Содержание

[Термины и определения 1](#_Toc64690220)

[Введение 2](#_Toc64690221)

[1 Обзор предметной области 2](#_Toc64690222)

[1.1 OLAP-системы 2](#_Toc64690223)

[1.1.1 Многомерная модель данных 2](#_Toc64690224)

[1.1.2 Основные операции манипулирования данными 4](#_Toc64690225)

[1.1.3 Архитектура OLAP-систем 8](#_Toc64690226)

[1.1.3.1 MOLAP 8](#_Toc64690227)

[1.1.3.2 ROLAP 10](#_Toc64690228)

[1.1.3.3 HOLAP 12](#_Toc64690229)

[2 Список использованных источников 12](#_Toc64690230)

Термины и определения

В настоящем отчете о НИР применяют следующие термины с соответствующими определениями.

Введение

Обработка и анализ накопленной информации актуальна для большого спектра исследовательских и прикладных задач. Наиболее популярным подходом к решению этой проблемы в настоящее время является технология оперативной аналитической обработки данных OLAP (online analytical processing) и Data Mining. [1]

В последние десятилетия появилось множество программных комплексов, реализующих те или иные подходы к работе с многомерными данными.

Актуальность:

Целью исследования будет являться анализ методов и алгоритмов поиска и извлечения информации из разреженного гиперкуба данных.

Для достижения цели исследования поставлены следующие задачи:

1. произвести дополнительный поиск источников;
2. найти и выбрать подходящие методы и алгоритмы;
3. разработать программную модель разреженного гиперкуба данных;
4. реализовать и применить к программной модели выбранные методы и алгоритмы;
5. выполнить анализ и оценку по поставленным критериям;
6. сделать выводы по полученным результатам.

# Обзор предметной области

## OLAP-системы

### Многомерная модель данных

**Хранилище данных** (Data Warehouse) — это предметно-ориентированное, привязанное ко времени и неизменяемое собрание данных для поддержки процесса принятия управляющих решений. Задача хранилища — предоставить «сырье» для анализа в одном месте и в простой, понятной структуре [5].

**OLAP** (On-Line Analytical Processing, оперативная аналитическая обработка) служит для описания модели представления данных и соответственно технологии их обработки в хранилищах данных. [3]

Многомерная модель данных, лежащая в основе построения многомерных хранилищ данных, опирается на концепцию многомерных кубов, или гиперкубов. Они представляют собой упорядоченные многомерные массивы, которые также часто называют OLAP-кубами.

**Гиперкуб** данных (data hypercube) содержит одно или более *измерений* и представляет собой упорядоченный набор *ячеек* (рисунок Рисунок 1). Каждая ячейка определяется одним и только одним набором значений измерений — *меток*. Ячейка может содержать данные — *меру* или быть пустой. [4]

**Измерение** (dimension) — множество меток, образующих одну из граней гиперкуба. Примером временного измерения является список дней, месяцев, кварталов. Примером географического измерения может быть перечень территориальных объектов: населенных пунктов, районов, регионов, стран и т.д. [4]

**Метка** (member) — значение, «откладываемое» вдоль измерений. Значения меток отображаются в двумерном представлении куба как заголовки строк и столбцов. [5]

**Факт** (мера, measure) — это числовая величина, которая располагается в ячейках гиперкуба. Один OLAP-куб может обладать одним или несколькими показателями. [3]

**Ячейка** (cell) — атомарная структура куба, соответствующая полному набору конкретный значений измерений. [3]

Для получения доступа к данным пользователю необходимо указать одну или несколько ячеек путем выбора значений измерений, которым соответствуют необходимые ячейки. Процесс выбора значений измерений будем называть фиксацией меток, а множества выбранных значений измерений – множеством фиксированных меток. [4]

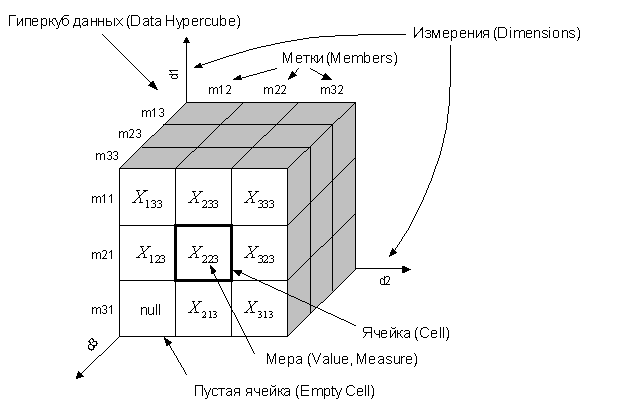


Рисунок 1 — Изображение модели данных разреженного гиперкуба

Можно выделить следующие возможности работы с данными при многомерной модели [6]:

* быстрый доступ;
* преагрегация;
* иерархии;
* интеллектуальный анализ (Data Mining);
* многоуровневое кэширование.

### Основные операции манипулирования данными

**Срез** (сечение, slice) (рисунок Рисунок 2) — формирование подмножества многомерного массива данных, соответствующего единственному значению одного или нескольких элементов измерений, не входящих в это подмножество. [2]

Например, при выборе элемента «Факт», измерения «Сценарий» срез данных представляет собой подкуб, в который входят все остальные измерения. Данные, что не вошли в сформированный срез, связаны с теми элементами измерения «Сценарий», которые не были указаны в качестве определяющих (например, «План», «Отклонение», «Прогноз» и т. п.). Если рассматривать термин «срез» с позиции конечного пользователя, то наиболее часто его роль играет двумерная проекция куба. [2]

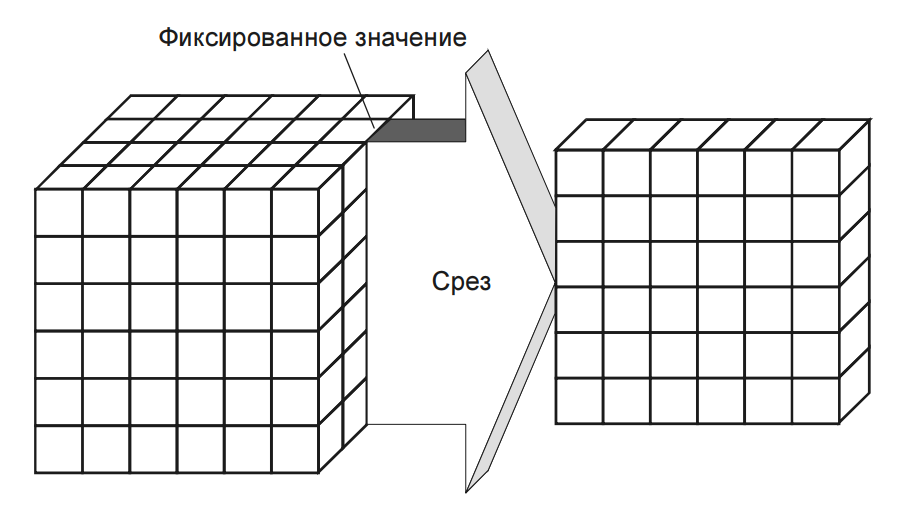


Рисунок 2 – Изображение операции среза

**Вращение** (поворот, rotate) (рисунок Рисунок 3) — изменение расположения измерений, представленных в отчете или на отображаемой странице. [2]

Например, операция вращения может заключаться в перестановке местами строк и столбцов таблицы или перемещении интересующих измерений в столбцы или строки создаваемого отчета, что позволяет придавать ему желаемый вид. Кроме того, вращением куба данных является перемещение внетабличных измерений на место измерений, представленных на отображаемой странице, и наоборот (при этом внетабличное измерение становится новым измерением строки или измерением столбца). В качестве примера первого случая может служить отчет, для которого элементы измерения «Время» располагаются поперек экрана (являются заголовками столбцов таблицы), а элементы измерения «Продукция» – вдоль экрана (заголовки строк таблицы). [2]

После применения операции вращения отчет будет иметь следующий вид: элементы измерения «Продукция» будут расположены по горизонтали, а элементы измерения «Время» – по вертикали. Примером второго случая может служить преобразование отчета с измерениями «Меры» и «Продукция», расположенными по вертикали, и измерением «Время», расположенным по горизонтали, в отчет, у которого измерение «Меры» располагается по вертикали, а измерения «Время» и «Продукция» – по горизонтали. При этом элементы измерения «Время» располагаются над элементами измерения «Продукция». Для третьего случая применения операции вращения можно привести пример преобразования отчета с расположенными по горизонтали измерением «Время» и по вертикали измерением «Продукция» в отчет, у которого по горизонтали представлено измерение «Время», а по вертикали – измерение «География». [2]

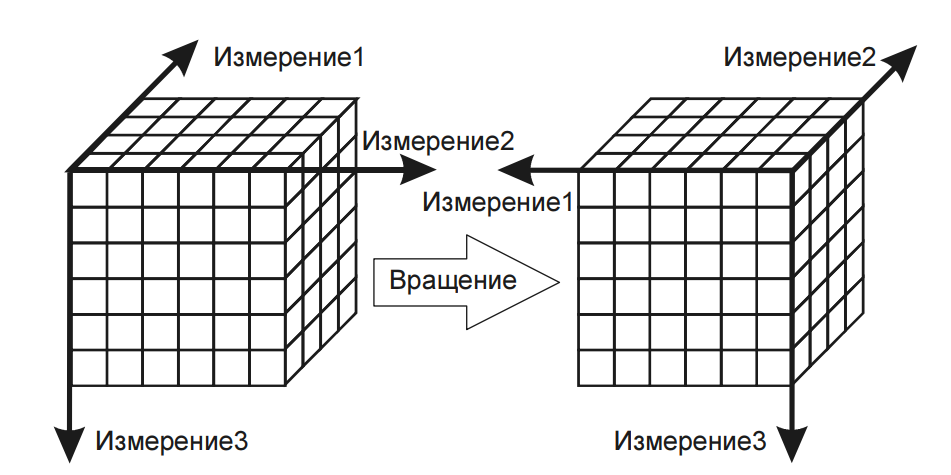


Рисунок 3 – Изображение операции вращения

**Консолидация** (свёртка, roll-up/drill-up) и **детализация** (раскрытие, drill-down) (рисунок Рисунок 4) — операции, которые определяют переход вверх по направлению от детального (down) представления данных к агрегированному (up) и наоборот, соответственно. Направление детализации (обобщения) может быть задано как по иерархии отдельных измерений, так и согласно прочим отношениям, установленным в рамках измерений или между измерениями. [2]

Например, если при анализе данных об объемах продаж в Северной Америке выполнить операцию детализации для измерения «Регион», то на экране будут отображены такие его элементы, как «Канада», «Восточные Штаты Америки» и «Западные Штаты Америки». В результате дальнейшей детализации элемента «Канада» будут отображены элементы «Торонто», «Ванкувер», «Монреаль» и т. д. [2]

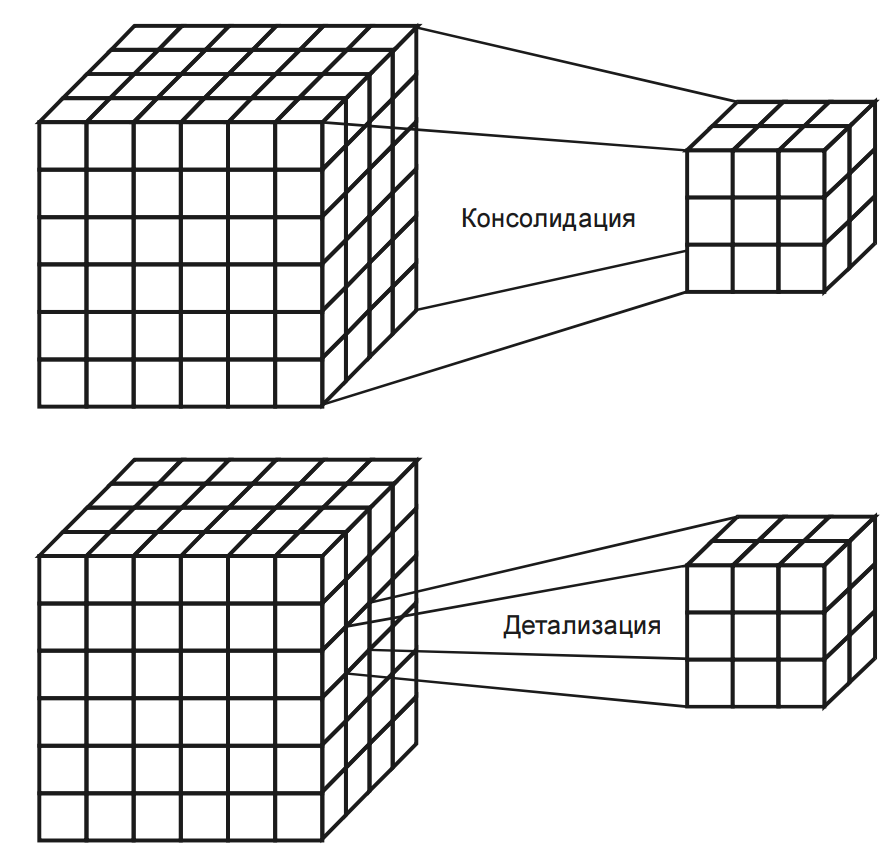


Рисунок 4 — Изображение операций консолидации и детализации

### Архитектура OLAP-систем

OLAP-серверы скрывают от конечного пользователя способ реализации многомерной модели. Они формируют гиперкуб, с которым пользователи посредством OLAP-клиента выполняют все необходимые манипуляции, анализируя данные. Между тем способ реализации очень важен, т. к. от него зависят такие характеристики, как производительность и занимаемые ресурсы. [2]

Выделяют три основных способа реализации:

1. MOLAP — многомерный (multivariate) OLAP. Для реализации многомерной модели используют многомерные БД;
2. ROLAP — реляционный (relational) OLAP. Для реализации многомерной модели используют реляционные БД;
3. HOLAP — гибридный (hybrid) OLAP. Для реализации многомерной модели используют и многомерные, и реляционные БД.

#### MOLAP

Детальные и агрегированные данные хранятся в многомерной базе данных. Хранение данных в многомерных структурах позволяет манипулировать данными как многомерным массивом, благодаря чему скорость вычисления агрегатных значений одинакова для любого из измерений. Однако в этом случае многомерная база данных оказывается избыточной, так как многомерные данные полностью содержат детальные реляционные данные. [3]

Физически данные, представленные в многомерном виде, хранятся в «плоских» файлах. При этом куб представляется в виде одной плоской таблицы, в которую построчно вписываются все комбинации членов всех измерений с соответствующими им значениями мер (рисунок Рисунок 5). [2]



Рисунок 5 — Изображение реализации многомерного представления данных с помощью плоской таблицы

Можно выделить следующие преимущества MOLAP:

1. поиск и выборка данных осуществляются значительно быстрее, чем при многомерном концептуальном взгляде на реляционную БД, т. к. многомерная база данных денормализована и содержит заранее агрегированные показатели, обеспечивая оптимизированный доступ к запрашиваемым ячейкам и не требуя дополнительных преобразований при переходе от множества связанных таблиц к многомерной модели; [2]
2. многомерные БД легко справляются с задачами включения в информационную модель разнообразных встроенных функций, тогда как объективно существующие ограничения языка SQL делают выполнение этих задач на основе реляционных БД достаточно сложным, а иногда и невозможным; [2]
3. структура и интерфейсы наилучшим образом соответствуют структуре аналитических запросов. [3]

Недостатки:

1. в подавляющем большинстве случаев информационный гиперкуб является сильно разреженным, а поскольку данные хранятся в упорядоченном виде, неопределенные значения удается удалить только за счет выбора оптимального порядка сортировки, позволяющего организовать данные в максимально большие непрерывные группы; [2]
2. многомерные БД чувствительны к изменениям в многомерной модели. Так при добавлении нового измерения приходится изменять структуру всей БД, что влечет за собой большие затраты времени; [2]
3. по сравнению с реляционными, очень неэффективно используют внешнюю память, обладают худшими по сравнению с реляционными БД механизмами транзакций; [3]
4. отсутствуют единые стандарты на интерфейс, языки описания и манипулирования данными. [3]

#### ROLAP

ROLAP-серверы используют реляционные БД. В настоящее время распространены две основные схемы реализации многомерного представления данных с помощью реляционных таблиц: схема «звезда» (рисунок Рисунок 6) и схема «снежинка» (рисунок Рисунок 7). [2]

Можно выделить следующие преимущества такой архитектуры:

1. в случае переменной размерности задачи, когда изменения в структуру измерений приходится вносить достаточно часто, ROLAP-системы с динамическим представлением размерности являются оптимальным решением, т. к. в них такие модификации не требуют физической реорганизации БД; [2]
2. реляционные СУБД обеспечивают значительно более высокий уровень защиты данных и хорошие возможности разграничения прав доступа; [2]
3. реляционные СУБД имеют реальный опыт работы с очень большими БД и развитые средства администрирования; [3]
4. системы ROLAP могут функционировать на гораздо менее мощных клиентских станциях, чем системы MOLAP, поскольку основная вычислительная нагрузка в них ложится на сервер, где выполняются сложные аналитические SQL-запросы, формируемые системой. [3]

Главный недостаток ROLAP по сравнению с многомерными СУБД — меньшая производительность. Для обеспечения производительности, сравнимой с MOLAP, реляционные системы требуют тщательной проработки схемы базы данных и настройки индексов, т. е. больших усилий со стороны администраторов БД. [2]

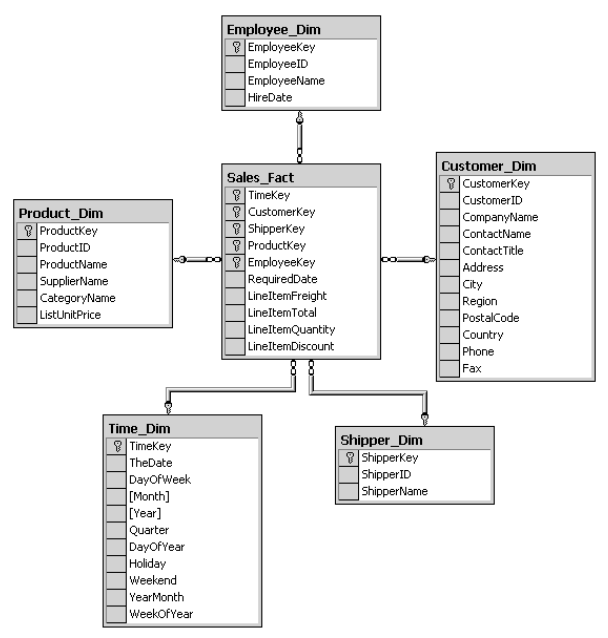


Рисунок 6 — Изображение реализации многомерного представления данных с помощью схемы «звезда»

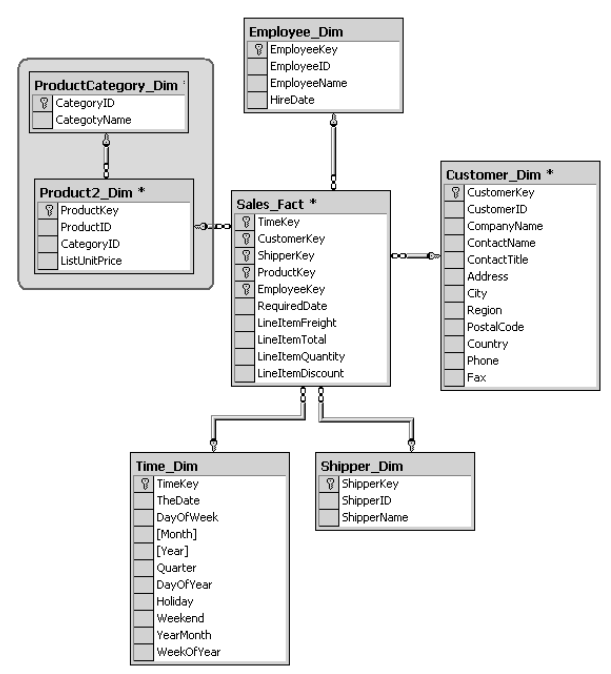


Рисунок 7 — Изображение реализации многомерного представления данных с помощью схемы «снежинка»

#### HOLAP

## Интеллектуальный анализ данных

# Обзор средств разработки

## Analysis Services

Службы Analysis Services предоставляют множество встроенных функций для использования с языками многомерных выражений (MDX) и расширений интеллектуального анализа данных (DMX), предназначенных для выполнения всех операций, от стандартных статистических вычислений до обхода элементов в иерархии.

Службы Analysis Services позволяет добавлять сборки в экземпляр Службы Analysis Services или базу данных. Сборки позволяют создавать внешние пользовательские функции при помощи любого языка среды CLR, например Microsoft Visual Basic .NET или Microsoft Visual C#. Также можно использовать языки автоматизации COM, например Microsoft Visual Basic или Microsoft Visual C++.

Сборки позволяют расширять бизнес-функции многомерных выражений и расширений интеллектуального анализа данных. Требуемые функции формируются в виде библиотеки, например динамически подключаемой библиотеки (DLL), которая затем добавляется в качестве сборки к экземпляру служб Службы Analysis Services или к базе данных служб Службы Analysis Services. Общие методы в библиотеке затем открываются в виде пользовательских функций для многомерных выражений и выражений расширений интеллектуального анализа данных, их процедур, вычислений, действий и клиентских приложений

# Список использованных источников

1. Мосин, С. В. Методы и алгоритмы формирования многомерных данных с использованием промежуточных представлений: автореферат дис. … кандидата физико-математических наук: 05.13.11. – Омск, 2017;
2. Барсегян, А. А. Анализ данных и процессов: учеб. пособие [Текст] / А. А. Барсегян, М. С. Куприянов, И. И. Холод, М. Д. Тесс, С. И. Елизаров. – 3-е изд., перераб. и доп. – СПб.: БХВ-Петербург, 2009. – 512 с.;
3. Распределенные базы и хранилища данных [Электронный ресурс]. – Режим доступа: <https://intuit.ru/studies/courses/1145/214/info>, свободный (дата обращения: 01.11.2020);
4. Методы представления информации в разреженных гиперкубах данных [Электронный ресурс]. — Режим доступа: <http://www.olap.ru/basic/theory.asp>, свободный (дата обращения: 01.11.2020);
5. Введение в OLAP-технологии и многомерные базы данных. [Электронный ресурс]. — Режим доступа: <http://www.olap.ru/basic/alpero2i.asp>, свободный (дата обращения: 01.11.2020);
6. Многомерные кубы, OLAP и MDX [Электронный ресурс]. — Режим доступа: <https://habr.com/ru/post/66356/>, свободный (дата обращения: 01.11.2020);
7. IBM Cloud Education: Data Warehouse [Электронный ресурс]. — Режим доступа: <https://www.ibm.com/cloud/learn/data-warehouse>, свободный (дата обращения: 01.11.2020);
8. IBM Cloud Education: OLAP [Электронный ресурс]. — Режим доступа: <https://www.ibm.com/cloud/learn/olap>, свободный (дата обращения: 01.11.2020);