МИНОБРНАУКИ РОССИИ
Федеральное государственное бюджетное образовательное учреждение
высшего образования
«Московский технологический университет»
МИРЭА

Институт **Информационных технологий**
наименование института (полностью)

Кафедра **Инструментального и прикладного программного обеспечения**
наименование кафедры (полностью)

УТВЕРЖДАЮ
Зав. кафедрой ИППО
_____/ **В.А. Мордвинов** /
«___»_____20__

КОНСПЕКТ ЛЕКЦИЙ ПО ДИСЦИПЛИНЕ

Б1.В.ДВ.11 «Хранилища данных и OLAP-технологии»
(наименование дисциплины)

Направление подготовки **09.03.04 «Программная инженерия»**
(код и наименование)

Профиль **«Интеллектуальные программные системы и комплексы»**
(код и наименование)

Форма обучения **очная**
(очная, очно-заочная, заочная)

Программа подготовки **Бакалавр**
(академический, прикладной бакалавриат)

Квалификация выпускника **Бакалавр**

Москва 2017

Тема 1

ОБЩАЯ ХАРАКТЕРИСТИКА ХРАНИЛИЩ ДАННЫХ

1. ВВЕДЕНИЕ

В настоящее время информационные технологии находятся на следующем этапе:

1. Созданы для многих предприятий , технологических процессов информационные системы с использованием СУБД, локальных сетей, распределенных СУБД.

   Идет создание информационных систем для распределенных структур, фирм располагающихся в различных городах или даже странах с помощью облачных технологий

2. В созданных информационных системах идет непрерывный рост объема накопленных данных. в области информационных технологий в настоящее время акцент смещается в область аналитической обработки этих данных. В целом, человечество вошло в область создания информационного общества. С появлением первых ЭВМ наступил этап информатизации разных сторон человеческой деятельности. Если раньше человек основное внимание уделял веществу, затем энергии (рис. 1.1), то сегодня можно без преувеличения сказать, что наступил этап осознания процессов, связанных с информацией. Вычислительная техника создавалась, прежде всего, для обработки данных. В настоящее время современные вычислительные системы и компьютерные сети позволяют накапливать большие массивы данных для решения задач обработки и анализа.

Рисунок 1.1 – Уровень использования человеком различных объектов материального мира

В информационной области наступил этап, когда начали использовать Хранилища Данных и OLAP-технологии.

OLAP (On-Line Analytical Processing) — технологии комплексного многомерного анализа данных.

Концепция OLAP была описана в 1993 году Эдгаром Коддом, известным исследователем баз данных и автором реляционной модели данных.

В OLAP-технологии используются агрегатные данные, которые образуют **многомерный набор данных, называемый OLAP-куб**. Благодаря такой модели данных пользователи могут формулировать сложные запросы, генерировать отчеты, получать подмножества данных.

Целью OLAP-анализа является проверка возникающих гипотез. Аналитик выдвигает гипотезы, основываясь на своих знаниях и опыте.

Технологии OLAP тесно связаны с технологиями построения Хранилищ данных (Data Warehouse) и методами их интеллектуальной обработки - Data Mining.

Термин Data Mining – добыча данных появился в 1996 году. Одним из основателей метода является Пятецкий-Шапиро

Идея метода Data Mining заключается в обнаружении знаний, содержащихся в **терабайтах** накопленной информации. Эти знания человек-аналитик не в состоянии исследовать самостоятельно. Data Mining относится к методам интеллектуального анализа данных с использованием хранилищ данных.

В 2008 г. был введен термин "Большие данные (Big Data)" Автором этого термина является Клиффорд Линчем, редактором журнала «Nature» .

С « big data» связывают задачи:

- создания хранилищ с использованием новых технологий для хранения больших объёмов данных;
- управления;
- работы с неструктурированной информацией;
- разработки методов анализа big data.

2. Хранилища данных

1.1. Общая информация о хранилищах данных

Понятие хранилище данных (ХД) имеет свою эволюцию в истории. Первые работы IBM связанные с концепцией информационного хранилища проводились ещё до создания реляционной модели.

1993 г. Бил Инмон получил титул отца-основателя ХД. Концепция ХД была запущена в область информационных технологий.

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

Рассмотрим определение "хранилище данных" подробнее.

- **Предметная ориентированность**

Хранилище данных организовано вокруг основных предметов или субъектов организации (например, клиенты, товары, продажи), а не вокруг прикладных областей деятельности (выписки счетов клиенту, контроль запасов и т.д.). Это свойство отражает необходимость хранения данных предназначенных для поддержки принятия решений.

- **Интегрированность**

Смысл этой характеристики состоит в том, что оперативно-прикладные данные обычно поступают из различных источников, которые часто имеют несогласованное представление одних и тех же данных, например, используют разный формат. Для представления пользователю единого обобщённого представления данных необходимо создать интегрированный источник, обеспечивающий согласованность хранимой информации.

- **Привязка ко времени.**

Данные в хранилище точны и корректны в том случае, когда они привязаны к моменту или промежутку времени. Хранимая информация представляет собой набор моментальных снимков состояния данных

- **Неизменяемость.**

Данные не обновляются в оперативном режиме, а лишь регулярно пополняются за счёт информации из оперативных систем обработки. При этом новые данные никогда не заменяются прежними. Таким образом, базы данных хранилища постоянно пополняются новыми данными, последовательно интегрируемой с уже накопленной информации.

1.2. Сравнение OLTP-систем и ХД.

OLTP (On Line Transaction Processing) - обработка информации (транзакций) в реальном времени.

СУБД уже поддерживает оперативную обработку транзакций, но обычно оно рассматривается как непригодное для организации ХД, т.к. к этим двум типам систем предъявляются совершенно разные требования. OLTP-системы проектируются с целью обеспечивать максимально интенсивную обработку фиксированных транзакций, тогда как ХД проектируются для обработки единичных произвольных запросов.

OLTP и ХД. Сходства и различия.

	OLTP	**ХД**
№1	Содержит текущие данные	Содержит исторические данные
№2	Хранит подробные сведения	Хранит подробные сведения, а также значительные обобщённые данные
№3	Данные являются динамическими	Данные в основном являются статическими
№4	Используются повторяющийся способ обработки данных	Используется нерегламентируемый, неструктурированный и эвристический способ обработки данных
№5	Высокая эффективность обработки транзакций	Средняя и низкая эффективность обработки транзакций
№6	Предсказуемы способ использования данных	Непредсказуемый способ использования данных
№7	Предназначено для обработки	Предназначено для проведения

	транзакций	анализа
№8	Поддержка принятия поведенческих решений	Поддерживает принятие стратегических решений
№9	Обслуживает большое количество работников	Обслуживает относительно малое количество работников руководящего звена

Характерный запрос для OLTP-систем:

Какова средняя цена объектов недвижимости в крупнейших городах страны?

Характерные запросы для ХД:

- Какие типы объектов недвижимости продаются по ценам выше средней цены объектов недвижимости в крупнейших городах страны, и как эти объекты коррелируются с демографическими данными?

- Какие три района в обслуживаемых городах были более популярны с точки зрения аренды объектов недвижимости в 2010 г., и как эти данные связаны с данными за предыдущие два года?

3. Архитектура хранилищ данных

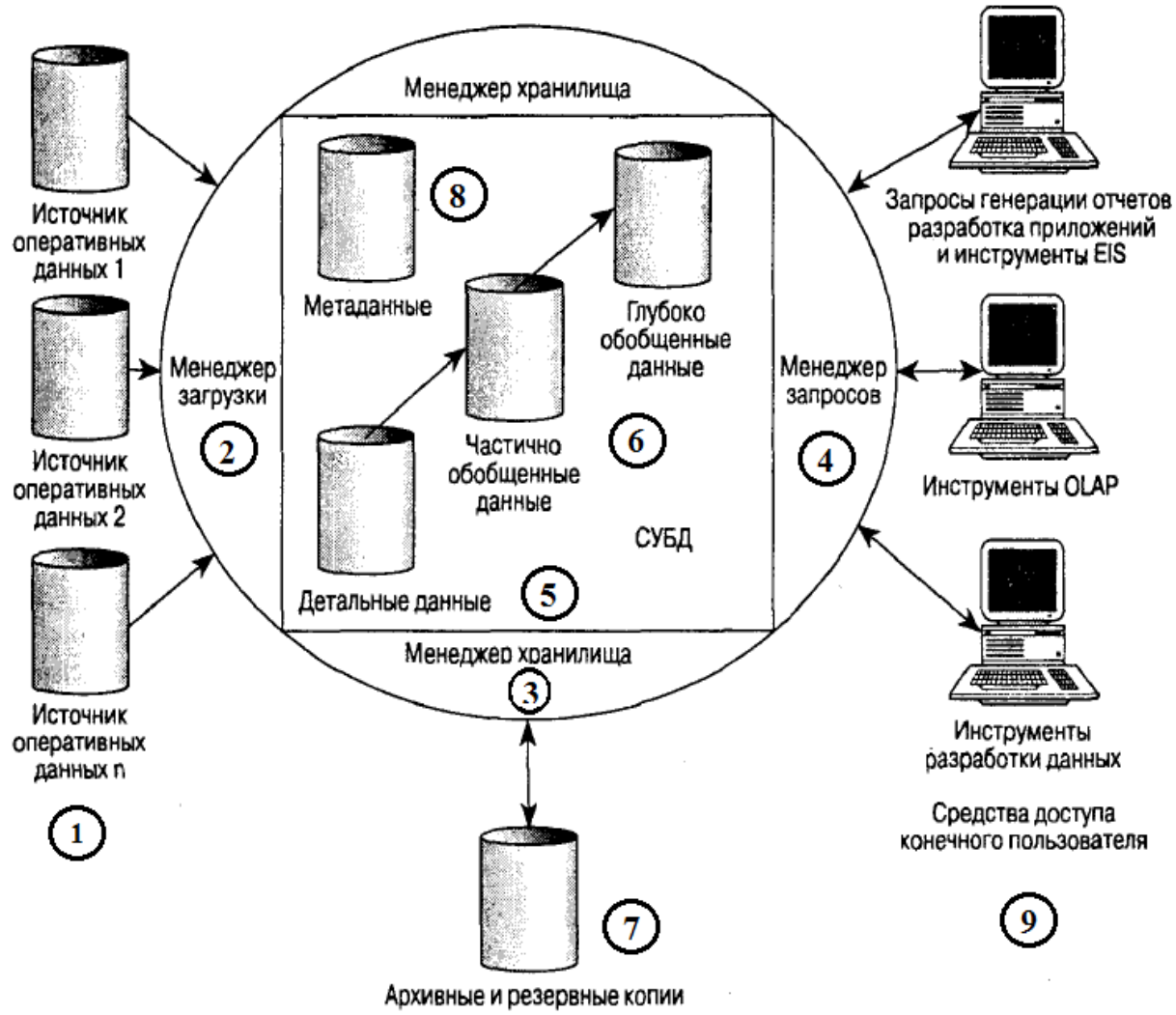

Рисунок 1.2 – Архитектура хранилища данных

На рисунке 1.2. цифрами отмечено:

1. Оперативные данные
2. Менеджер загрузки
3. Менеджер хранилища
4. Менеджер запросов
5. Детальные данные

6.1. Частично-обобщённые данные

6.2. Глубоко-обобщённые данные

7. Архивная и резервные копии

8. Метаданные

9. Средства доступа к данным конечного пользователя

1П – входной поток

2П – восходящий поток

3П – нисходящий поток

4П – исходящий поток

5П – метапоток

1. Оперативные данные

Исходящие данные, помещённые в хранилище поступают из следующих источников:

1. Оперативные данные, содержащиеся в БД первого поколения (иерархические и сетевые);
2. Данные различных подразделений, сохраняемые в БД и реляционных СУБД;
3. БД внешних систем, принадлежащих поставщикам или клиентам организаций.

2. Менеджер загрузки

Выполняет все операции, связанные с извлечением и загрузкой данных в хранилище. Эти операции включают простые преобразования данных, необходимые для их подготовки к вводу в хранилище. В состав менеджера загрузки обычно входят:

- Инструментарий, поставляемый сторонними организациями.

3. Менеджер хранилища

Менеджер хранилища выполняет все операции, связанные с управлением информацией, помещённой в ХД. Менеджер хранилища выполняет следующие операции:

- Анализ непротиворечивости данных;
- Преобразование и перемещение исходных данных из временного хранилища в основные таблицы ХД;
- Создание представлений для базовых таблиц;
- Резервное копирование и архивирование данных;
- Денормализация данных

4. менеджер запросов

Менеджер запросов выполняет все операции, связанные с управлением пользовательскими запросами.

5. Детальные данные

В этот блок загружаются все данные, описанные в концептуальной модели данных хранилища.

6. Частично обобщённые данные и глубоко-обобщённые данные

Здесь размещаются все данные, предварительно обработанные менеджером хранилища с целью частичного или глубоко обобщения. Эта часть хранилища является временной, так как постоянно подвергается изменениям в соответствии с вопросами, возникающими у пользователей.

7. Архивная и резервная копия

Этот компонент хранилища отвечает за подготовку детальной и обобщённой информации к помещению в резервные и архивные копии.

В архивные данные могут быть записаны заранее обобщённые данные, которые требуются для часто повторяемых запросов.

8. Метаданные (данные о данных)

В этой области ХД хранятся все метаданные, которые используются любыми процессами хранилища. Метаданные могут применяться для изменения и загрузки данных, а также обслуживания хранилища при автоматизации подготовки таблиц с обобщёнными значениями.

9. Средства доступа к данным конечного пользователя

Основным назначением ХД является представление конечным пользователям информации, необходимой им для принятия стратегических решений.

Пользователи взаимодействуют с хранилищем с помощью специальных инструментов доступа к данным.

ХД должно обеспечивать эффективное выполнение произвольных запросов и представлять средства для проведения анализов.

Высокая производительность хранилища достигается за счёт тщательного предварительного планирования и составления периодических таблиц, которые могут потребоваться конечным пользователям.

Выделяют 5 основных групп в этих средствах доступа к данным:

- Инструмент создания отчётов и запросов;
- Инструменты разработки приложений;
- Инструменты информационной системы руководителя;
- Инструменты оперативно-аналитической разработки (OLAP-инструменты);
- Инструменты разработки данных (data mining и big data).

Инструменты создания отчётов подразделяются на:

- Итоговые отчёты
- Редакторы отчётов

Информационная система руководителя разработана для поддержки принятия высокоуровневых стратегических решений.

OLAP – инструменты – инструменты оперативной системы обработки данных. Создаются на основе концепции многомерной базы данных.

Эти модели позволяют квалифицированным пользователям анализировать данные с помощью сложных многомерных представлений.

Инструменты разработки данных

Методы разработки данных превосходят возможности OLAP-технологий, так как они создают предсказательные модели в отличие от ретроспективных моделей OLAP-технологий.

1.4. Информационные потоки в ХД

В технологии хранилищ данных основное внимание уделяется управлению пятью основными информационными потоками: входным, восходящим, нисходящим, выходным и метапотоком (Hackathorn, 1995). Место этих потоков в структуре хранилища данных схематически показано на рис. 25.2.

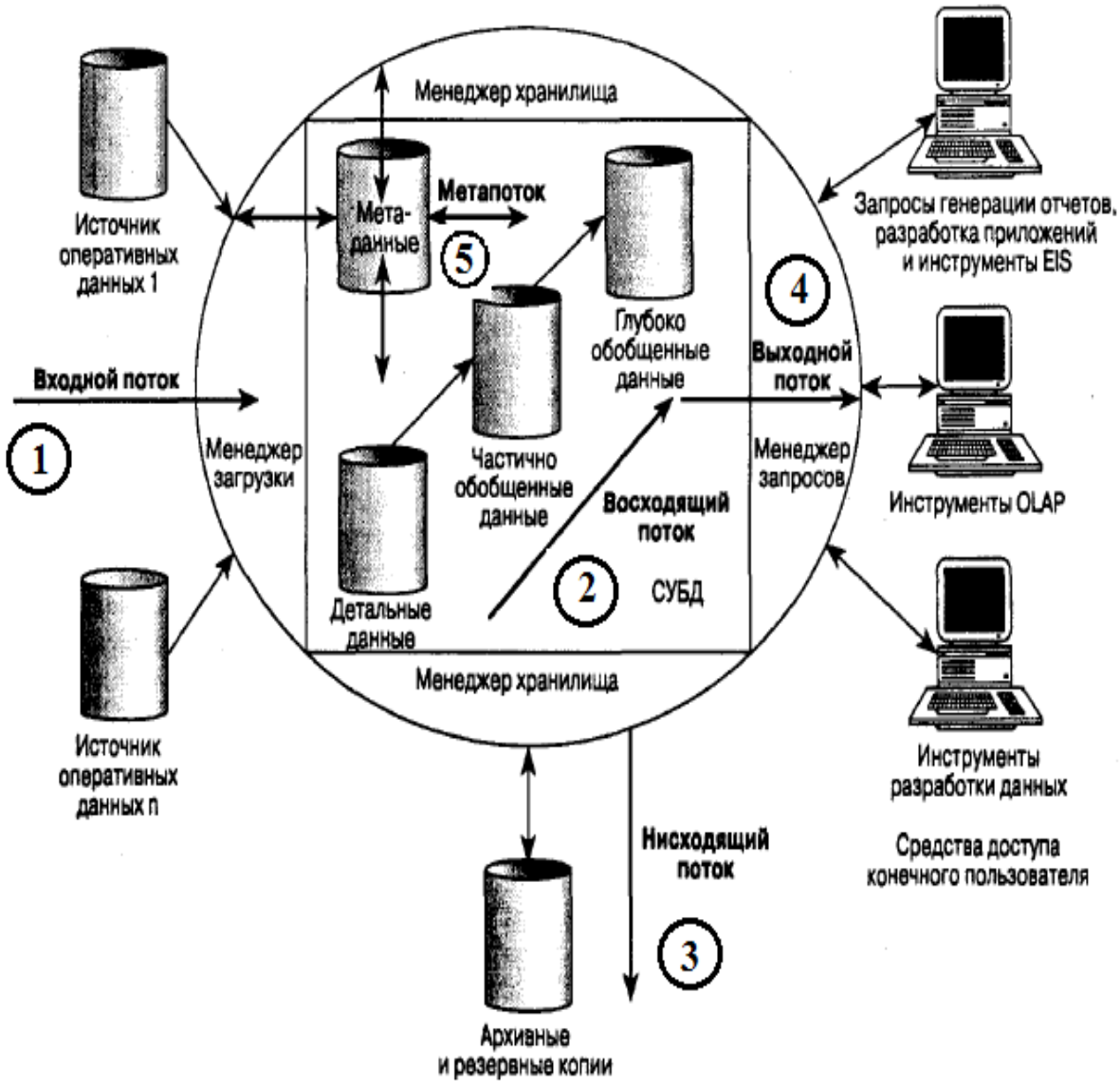

Рис. 1.3 Информационные потоки в хранилище данных

1. Входной поток

В нём происходят процессы:

- Извлечения
- Очистки
- Загрузки

2. Восходящий поток

В нём происходят процессы:

- Повышение ценности сохраняемых в хранилище данных путём
  - Обобщения
  - Упаковки
  - Распределения

3. Нисходящий поток – архивирование данных

В нём происходят процессы:

- Резервирования
- Копирования

4. ВЫХОДНОЙ ПОТОК

Представление данных пользователей

5. Поток метаданных

В нём происходят процессы управления метаданными

Рассмотрим содержание потоков.

1. Входной поток ХД

Процессы, связанные с извлечением, очисткой и загрузкой информации из источника данных в хранилище информации. Входной поток связан с

выборкой данных из источника данных с целью их последующей загрузки в ХД.

Необходима реализация операций:

1. Очистки данных
2. Преобразования данных в соответствии с требованием хранилища (добавление, удаление полей, процесс денормализации данных)
3. Проверки внутренней непротиворечивости данных и данных уже загруженных в хранилище.

1. Восходящий поток

Процессы, связанные с повышением ценности сохраняемых в хранилище данных посредством обобщения, упаковки и распределения исходных данных. Обслуживание входного потока включает в себя выполнение следующих действий:

1. Обобщение данных посредством операций выборки, проекции, соединения и группировки данных.
2. Упаковка данных с преобразованием обобщенных данных в более удобный формат представления (в виде электронных таблиц, текстовых документов и графических представлений различного типа).
3. Распределение данных в группы для повышения доступности.

3.Нисходящий поток

Нисходящий поток связан с процессами архивирования и резервного копирования информации в ХД. Архивирование устаревших данных играет важную роль при обеспечении эффективности и производительности ХД. Нисходящий поток включает процедуры, обеспечивающие восстановление текущего состояния ХД в случае потери данных из-за сбоя программы.

4.Выходной поток

Выходной поток связан с представлением информации для пользователя. Важное значение – создание дружелюбного интерфейса для взаимодействия с пользователем.

5.Метапоток

Метапоток – это процесс, связанный с перемещением метаданных. Метаданные – это описание информационного содержания ХД, то есть что содержится в ХД, какие операции выполняются внутри ХД, как производится интеграция и обобщение.

1.5 Магазины данных

Магазины данных – это подмножество ХД, которое поддерживает требования отдельной сферы организации.(рис 1.3)

Магазин данных обычно не содержит детальных оперативных сведений, поэтому информация магазинов данных более проста и понятна в управлении.

Существует несколько подходов создания магазинов данных:

1. Создание хранилища для предприятия или фирмы, и он может быть использован предприятием, а также использовать сведения магазина данных.
2. Создание нескольких магазинов данных с возможностью их интеграции в единое ХД.

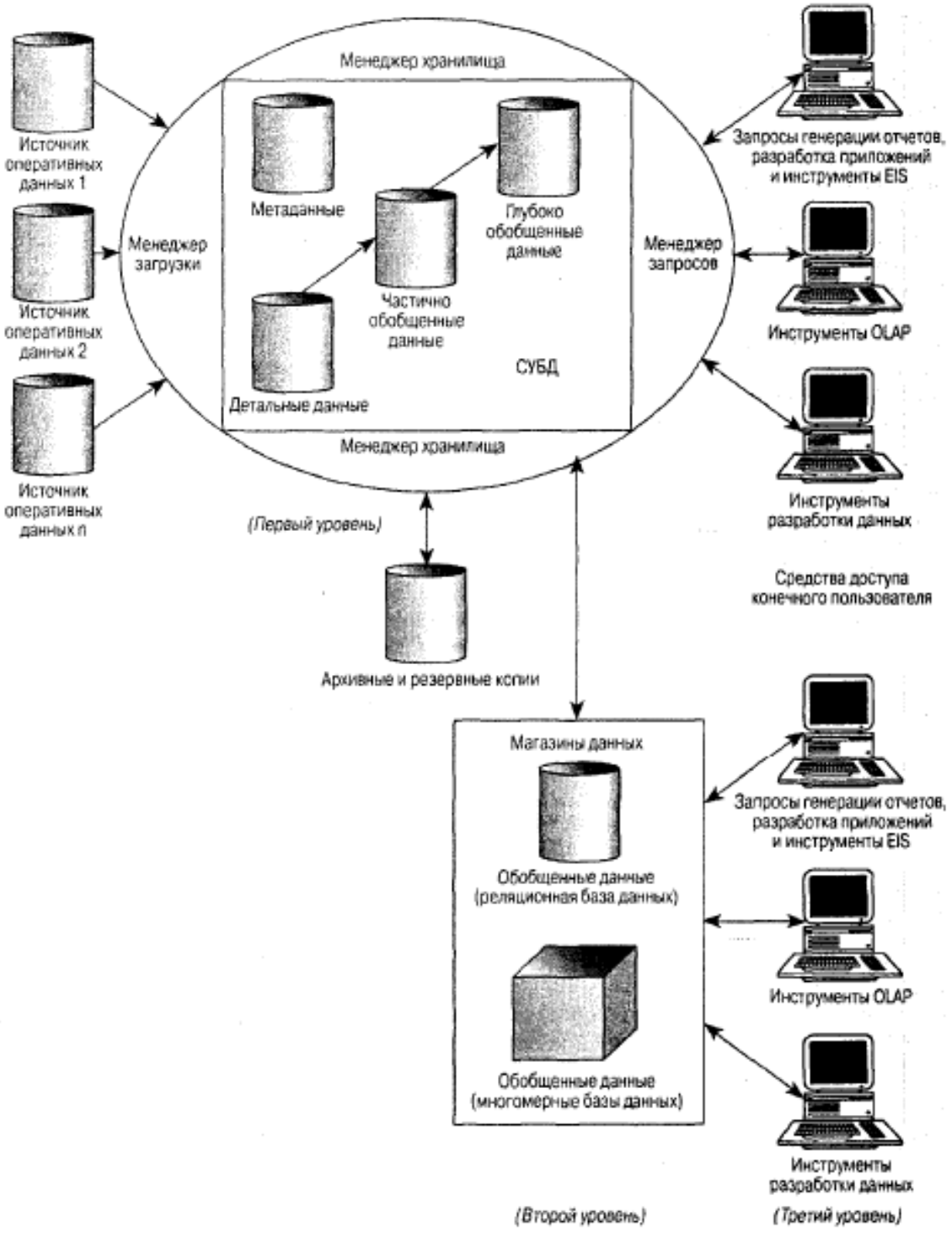

***Рис1.4** Типичная архитектура хранилища данных и магазины данных*