Минский инновационный университет
Кафедра информационных технологий
Гедранович О.Б.

Модели и технологии оперативного анализа данных

Лекция 2 Технологии оперативного анализа данных

Многомерная модель данных. Определение OLAP-систем. Архитектура OLAP-систем: ROLAP, MOLAP, HOLAP. Основные операции OLAP-систем.

Многомерная модель данных

В процессе принятия решений пользователь генерирует некоторые гипотезы. Для превращения их в законченные решения эти гипотезы должны быть проверены. Проверка гипотез осуществляется на основании информации об анализируемой предметной области. Как правило, наиболее удобным способом представления такой информации для человека является зависимость между некоторыми параметрами. Например, зависимость объемов продаж от региона, времени, категории товара и т. п. Другим примером может служить зависимость количества выздоравливающих пациентов от применяемых средств лечения, возраста и т. п.

В процессе анализа данных, поиска решений часто возникает необходимость в построении зависимостей между различными параметрами. Кроме того, число таких параметров может варьироваться в широких пределах. Как уже отмечалось ранее, традиционные средства анализа, оперирующие данными, которые представлены в виде таблиц реляционной БД, не могут в полной мере удовлетворять таким требованиям. В 1993 г. Е. Кодд — основоположник реляционной модели БД — рассмотрел ее недостатки, указав в первую очередь на невозможность "объединять, просматривать и анализировать данные с точки зрения множественности измерений, т. е. самым понятным для аналитиков способом".

Измерение — это последовательность значений одного из анализируемых параметров. Например, для параметра "время" это последовательность календарных дней, для параметра "регион" это может быть список городов.

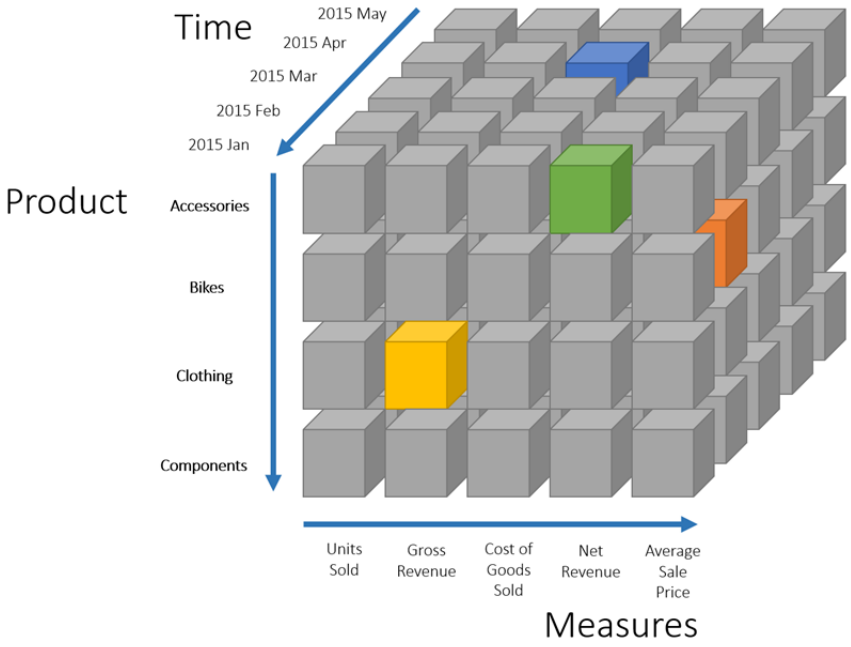
Множественность измерений предполагает представление данных в виде многомерной модели. По измерениям в многомерной модели откладывают параметры, относящиеся к анализируемой предметной области.

По Кодду, многомерное концептуальное представление (multi-dimensional conceptual view) — это множественная перспектива, состоящая из нескольких независимых измерений, вдоль которых могут быть проанализированы определенные совокупности данных. Одновременный анализ по нескольким измерениям определяется как многомерный анализ.

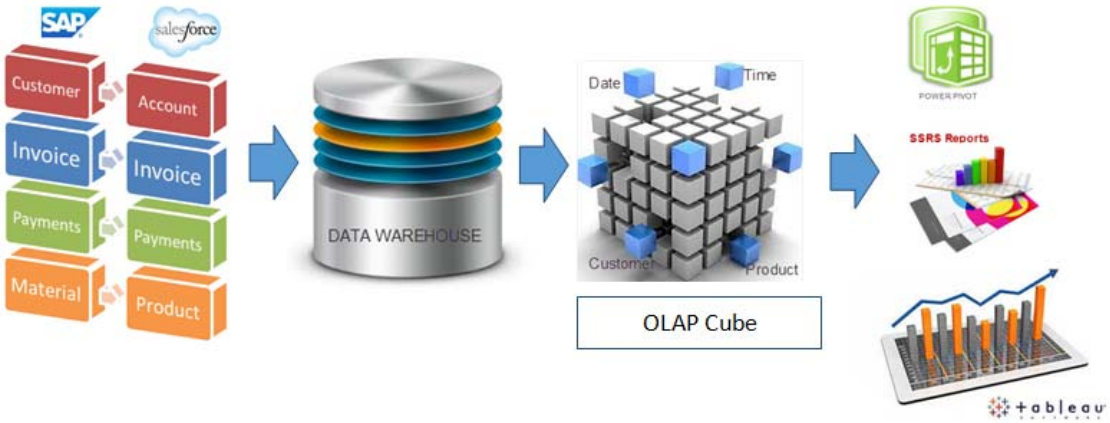
Каждое измерение может быть представлено в виде иерархической структуры. Например, измерение "Исполнитель" может иметь следующие иерархические уровни: "предприятие — подразделение — отдел — служащий". Более того,

некоторые измерения могут иметь несколько видов иерархического представления. Например, измерение "Время" может включать две иерархии со следующими уровнями: "год — квартал — месяц — день" и "неделя — день".

На пересечениях осей измерений (Dimensions) располагаются данные, количественно характеризующие анализируемые факты, — меры (Measures). Это могут быть объемы продаж, выраженные в единицах продукции или в денежном выражении, остатки на складе, издержки и т. п.



Таким образом, многомерную модель данных можно представить как гиперкуб (конечно, название не очень удачное, поскольку под кубом обычно понимают фигуру с равными ребрами, что в данном случае далеко не так). Ребрами такого гиперкуба являются измерения, а ячейками — меры.



Определение OLAP-систем

С концепцией многомерного анализа данных тесно связывают оперативный анализ, который выполняется средствами OLAP-систем.

OLAP (от англ. OnLine Analytical Processing — оперативная аналитическая обработка данных, также: аналитическая обработка данных в реальном времени, интерактивная аналитическая обработка данных) — подход к аналитической

обработке данных, базирующийся на их многомерном иерархическом представлении, являющийся частью более широкой области информационных технологий — бизнес-аналитики (BI).

Основное назначение OLAP-систем — поддержка аналитической деятельности, произвольных (ad-hoc) запросов пользователей-аналитиков. Цель OLAP-анализа — проверка возникающих гипотез.

OLAP является ценным из-за своей гибкости. После того, как факты и измерения определяются в пределах сервера OLAP, средства OLAP обеспечивают простой способ анализа данных, например, простым перетаскиванием измерений и фактов, если это обеспечено в приложении аналитика.

OLAP помогает ответить на вопрос «что происходит?», но не предсказывает, что может произойти, и не объясняет, почему так происходит.

У истоков технологии OLAP стоит основоположник реляционного подхода Э. Кодд. В 1993 г. он опубликовал статью под названием «OLAP для пользователей-аналитиков: каким он должен быть». В данной работе изложены основные концепции оперативной аналитической обработки и определены следующие 12 требований, которым должны удовлетворять продукты, позволяющие выполнять оперативную аналитическую обработку.

Далее перечислены 12 правил, изложенных Коддом и определяющих OLAP:

1. Многомерность. Multidimensional conceptual view

   OLAP-система на концептуальном уровне должна представлять данные в виде многомерной модели, что упрощает процессы анализа и восприятия информации.

2. Прозрачность. Transparency

   OLAP-система должна скрывать от пользователя реальную реализацию многомерной модели, способ организации, источники, средства обработки и хранения.

3. Доступность. Accessibility

   OLAP-система должна предоставлять пользователю еди-

   ную, согласованную и целостную модель данных, обеспечивая доступ

   к данным независимо от того, как и где они хранятся.

4. Постоянная производительность при разработке отчетов. Consistent reporting performance

   Производительность OLAP-систем не должна значительно уменьшаться при увеличении количества измерений, по которым выполняется анализ.

5. Клиент-серверная архитектура. Client/server architecture

   OLAP-система должна быть способна работать в среде "клиент-сервер", т. к. большинство данных, которые сегодня требуется подвергать оперативной аналитической обработке, хранятся распределенно. Главной идеей здесь является то, что серверный компонент инструмента OLAP должен быть достаточно интеллектуальным и позволять строить общую концептуальную схему на основе обобщения и консолидации различных логических и физических схем корпоративных БД для обеспечения эффекта прозрачности.

6. Равноправие измерений. Generic Dimensionality

   OLAP-система должна поддерживать многомерную модель, в которой все измерения равноправны. При необходимости дополнительные характеристики могут быть предоставлены отдельным измерениям, но такая возможность должна быть у любого измерения.

7. Динамическое управление разреженными матрицами. Dynamic sparse matrix handling

   OLAP-система должна обеспечивать оптимальную обработку разреженных матриц. Скорость доступа должна сохраняться вне зависимости от расположения ячеек данных и быть постоянной величиной для моделей, имеющих разное число измерений и различную степень разреженности данных.

8. Поддержка многопользовательского режима. Multi-user support

   OLAP-система должна предоставлять возможность нескольким пользователям работать совместно с одной аналитической моделью или должна создавать для них различные модели из единых данных. При этом возможны как чтение, так и запись данных, поэтому система должна обеспечивать их целостность и безопасность.

9. Неограниченные перекрестные операции. Unrestricted cross-dimensional operations

   OLAP-система должна обеспечивать сохранение функциональных отношений, описанных с помощью определенного формального языка между ячейками гиперкуба при выполнении любых операций среза, вращения, консолидации или детализации. Система должна самостоятельно (автоматически) выполнять преобразование установленных отношений, не требуя от пользователя их переопределения.

10. Интуитивная манипуляция данными. Intuitive data manipulation

    OLAP-система должна предоставлять способ выполнения операций среза, вращения, консолидации и детализации над гиперкубом без необходимости пользователю совершать множество действий с интерфейсом. Измерения, определенные в аналитической модели, должны содержать всю необходимую информацию для выполнения вышеуказанных операций.

11. Гибкие возможности получения отчетов. Flexible reporting

    OLAP-система должна поддерживать различные способы визуализации данных, т. е. средства формирования отчетов должны представлять синтезируемые данные или информацию, следующую из модели данных, в ее любой возможной ориентации. Это означает, что строки, столбцы или страницы должны показывать одновременно от 0 до N измерений, где N — число измерений всей аналитической модели. Кроме того, каждое измерение содержимого, показанное в одной записи, колонке или странице, должно позволять показывать любое подмножество элементов (значений), содержащихся в измерении, в любом порядке.

12. Неограниченная размерность и число уровней агрегации. Unlimited Dimensions and aggregation levels.

    Исследование о возможном числе необходимых измерений, требующихся в аналитической модели, показало, что одновременно могут использоваться до 19 измерений. Отсюда вытекает настоятельная рекомендация, чтобы аналитический инструмент мог одновременно предоставить хотя бы 15, а предпочтительнее — и 20 измерений. Более того, каждое из общих измерений не должно быть ограничено по

числу определяемых пользователем-аналитиком уровней агрегации и путей консолидации.

Набор правил Кодда, послуживших де-факто определением OLAP, достаточно часто вызывает различные нарекания, например, правила 1, 2, 3, 6 являются требованиями, а правила 10, 11 — неформализованными пожеланиями. Таким образом, перечисленные 12 правил Кодда не позволяют точно определить OLAP.

В 1995 г. Кодд к приведенному перечню добавил следующие шесть правил:

13. Пакетное извлечение против интерпретации.

    OLAP-система должна в равной степени эффективно обеспечивать доступ как к собственным, так и к внешним данным.

14. Поддержка всех моделей OLAP-анализа.

    OLAP-система должна поддерживать все четыре модели анализа данных, определенные Коддом: категориальную, толковательную, умозрительную и стереотипную.

15. Обработка ненормализованных данных.

    OLAP-система должна быть интегрирована с ненормализованными источниками данных. Модификации данных, выполненные в среде OLAP, не должны приводить к изменениям данных, хранимых в исходных внешних системах.

16. Сохранение результатов OLAP: хранение их отдельно от исходных данных.

    OLAP-система, работающая в режиме чтения-записи, после модификации исходных данных должна сохранять результаты отдельно. Иными словами, должна обеспечиваться безопасность исходных данных.

17. Исключение отсутствующих значений.

    OLAP-система, представляя данные пользователю, должна отбрасывать все отсутствующие значения. Другими словами, отсутствующие значения должны отличаться от нулевых значений.

18. Обработка отсутствующих значений.

    OLAP-система должна игнорировать все отсутствующие значения без учета их источника. Эта особенность связана с 17-м правилом.

Кроме того, Е. Кодд разбил все 18 правил на следующие четыре группы, назвав их особенностями. Эти группы получили названия B, S, R и D.

Основные особенности (B) включают следующие правила:

1, 2, 3, 5, 8, 10,13, 14

Специальные особенности (S):

15, 16, 17, 18

Особенности представления отчетов (R):

4, 7, 11

Управление измерениями (D):

6, 9, 12

Найджел Пендс (Nigel Pendse) предложил использовать взамен предложенных Коддом 12-ти правил OLAP так называемый **тест FASMI** (от англ. Fast Analysis of Shared Multidimensional Information — быстрый анализ доступной многомерной информации), более точно харакетеризующую требования к этим системам.

Fast (быстрый) в отражает упомянутое выше требование к скорости реакции системы. По Пендсу, интервалы с момента инициации запроса до получения результата должен измеряться секундами. Важность этого требования возрастает при использовании таких систем в качестве инструмента оперативного представления данных для аналитика, так как длительное время ожидания может пагубно влиять на цепочку рассуждений аналитика.

Analysis (анализ) предполагает приспособленность системы к использованию в релевантной для задачи и пользователя бизнес-логике с сохранением доступной «обычному» пользователю легкости оперирования данными без использования низкоуровневого специального инструментария.

Shared (доступность, общедоступность) описывает очевидное требование к возможности одновременного многопользовательского доступа к информации с интегрированной системой разграничения прав доступа вплоть до уровня конкретной ячейки данных.

Multidimensional (многомерность) является ключевым требованием концепции. Предполагается, что система должна обеспечивать полную поддержку многомерного иерархического представления как «наиболее логичного пути анализа бизнеса и организаций». Отметим, что многомерность указывает на модель концептуального представления данных, то есть на то, как пользователь должен представлять организацию данных при формулировании запросов, а не на то, в каких структурах хранятся данные физически.

Многомерность в рамках OLAP предполагает концептуальное представление данных в виде многомерной структуры данных — гиперкуба (OLAP-куба), рёбрами в котором выступают измерения (dimension), а данные (facts — факты; measures — меры, показатели) расположены на пересечении осей измерений.

Information (информация) — это все релевантные целям пользователя данные, при этом наличие «лишних» данных негативно сказывается на требовании к скорости реакции системы.

Архитектура OLAP-систем: ROLAP, MOLAP, HOLAP

OLAP-система включает в себя два основных компонента:

- OLAP-сервер — обеспечивает хранение данных, выполнение над ними необходимых операций и формирование многомерной модели на

концептуальном уровне. В настоящее время OLAP-серверы объединяют с ХД или ВД;

- OLAP-клиент — представляет пользователю интерфейс к многомерной модели данных, обеспечивая его возможностью удобно манипулировать данными для выполнения задач анализа.

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.

Выделяют три основных способа реализации:

MOLAP — многомерный (multivariate) OLAP. Для реализации многомерной модели используют многомерные БД;

ROLAP — реляционный (relational) OLAP. Для реализации многомерной модели используют реляционные БД;

HOLAP — гибридный (hybrid) OLAP. Для реализации многомерной модели используют и многомерные, и реляционные БД.

Часто в литературе по OLAP-системам можно встретить аббревиатуры DOLAP и JOLAP:

DOLAP — настольный (desktop) OLAP. Является недорогой и простой в использовании OLAP-системой, предназначенной для локального анализа и представления данных, которые загружаются из реляционной или многомерной БД на машину клиента;

JOLAP — новая, основанная на Java коллективная OLAP-API-инициатива, предназначенная для создания и управления данными и метаданными на серверах OLAP. Основной разработчик — Hyperion Solutions. Другими членами группы, определяющей предложенный API, являются компании IBM, Oracle и др.

MOLAP

MOLAP-серверы используют для хранения и управления данными многомерных БД. При этом данные хранятся в виде упорядоченных многомерных массивов. Такие массивы подразделяются на гиперкубы и поликубы.

В гиперкубе все хранимые в БД ячейки имеют одинаковую мерность, т. е. находятся в максимально полном базисе измерений.

В поликубе каждая ячейка хранится с собственным набором измерений, и все связанные с этим сложности обработки перекладываются на внутренние механизмы системы.

Очевидно, что 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 (табл. 3.1).

Измерения				Меры	
Клиент	Время	Продавец	Продукт	Сумма сделки	Объем сделки
Школа №25	20.08.2016	Юрий Т.	Карандаши	690	30
Школа №25	20.08.2016	Юрий Т.	Ручки	830	40
Школа №25	20.08.2016	Юрий Т.	Тетради	500	25
Школа №25	20.08.2016	Юрий Т.	Фломастеры	700	35
Школа №25	20.08.2016	Юрий Т.	Краски	600	15
Школа №25	20.08.2016	Юрий Т.	Маркеры	1 500	100
Школа №25	20.08.2016	Дмитрий А.	Карандаши	690	30
Школа №25	20.08.2016	Дмитрий А.	Ручки	830	40
Школа №25	20.08.2016	Дмитрий А.	Тетради	500	25
Школа №25	20.08.2016	Дмитрий А.	Фломастеры	700	35
Школа №25	20.08.2016	Дмитрий А.	Краски	2 000	50
Школа №25	20.08.2016	Дмитрий А.	Маркеры	2 250	150
Школа №25	20.08.2016	Алексей Ш.	Карандаши	230	10
Школа №25	20.08.2016	Алексей Ш.	Ручки	1 000	0

Можно выделить следующие преимущества использования многомерных БД в OLAP-системах:

- 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;
- 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.

С другой стороны, имеются также существенные недостатки многомерных БД:

- за счет денормализации и предварительно выполненной агрегации объем данных в многомерной БД, как правило, соответствует (по оценке Кодда) в 2,5 ÷ 100 раз меньшему объему исходных детализированных данных;
- 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. Но даже в этом случае проблема решается только частично. Кроме того, оптимальный с точки зрения хранения разреженных данных порядок сортировки, скорее всего, не

будет совпадать с порядком, который чаще других используется в запросах. Поэтому в реальных системах приходится искать компромисс между быстродействием и избыточностью дискового пространства, занятого базой данных;

- 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.

На основании анализа достоинств и недостатков многомерных БД можно выделить следующие условия, при которых их использование является эффективным:

- объем исходных данных для анализа не слишком велик (не более нескольких гигабайт), т. е. уровень агрегации данных достаточно высок;
- набор информационных измерений стабилен;
- время ответа системы на нерегламентированные запросы является наиболее критичным параметром;
- требуется широкое использование сложных встроенных функций для выполнения кроссмерных вычислений над ячейками гиперкуба, в том числе необходима возможность написания пользовательских функций.

ROLAP

ROLAP-серверы используют реляционные БД. По словам Кодда, "реляционные БД были, есть и будут наиболее подходящей технологией для хранения данных. Необходимость существует не в новой технологии БД, а скорее в средствах анализа, дополняющих функции существующих СУБД, и достаточно гибких, чтобы предусмотреть и автоматизировать разные виды интеллектуального анализа, присущие OLAP".

В настоящее время распространены две основные схемы реализации многомерного представления данных с помощью реляционных таблиц: схема "звезда" и схема "снежинка".

Использование реляционных БД в OLAP-системах имеет следующие достоинства:

- в большинстве случаев корпоративные хранилища данных реализуются средствами реляционных СУБД, и инструменты ROLAP позволяют производить анализ непосредственно над ними. При этом размер хранилища не является таким критичным параметром, как в случае MOLAP;
- 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;
- реляционные СУБД обеспечивают значительно более высокий уровень защиты данных и хорошие возможности разграничения прав доступа.

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. Только при использовании схем типа "звезда" производительность хорошо настроенных реляционных систем может быть приближена к производительности систем на основе многомерных баз данных.

HOLAP

HOLAP-серверы используют гибридную архитектуру, которая объединяет технологии ROLAP и MOLAP. В отличие от MOLAP, которая работает лучше, когда данные более-менее плотные, серверы ROLAP показывают лучшие параметры в тех случаях, когда данные сильно разрежены. Серверы HOLAP применяют подход ROLAP для разреженных областей многомерного пространства и подход MOLAP для плотных областей. Серверы HOLAP разделяют запрос на несколько подзапросов, направляют их к соответствующим фрагментам данных, комбинируют результаты, а затем предоставляют результат пользователю.

OLAP vs. OLTP

Data Warehouse (OLAP)	Operational Database (OLTP)
Involves historical processing of information.	Involves day-to-day processing.
OLAP systems are used by knowledge workers such as executives, managers and analysts.	OLTP systems are used by clerks, DBAs, or database professionals.
Useful in analyzing the business.	Useful in running the business.
It focuses on Information out.	It focuses on Data in.
Based on Star Schema, Snowflake, Schema and Fact Constellation Schema.	Based on Entity Relationship Model.
Contains historical data.	Contains current data.
Provides summarized and consolidated data.	Provides primitive and highly detailed data.
Provides summarized and multidimensional view of data.	Provides detailed and flat relational view of data.
Number or users is in hundreds.	Number of users is in thousands.
Number of records accessed is in millions.	Number of records accessed is in tens.
Database size is from 100 GB to 1 TB	Database size is from 100 MB to 1 GB.
Highly flexible.	Provides high performance.

Основные операции OLAP-систем

Над таким гиперкубом могут выполняться следующие операции:

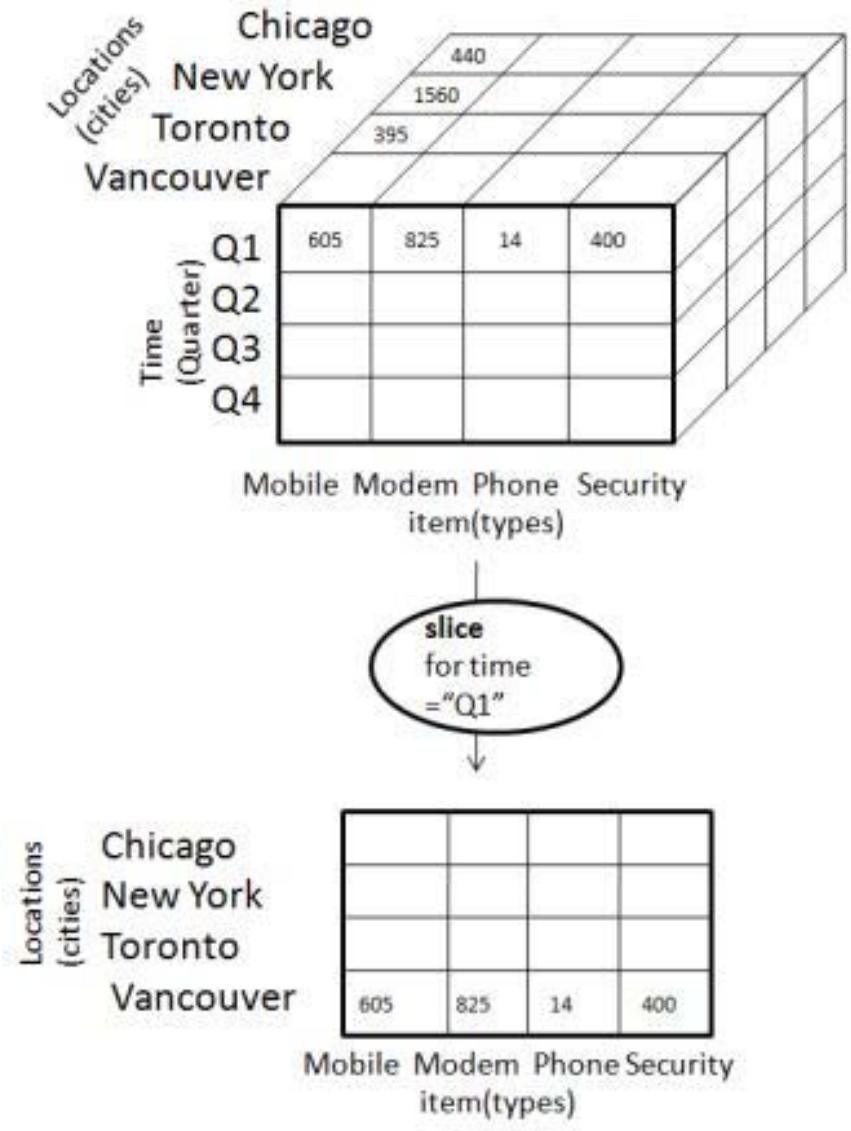
- Slice (двумерный (плоскостной) срез);
- Dice (многомерный подкуб);
- Roll-Up (консолидация, агрегация, обобщение);
- Drill down (детализация);
- Pivot (вращение).

Slice

The slice operation selects one particular dimension from a given cube and provides a new sub-cube.

Here Slice is performed for the dimension "time" using the criterion time = "Q1".

It will form a new sub-cube by selecting one or more dimensions.

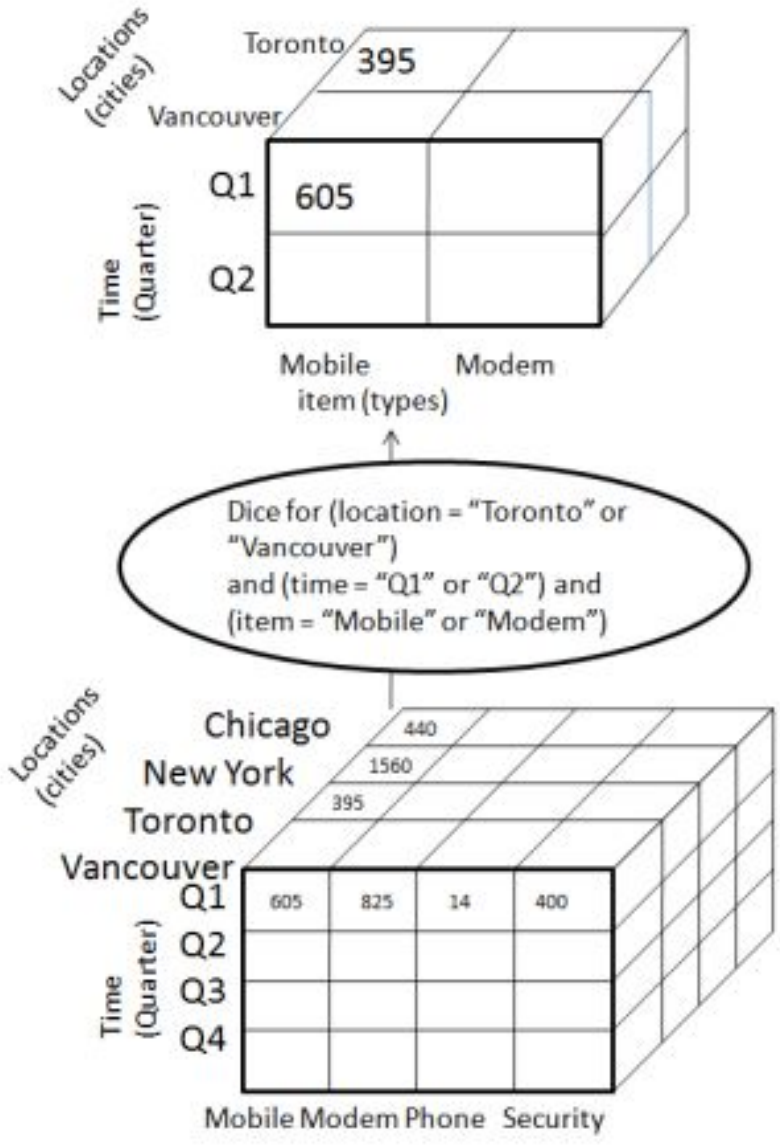


Dice

Dice selects two or more dimensions from a given cube and provides a new sub-cube.

The dice operation on the cube based on the following selection criteria involves three dimensions.

- (location = "Toronto" or "Vancouver")
- (time = "Q1" or "Q2")
- (item =" Mobile" or "Modem")



Roll-up

Roll-up performs aggregation on a data cube in any of the following ways:

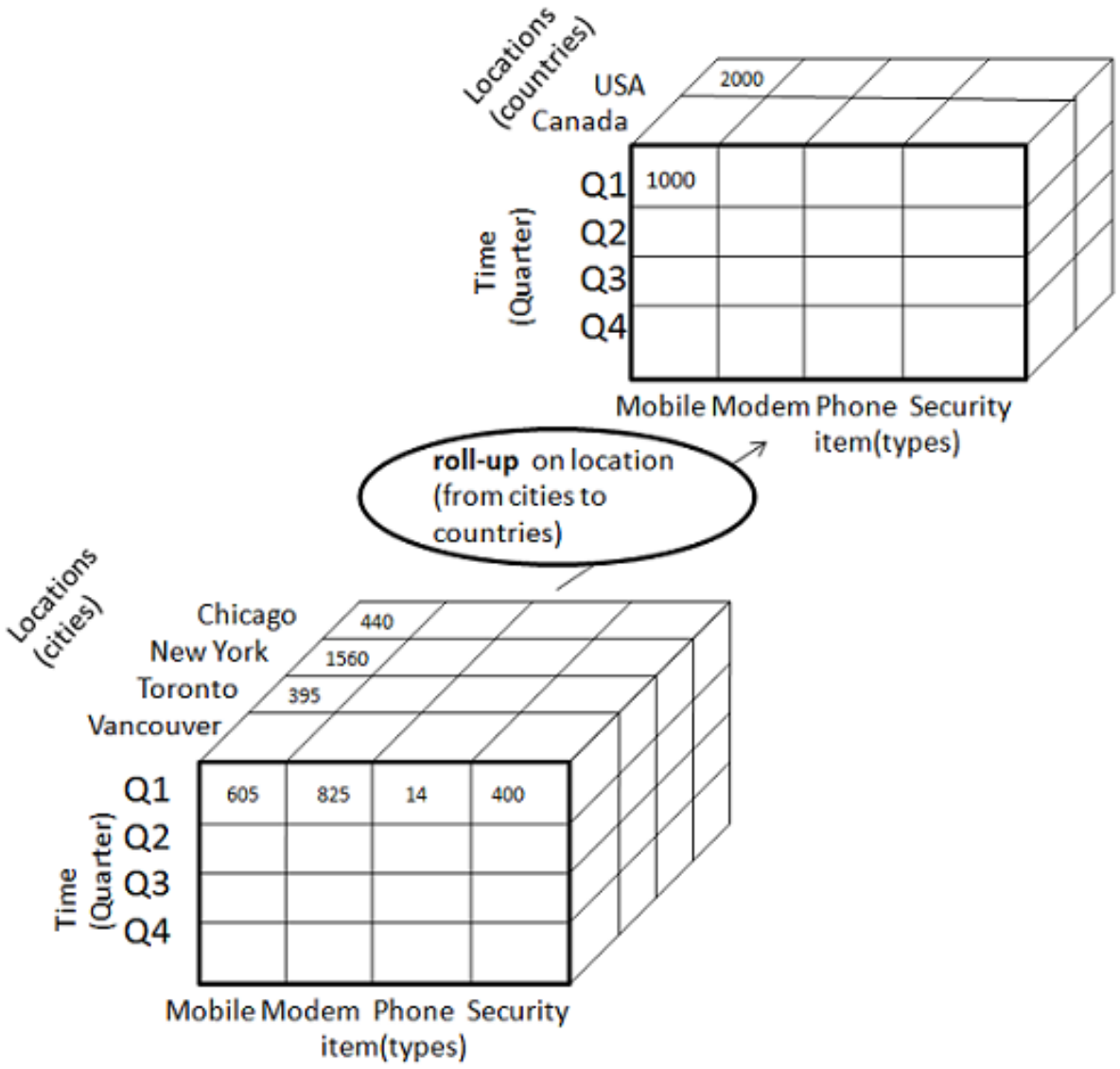
- By climbing up a concept hierarchy for a dimension
- By dimension reduction

Roll-up is performed by climbing up a concept hierarchy for the dimension location.

Initially the concept hierarchy was "street < city < province < country".

On rolling up, the data is aggregated by ascending the location hierarchy from the level of city to the level of country.

The data is grouped into cities rather than countries. But when roll-up is performed, one or more dimensions from the data cube are removed.



Drill-down

Drill-down is the reverse operation of roll-up. It is performed by either of the following ways:

- By stepping down a concept hierarchy for a dimension
- By introducing a new dimension.

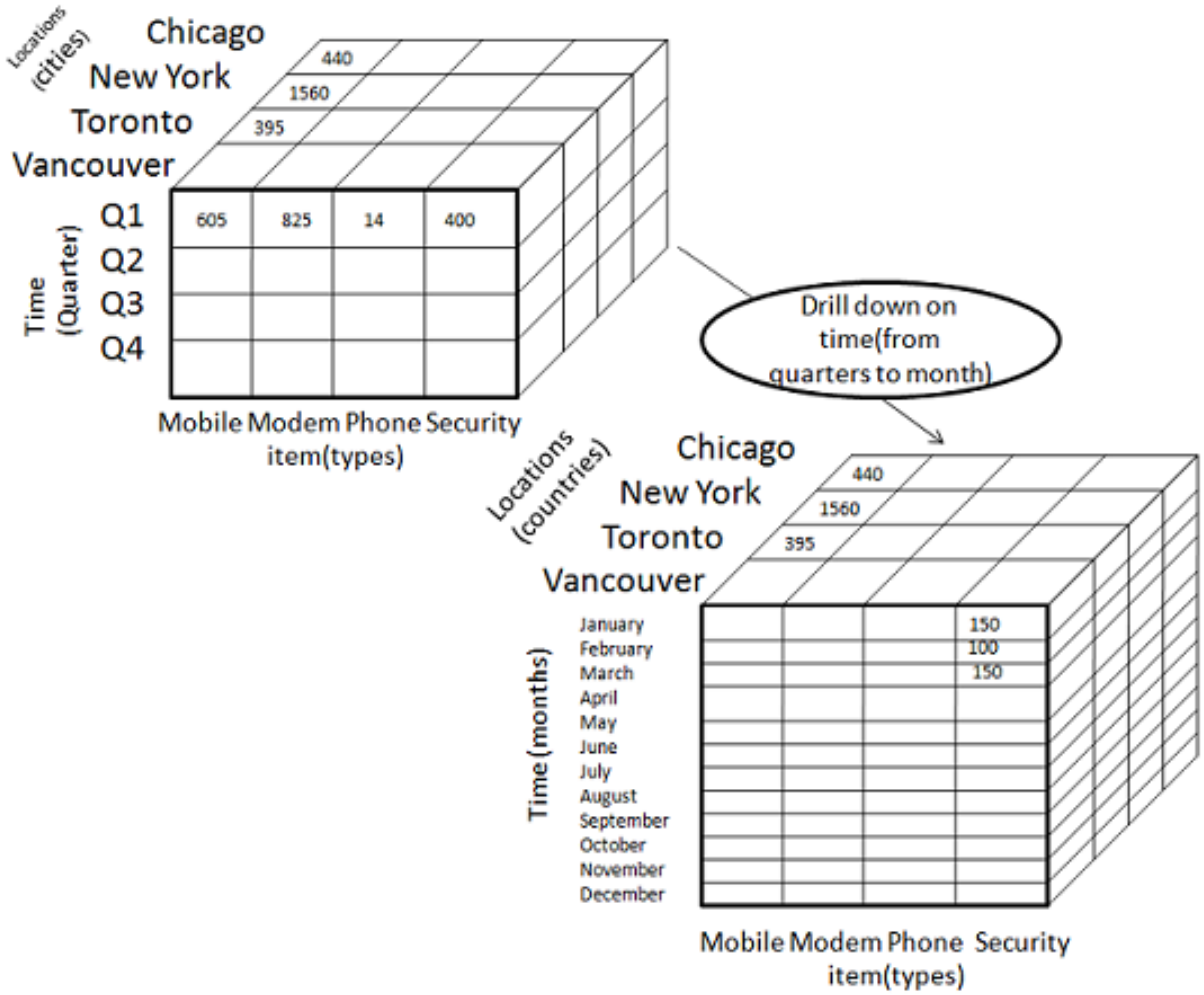
Drill-down is performed by stepping down a concept hierarchy for the dimension time.

Initially the concept hierarchy was "day < month < quarter < year".

On drilling down, the time dimension is descended from the level of quarter to the level of month.

When drill-down is performed, one or more dimensions from the data cube are added.

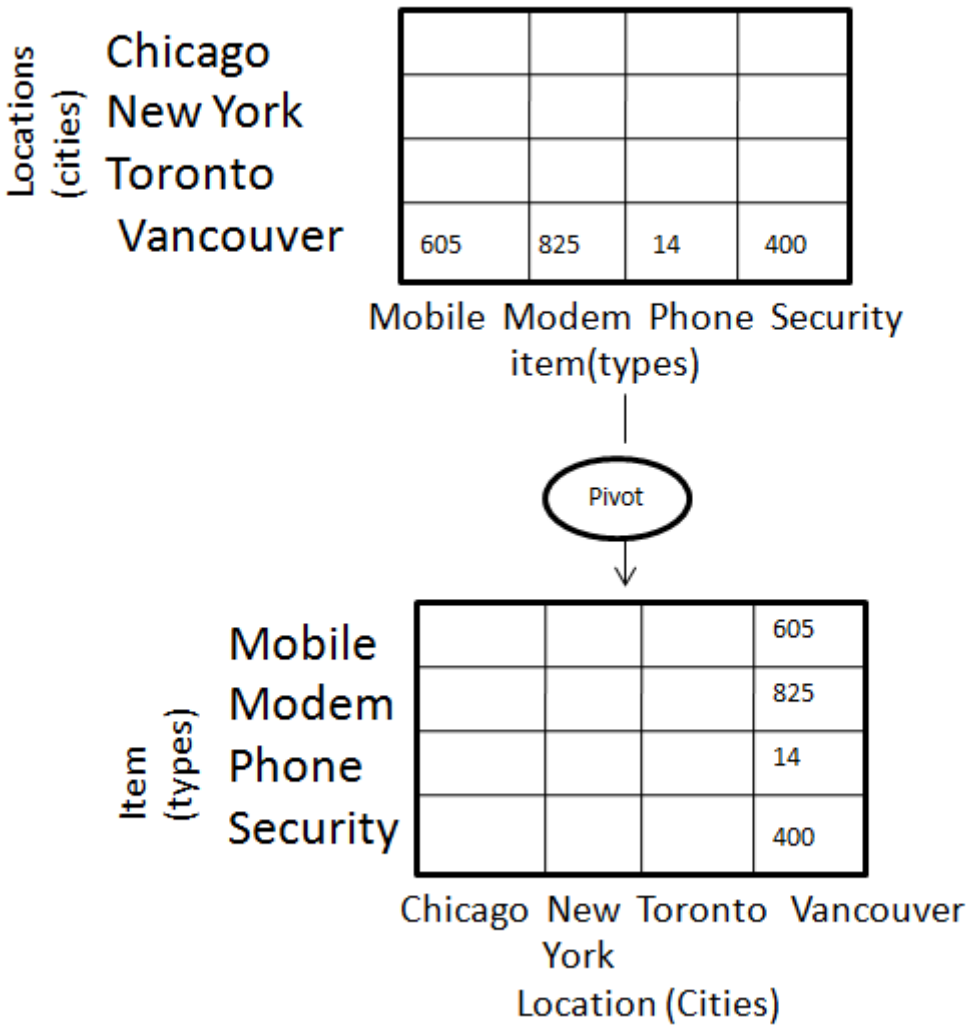
It navigates the data from less detailed data to highly detailed data.



Pivot

The pivot operation is also known as rotation. It rotates the data axes in view in order to provide an alternative presentation of data.

In this the item and location axes in 2-D slice are rotated.



Литература:

Анализ данных и процессов : учеб. пособие / 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.

Data Warehousing – OLAP [Electronic resource] / TutorialsPoint. – Mode of access: https://www.tutorialspoint.com/dwh/dwh_olap.htm. – Date of access: 20.01.2017.

OLAP-системы [Electronic resource] / TAdviser. – Mode of access: http://www.tadviser.ru/-index.php/Статья:OLAP-системы. – Date of access: 20.01.2017.