

Фактографические системы

Отличительной особенностью этих подсистем является использование данных, как способа представления фактов, концепций или инструкций. Данные могут быть представлены в виде текста. При этом не ставится задача автоматического структурирования этого текста, т.е. если описание факта и имеет структуру, то автоматический поиск этой структуры не является задачей фактографической системы.

Фактографические системы берут свое начало с систем обработки данных. **Системы обработки данных (СОД) - комплекс взаимосвязанных методов и средств сбора и обработки данных, необходимых для организации управления объектами.** СОД основываются на применении ЭВМ и других современных средств информационной техники, поэтому их также называют автоматизированными системами обработки данных (АСОД)

Развитие фактографических систем связано с развитием технологий хранения и обработки информации.

Из истории развития **технологии хранения** можно выделить следующие этапы.

Первый этап - хранение данных в файлах. В том случае использовались файлы последовательного доступа и файлы с произвольным доступом. Отличительной особенностью этого этапа - работа программиста с низкоуровневыми операциями ввода, вывода, поиска и т.д.

Второй этап - использования баз данных. Используются также файлы, хранящие кроме данных еще и структуру данных. Развитие баз данных и систем управления базами данных (СУБД) является одним из основных факторов развития современных фактографических систем.

Появление OLAP – систем и хранилищ данных (Data Warehouse) связано с появлением технологий работы с многомерными таблицами данных. Вместе с тем, развиваются не только технологии программного уровня, но и технологии аппаратного уровня.

Функциональное развитие автоматизированных систем продолжается с появлением новых концепций обработки и представления информации, использованных в фактографических информационных системах. С появлением концепций информационно-управляющих систем, была добавлена **функция, направленная на обеспечение менеджеров необходимыми для принятия управленческих решений отчетами, составленными на основе собранных о процессе данных. Очень часто к информационно-управляющим системам относят системы принятия решений (СППР), экспертные системы (ЭС) и управленческие информационные системы.**

СППР

Концепция **систем поддержки принятия решений** (СППР, **decision support systems** - DDS) отражала потребность менеджеров в специализированном инструменте, обеспечивающего интерактивную поддержку процессов принятия уникальных решений.

Современные системы поддержки принятия решения (СППР), возникшие как естественное развитие и продолжение управленческих информационных систем и систем управления базами данных, представляют собой системы, максимально приспособленные к решению задач повседневной управленческой деятельности. Они являются инструментом, призванным оказать помощь **лицам, принимающим решения (ЛПР)**.

Обобщенная архитектура СППР состоит из 5 различных частей:

- 1.система управления данными,
- 2.система управления моделями,
- 3.машина знаний,
- 4.интерфейс пользователя,
- 5.пользователь.

Система поддержки принятия решений обладает следующими четырьмя основными характеристиками:

- 1.использование и данных, и моделей;
- 2.помощь менеджерам в принятии решений для слабоструктурированных и неструктурированных задач;
- 3.поддержка, а не замена, выработки решений менеджерами (лицо, принимающее решение - ЛПР);
- 4.применение с целью улучшения эффективности решений.

В первых системах СППР требовалось хранение больших массивов данных и выполнение с большой скоростью транзакций в распределенных системах. Это направление развития СППР привело к созданию систем оперативной обработки транзакции (**OLTP – On-line Transaction Processing**).

OLTP-системы проектируются для управления большим потоком транзакций, каждый из которых сопровождался внесением незначительных изменений в оперативные данные предприятий. **Данные системы должны иметь инструмент обработки информации** (операций, документов) в режиме реального времени. Объем данных может колебаться от нескольких мегабайт до терабайт и петабайт.

В тоже время, **такие системы трудно использовать для анализа хранимых данных.**

OLAP

В 1993 году Е. Коддом была предложена концепция инструментов, **реализующая оперативную аналитическую обработку данных (On-Line Analysis Processing - OLAP).**

По Кодду - **OLAP-технология — это технология комплексного динамического синтеза, анализа и консолидации больших объемов многомерных данных.**

Он же сформулировал 12 принципов OLAP, которые позже были переработаны в, так называемый, **тест Быстрый Анализ Разделяемой Многомерной Информации** - или кратко – **тест FASMI (Fast Analysis of Shared Multidimensional Information) :**

Fast (быстрый) — предоставление пользователю результатов анализа за приемлемое время (обычно не более 5 с), пусть даже ценой менее детального анализа;

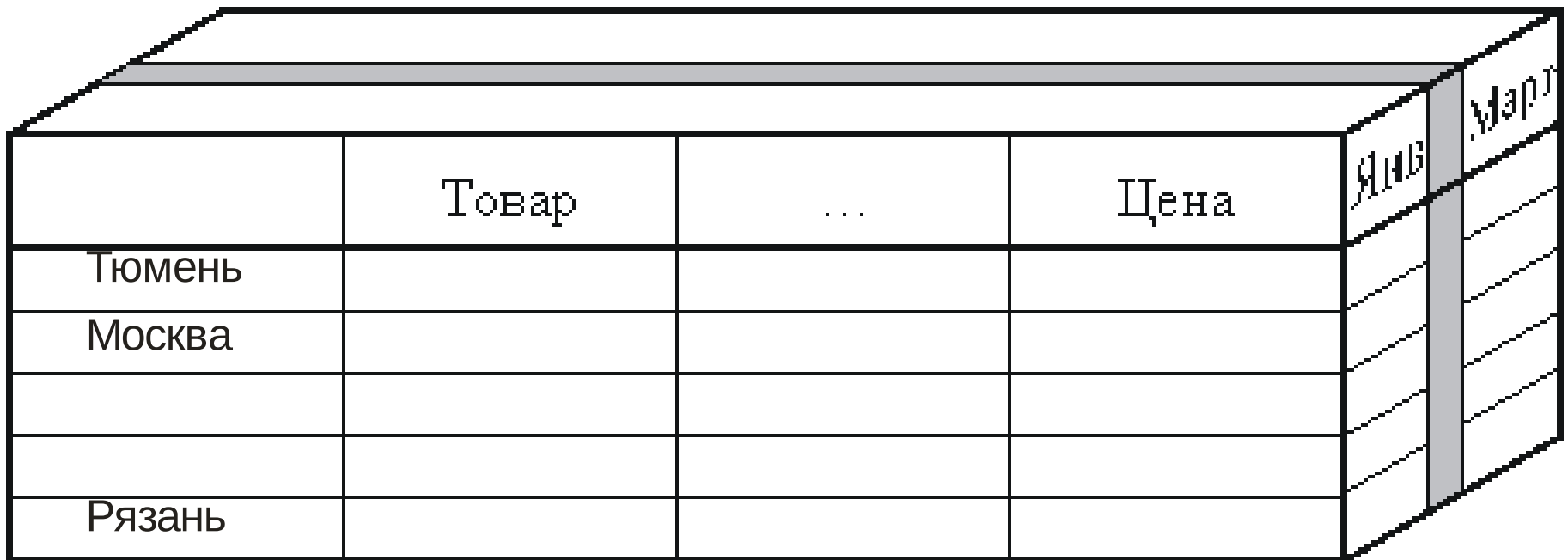
Analysis (анализ) — возможность осуществления любого логического и статистического анализа, характерного для данного приложения и его сохранения в доступном для конечного пользователя виде;

Shared (разделяемой) — многопользовательский доступ к данным с поддержкой соответствующих механизмов блокировок и средств авторизованного доступа;

Multidimensional (многомерной) — многомерное концептуальное представление данных, включая полную поддержку для иерархий и множественных иерархий (ключевое требование OLAP);

Information (информации) — возможность обращаться к любой нужной информации независимо от ее объема и места хранения.

В силу принципа «Multidimensional», **OLAP** ассоциируется с многомерным кубом или гиперкубом (рисунок). Гиперкуб является концептуальной логической моделью организации данных, а не физической реализацией их хранения, поскольку храниться такие данные могут и в реляционных таблицах.



	Товар	...	Цена
Тюмень			
Москва			
Рязань			

Над гиперкубом производятся аналитические OLAP-операции:

1.Сечение. При выполнении операции сечения формируется подмножество гиперкуба, в котором значение одного или более измерений фиксировано (например, год).

2.Вращение (rolling). Операция вращения изменяет порядок представления измерений, обеспечивая представление метакуба в более удобной для восприятия форме.

3.Drill Down/Up - Drill Down (Углубление в данные) или Drill Up (Консолидация (обобщение) как отдельных измерений, так и выбранных элементов измерений) - это специальная техника анализа, используемая при изучении данных. Направление детализации может быть задано как по иерархии отдельных измерений, так и прочим отношениям.

4. **Разбиение с поворотом (slicing and dicing).** Термин, используемый для описания функции сложного анализа данных: выборка данных из многомерного куба с заданными значениями и заданным взаимным расположением измерений. Пользователь при этом обычно использует операции вращения концептуального куба данных и детализации/агрегирования данных.

**Физическая организация концептуальной модели
возможна в трех вариантах :**

1. **MOLAP (Multidimensional OLAP);**
2. **ROLAP (Relational OLAP);**
3. **HOLAP (Hybrid OLAP).**

MOLAP (Multidimensional OLAP)

В MOLAP-модели многомерное представление данных реализуется физически. В специализированных СУБД, основанных на многомерном представлении данных, данные организованы не в форме реляционных таблиц, а в виде упорядоченных многомерных массивов.

Достоинства:

- 1. высокая производительность, т.к. в случае использования многомерных СУБД поиск и выборка данных осуществляется значительно быстрее, чем при использовании реляционных баз данных;**
- 2. в MOLAP легко встраиваются различные функции, которые трудно реализовать в SQL.**

Недостатки MOLAP-модели:

- 1.не позволяют работать с большими массивами данных;**
- 2.трудности хранения и обработки разреженных данных (данные или неизвестны или нулевые);**

Область применения:

- 1.объем исходных данных для анализа не слишком велик (не более нескольких гигабайт);**
- 2.набор информационных измерений стабилен**
(поскольку любое изменение в их структуре почти всегда требует полной перестройки гиперкуба);
- 3.время ответа системы на нерегламентированные запросы является наиболее критичным параметром.**

ROLAP (Relational OLAP)

Системы оперативной аналитической обработки реляционных данных (ROLAP) позволяют представлять данные, хранимые в реляционной базе в многомерной форме, обеспечивая преобразование информации в многомерную модель через промежуточный слой метаданных.

В этом случае гиперкуб эмулируется СУБД на логическом уровне.

Для большинства хранилищ данных наиболее эффективным способом моделирования N-мерного куба фактов является схема «звезда» (рисунок).

Таблица покупателей

Ключ "Покупатель"
Описание покупателя
Ключ "Населенный пункт"
Описание нас. пункта
Ключ "Регион"
Описание региона
Ключ "Государство"
Описание государства
Код уровня

Таблица фактов

Ключ "Поставщик"
Ключ "Покупатель"
Ключ "Продукт"
Ключ "Период"
Количество
Цена за единицу
Стоимость

Таблица поставщиков

Ключ "Поставщик"
Описание поставщика

Таблица периодов

Ключ "Период"
Описание
Год
Квартал
Месяц
День
Код уровня

Таблица продуктов

Ключ "Продукт"
Описание продукта
Ключ "Категория"
Описание категории
Ключ "Производитель"
Описание производителя
Код уровня

Основными составляющими структуры хранилищ данных являются *таблица фактов* (fact table) и *таблицы измерений* (dimension tables). В сложных задачах с многоуровневыми измерениями используются различные расширения схемы «звезда» — схема «*снежинка*».

Достоинства:

1. размер хранилища не является критичным параметром, как в случае MOLAP;
2. внесение изменений в структуру измерений не требует физической реорганизации базы данных, как в случае MOLAP;
3. реляционные СУБД обеспечивают высокий уровень защиты данных.

Недостаток:

1. меньшая производительность.

HOLAP (Hybrid OLAP)

Гибридные системы (Hybrid OLAP, HOLAP) разработаны с целью совмещения достоинств и минимизации недостатков, присущих предыдущим классам.

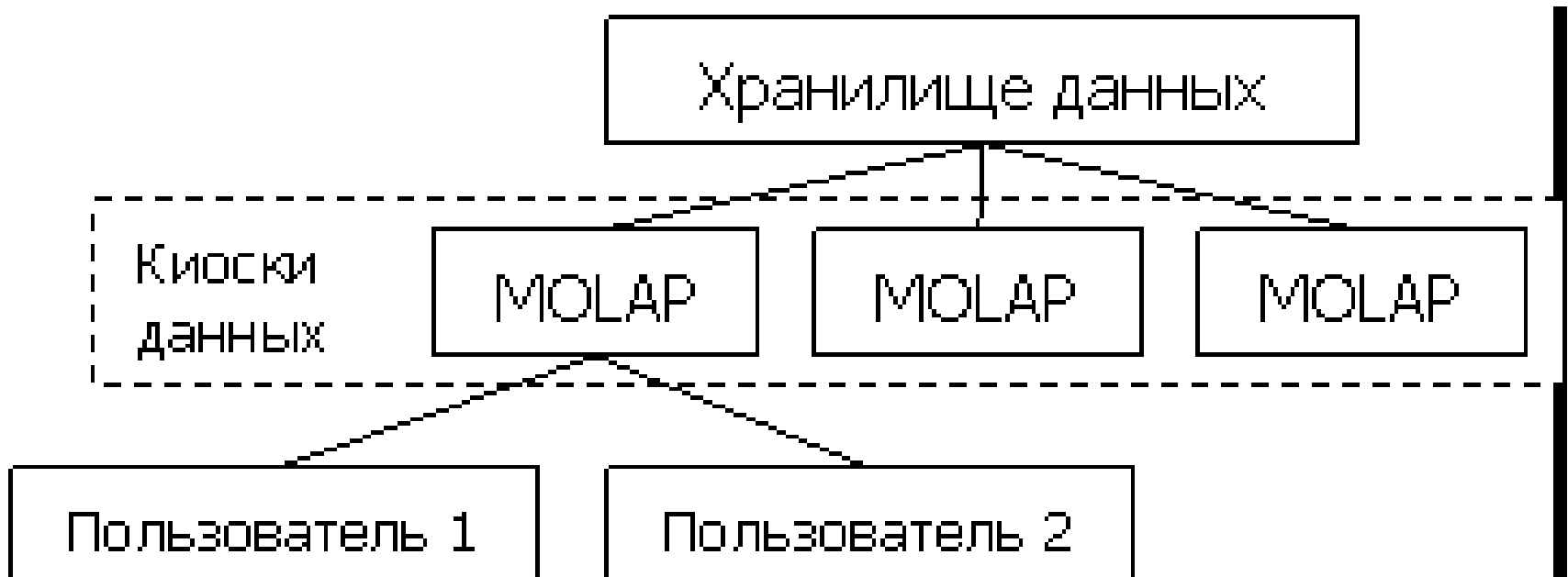
Основные данные хранятся в реляционной базе (ROLAP), агрегированные — в многомерной структуре (кубе MOLAP), так как ситуация, когда для анализа нужны все данные, возникает достаточно редко.

Многомерные данные представляются в виде киосков данных (рисунок). Построенный куб данных анализируется средствами многомерного OLAP.

HOLAP (Hybrid OLAP)

Достоинства:

- 1. относительная простота инсталляции, администрирования и сопровождения;**
- 2. способность каждого пользователя создавать свои собственные кубы данных.**



С OLAP - технологией тесно связано понятие хранилищ данных. Изначально хранилища данных были предложены фирмой IBM как решение, обеспечивающее доступ к данным, накопленным в нереляционных системах.

В настоящее время хранилища данных являются рабочей средой для СППР, которая включает не только технологию управления данными, но и их анализ.

В таблице приведены данные сравнения OLTP - систем и хранилищ данных.

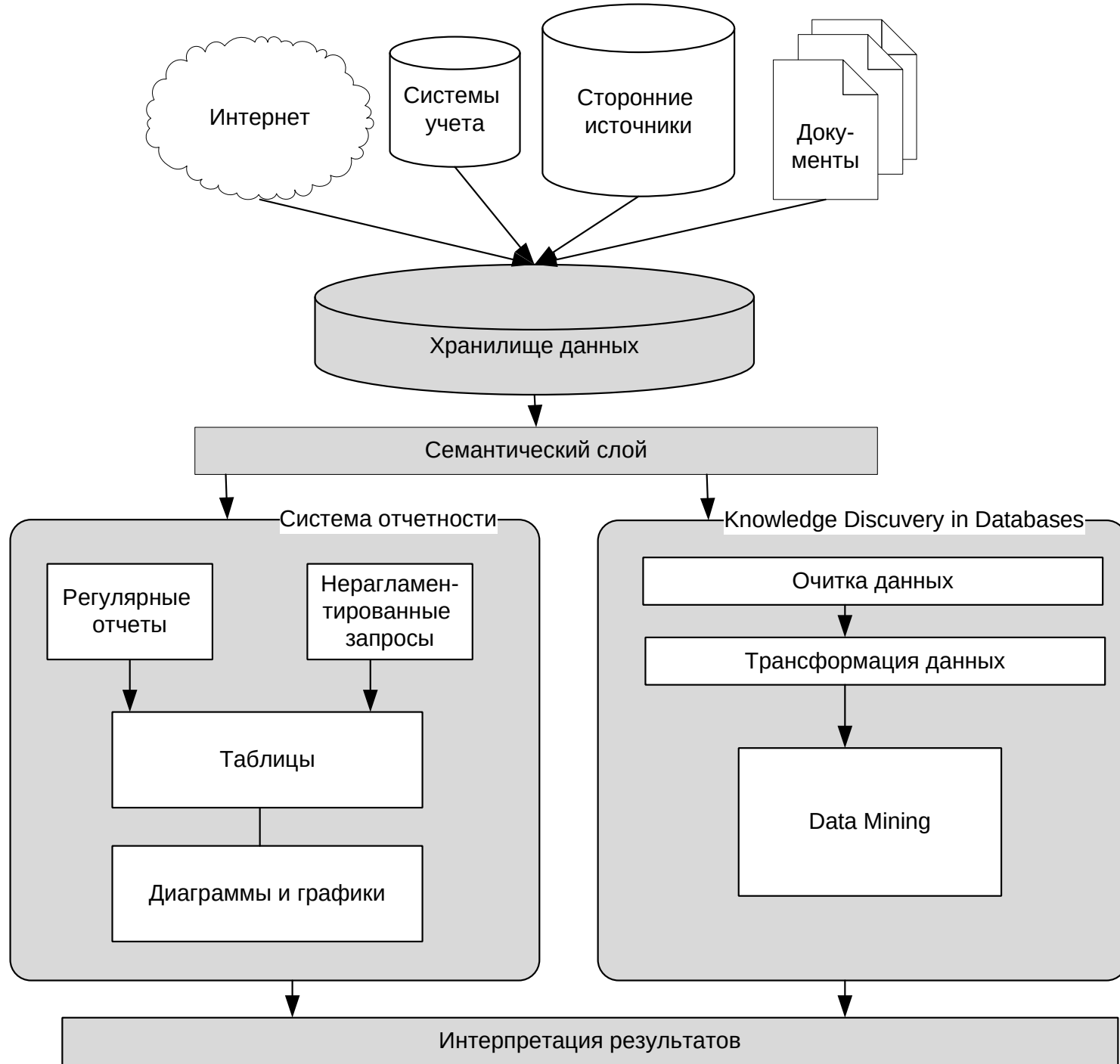
Хранилище данных (data warehouse) - предметно-ориентированная информационная корпоративная база данных, предназначенная для подготовки отчетов, анализа бизнес-процессов и поддержки принятия решений.

OLTP система	Хранилище данных
Содержит текущие данные	Содержит исторические данные
Хранит подробные сведения	Хранит подробные сведения, а также частично и значительно обобщенные данные
Данные являются динамическими	Данные в основном являются статическими
Повторяющийся способ обработки данных	Нерегламентированный, неструктурированный и эвристический способ обработки данных
Высокая интенсивность обработки транзакций	Средняя и низкая интенсивность обработки транзакций
Предсказуемый способ использования данных	Непредсказуемый способ использования данных
Предназначена для обработки транзакций	Предназначена для проведения анализа
Ориентирована на прикладные области	Ориентирована на предметные области
Поддержка принятия повседневных решений	Поддержка принятия стратегических решений
Обслуживает большое количество работников исполнительного звена	Обслуживает малое количество работников руководящего звена

Логическим продолжением применения хранилища данных являются системы бизнес - анализа (*Business Intelligence*). *Business Intelligence* - интеллектуальный инструментарий, позволяющий решать проблемы доступа к разнородным данным, построению отчетов пользователей и анализу данных.

Данные системы могут включать: хранилища данных, запросы конечного пользователя и инструмент для создания отчетов, OLAP инструменты, Data Mining инструменты.

Архитектура Business Intelligence представлена на рисунке.



Knowledge Discovery in Databases (KDD) - процесс поиска полезных знаний в «сырых данных».

KDD включает в себя процессы: подготовки данных, выбора информативных признаков, очистки данных, применения методов Data Mining, постобработки данных.

KDD не задает набор методов обработки и алгоритмов, он определяет последовательность действий, которые необходимо сделать для того, чтобы из исходных данных получить знания.

Data mining (Добыча данных) - процесс аналитического исследования больших массивов информации (обычно экономического характера) с целью выявления определенных закономерностей и систематических взаимосвязей между переменными, которые затем можно применить к новым совокупностям данных.

Этот процесс отнюдь не является аналогом поиска отклонений в данных при помощи OLAP-инструментария и включает три основных этапа: исследование, построение прогнозирующей модели или структуры и ее проверку.

Как правило, приложения для «добычи данных» существенно отличаются от OLAP-продуктов и в большей степени предназначены непосредственно для специалистов.

В системах «добычи данных» реализованы совершенно другие инструментальные средства от производителей ПО иного, чем средства OLAP.

В основу data mining заложены готовые фрагменты, отражающие фрагменты данных (*паттерны данных*).

Он позволяет определить заранее неизвестные типы закономерности из известных. При этом применяются различные алгоритмы для нахождения знаний:

- 1.нейронные сети,**
- 2.деревья решений,**
- 3.алгоритмы кластеризации,**
- 4.установления ассоциаций,**
- 5.фильтрации,**
- 6.нечеткая логика,**
- 7.ассоциативные правила и т.д.**

Алгоритм Шеннона — Фано

Алгоритм использует коды переменной длины: часто встречающийся символ кодируется кодом меньшей длины, редко встречающийся — кодом большей длины.

Коды Шеннона — Фано **префиксные**, то есть никакое кодовое слово не является префиксом любого другого. Это свойство позволяет однозначно декодировать любую последовательность кодовых слов.