. Панель инструментов – предназначена для настройки параметров отчета измерений.

2. Отчет изменений – таблица, отображающая итоговые показатели за период и изменения между ними.

3. Темпы изменений – таблица, отображающая динамику изменений показателей (первая производная).

4. Графическое представление данных отчета.

В графическом интерфейсе тренда отображается три графика:

1. Первый график строится по абсолютным значениям показателя в определенные моменты времени и показывает, как изменялось значение показателя – увеличивалось или уменьшалось.

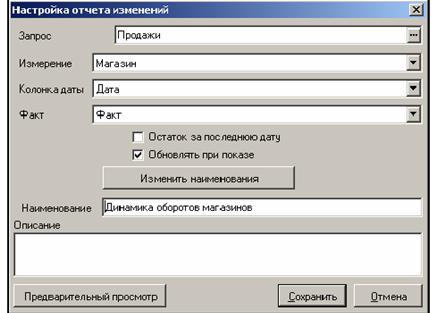
2. Второй график строится по значениям изменений показателя за периоды между двумя соседними моментами времени и определяет направление развития показателя. С точки зрения математики график – это первая дискретная производная, а с точки зрения физики – скорость изменений показателя от периода к периоду.

3. Третий график строится по величине изменений скорости и показывает изменение интенсивности развития показателя. С точки зрения математики график – это вторая дискретная производная от значений показателя, а с точки зрения физики – ускорение изменения на каждом интервале времени. Именно этот график и является трендом.

Для настройки отчета изменений укажите (рис. 15):

1. Запрос, на основании которого будут поступать данные из источников.

2. Поле запроса, которое будет являться измерением



**Рисунок 15.** Настройка тренда.

3. Поле запроса, которое будет использовано в качестве измерения Дата.

4. Поле запроса, которое будет являться фактом, и изменения значений которого будут анализироваться.

5. Алгоритм агрегации факта.

6. Название отчета изменений.

7. Краткое описание отчета изменений.

При необходимости можно откорректировать названия элементов интерфейса. Для этого нужно нажать кнопку «Изменить наименование».

 Готовое приложение вы можете передать конечному пользователю. Это делается путем передачи пользователю caf-файла и инструкции по настройке приложения на источники данных.

Приложение может использоваться  локально, одним пользователем – в этом случае его следует разместить на компьютере пользователя. Чтобы сделать приложение доступным группе пользователей, файл приложения надо разместить на каком-либо разделяемом ресурсе.

У клиента может возникнуть задача разграничения прав доступа для различных пользователей приложения.

Разграничение прав доступа может быть реализовано несколькими способами.

Если приложение использует локальные таблицы, то возможны два варианта.

Первый вариант – ограничить доступ пользователей к локальным таблицам средствами операционной системы. При этом при запуске отчета, использующего закрытые для пользователя таблицы, пользователь получит сообщение о том, что источник данных недоступен.

Второй вариант – сделать копии приложения для различных групп пользователей с разными правами, оставив  для каждой группы пользователей свой набор отчетов.

Если же приложение настроено на реляционную базу данных, доступ пользователя к базе будет контролироваться средствами СУБД. При первом обращении пользователя к базе данных из системы Контур Стандарт можно запрашивать login и пароль пользователя, и далее доступ к данным базы предоставлять в соответствии с правами этого пользователя.

**6. Пример анализа продаж по данным БД “Northwind” с помощью «Контур Стандарт»**

Проиллюстрируем теорию примером. Создадим отчет для  анализа продаж с помощью “Контур Стандарт”. Данные о продажах представим в разрезе товаров, покупателей и менеджеров. Учтем также географический и временной факторы. В качестве источника данных используем БД “Northwind”, предлагаемую в составе MS SQL Server и MS Access.

На первом этапе настройки OLAP-приложения определим путь к БД “Northwind” и в терминах предметной области опишем ее структуру (таблицы и отдельные поля этих таблиц). Описывая таблицы, выберем только те поля, которые нам необходимы для анализа. Для удобства российских пользователей при описании источников данных (таблиц и их полей) дадим им русскоязычные имена. Далее в отчетах пользователей будут фигурировать именно эти наименования.

Для анализа продаж воспользуемся таблицами:

 Products (Товары):  ProductID (Код товара), ProductName (Наименование товара), CategoryID (Группа товаров).

 Categories (Группы товаров):  CategoryID (Код группы товаров), CategoryName (Группа товаров).

 Employees (Сотрудники):  EmployeeID (Код сотрудника), LastName (Сотрудник).

 Customers (Покупатели):  CustomerID (Код покупателя), CompanyName (Компания), Region (Регион), Country (Страна), City(Город).

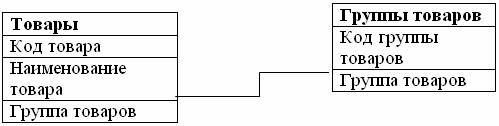
 Orders (Договора):  OrderID (Договор), OrderDate (Дата), EmployeeID (Сотрудник), CustomerID (Покупатель).

 Order Details (Детали договора):  OrderID (Договор), ProductID (Товар), Quantity (Количество), Discount (Скидка).

Далее создаем запрос – “выборку” в терминах “Контур Стандарт”. Здесь описываются отображаемые поля, связи таблиц по ключевым полям, условия фильтрации. Таким образом, запрос описывает алгоритмы генерации SQL-запросов к источникам данных.

Перечислим отображаемые в запросе поля таблиц: “Группа товаров”,  “Наименование товара”,  “Сотрудник”,  “Регион”,  “Страна”,  “Город”, “Компания”, “Дата”, “Скидка” и “Количество”. В этот список входят все информативные поля, используемые при анализе. Поля-идентификаторы не включены в список, т.к. в отчетах они фигурировать не будут.

Запрос объединяет данные нескольких таблиц. Поэтому необходимо указать связи таблиц по ключевым полям, например (рис. 16).



**Рисунок 16.**   Настройка выборки данных – добавление связей между таблицами.

Соединим таблицы **Покупатели** и **Договора** по полю «Код покупателя», таблицы **Договора** и **Сотрудники** по полю «Код сотрудника», таблицы **Товары** и **Детали договора** по полю «Код товара», таблицы **Договора** и **Детали договора** по полю «Код договора», таблицы **Товары** и **Группы товара** по полю «Код  группы товара».

Вместе с описанием источников данных запросы образуют семантический слой, скрывающий от пользователя реальную структуру БД. Этот семантический слой (метаданные) используется оптимизированным генератором SQL-запросов для создания первичной агрегированной плоской выборки записей. Если SQL-запрос будет выполняться сервером, то в запрос будет автоматически включено выражение GROUP BY, чтобы заставить сервер выполнить предварительную агрегацию. Если это запрос к локальным таблицам, то предварительная  агрегация выполнится самой OLAP-машиной.

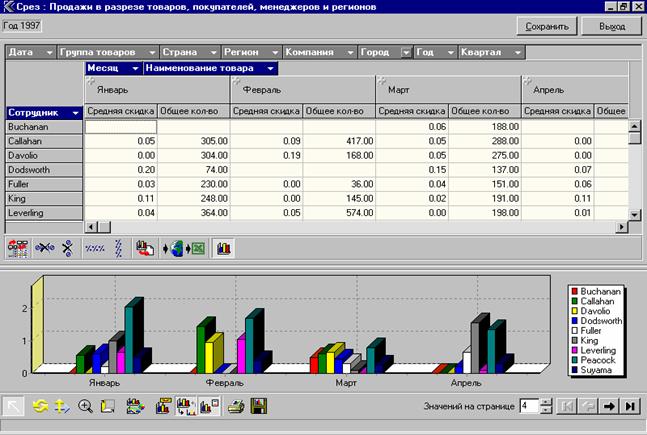
Далее на базе сформированной выборки создадим отчет. Он представляет собой пользовательский интерфейс, который позволяет интерактивно управлять структурой данных динамической таблицы, графической иллюстрацией и печатью отчетов.

В терминах "Контур Стандарт" отчет, содержащий динамическую таблицу, называется «срез». При настройке среза из списка полей выборки добавляем необходимые для отчета «факты» и «измерения». При добавлении поля «факта» пользователь выбирает один или несколько алгоритмов его агрегации: сумма, количество и т.д. При добавлении поля даты как «измерения» данные можно разложить на выбранные временные периоды. То есть «измерений» даты будет несколько: год, квартал, день и т.д.

В этом примере в отчете будут использоваться все поля выборки (рис. 17). Измерение «Дата» разложим на периоды: год, квартал, месяц, дата.

Результат настройки отчета можно увидеть в режиме «предварительный просмотр». На практике отчеты обычно оперируют меньшим количеством измерений. При анализе более 7-8 измерений человек с трудом воспринимает информацию. Поэтому, в режиме «предварительный просмотр» настроим внешний вид отчета, удобный для восприятия и оценки информации.

Например, перенесем в область «неактивных» измерений – вверху таблицы – измерения, которые будем использовать для фильтрации таблицы, а не для вычисления итогов продаж по ним (рис. 17): “Страна”, ”Регион”, ”Город”, ”Компания”, “Год”, ”Квартал”, ”Дата” и ”Группа товаров”.



**Рисунок 17.**   Редактирование отчета в режиме «предварительный просмотр».

Таким образом, мы получили отчет о продажах товаров по месяцам в разрезе менеджеров. Для изменения ракурса анализа пользователь в процессе работы с отчетом может вернуть любое измерение в таблицу – «открыть» измерение.

На этом процесс создания приложения закончен. Все настройки сохраняются в файле OLAP-приложения.