

Лекция-3

Основные понятия системы баз данных.
Концепция хранилищ данных.

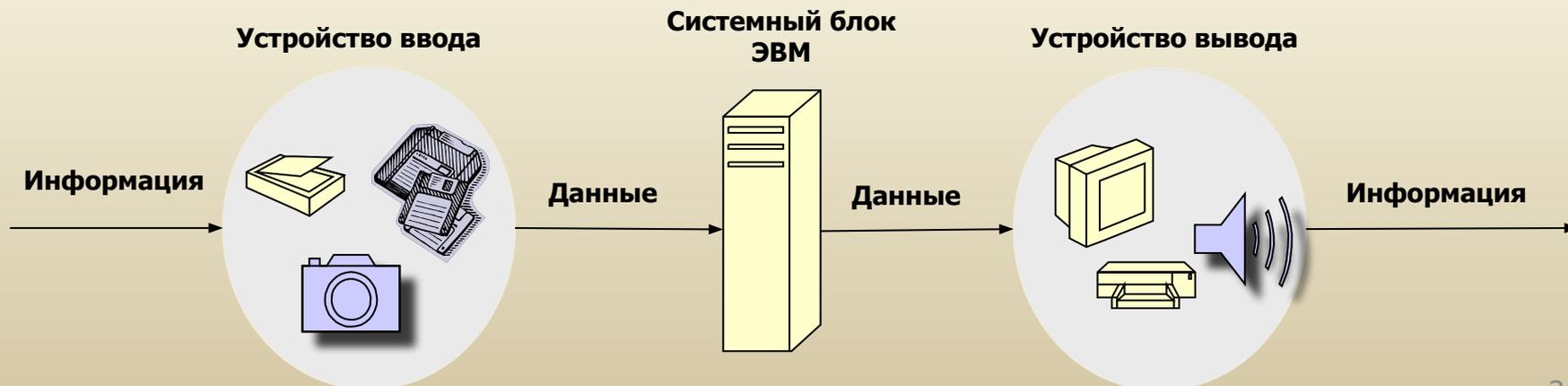
Изучаемые вопросы

1. Основные понятия и определения системы баз данных.
Предметная область ИС.
2. Системы класса OLTP (On-Line Transaction Processing).
Хранилища данных (DWH, Data Warehouse).

1. Основные понятия и определения системы баз данных. Предметная область ИС.

Автоматизированные информационные процессы оперируют машинным представлением информации – **данными**.

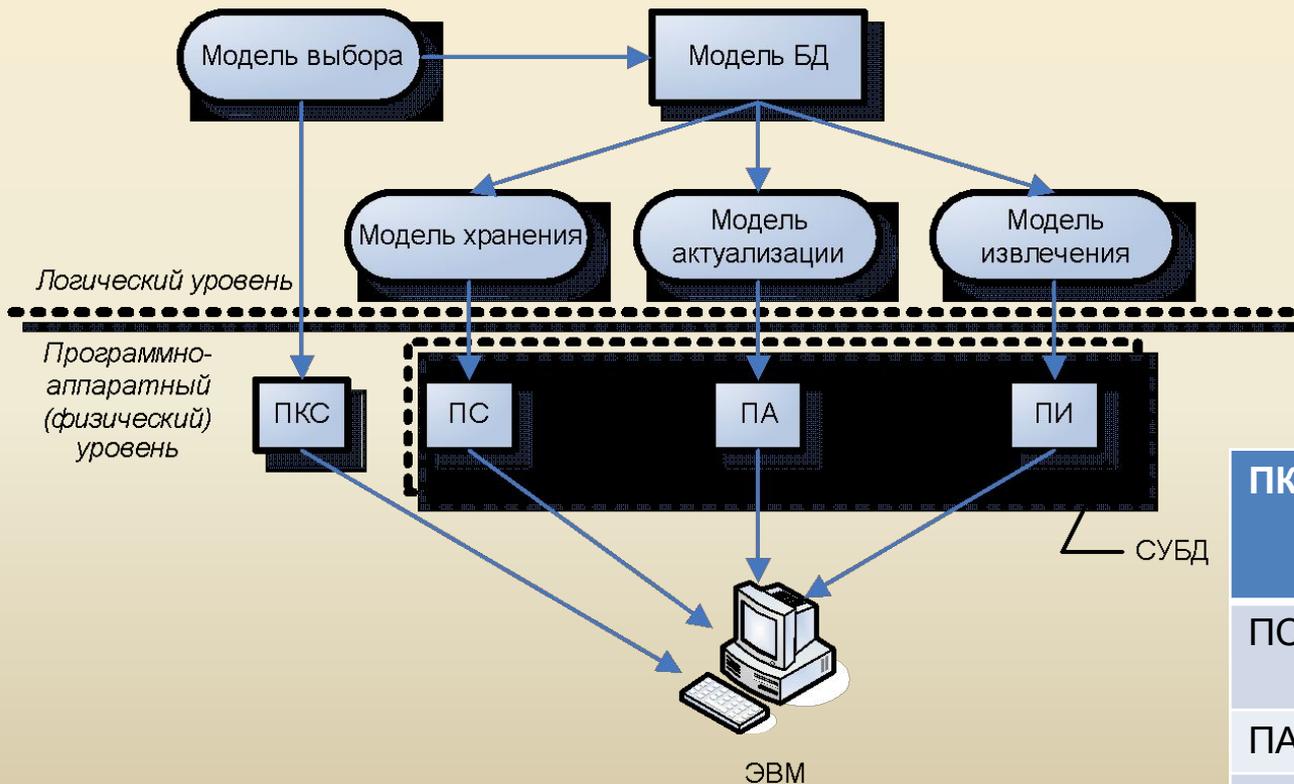
Информация в машинном виде, т. е. в форме электрических, магнитных и тому подобных сигналов и состояний, носит название **данных**.



Процесс накопления данных

Назначение технологического процесса накопления данных состоит в создании, хранении и поддержании в актуальном состоянии информационного фонда, (информационных ресурсов) необходимого для выполнения функциональных задач системы управления, для которой построен контур информационной технологии. Процесс накопления данных состоит из ряда основных процедур, таких, как **выбор хранимых данных, хранение данных, их актуализация и извлечение.**

Логический (модельный) уровень процесса накопления связан с физическим через программы, осуществляющие создание канонической структуры БД, схемы ее хранения и работу с данными.



PKS	Программа создания канонической структуры БД
PS	Программа создания структуры хранения БД
PA	Программа актуализации
PI	Программа извлечения

Базы данных

База данных – совместно используемый набор логически связанных данных (и описание этих данных), предназначенный для удовлетворения информационных потребностей организации.

Признаки баз данных:

- Это единое, большое информационное хранилище, которое однократно определяется, а затем используется одновременно многими пользователями.
- Все данные собраны вместе с минимальной долей избыточности (в отличие от разрозненных файлов с избыточными данными).
- Хранит не только рабочие данные, но и их описания. По этой причине базу данных еще называют набором интегрированных записей с самоописанием.

Файловые системы

Файловые системы - набор программ, которые выполняют для пользователей некоторые операции.

При этом каждая программа определяет свои собственные данные и управляет ими.

Ограничения, присущие файловым системам:

- Разделение и изоляция данных.
- Дублирование данных.
- Несовместимость файлов.

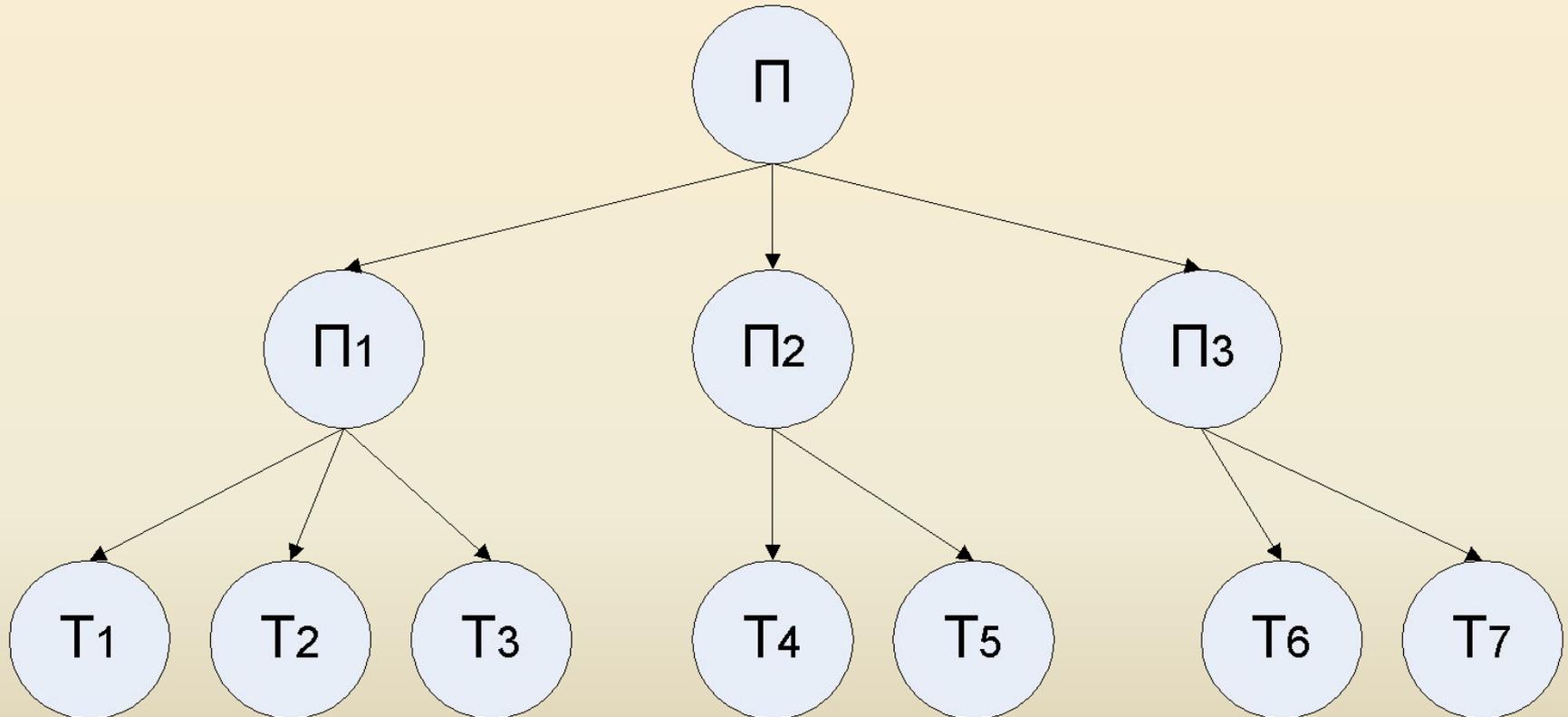
Модели баз данных

Модели баз данных базируются на предположении, что структуры данных обладают относительной устойчивостью. Поэтому возможно построение базы данных с постоянной структурой и изменяемыми значениями данных.

№	Название модели баз данных	Краткая характеристика модели
1.	Иерархическая модель	Представляет данные в виде древовидного графа, в котором объекты располагаются по уровням соподчиненности (иерархии) объектов.
2.	Сетевая модель	Представляет данные в виде диаграммы связей между основным и зависимым объектами (без ограничения на число обратных связей).
3.	Реляционная модель	Представляет данные в виде двумерных таблиц и связей (отношений) между ними.

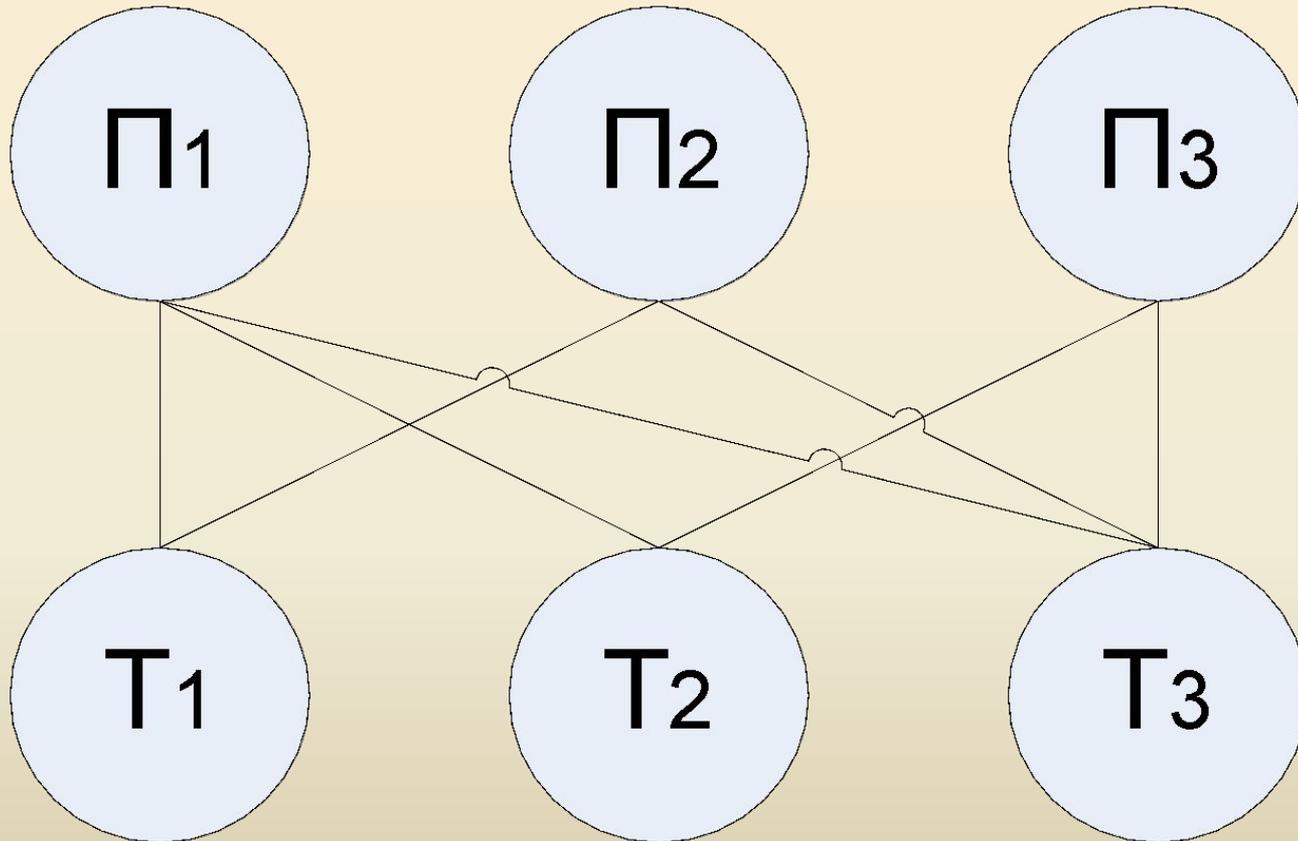
Иерархическая модель базы данных

Иерархическая модель представляет данные в виде древовидного графа, в котором объекты располагаются по уровням соподчиненности (иерархии) объектов.



Сетевая модель базы данных

Сетевая модель представляет данные в виде диаграммы связей между основным и зависимым объектами.



Реляционная модель базы данных

В **реляционной модели** базы данных отсутствуют различия между объектами и взаимосвязями. Данные представляются в виде двумерных таблиц и связей (отношений) между ними – тоже в виде таблиц.

R1 (поставщики)

П1
П2
П3

R2 (товары)

Т1
Т2
Т3

R3 (поставка товаров)

П1	Т1
П1	Т2
П1	Т3
П2	Т1
П2	Т3
П3	Т2
П3	Т3

Системы управления базами данных

Система управления базами данных (СУБД) - программное обеспечение, с помощью которого пользователи могут создавать и поддерживать (актуализировать) базу данных, а также осуществлять к ней контролируемый доступ.

СУБД обеспечивает взаимодействие прикладных программ пользователя с базой данных.



Функции СУБД

№	Функция	Описание
1	Хранение, извлечение и обновление данных	СУБД должна предоставлять пользователям возможность сохранять, извлекать и обновлять данные в базе данных.
2	Каталог, доступный конечным пользователям	СУБД должна иметь доступный конечным пользователям каталог, в котором хранится описание элементов данных.
3	Поддержка транзакций	СУБД должна иметь механизм, который гарантирует выполнение либо всех операций обновления данной транзакции, либо ни одной из них.
4	Сервисы управления параллельностью	СУБД должна иметь механизм, который гарантирует корректное обновление базы данных при параллельном выполнении операций обновления многими пользователями.
5	Сервисы восстановления	СУБД должна предоставлять средства по восстановлению базы данных на случай какого-либо ее повреждения или разрушения.
6	Сервисы контроля доступа к данным	СУБД должна иметь механизм, гарантирующий возможность доступа к базе данных только санкционированных пользователей.
7	Поддержка обмена данными	СУБД должна обладать способностью к интеграции с коммуникационным программным обеспечением.
8	Службы поддержки целостности данных	СУБД должна обладать инструментами контроля за тем, чтобы данные и их изменения соответствовали заданным правилам.
9	Службы поддержки независимости от данных	СУБД должна обладать инструментами поддержки независимости программ от фактической структуры базы данных
10	Вспомогательные службы	СУБД должна предоставлять вспомогательные службы: <ul style="list-style-type: none">■ утилиты импорта и экспорта;■ средства мониторинга характеристик функционирования и производительности;■ и другие

Понятие транзакции

Транзакция - это последовательность операторов манипулирования данными, выполняющаяся как единое целое и переводящая базу данных из одного целостного состояния в другое целостное состояние.

Транзакция обладает четырьмя важными свойствами, известными как свойства ACID:

№	Свойство	Описание
1	Атомарность	Транзакция выполняется как атомарная операция - либо выполняется вся транзакция целиком, либо она целиком не выполняется.
2	Согласованность	Транзакция переводит базу данных из одного согласованного (целостного) состояния в другое согласованное (целостное) состояние. Внутри транзакции согласованность базы данных может нарушаться.
3	Изоляция	Транзакции разных пользователей не должны мешать друг другу (например, как если бы они выполнялись строго по очереди).
4	Долговечность	Если транзакция выполнена, то результаты ее работы должны сохраниться в базе данных, даже если в следующий момент произойдет сбой системы.

2. Системы класса OLTP (On-Line Transaction Processing). Хранилища данных (DWH, Data Warehouse).

OLTP (Online Transaction Processing) — онлайн-обработка транзакций.

OLTP-системы предназначены для ввода, структурированного хранения и обработки информации (операций, документов) в режиме реального времени и обслуживают в первую очередь текущую, повседневную деятельность отдельных подразделений компании: склад, бухгалтерию, кадры и пр.

OLTP-системы автоматизируют структурированные, повторяющиеся задачи обработки данных, такие как ввод заказов и банковские транзакции.

Характеристики OLTP-систем:

- Проектируются, настраиваются и оптимизируются для выполнения максимального количества транзакций за короткие промежутки времени.
- Относительно простые алгоритмически, чаще всего используется фиксированный набор надежных и безопасных методов ввода, модификации, удаления данных и выпуска оперативной отчетности.
- Относительно невысокая компьютерная квалификация персонала (пользователей).
- Малое время отклика на запрос.
- Относительно короткие запросы, участие в запросах небольшого числа таблиц.
- Показателем эффективности является количество транзакций, выполняемых за секунду.
- Обычно аналитические возможности OLTP-систем сильно ограничены (либо вообще отсутствуют).

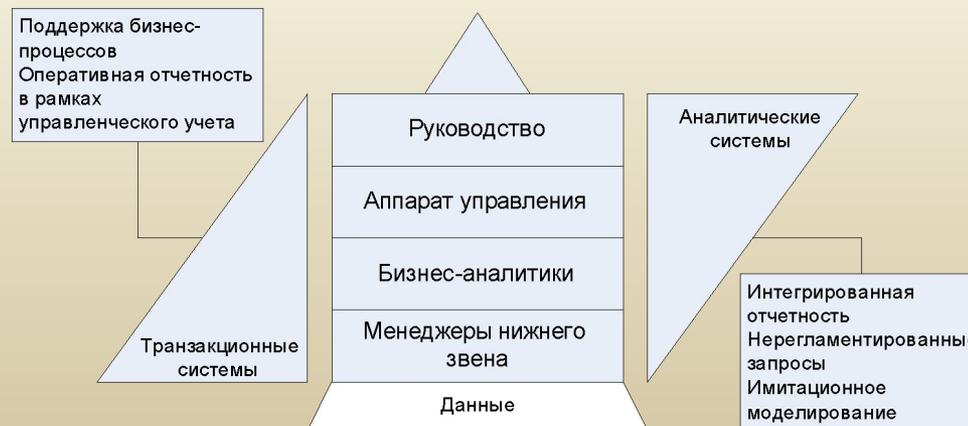
Хранилища данных (Data Warehouse)

Хранилище данных — предметно-ориентированный, интегрированный, привязанный ко времени и неизменяемый набор данных, предназначенный для поддержки принятия решений.

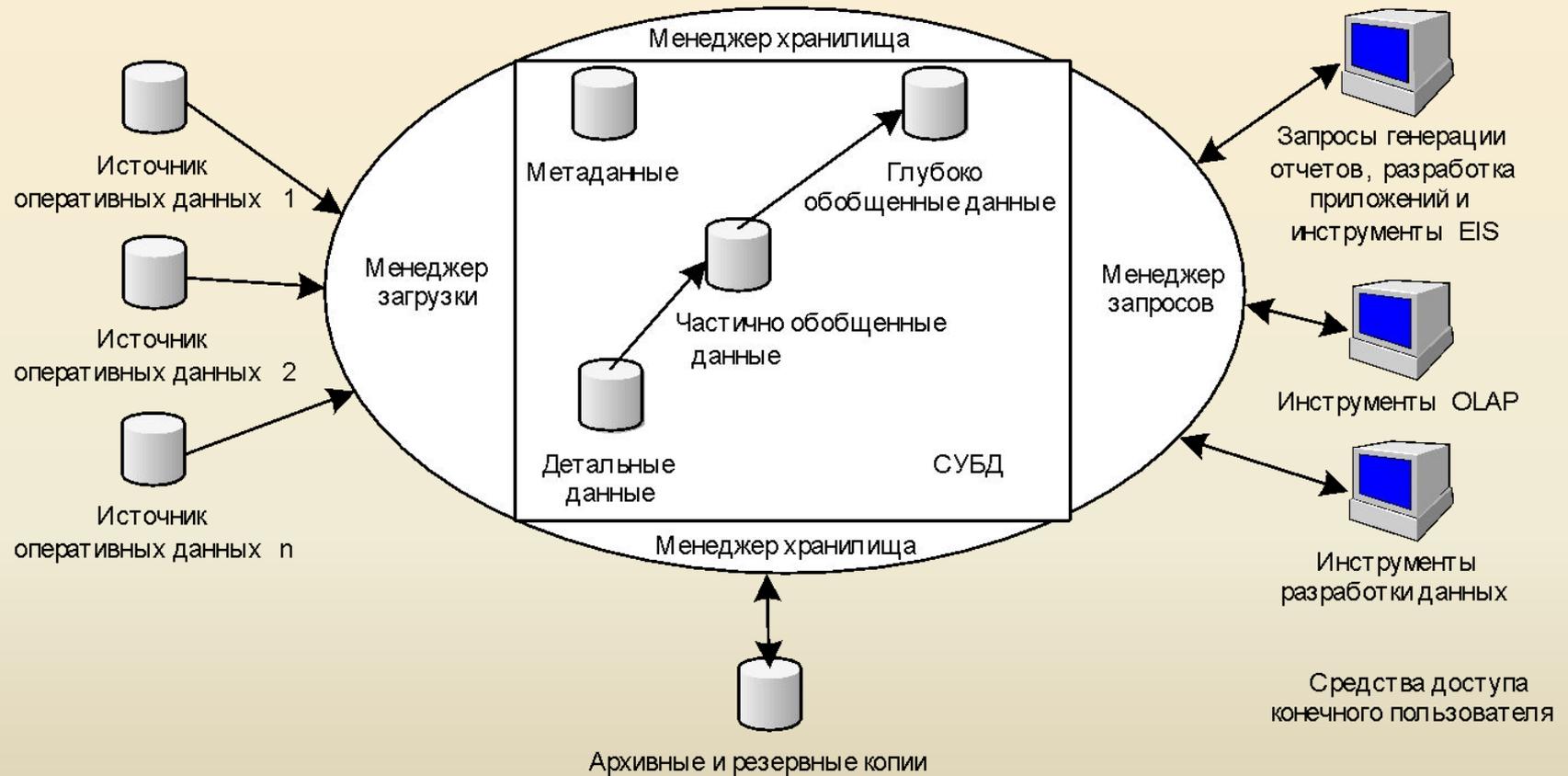
№	Характеристика	Комментарий
1	Предметная ориентированность	Хранилище данных организовано вокруг основных предметов (или субъектов) организации (например, клиенты, товары и продажи), а не вокруг прикладных областей деятельности (выписка счета клиенту, контроль товарных запасов и продажа товаров). Это свойство отражает необходимость хранения данных, предназначенных для поддержки принятия решений, а не обычных оперативно-прикладных данных.
2	Интегрированность	Оперативно-прикладные данные обычно поступают из разных источников, которые часто имеют несогласованное представление одних и тех же данных, например, используют разный формат. Для предоставления пользователю единого обобщенного представления данных необходимо создать интегрированный источник, обеспечивающий согласованность хранимой информации.
3	Привязка ко времени	Данные в хранилище точны и корректны только в том случае, когда они привязаны к некоторому моменту или промежутку времени. Привязанность хранилища данных ко времени следует из большой длительности того периода, за который была накоплена сохраняемая в нем информация, из явной или неявной связи временных отметок со всеми сохраняемыми данными, а также из того факта, что хранимая информация фактически представляет собой набор моментальных снимков состояния данных.
4	Неизменяемость	Это означает, что данные не обновляются в оперативном режиме, а лишь регулярно пополняются за счет информации из оперативных систем обработки. При этом новые данные никогда не заменяют прежние, а лишь дополняют их. Таким образом, база данных хранилища постоянно пополняется новыми данными, последовательно интегрируемыми с уже накопленной информацией.

Сравнение OLTP-систем и хранилищ данных

OLTP-система	Аналитическая система на основе хранилища данных
Содержит текущие данные	Содержит исторические, текущие и прогнозные данные
Хранит подробные сведения	Хранит подробные сведения, а также частично и значительно обобщенные данные
Ориентирована на прикладные области	Ориентирована на предметные области
Предсказуемый способ использования данных	Непредсказуемый способ использования данных
Предназначена для обработки транзакций	Предназначена для проведения анализа
Высокая интенсивность обработки транзакций	Средняя и низкая интенсивность обработки транзакций
Повторяющийся способ обработки данных	Нерегламентированный и неструктурированный способ обработки данных
Поддержка принятия повседневных решений	Поддержка принятия стратегических решений
Обслуживает работников исполнительного звена	Обслуживает работников руководящего звена



Архитектура хранилищ данных



Архитектура хранилищ данных

Менеджер хранилища

Менеджер хранилища (warehouse manager) выполняет все операции, связанные с управлением информацией, помещенной в хранилище данных.

Этот компонент может включать программы собственной разработки и инструменты, предоставленные сторонними фирмами.

Менеджер хранилища выполняет такие операции, как:

- Управление объектами хранилища данных (таблицами, индексами, представлениями, процедурами и др.).
- Анализ непротиворечивости данных.
- Преобразование и перемещение исходных данных из временного хранилища в основные таблицы хранилища данных.
- Обобщение данных (в случае необходимости).
- Резервное копирование и архивирование данных.

Архитектура хранилищ данных

Менеджер загрузки

Менеджер загрузки (load manager), который часто называют внешним (front-end) компонентом, выполняет все операции, связанные с извлечением и загрузкой данных хранилище. Эти операции включают преобразования данных, необходимостью для их подготовки к вводу в хранилище.

Размеры и сложность данного компонента могут варьироваться в значительной степени, поскольку в его состав обычно входят не только программы собственной разработки, но и инструменты, созданные сторонними поставщиками.

Менеджер запросов

Менеджер запросов (query manager), который часто называют внутренним (back-end) компонентом, выполняет все операции, связанные с управлением пользовательскими запросами. Этот компонент обычно создается на базе предоставляемых разработчиком СУБД инструментов доступа к данным, инструментов мониторинга хранилища и программ собственной разработки, использующих весь набор функциональных возможностей СУБД.

К числу выполняемых этим компонентом операций относятся управление запросами к соответствующим таблицам и составление графиков выполнения этих запросов.

Архитектура хранилищ данных

Детальные данные

В этой части хранилища данных хранятся все детальные данные. Как правило, детальные данные периодически добавляются в хранилище с автоматическим выполнением обобщения исходной информации до необходимого уровня.

Частично и глубоко обобщенные данные

В этой области хранилища размещаются все данные, предварительно обработанные менеджером хранилища с целью их частичного или глубокого обобщения (aggregate).

Назначение обобщенных данных состоит в повышении производительности запросов. Хотя предварительное обобщение информации связано с некоторым повышением расходов на обслуживание, однако эти дополнительные затраты компенсируются за счет исключения необходимости многократно выполнять обобщающие операции (например, сортировку или группирование) при обработке каждого из запросов пользователей. Хранимые обобщенные данные обновляются по мере загрузки новых порций детальных данных в хранилище.

Архивные и резервные копии

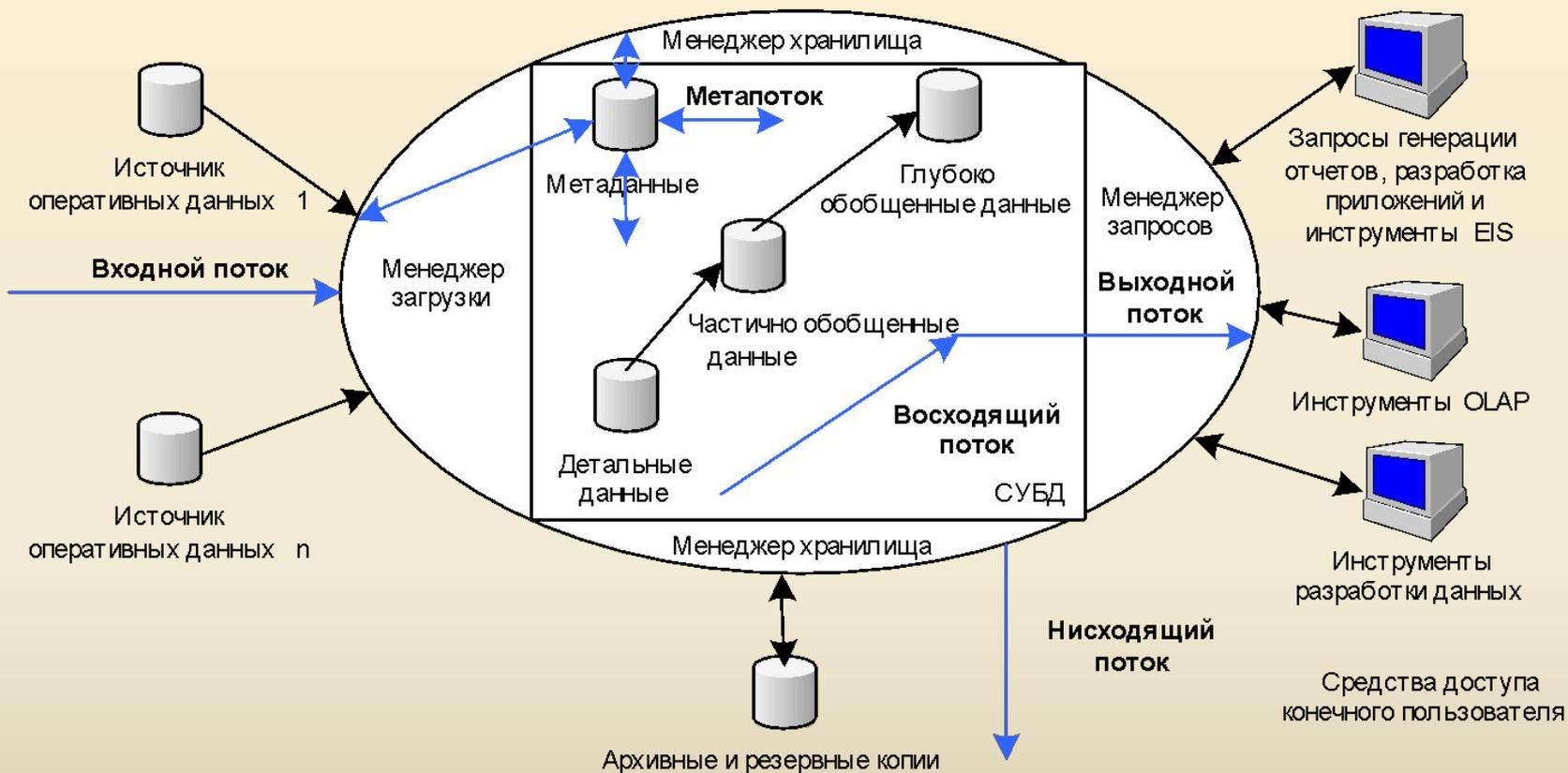
Этот компонент хранилища данных отвечает за подготовку детальной и обобщенной информации к помещению в резервные и архивные копии. Хотя обобщенные данные генерируются на основе детальных, может потребоваться помещать в резервную заранее обобщенные данные, если предполагаемый период их хранения превышает срок хранения тех детальных данных, на основе которых они были созданы. Как правило, резервные и архивные копии размещаются на таких носителях, как магнитная лента или оптический диск.

Метаданные

Метаданные — это описание информационного содержания хранилища данных: что в нем содержится, откуда что поступает, какие операции выполнялись во время очистки, как осуществлялись интеграция и обобщение.

Средства доступа конечных пользователей к данным используют метаданные для выбора способа построения запроса.

Информационные потоки в хранилище данных



Архитектура хранилищ данных

Входной поток - процессы, связанные с извлечением, очисткой и загрузкой информации из источников данных в хранилище данных.

Поскольку исходные данные генерируются преимущественно OLTP-системами, эти данные должны быть перестроены в соответствии с требованиями хранилища данных. Перестройка данных включает такие операции, как:

- очистка данных;
- преобразование данных в соответствии с требованиями хранилища данных;
- проверка внутренней непротиворечивости данных и их непротиворечивости по отношению к данным, уже загруженным в хранилище.

Непосредственно после извлечения из источника данные обычно загружаются во временное хранилище с целью выполнения очистки и проверки непротиворечивости.

Архитектура хранилищ данных

Восходящий поток - процессы, связанные с повышением ценности сохраняемых в хранилище данных посредством обобщения, упаковки и распределения исходных данных.

Обслуживание восходящего потока включает выполнение приведенных ниже действий.

- Обобщение данных посредством операций выборки, проекции, соединения и группирования связанных данных, выполняемое для получения более удобных и полезных для пользователей представлений информации. Обобщение может включать выполнение не только простых реляционных операций, но и проведение сложного статистического анализа, включая вычисление трендов, кластеризацию и подбор типичных значений.
- Упаковка данных с преобразованием подробных исходных или обобщенных данных в более удобный формат представления, например в виде электронных таблиц, текстовых документов, диаграмм, других графических представлений, закрытых баз данных и анимированных материалов.
- Распределение исходных данных на соответствующие группы для повышения их подготовленности к использованию и доступности.

Архитектура хранилищ данных

Нисходящий поток - процессы, связанные с архивированием и резервным копированием информации в хранилище данных.

Нисходящий поток информации включает процедуры, обеспечивающие возможность восстановления текущего состояния хранилища в случае потери данных из-за сбоев в программном или аппаратном обеспечении. Архивные данные следует хранить таким образом, чтобы в случае необходимости они снова могли быть восстановлены в хранилище данных.

Выходной поток - процессы, связанные с предоставлением данных пользователям.

В качестве основных действий, связанных с выходным потоком, следует упомянуть следующие.

- ***Доступ к данным;***
- ***Доставка данных.***

Метапоток - процессы, связанные с управлением метаданными.

Предыдущие потоки характеризуют управление хранилищем данных в отношении перемещения данных в хранилище и из него.

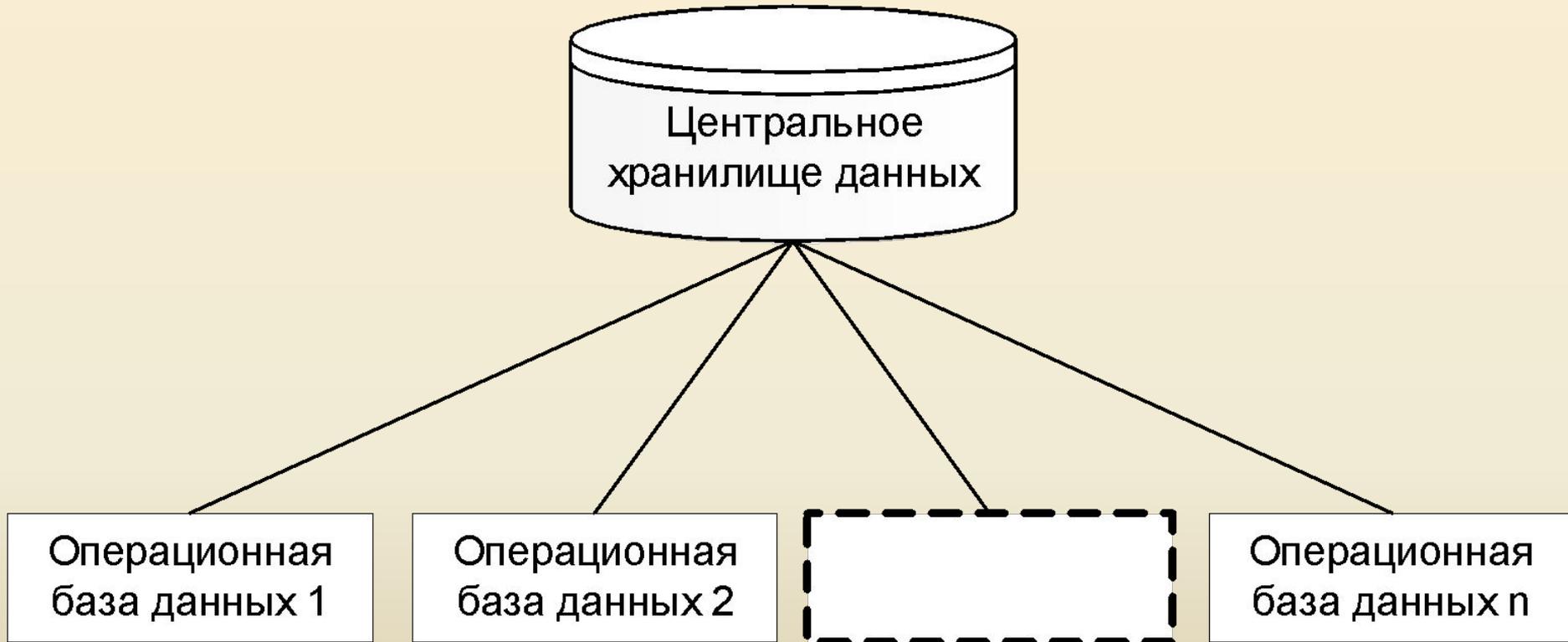
Метапоток — это процесс, связанный с перемещением метаданных, т. е. данных о других потоках.

Архитектура хранилищ данных

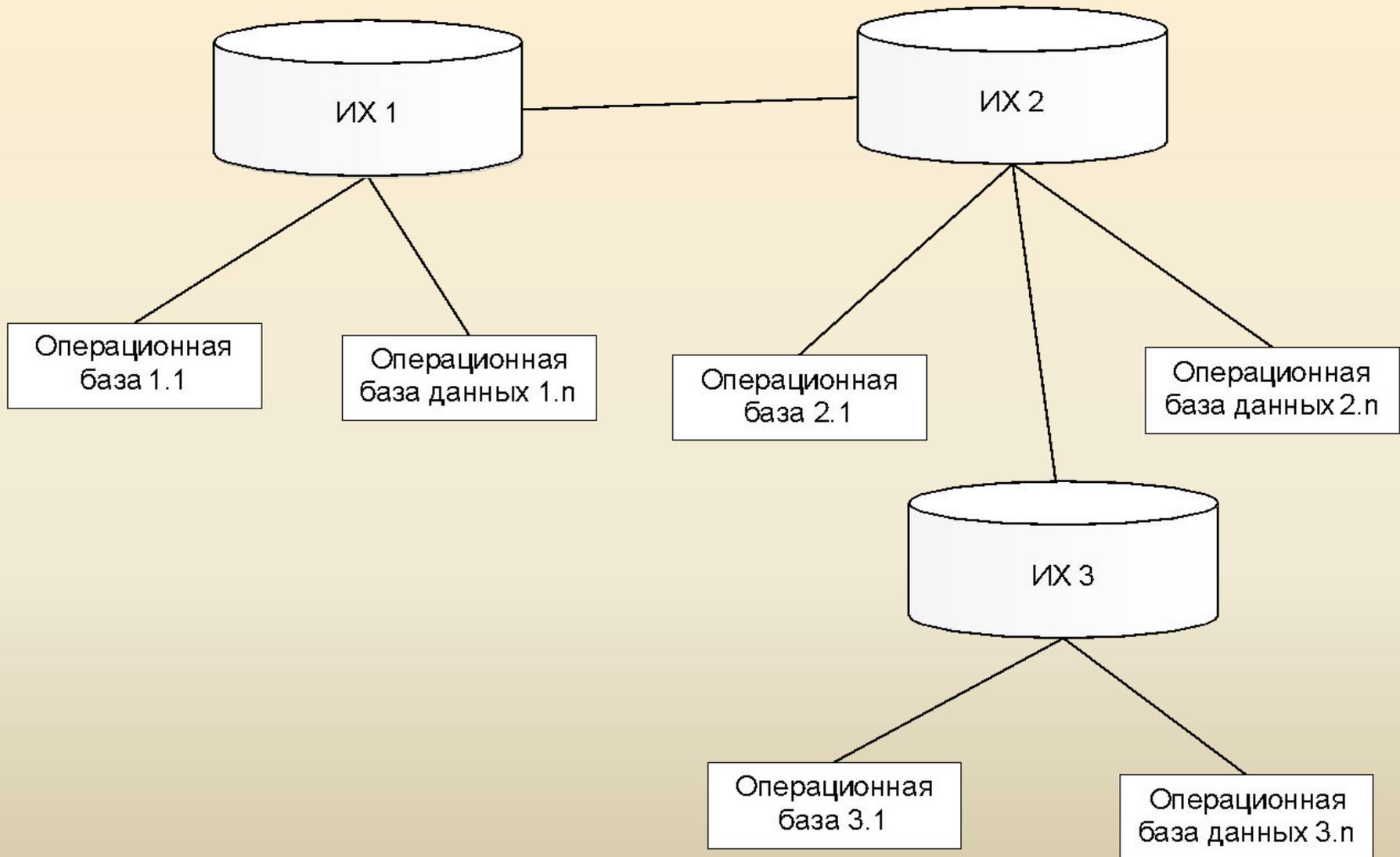
Метапоток - процессы, связанные с управлением метаданными.

Предыдущие потоки характеризуют управление хранилищем данных в отношении перемещения данных в хранилище и из него. Метапоток — это процесс, связанный с перемещением метаданных, т.е. данных о других потоках.

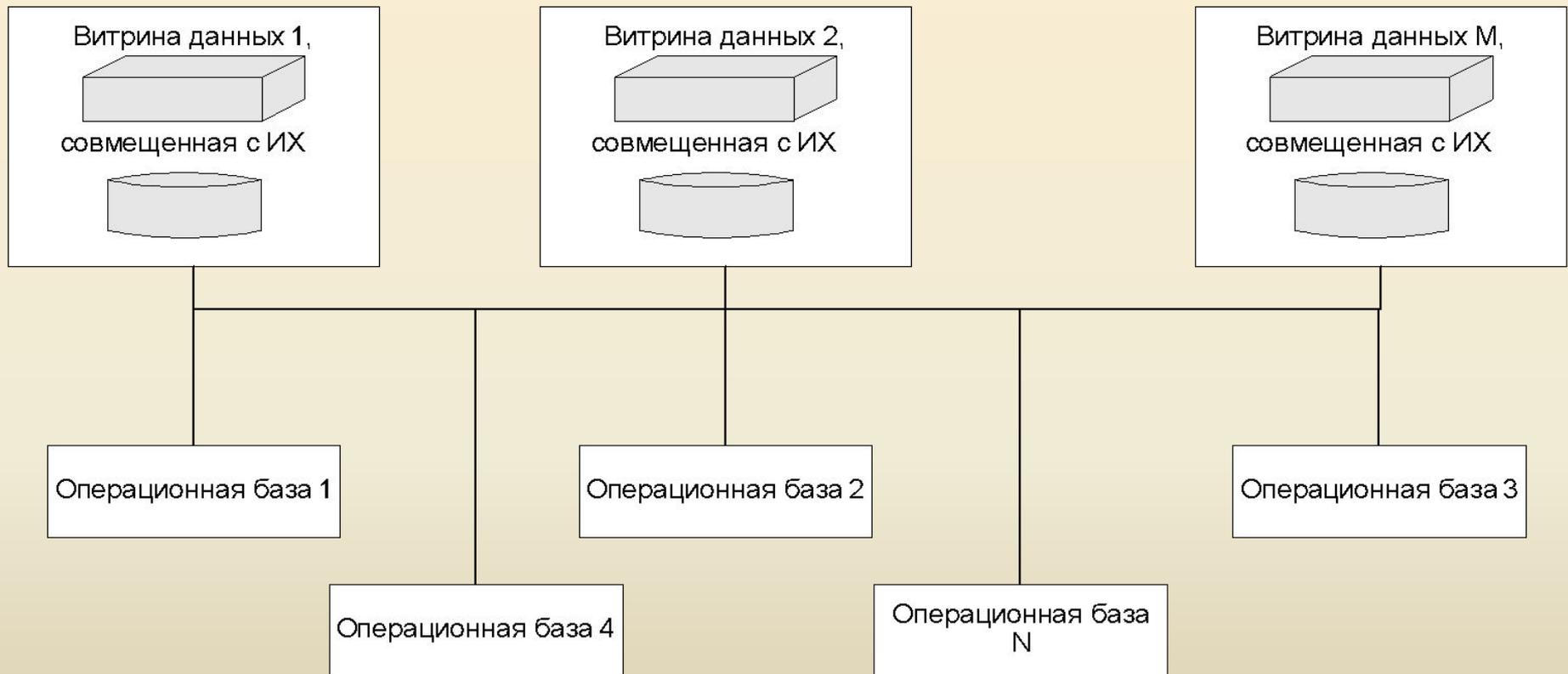
Подходы к организации хранилища данных. Централизованное хранилище данных



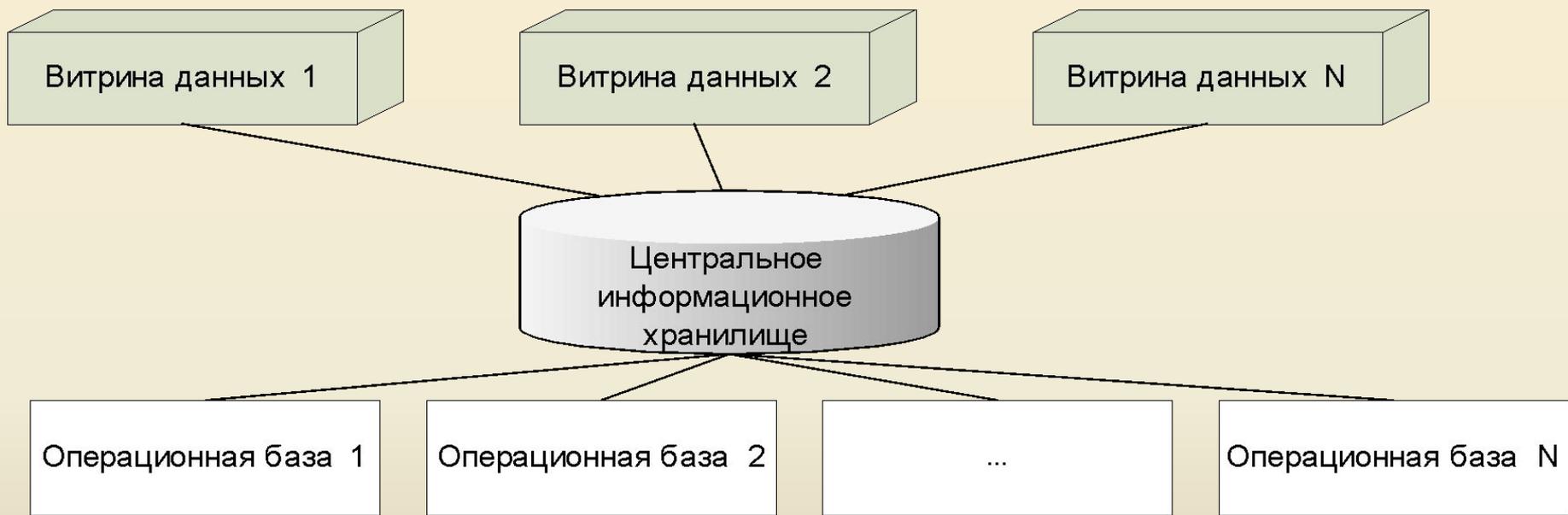
Подходы к организации хранилища данных. Распределенное хранилище данных



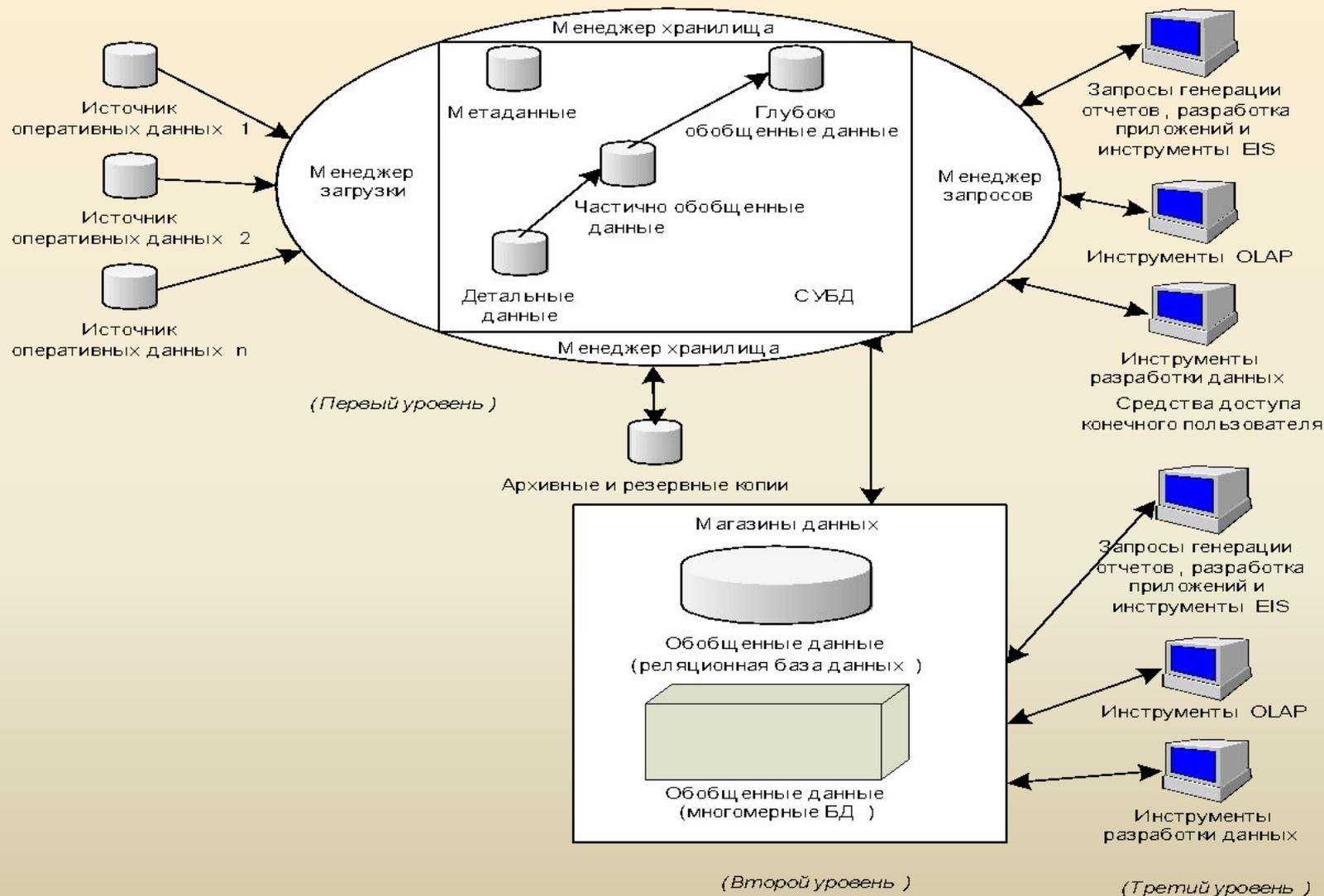
Подходы к организации хранилища данных. Автономные витрины данных



Подходы к организации хранилища данных. Единое интегрированное хранилище и многие витрины данных



Пример архитектуры хранилища и витрины данных



Литература

- Семенов М.И., Трубилин И.Т., Лойко В.И., Барановская Т.П. Автоматизированные информационные технологии в экономике: Учебник. – М.: Финансы и статистика, 2000. – 416 с.
- Конноли Т., Бегг К., Страчан А. Базы данных: проектирование, реализация и сопровождение. Теория и практика. 2-е изд.: Пер. с англ.: Учебное пособие. – М.: Издательский дом «Вильямс», 2000. – 1120 с.
- Архипенков С.Я., Голубев Д.В., Максименко О.Б. Хранилища данных. от концепции до внедрения. – М.: Диалог-Мифи, 2002. – 528 с.
- Лядова Л.Н., Мызникова Б.И., Фролова Н.В. Основы информатики и информационных технологий: Учебное пособие для студентов экономических специальностей – Пермь.: Пермский университет, 2004. – 311 с.
- Белов В.С. Информационно-аналитические системы. Основы проектирования и применения: учебное пособие, руководство, практикум/ Московский государственный университет экономики, статистики и информатики – М., 2004. – 116 с.