OLAP системы и технологии

OLTP и OLAP

- Значительная часть корпоративной информации ~ 90% - лежит невостребованной и никак не анализируется.
- => Необходимы технологии, которые бы позволили анализировать накопленную информацию и предоставили бы возможность оперативно принимать решения.
- Зачастую имеет место серьезное недопонимание различий в возможностях, назначении и роли технологий, предназначенных для сбора данных, - OLTP-систем и технологий анализа данных.

- Задачи OLTP-системы – это быстрый сбор и оптимальное размещение данных в БД, а также обеспечение их полноты, актуальности и согласованности.
- Однако такие системы не предназначены для эффективного, быстрого и многоаспектного анализа.
- По собранным данным можно строить отчеты, но это требует от бизнес-аналитика или постоянного взаимодействия с IT-специалистом, или специальной подготовки в области программирования и вычислительной техники.

- Традиционный процесс принятия решений в российской компании, использующей информационную систему, построенную на OLTP-технологии:
  1) Менеджер дает задание специалисту информационного отдела в соответствии со своим пониманием вопроса.
  2) Специалист информационного отдела, по-своему осознав задачу, строит запрос оперативной системе, получает электронный отчет и доводит его до сведения руководителя.

Недостатки такой схемы принятия решений:

- используется малое количество данных;
- процесс занимает длительное время;
- требуется повторение цикла в случае необходимости уточнения данных или рассмотрения данных в другом разрезе, а также при возникновении дополнительных вопросов;
- ИТ специалист и руководитель мыслят разными категориями => непонимание
- сложность электронных отчетов (в цифровом виде) для восприятия => ИТ специалист вынужден отвлекаться на рутинную работу по составлению таблиц, диаграмм и т.д.

- **Выход из этой ситуации** – исходная информация должна быть доступна ее непосредственному потребителю – аналитику (Билл Гейтс – «Информация на кончиках пальцев»).
- OLAP-технология и предназначена для этого.
- Инструменты OLAP-технологии позволяют бизнес-аналитикам даже без специальной подготовки самостоятельно (**непосредственно**) и оперативно получать всю необходимую для исследования закономерностей бизнеса информацию в различных комбинациях и срезах.
- При этом максимальный отклик любого отчета не превышает ~5 секунд.

Основы OLAP

- **OLAP** – технологии интерактивной аналитической обработки данных в системах БД, предназначенные для поддержки принятия решений и ориентированные гл. образом на нерегламентированные интерактивные запросы.
- Термин OLAP был введен Э. Коддом в 1993г.
- По способам организации источников данных систем OLAP различают технологии:
  - **ROLAP** (Relational OLAP),
  - **MOLAP** (Multi-Dimensional OLAP),
  - **HOLAP** (Hybrid OLAP).

- В качестве источников данных часто используют *хранилища данных*.
- Обеспечивает *многомерный анализ данных* (с т. зр. их концептуального представления).
- Основная структура – *N-мерный куб данных.*
- Куб данных обладает 2-мя или более независимыми *измерениями* (атрибутами) => система координат пространства данных.
- Совокупности координат соответствуют значения данных в точках куба, называемые *элементами* (Item) или *ячейками* (Cell).
- Для анализа на многомерном кубе делают *«срезы»* (обычные двумерные таблицы)

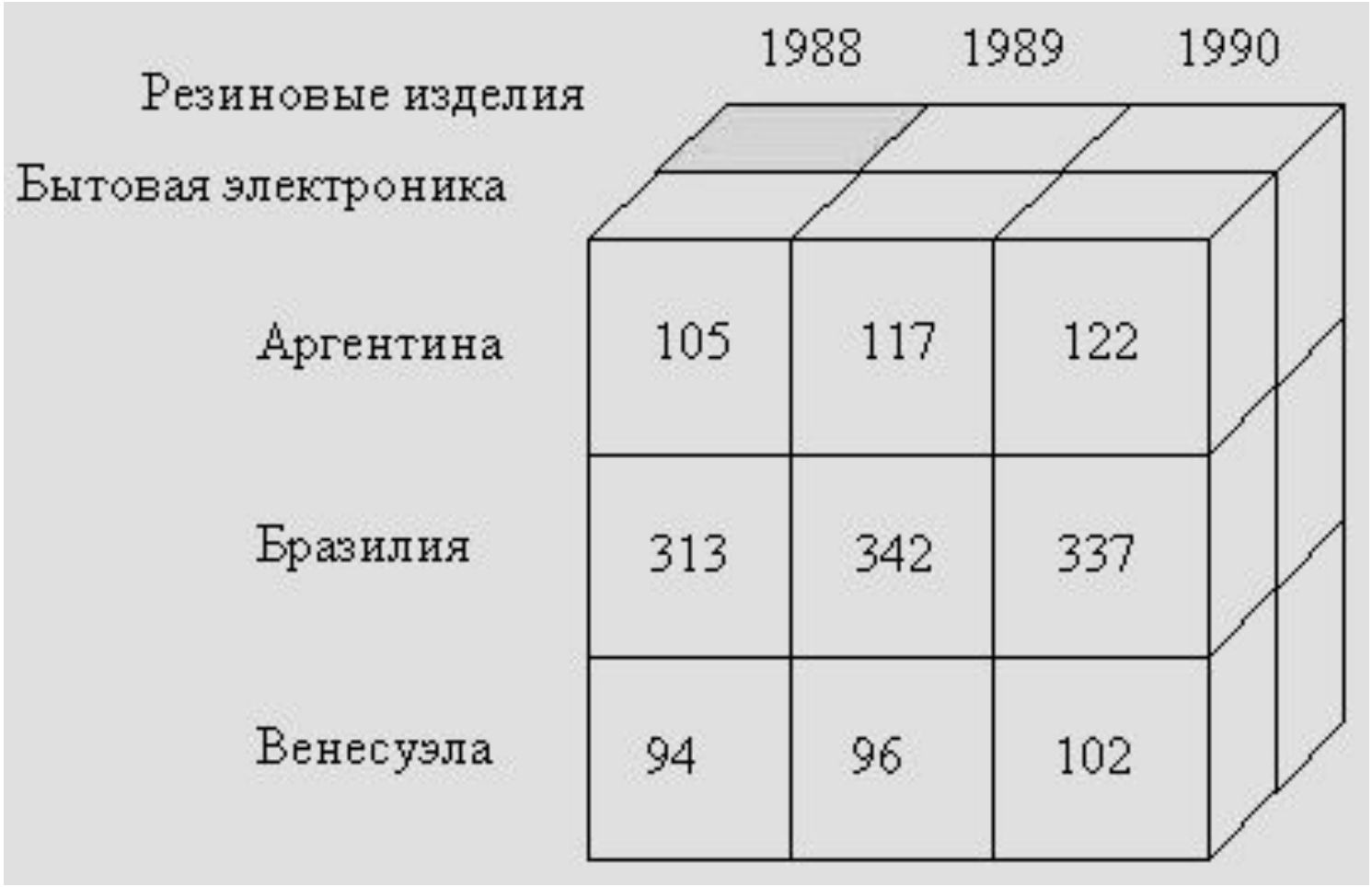
OLAP (On-Line Analytical Processing)

- OLAP – это совокупность концепций, принципов и требований, лежащих в основе программных продуктов, облегчающих аналитикам доступ к данным.
- Аналитика не интересует одиночный факт - ему нужна информация о сотнях и тысячах подобных событий (причем, без лишних подробностей).
- Задача аналитика – находить закономерности в больших массивах данных.
- Данные, которые требуются аналитику, обязательно содержат числовые значения.

Итак, аналитику нужно много данных, эти данные являются выборочными, а также носят характер «набор атрибутов – число»:

Страна	Товар	Год	**Объем продаж**
Аргентина	Бытовая электроника	1988	105
Аргентина	Бытовая электроника	1989	117
Аргентина	Бытовая электроника	1990	122
Аргентина	Резиновые изделия	1989	212
Аргентина	Резиновые изделия	1990	217
Бразилия	Бытовая электроника	1988	313
Бразилия	Бытовая электроника	1989	342
Бразилия	Бытовая электроника	1990	337
Бразилия	Резиновые изделия	1988	515
Бразилия	Резиновые изделия	1989	542
Бразилия	Резиновые изделия	1990	566
Венесуэла	Бытовая электроника	1988	94
Венесуэла	Бытовая электроника	1989	96
Венесуэла	Бытовая электроника	1990	102
Венесуэла	Резиновые изделия	1988	153
Венесуэла	Резиновые изделия	1989	147
Венесуэла	Резиновые изделия	1990	162

Трехмерное представление таблицы (**куб** OLAP):



- В общем случае куб может быть многомерным (~ до 20 измерений) – «система координат»
- В принципе, все измерения равноправны

- Измерения OLAP-кубов (например: *страна, товар, год*) состоят из т.н. **меток** или **членов** (members). Например: измерение "Страна" состоит из меток "Аргентина", "Бразилия", "Венесуэла" и так далее.
- Элементы куба м.б. не заполнены (нет данных) – «вакуум».
- Куб (гиперкуб) – это логическое представление данных (для пользователя). Данные физически не обязательно хранятся в многомерной структуре. Благодаря спец. способам компактного хранения многомерных данных решается проблема «вакуума» (бесполезной траты памяти)

- Куб сам по себе не пригоден для восприятия и анализа человеком (нельзя адекватно представить более 3-х измерений).
- Перед употреблением из n-мерного куба извлекают обычные двумерные таблицы. Эта операция называется **«разрезанием»** (slice) куба.
- При «разрезании» куба оставляются только необходимые измерения (обычно не больше двух), остальные измерения – фиксируются на интересующих аналитика метках.
- **Пример**: фиксируем измерение «Товары» на метке «Бытовая электроника» и анализируем объемы продаж по странам и годам.

Страна	Товар	Год	Объем продаж
Аргентина	Бытовая электроника	1988	105
Аргентина	Бытовая электроника	1989	117
Аргентина	Бытовая электроника	1990	122
Аргентина	Резиновые изделия	1989	212
Аргентина	Резиновые изделия	1990	217
Бразилия	Бытовая электроника	1988	313
Бразилия	Бытовая электроника	1989	342
Бразилия	Бытовая электроника	1990	337
Бразилия	Резиновые изделия	1988	515
Бразилия	Резиновые изделия	1989	542
Бразилия	Резиновые изделия	1990	566

- Данные в таблице не являются первичными, а получены в результате агрегирования более мелких элементов:
  - Год => кварталы => месяцы => недели => дни.
  - Страна => регионы => населенные пункты =>. районы => конкретные торговые точки.

- Такие многоуровневые объединения значений атрибутов-измерений называется **иерархиями**

Пример иерархии:

- Исходные данные берутся из нижних уровней иерархий, а затем суммируются для получения значений более высоких уровней.
- Средства OLAP дают возможность в любой момент перейти на нужный уровень иерархии с помощью операций ***агрегации*** (aggregation) и ***детализации*** (drill-down).
- Для ускорения процесса перехода, просуммированные значения для разных уровней хранятся в кубе.
- Операция ***поворота*** (rotation) позволяет изменить порядок измерений в кубе данных нужным для пользователя образом.

- Средства OLAP позволяют значительно повысить эффективность работы аналитика с данными по сравнению с OLTP-системами.
- Аналитик непосредственно работает с заранее подготовленными (загруженными из OLTP БД) данными, оптимизированными для быстрой аналитической обработки (нет необходимости каждый раз обрабатывать тысячи и миллионы первичных данных).
- Кубы OLAP представляют собой, по сути, многомерные отчеты. Разрезая многомерные кубы по измерениям, аналитик получает интересующие его "обычные" двумерные отчеты.

Тест FASMI (требования к продуктам OLAP):

- **Fast** (Быстрый) - время доступа к аналитическим данным - порядка 5 секунд;
- **Analysis** (Анализ) - возможность осуществлять числовой и статистический анализ;
- **Shared** (Разделяемый доступ) - возможность работы с информацией многим пользователям одновременно;
- **Multidimensional** (Многомерность) - см. выше;
- **Information** (Информация) - возможность получать нужную информацию, в каком бы электронном хранилище данных она не находилась.

Хранилища данных (Data Warehouse)

- Хранилище данных (ХД) и OLAP - две разные технологии. Однако, в комплексных решениях обе технологии применяются совместно.
- **Задача ХД** – интеграция, актуализация и согласование оперативных данных из разнородных источников для формирования единого непротиворечивого взгляда на объект управления в целом.
- ХД используются для составления отчетности, проведения оперативной аналитической обработки и глубинного анализа данных (Data Mining).

Понятие хранилища данных:

- **Хранилище данных** — система, содержащая непротиворечивую интегрированную предметно-ориентированную совокупность исторических данных крупной корпорации или иной организации с целью поддержки принятия стратегических решений.
- Хранилище:

  (1) собирает, (2) очищает, (3) загружает, (4) агрегирует, (5) хранит данные и (6) предоставляет к ним быстрый доступ.
- Основной источник данных - учетные системы (OLTP)

Билл Инмон («отец» хранилищ данных):

- **Хранилища данных** - "предметно ориентированные, интегрированные, неизменчивые, поддерживающие хронологию наборы данных, организованные с целью поддержки управления" и призванные выступать в роли "единого и единственного источника истины", который обеспечивает менеджеров и аналитиков достоверной информацией, необходимой для оперативного анализа и принятия решений.

- **Предметная ориентация** – данные объединены в категории и сохраняются соответственно областям, которые они описывают, а не применениям, их использующим.
- **Интегрированность** – данные удовлетворяют требованиям всего предприятия, а не одной функции бизнеса (одинаковые отчеты, сгенерированные для разных аналитиков, будут содержать одинаковые результаты).
- **Неизменность** – попав один раз в хранилище, данные там сохраняются и не изменяются. Данные могут лишь добавляться.

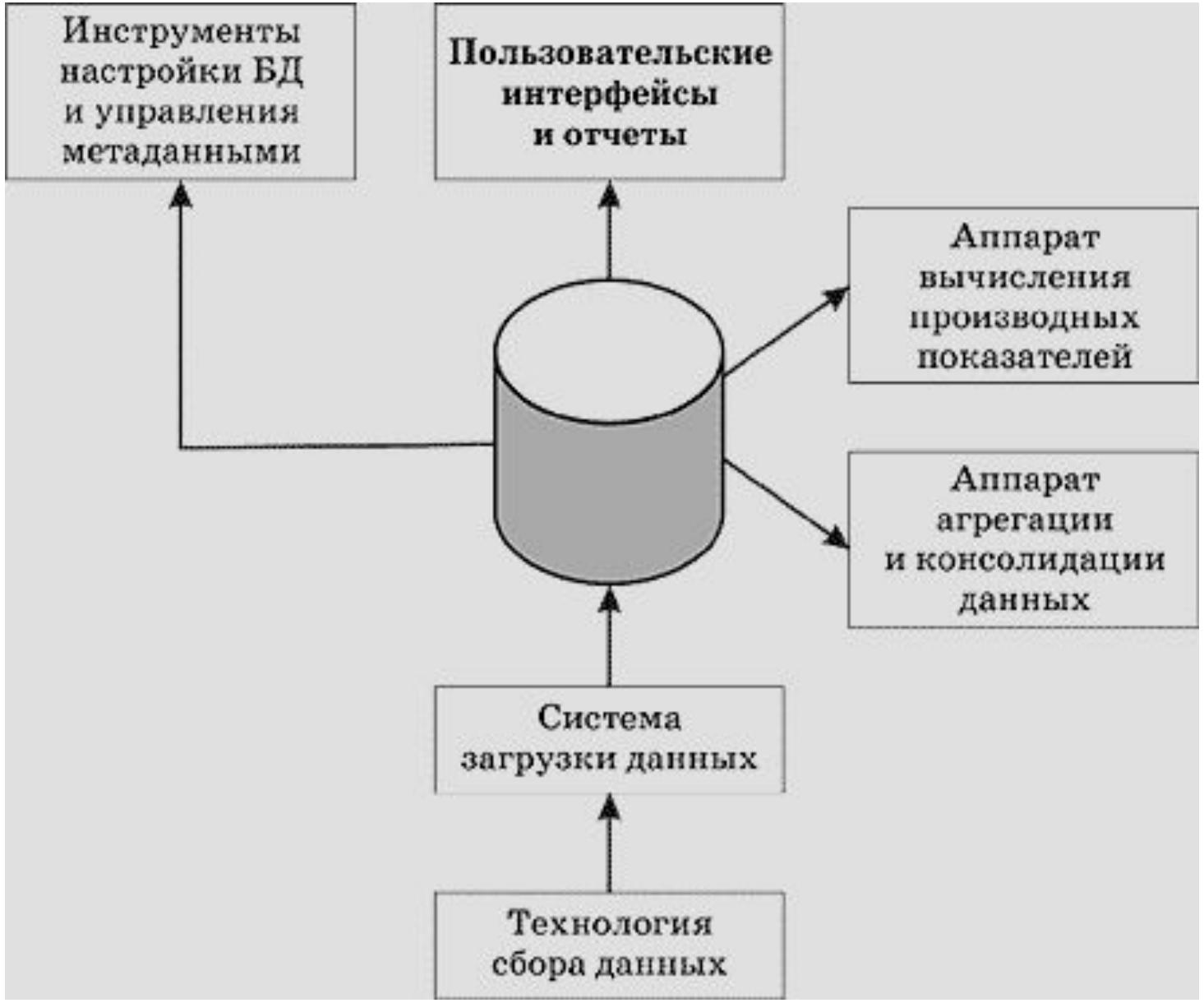
- **Привязка ко времени** – хранилище можно рассматривать как совокупность "исторических" данных: возможно восстановление данных на любой момент времени. Атрибут времени явно присутствует в структурах хранилища данных.
  - Т.о., хранилище данных представляет собой своеобразный накопитель информации о деятельности предприятия.
  - ХД изначально технологически оптимизированы не для ввода, а для быстрого поиска и анализа информации => имеют другую архитектуру БД (структура часто денормализована)

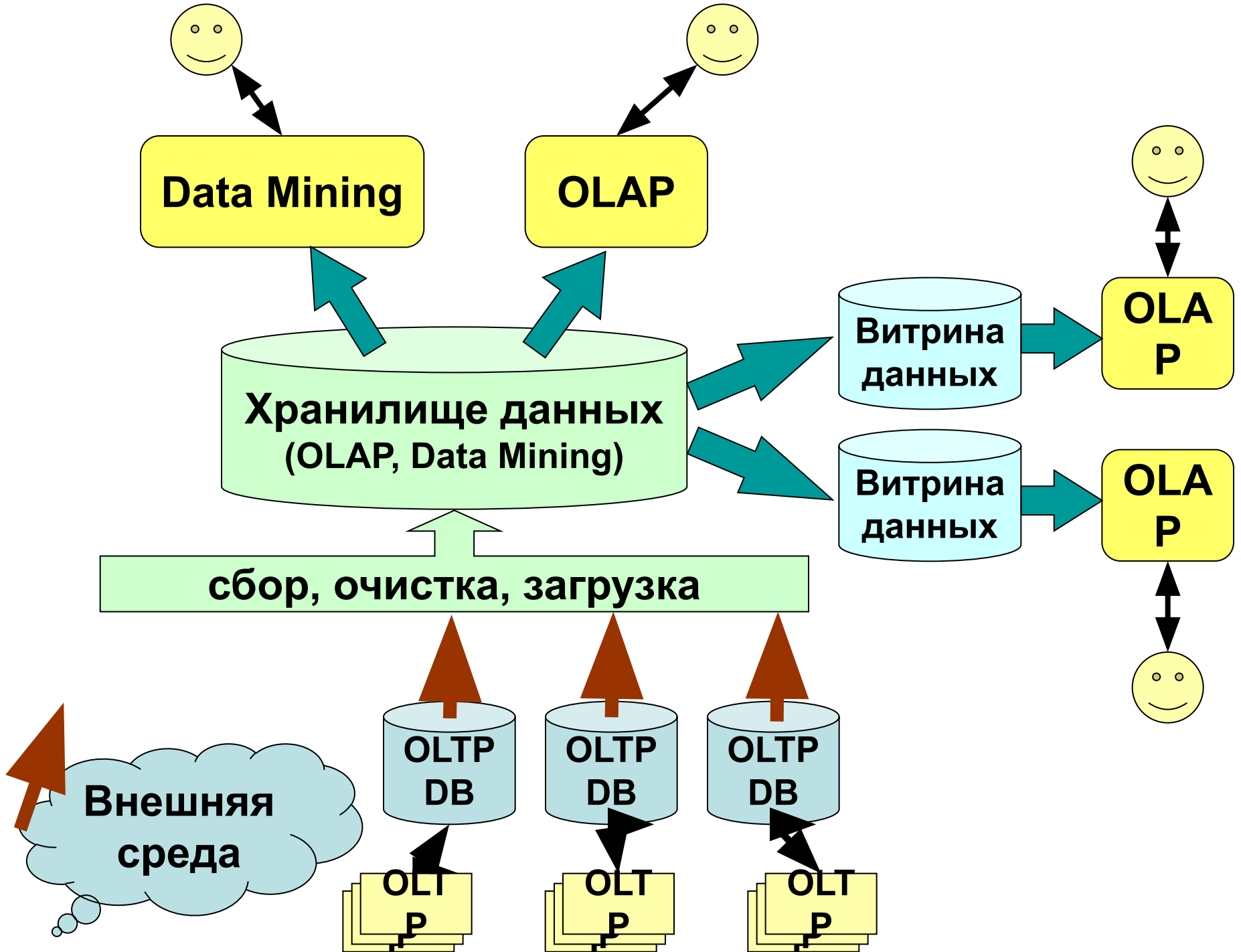
В дополнение к единому ХД могут создаваться т.н. ***витрины данных***

Витрина данных (Data Mart) – хранилище данных, связанных с какими-либо *конкретными аспектами деятельности* организации.

- ✔ Используется для поддержки принятия решений в интересах какого-либо подразделения организации или обеспечения какой-либо сферы ее деятельности.
- ✔ Источником данных может быть общее хранилище данных организации.

Архитектура Хранилища данных




Data Mining
OLAP
Хранилище данных
(OLAP, Data Mining)
Витрина данных
Витрина данных
OLA P
OLA P
сбор, очистка, загрузка
Внешняя среда
OLTP DB
OLTP DB
OLTP DB
OLT P
OLT P
OLT P

Контрольные вопросы:

1. Сущность и назначение операции *разрезания* (slice) куба OLAP
2. Сущность и назначение *иерархий* значений в измерениях куба OLAP
3. Сущность и назначение *Хранилищ данных*
4. *Приведите схемы реализации многомерного представления данных с помощью реляционных таблиц (использовать доп. литературу)*