Интеллектуальный анализ данных

Понятие и модель данных OLAP

OLAP (Online Analytical Processing) - технология оперативной аналитической обработки данных, использующая методы и средства для сбора, хранения и анализа многомерных данных в целях бизнес-анализа (BI, Business Intelligence) или , в русскоязычной терминологии , поддержки процессов принятия решений(СППР).

Основное назначение OLAP-систем - поддержка аналитической деятельности, произвольных запросов пользователей - аналитиков. Цель OLAP-анализа - проверка возникающих гипотез.

Сбор и хранение информации, а также решение задач информационно-поискового запроса эффективно реализуются средствами систем управления базами данных (СУБД). В OLTP (*Online Transaction Processing*)-подсистемах реализуется *транзакционная обработка* данных. Непосредственно OLTP-системы не подходят для полноценного анализа информации в силу противоречивости требований, предъявляемых к OLTP-системам и *СППР*.

Для предоставления необходимой для принятия решений информации обычно приходится собирать данные из нескольких *транзакционных баз данных* различной структуры и содержания. Основная проблема при этом состоит в несогласованности и противоречивости этих баз-источников, отсутствии единого логического взгляда на корпоративные данные.

Поэтому для объединения в одной системе OLTP и *СППР* для реализации подсистемы хранения используются концепция хранилищ данных (ХД). В основе концепции ХД лежит идея разделения данных, используемых для оперативной обработки и для решения задач анализа, что позволяет оптимизировать структуры хранения. ХД позволяет интегрировать ранее разъединенные детализированные данные, содержащиеся в исторических архивах, накапливаемых в традиционных OLTP-системах, поступающих из внешних источников, в единую базу данных, осуществляя их предварительное согласование и, возможно, агрегацию.

Подсистема анализа может быть построена на основе:

- подсистемы информационно-поискового анализа на базе реляционных СУБД и статических запросов с использованием языка SQL;
- подсистемы оперативного анализа. Для реализации таких подсистем применяется технология оперативной аналитической обработки данных OLAP, использующая концепцию многомерного представления данных;

- подсистемы интеллектуального анализа, реализующие методы и алгоритмы Data Mining.

Все данные в ХД делятся на следующие категории :

1. детальные данные - данные, переносимые непосредственно из OLTP-подсистем. Соответствуют *элементарным событиям*, фиксируемым в OLTP-системах. Подразделяются на:
    - измерения - наборы данных, необходимые для описания событий (товар, продавец, покупатель, магазин, …);
    - факты - данные, отражающие сущность события (количество проданного товара, сумма продаж, …);
2. агрегированные (обобщенные) данные - данные, получаемые на основании детальных путем суммирования по определенным измерениям;
3. метаданные - данные о данных, содержащихся в ХД. Могут описывать:
    - объекты предметной области, информация о которых содержится в ХД;
    - категории пользователей, использующих данные в ХД;
    - места и способы хранения данных;
    - действия, выполняемые над данными;
    - время выполнения различных действий над данными;
    - причины выполнения различных действий над данными.

Информационные потоки в ХД

Данные в ХД образуют следующие информационные потоки:

входной поток - образуется данными, копируемыми из OLTP-систем в ХД; данные при этом часто очищаются и обогащаются путем добавления новых атрибутов;

поток обобщения - образуется агрегированием детальных данных и их сохранением в ХД;

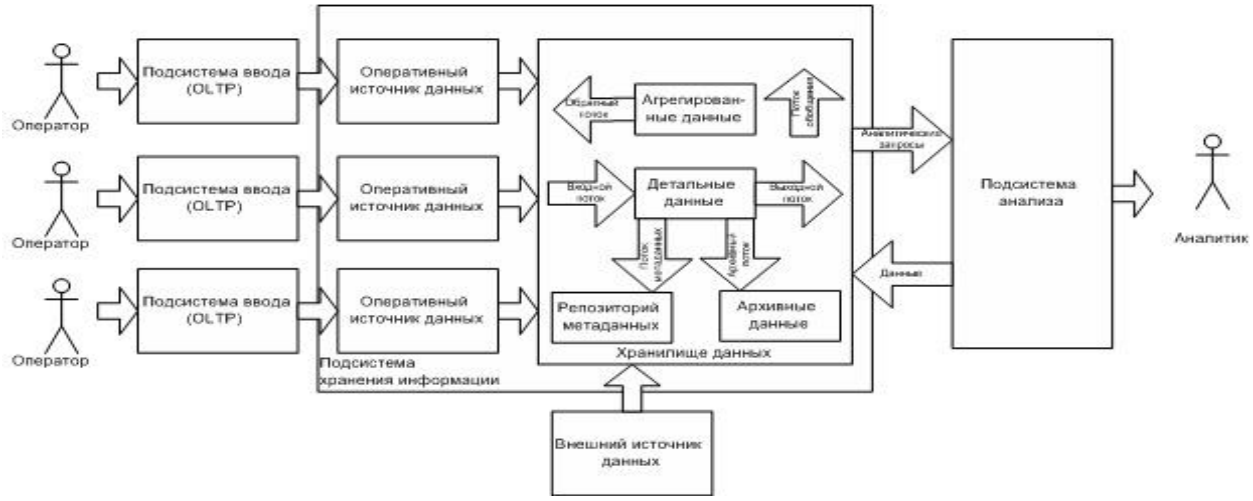
архивный поток - образуется перемещением детальных данных, количество обращений к которым снизилось;

поток метаданных - образуется потоком информации о данных в репозиторий данных;

выходной поток - образуется данными, извлекаемыми пользователями;

обратный поток - образуется очищенными данными, записываемыми обратно в OLTP-системы.

Категории данных в ХД

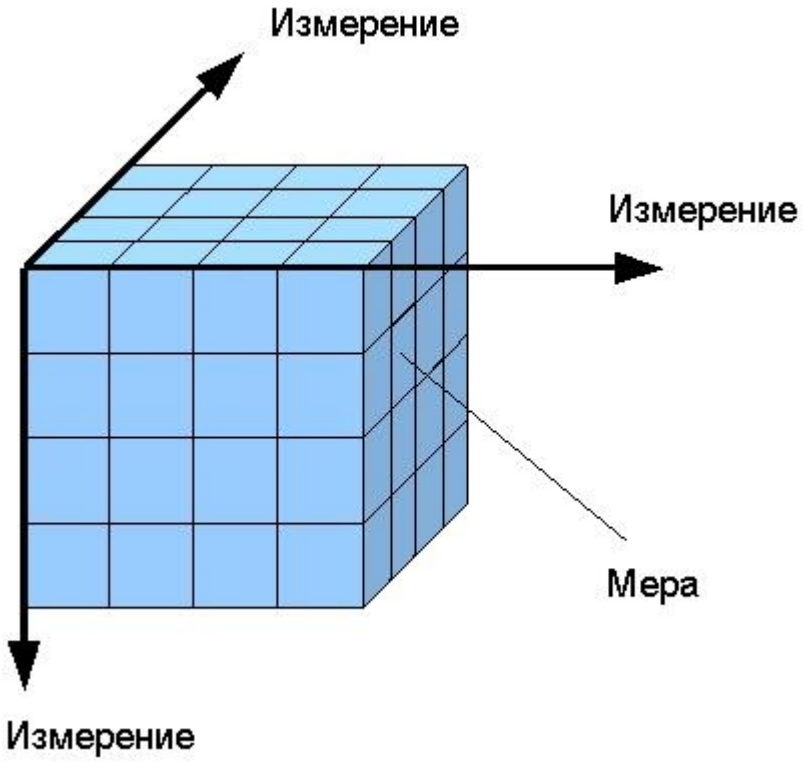


Структура OLAP-куба

В процессе анализа данных часто возникает необходимость построения зависимостей между различными параметрами, число которых может быть значительным.

Под измерением будем понимать последовательность значений одного из анализируемых параметров. Например, для параметра "время" это - последовательность дней, месяцев, кварталов, лет.

Возможность анализа зависимостей между различными параметрами предполагает возможность представления данных в виде многомерной модели - гиперкуба или OLAP-куба.



Гиперкуб

Оси куба представляют собой измерения, по которым откладывают параметры, относящиеся к анализируемой предметной области, например, названия товаров и названия месяцев года.

На пересечении осей измерений располагаются данные, количественно характеризующие анализируемые факты - меры, например, объемы продаж, выраженные в единицах продукции.

В простейшем случае двумерного куба получается таблица, показывающая значения уровней продаж по товарам и месяцам.

Дальнейшее усложнение модели данных возможно по нескольким направлениям:

- увеличение числа измерений данные о продажах не только по месяцам и товарам, но и по регионам. В этом случае куб становится трехмерным;
- усложнение содержимого ячейки например, нас может интересовать не только уровень продаж, но и *чистая прибыль* или остаток на складе. В этом случае в ячейке будет несколько значений;
- введение иерархии в пределах одного измерения общее понятие "время" связано с иерархией значений: год состоит из кварталов, квартал из месяцев и т.д.

Иерархия измерений OLAP-кубов

Каждое из измерений OLAP-куба может быть представлено в виде иерархической структуры. Например, измерение "Регион" может иметь следующие уровни иерархии: "страна - федеральный округ - область - город - район".

Некоторые измерения могут иметь несколько уровней иерархического представления, например измерение "время" - представление "год - квартал - месяц - день" и представление "год - неделя - день".

Точно так же в рамках измерения "География" можно ввести уровни "Страна", "Регион", "Область" и "Город".

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.

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.

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 (обычно это суррогатный ключ) 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.

Архитектура OLAP-систем

Полномасштабная OLAP-система должна выполнять сложные и разнообразные функции, включающие сбор данных из различных источников, их согласование, преобразование и загрузку в хранилище, хранение аналитической информации, регламентную отчетность, поддержку произвольных запросов, многомерный анализ и др.

В настоящее время существуют фактические стандарты построения OLAP-систем, основанных на концепции ХД. Эти стандарты опираются на современные исследования и общемировую практику создания хранилищ данных и аналитических систем.

В общем виде архитектура корпоративной OLAP-системы описывается схемой с тремя выделенными слоями:

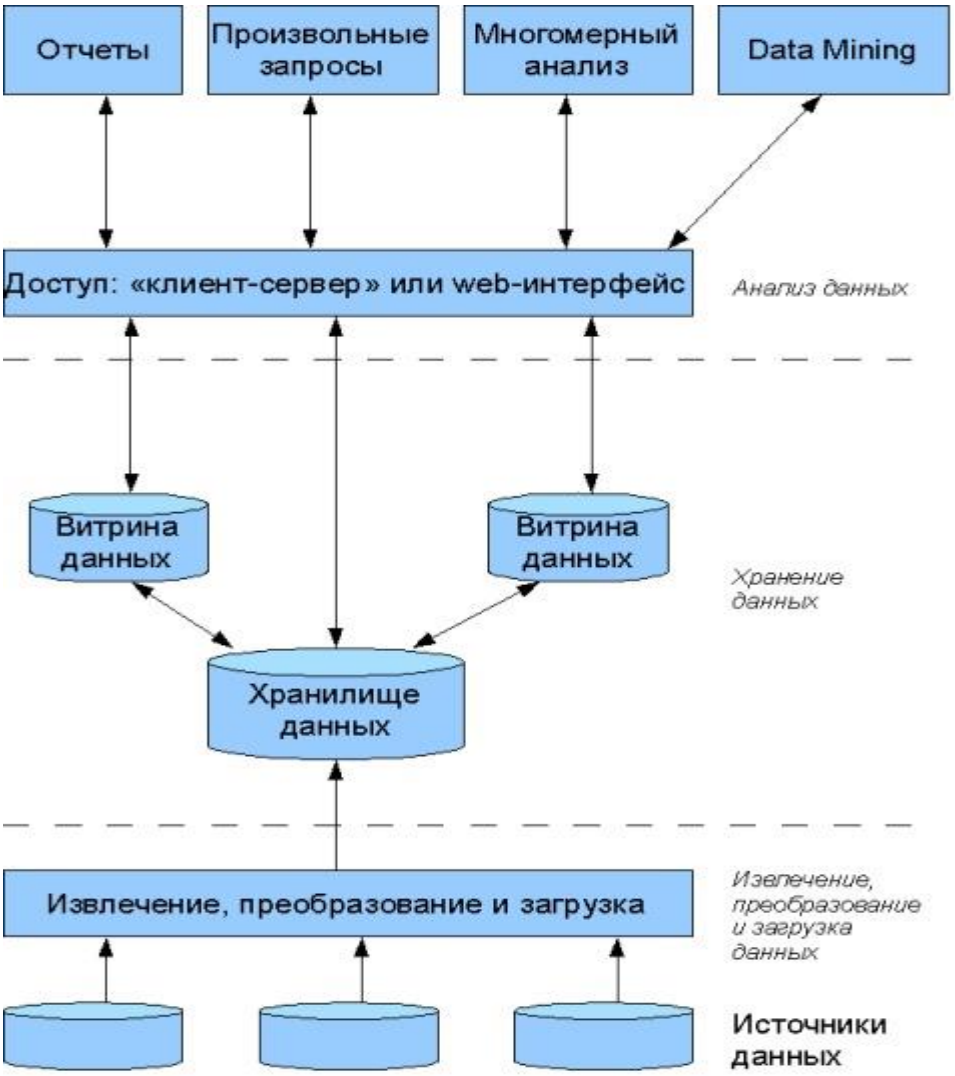


Рис. Архитектура корпоративной OLAP-системы

- извлечение, преобразование и загрузка данных;
- хранение данных;
- анализ данных.

Данные поступают из различных внутренних OLTP-систем, от подчиненных структур, от внешних организаций в соответствии с установленным регламентом, формами и макетами отчетности. Вся эта информация проверяется, согласуется, преобразуется и помещается в хранилище и витрины данных. После этого пользователи с помощью специализированных инструментальных средств получают необходимую им

информацию для построения различных табличных и графических представлений, прогнозирования, моделирования и выполнения других аналитических задач.

Слой извлечения, преобразования и загрузки данных

С организационной точки зрения, данный слой включает подразделения и структуры организации всех уровней, поддерживающие базы данных оперативного доступа. Он представляет собой низовой уровень генерации информации, уровень внутренних и внешних информационных источников, вырабатывающих "сырую" информацию. Эта информация является рабочей для повседневной деятельности различных подразделений, которые ее вырабатывают и используют.

С системно-технической точки зрения данный слой представлен ЛВС всех подразделений всех уровней, к которым подключены специализированные технические комплексы, хранящие информацию, чаще всего реализованные в виде реляционных СУБД.

Из источников данных информация перемещается на основе некоторого регламента в *централизованное хранилище*. Как правило, необходимые для хранилища данные не хранятся в окончательном виде ни в одной из OLTP-систем. Эти данные обычно можно получить из исходных баз данных путем специальных преобразований, вычислений и агрегирования.

Кроме того, несмотря на различную функциональную направленность, исходные транзакционные системы часто "пересекаются" по данным, т.е. их локальные базы данных содержат однотипную по смыслу информацию. Это, прежде всего, касается нормативно-справочной информации, которая используется в том или ином виде в любой OLTP-системе. При этом существенно, что одинаковые по смыслу данные обычно имеют в разных системах различный формат, вид представления, идентификацию, единицы измерения и т.п. Перед загрузкой в хранилище вся эта информация должна быть согласована, чтобы обеспечить целостность и непротиворечивость аналитических данных.

Согласование данных необходимо и при загрузке данных из одного источника. Дело в том, что в хранилище хранятся исторические данные, т.е. данные за достаточно большой промежуток времени. В оперативной системе данные хранятся в целостном виде за ограниченный промежуток, после чего они отправляются в архив. При изменениях в структуре или собственно данных архивы не подвергаются никакой дополнительной обработке, а хранятся в исходном виде. Следовательно, при необходимости иметь данные за достаточно большой период времени необходимо согласовывать архивную информацию с текущей.

Таким образом, загрузка данных из источников в хранилище осуществляется специальными процедурами, позволяющими:

- извлекать данные из различных баз данных, текстовых файлов;
- выполнять различные типы согласования и очистки данных;
- преобразовывать данные при перемещении их от источников к хранилищу;
- загружать согласованные и "очищенные" данные в структуры хранилища.

Слой хранения данных

Слой хранения данных предназначен непосредственно для хранения значимой, проверенной, согласованной, непротиворечивой и хронологически целостной информации, которую с достаточно высокой степенью уверенности можно считать достоверной.

Собственно ХД не ориентировано на решение какой-либо определенной функциональной аналитической задачи. Цель ХД - обеспечить целостность и поддерживать хронологию всевозможных корпоративных данных, и с этой точки зрения оно нейтрально по отношению к приложениям. В связи с этим в большинстве случаев для выполнения определенного комплекса функционально замкнутых аналитических задач рационально создавать витрины данных, в основе которых может быть как многомерная, так и реляционная модель данных. По существу витрина представляет собой относительно небольшое, но что самое важное, функционально-ориентированное ХД, в котором информация хранится специальным образом, оптимизированным с точки зрения решения конкретных аналитических задач некоторого подразделения или группы аналитиков.

ХД чаще всего реализуется в виде реляционной БД, работающей под управлением достаточно мощной реляционной СУБД. Такая СУБД должна поддерживать эффективную работу с терабайтными объемами информации, иметь развитые средства ограничения доступа, обеспечивать повышенный уровень надежности и безопасности, соответствовать необходимым требованиям по восстановлению и архивации.

Слой анализа данных

Для организации доступа аналитиков к данным ХД и ВД используются специализированные рабочие места, поддерживающие необходимые технологии как оперативного, так и долговременного анализа. Результаты работы аналитиков оформляются в виде отчетов, графиков, рекомендаций и сохраняются как на локальном компьютере, так и в общедоступном узле локальной сети.

Аналитическая деятельность в рамках корпорации достаточно разнообразна и определяется характером решаемых задач, организационными особенностями компании, уровнем и степенью подготовленности аналитиков.

В связи с этим современный подход к инструментальным средствам анализа не ограничивается использованием какой-то одной технологи. В настоящее время принято различать следующие основные виды аналитической деятельности:

- стандартная отчетность;
- нерегламентированные запросы;
- многомерный анализ (OLAP);
- извлечение знаний (data mining).

Каждая из этих технологий имеет свои особенности, определенный набор типовых задач и должна поддерживаться специализированной инструментальной средой.

Клиентские OLAP-средства

Клиентские OLAP-средства представляют собой приложения, осуществляющие вычисление агрегатных данных (сумм, средних величин, максимальных или минимальных значений) и их отображение, при этом сами агрегатные данные содержатся в кэше внутри адресного пространства такого OLAP-средства.

Если исходные данные содержатся в настольной СУБД, вычисление агрегатных данных производится самим OLAP-средством. Если же источник исходных данных - серверная СУБД, многие из клиентских OLAP-средств посылают на сервер SQL-запросы, содержащие оператор GROUP BY, и в результате получают агрегатные данные, вычисленные на сервере.

Как правило, OLAP-функциональность реализована в средствах статистической обработки данных (из продуктов этого класса на российском рынке широко распространены продукты компаний StatSoft и SPSS) и в некоторых электронных таблицах. В частности, средствами многомерного анализа обладает Microsoft Excel. С помощью этого продукта можно создать и сохранить в виде файла небольшой локальный многомерный OLAP-куб и отобразить его двух- или трехмерные сечения.

Надстройки к пакету приложений Microsoft Office для извлечения и обработки данных представляют собой ряд функций, обеспечивающих доступ к возможностям извлечения и обработки данных из приложений Microsoft Office, и тем самым позволяющих осуществлять прогностический анализ на локальном компьютере. Благодаря тому, что встроенные в службы платформы Microsoft SQL Server алгоритмы извлечения и обработки данных

доступны из среды приложений Microsoft Office, бизнес-пользователи могут легко извлекать ценную информацию из сложных наборов данных всего несколькими щелчками мыши. Надстройки к пакету приложений Office для извлечения и обработки данных дают конечным пользователям возможность выполнять анализ непосредственно в приложениях Microsoft Excel и Microsoft Visio.

В состав Microsoft Office входят три отдельных OLAP-компонента:

- клиент извлечения и обработки данных для Excel позволяет создавать проекты извлечения и обработки данных на базе служб SSAS и управлять ими из Excel;
- средства анализа таблиц для приложения Excel позволяют использовать встроенные в службы SSAS функции извлечения и обработки информации для анализа данных, хранящихся в таблицах Excel;
- шаблоны извлечения и обработки данных для приложения Visio позволяют визуализировать деревья решений, деревья регрессии, кластерные диаграммы и сети зависимостей на диаграммах Visio.

Ниже изображена сводная таблица Excel, используемая для доступа клиентов к данным служб аналитики.

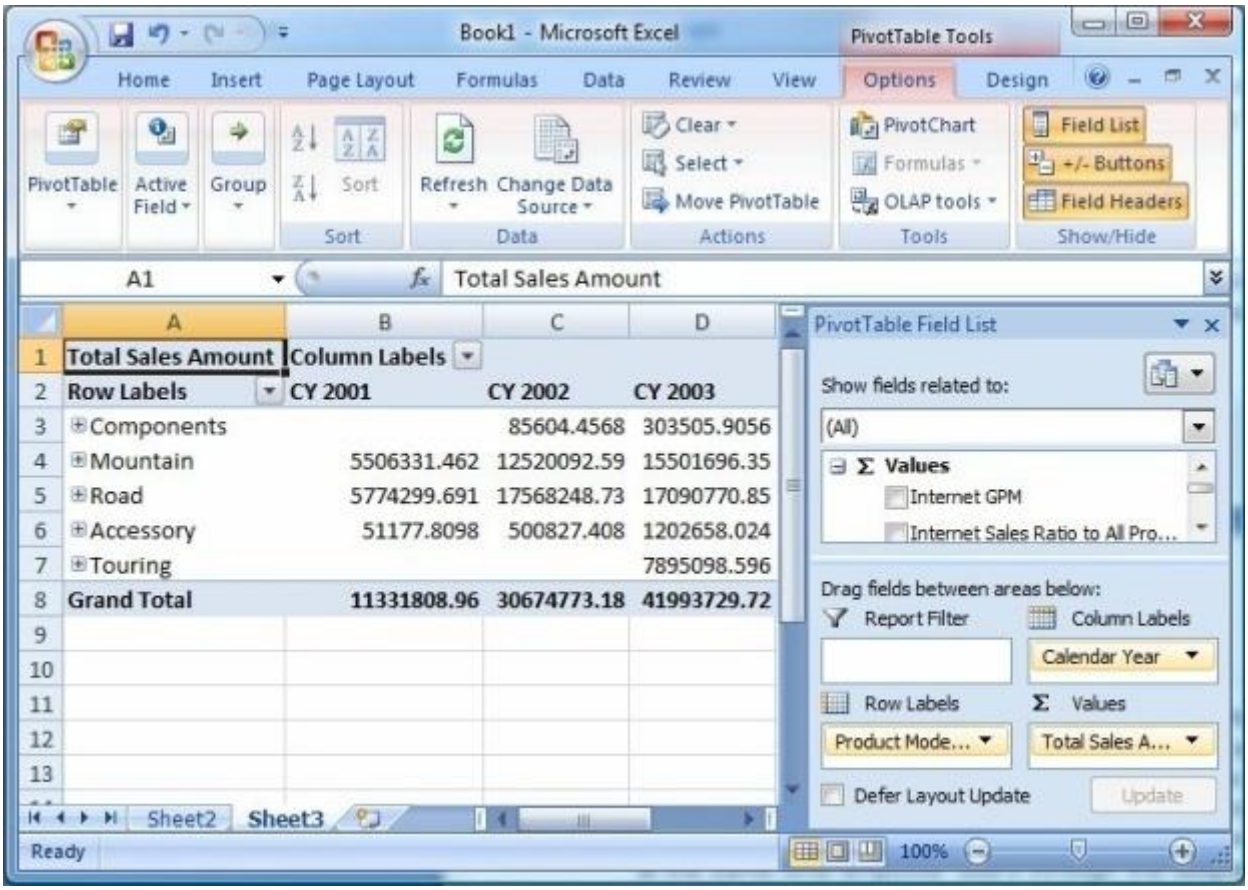


Рис. Сводная таблица Excel

С помощью приложения Microsoft Office Visio можно аннотировать, дополнять и отображать графические представления результатов извлечения

и обработки данных. Платформа SQL Server в сочетании с приложением Visio позволяет:

- визуализировать деревья решений, деревья регрессии, кластерные диаграммы и сети зависимостей;
- сохранять модели извлечения и обработки данных в виде документов Visio, внедренных в другие документы приложений Office или сохраненных в виде веб-страниц.

Клиентские OLAP-средства применяются, как правило, при малом числе измерений (обычно рекомендуется не более шести) и небольшом разнообразии значений этих параметров, - ведь полученные агрегатные данные должны умещаться в адресном пространстве подобного средства, а их количество растет экспоненциально при увеличении числа измерений. Поэтому даже самые примитивные клиентские OLAP-средства, как правило, позволяют произвести предварительный подсчет объема требуемой оперативной памяти для создания в ней многомерного куба.

Методы OLAP анализа

OLAP-сервис представляет собой инструмент для анализа больших объемов данных в режиме реального времени. Взаимодействуя с OLAP-системой, пользователь сможет осуществлять гибкий просмотр информации, получать произвольные срезы данных и выполнять аналитические операции детализации, свертки, сквозного распределения, сравнения во времени одновременно по многим параметрам. Вся работа с OLAP-системой происходит в терминах предметной области и позволяет строить статистически обоснованные модели деловой ситуации.

Программные средства OLAP - *это инструмент оперативного анализа данных*, содержащихся в хранилище. Главной особенностью является то, что эти средства ориентированы на использование не специалистом в области информационных технологий, не экспертом-статистиком, а профессионалом в прикладной области управления - менеджером отдела, департамента, управления, и, наконец, директором. Средства предназначены для общения аналитика *с проблемой, а не с компьютером*. На рисунке показан элементарный OLAP-куб, позволяющий производить оценки данных по трем измерениям.

Многомерный OLAP-куб и система соответствующих математических алгоритмов статистической обработки позволяет анализировать данные любой сложности на любых временных интервалах.

Имея в своем распоряжении гибкие механизмы манипулирования данными и визуального отображения, менеджер сначала рассматривает с разных сторон данные, которые могут быть (а могут и не быть) связаны с решаемой проблемой.

Далее он сопоставляет различные показатели бизнеса между собой, стараясь выявить скрытые взаимосвязи; может рассмотреть данные более пристально, детализировав их, например, разложив на составляющие по времени, по регионам или по клиентам, или, наоборот, еще более обобщить представление информации, чтобы убрать отвлекающие подробности. После этого с помощью модуля *статистического оценивания и имитационного моделирования* строится несколько вариантов развития событий, и из них выбирается наиболее приемлемый вариант.

У управляющего компанией, например, может зародиться гипотеза о том, что разброс роста активов в различных филиалах компании зависит от соотношения в них специалистов с техническим и экономическим образованием. Чтобы проверить эту гипотезу, менеджер может запросить из хранилища и отобразить на графике интересующее его соотношение для тех филиалов, у которых за текущий квартал рост активов снизился по сравнению с прошлым годом более чем на 10%, и для тех, у которых повысился более чем на 25%. Он должен иметь возможность использовать простой выбор из предлагаемого меню. Если полученные результаты ощутимо распадутся на две соответствующие группы, то это должно стать стимулом для дальнейшей проверки выдвинутой гипотезы.

В настоящее время быстрое развитие получило направление, называемое *динамическим моделированием* (Dynamic Simulation), в полной мере реализующее указанный выше принцип FASMI.

Используя динамическое моделирование, аналитик строит модель деловой ситуации, развивающуюся во времени, по некоторому сценарию. При этом результатом такого моделирования могут быть несколько новых бизнес - ситуаций, порождающих дерево возможных решений с оценкой вероятности и перспективности каждого.

Практически всегда задача построения аналитической системы для многомерного анализа данных - это задача построения *единой, согласованно функционирующей информационной системы, на основе неоднородных программных средств и решений*. И уже сам выбор средств для реализации ИС становится чрезвычайно сложной задачей. Здесь должно учитываться множество факторов, включая взаимную совместимость различных программных

компонент, легкость их освоения, использования и интеграции, эффективность функционирования, стабильность и даже формы, уровень и потенциальную перспективность взаимоотношений различных фирм производителей.

OLAP применим везде, где есть задача анализа многофакторных данных. Вообще, при наличии некоторой таблицы с данными, в которой есть хотя бы одна описательная колонка и одна колонка с цифрами, OLAP-инструмент будет эффективным средством анализа и генерации отчетов. В качестве примера применения OLAP-технологии рассмотрим исследование результатов процесса продаж.

Ключевые вопросы "Сколько продано?", "На какую сумму продано?" расширяются по мере усложнения бизнеса и накопления исторических данных до некоторого множества факторов, или разрезов: "..в Санкт-Петербурге, в Москве, на Урале, в Сибири…", "..в прошлом квартале, по сравнению с нынешним", "..от поставщика А по сравнению с поставщиком Б…" и т. д.

Ответы на подобные вопросы необходимы для принятия управленческих решений: об изменении ассортимента, цен, закрытии и открытии магазинов, филиалов, расторжении и подписании договоров с дилерами, проведения или прекращения рекламных кампаний и т. д.

Если попытаться выделить основные цифры (факты) и разрезы (аргументы измерений), которыми манипулирует аналитик, стараясь расширить или оптимизировать бизнес компании, то получится таблица, подходящая для анализа продаж как некий шаблон, требующий соответствующей корректировки для каждого конкретного предприятия.

Важный вопрос - наличие данных. Если они есть в каком-либо виде (Excel- или Access-таблица, данные из базы учетной системы, в виде структурированных отчетов филиалов), ИТ - специалист сможет передать их OLAP-системе напрямую или с промежуточным преобразованием. Для этого OLAP-системы имеют специальные инструменты конвертации данных.

После настройки OLAP-системы на данные пользователь получит возможность быстро получать ответы на ключевые вопросы путем простых манипуляций мышью над OLAP-таблицей и соответствующими меню. При этом будут доступны некоторые стандартные методы анализа, логически следующие из природы OLAP-технологии. К наиболее распространённым инструментам анализа можно отнести:

- Факторный (структурный) анализ
- Анализ динамики
- Анализ зависимостей

- Сопоставление
- Дисперсионный анализ

Факторный (структурный) анализ. Анализ структуры продаж для выявления важнейших составляющих в интересующем разрезе. Для этого удобно использовать, например, диаграмму типа "Пирог" в сложных случаях, когда исследуется сразу 3 измерения - "Столбцы". Например, в магазине "Компьютерная техника" за квартал продажи компьютеров составили $100000, фототехники -$10000, расходных материалов - $4500. Вывод: оборот магазина зависит в большой степени от продажи компьютеров (на самом деле, быть может, расходные материалы необходимы для продажи компьютеров, но это уже анализ внутренних зависимостей).

Анализ динамики (регрессионный анализ - выявление трендов). Выявление тенденций, сезонных колебаний. Наглядно динамику отображает график типа "Линия". Например, объемы продаж продуктов компании Intel в течение года падали, а объемы продаж Microsoft росли. Возможно, улучшилось благосостояние среднего покупателя, или изменился имидж магазина, а с ним и состав покупателей. Требуется провести корректировку ассортимента. Другой пример: в течение 3 лет зимой снижается объем продаж видеокамер.

Анализ зависимостей (корреляционный анализ). Сравнение объемов продаж разных товаров во времени для выявления необходимого ассортимента - "корзины". Для этого также удобно использовать график типа "Линия". Например, при удалении из ассортимента принтеров в течение первых двух месяцев обнаружилось падение продаж картриджей с порошком.

Сопоставление (сравнительный анализ). Сравнение результатов продаж во времени, или за заданный период, или для заданной группы товаров. В зависимости от количества анализируемых факторов (от 1 до 3-х) используется диаграмма типа "Пирог" или "Столбцы". Пример: сравнение результатов продаж однотипных магазинов для оценки качества работы менеджеров.

Дисперсионный анализ. Исследование распределения вероятностей и доверительных интервалов рассматриваемых показателей. Применяется для прогнозирования и оценки рисков.

Этими видами анализа возможности OLAP не исчерпываются. Например, применяя в качестве алгоритма вычисления промежуточных и окончательных итогов функции статистического анализа - дисперсию,

среднее отклонение, моды более высоких порядков, - можно получить самые изощренные виды аналитических отчетов.

OLAP-системы являются частью более общего понятия "интеллектуальные ресурсы предприятия" или "средства интеллектуального бизнес-анализа" (Business Intelligence - BI), которое включает в себя помимо традиционного OLAP-сервиса средства организации совместного использования данных и информации, возникающих в процессе работы пользователей хранилища. Технология Business Intelligence обеспечивает электронный обмен отчетными документами, разграничение прав пользователей, доступ к аналитической информации из Internet и Intranet.

Технологии Data Mining

В настоящее время элементы *искусственного интеллекта* активно внедряются в практическую деятельность менеджера. В отличие от традиционных систем искусственного интеллекта, технология интеллектуального поиска и анализа данных или *"добыча данных"* (Data Mining - DM), не пытается *моделировать* естественный интеллект, а *усиливает его возможности* мощностью современных вычислительных серверов, поисковых систем и хранилищ 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. Data Mining представляют большую ценность для руководителей и аналитиков в их повседневной деятельности. Деловые люди осознали, что с помощью методов Data Mining они могут получить ощутимые преимущества в конкурентной борьбе.

В основу современной технологии Data Mining (Discovery-driven Data Mining) положена концепция шаблонов (Patterns), отражающих фрагменты многоаспектных взаимоотношений в данных. Эти шаблоны представляют собой закономерности, свойственные выборкам 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. На рисунке показана схема преобразования данных с использованием технологии Data Mining.

Основой для всевозможных систем *прогнозирования* служит историческая информация, хранящаяся в БД в виде временных рядов. Если удается построить шаблоны, адекватно отражающие динамику поведения целевых показателей, есть вероятность, что с их помощью можно предсказать и поведение системы в будущем. На рисунке показан полный цикл применения технологии Data Mining.

Важное положение Data Mining - нетривиальность разыскиваемых шаблонов. Это означает, что найденные шаблоны должны отражать *неочевидные, неожиданные* (Unexpected) регулярности в данных, составляющие так называемые скрытые знания (Hidden Knowledge). К деловым людям пришло понимание, что *"сырые" данные* (Raw Data) содержат глубинный пласт знаний, и при грамотной его раскопке могут быть обнаружены настоящие самородки, которые можно использовать в конкурентной борьбе.

Сфера применения Data Mining *ничем не ограничена* - технологию можно применять всюду, где имеются огромные количества каких-либо "сырых" данных!

В первую очередь методы Data Mining заинтересовали коммерческие предприятия, развертывающие проекты на основе *информационных хранилищ данных* (Data Warehousing). Опыт многих таких предприятий показывает, что отдача от использования Data Mining может достигать 1000%. Известны

сообщения об экономическом эффекте, в 10-70 раз превысившем первоначальные затраты от 350 до 750 тыс. долларов. Есть сведения о проекте в 20 млн долларов, который окупился всего за 4 месяца. Другой пример - годовая экономия 700 тыс. долларов за счет внедрения Data Mining в одной из сетей универсамов в Великобритании.

Можно назвать пять стандартных типов закономерностей, выявляемых с помощью методов Data Mining:

- ассоциация,
- последовательность,
- классификация,
- кластеризация
- прогнозирование.

Ассоциация имеет место в том случае, если несколько событий связаны друг с другом. Например, исследование, проведенное в компьютерном супермаркете, может показать, что 55% купивших компьютер берут также и принтер или сканер, а при наличии скидки за такой комплект принтер приобретают в 80% 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.

Отдельно выделяется *задача прогнозирования* новых значений на основании имеющихся значений числовой последовательности (или нескольких последовательностей, между значениями в которых наблюдается корреляция). При этом могут учитываться имеющиеся тенденции (тренды), сезонность, другие факторы. Классическим примером является прогнозирование цен акций на бирже.

По способу решения задачи интеллектуального анализа можно разделить на два класса: обучение с учителем (от англ. supervisedlearning) и обучение без учителя (от англ. unsupervisedlearning). В первом случае

требуется обучающий набор данных, на котором создается и обучается модель интеллектуального анализа данных. Готовая модель тестируется и впоследствии используется для предсказания значений в новых наборах данных. Иногда в этом же случае говорят об управляемых алгоритмах интеллектуального анализа. Задачи классификации относятся как раз к этому типу.

Во втором случае целью является выявление закономерностей имеющихся в существующем наборе данных. При этом обучающая выборка не требуется. В качестве примера можно привести задачу анализа потребительской корзины, когда в ходе исследования выявляются товары, чаще всего покупаемые вместе. К этому же классу относится задача кластеризации.