Агрегация показателей в Olap-кубе

Чемидова А.Б., магистр
ФГБОУ ВО «Московский политехнический университет», Москва, Россия
Суворов С.В., к.э.н., профессор
ФГБОУ ВО «Московский политехнический университет», Москва, Россия
Царькова Н.И., к.п.н., доцент
ФГБОУ ВО «Московский политехнический университет», Москва, Россия
Жиляева И.А., к.э.н., доцент,
ФГБОУ ВО «Московский политехнический университет», Москва, Россия

Аннотация. В статье обсуждаются особенности обработки данных с использованием многомерного анализа. Обсуждаются особенности агрегации данных, представлено графическое представление гиперкуба и основные операции с кубом. Представлен обзор аналогов Olap-куба и сформулирована постановка задачи.

Ключевые слова: Olap-куб, многомерная модель данных, агрегация, гиперкуб, интеллектуальный анализ данных.

Aggregation of indicators in Olap-cube

Chemidova A.B., Magister
Moscow Polytechnic University, Moscow, Russia
Suvorov S.V., candidate of Economic Sciences, Professor
Moscow Polytechnic University, Moscow, Russia
Tsarkova N.I., Candidate of Pedagogical Sciences, associate Professor
Moscow Polytechnic University, Moscow, Russia
Zhilyaeva I.A., Candidate of economic Sciences, associate Professor
Moscow Polytechnic University, Moscow, Russia

Abstract. The article discusses the features of data processing using multivariate analysis. The features of data aggregation are discussed, a graphical representation of the hypercube and basic operations with the cube are presented. The review of analogs of Olap-cube is presented and the problem statement is formulated.

Keywords: Olap cube, multidimensional data model, aggregation, hypercube, data mining.

Введение

В современное время большинство компаний используют для хранения и записи всех транзакций оперативную обработку транзакций (OLTP) баз данных. Эти транзакции со временем только продолжают расти. Миллиарды строк данных могут лежать мертвым грузом статистики и оставаться отработанной информацией о продажах. Вместо этого, эти данные могут стать основой для контроля и выявления слабых мест бизнеса, открытия возможностей для его роста и развития. Для таких данных нужен инструмент, который помогал бы «крутить» данные: складывать разные показатели, формировать нешаблонную отчетность и показывать картину текущего состояния бизнеса и целиком, и детально. Базы данных OLAP специально предусмотрены для упрощения извлечения необходимых сведений. Эта статья будет посвящена обсуждению аналитической обработке данных (OLAP).

Основные понятия OLAP

Системы поддержки принятия решений обладают средствами предоставления пользователю агрегатных данных для различных выборок из исходного набора в удобном для восприятия и анализа виде, благодаря этому пользователи могут формулировать сложные запросы, генерировать отчеты, получать подмножества данных [3].

OLAP является аббревиатурой для On-Line Analytical Processing и является категорией программного обеспечения, которая позволяет пользователям

анализировать информацию из нескольких систем баз данных одновременно. OLAP системы позволяют обрабатывать большие объемы данных в более эффективном способе с аналитической точки зрения, так как оптимизированы для чтения задач. Это основная функция, которая позволяет использовать систему OLAP в технических кубах.

Главную роль в OLAP-системах занимает операции агрегирования, которые представляют собой процедуру автоматического формирования меньшего количества результирующих показателей из большого количества исходных значений.

Обычно операции с данными и анализ выполняются с использованием простой электронной таблицы, где значения данных располагаются в формате строк и столбцов. Это идеально подходит для двумерных данных. Однако OLAP содержит многомерные данные, причем данные обычно получают из другого и несвязанного источника. Использование электронной таблицы не является оптимальным вариантом. Куб может хранить и анализировать многомерные данные в логической и упорядоченной форме.

OLAP куб представляет собой многомерный набор данных. Несмотря на то, что куб представляется как трехмерный объект, однако он может иметь большое количество измерений. Кубы OLAP позволяют анализировать данные из разных размеров или углов, для того, чтобы принимать решения. Его также называют гиперкубом.

Структура данных в кубе OLAP напоминает звезду. Звездная схема состоит из двух типов таблиц: меры и измерения.

Измерения (Dimensions) содержат первичные ключи вместе с атрибутами измерения, которые предоставляют пути детализации управления анализом OLAP. На рисунке 1 представлено графическое изображение куба, состоящего из 3-х измерений: Date (Дата), Emp (Сотрудники), Products (Продукты). Каждое измерение состоит из множества значений. При пересечении координат по каждому измерению образуется ячейка, которая является частью данных на каждом измерении многомерного массива.

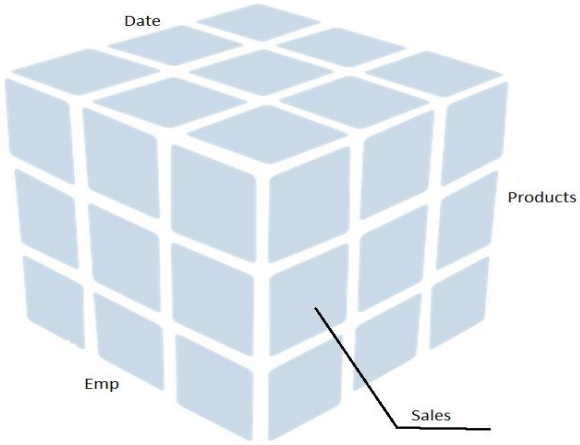

Рис. 1 – Гиперкуб

Иерархии (Hierarchy). Элементы измерения, как правило, упорядочены в иерархии, чтобы обеспечить структуру размерности. Показатели, объединяемые для одной иерархии должны быть количественно сопоставимы [1, с.106]. Широко используемое измерение Дата, которое может быть организовано в иерархию, представленное на рисунке 2, где день сгруппирован в месяц, месяц сгруппирован в квартал, а квартал в год. Это позволяет возможность проанализировать меры, основанные на различных уровнях в размерах.

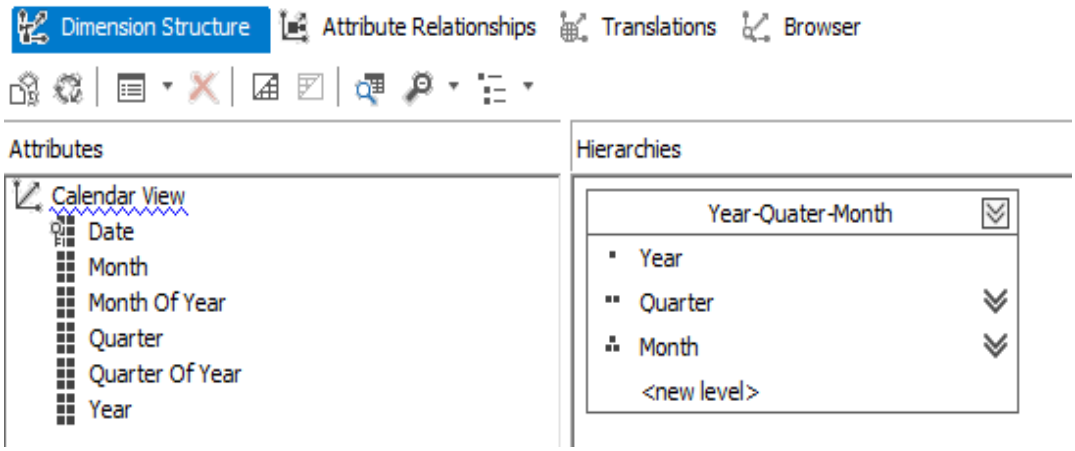

Рисунок 2 – Иерархии

Меры (Measures) содержат показатели, то есть числовые значения для сумм продаж, проданных единиц и т. д., а также внешние ключи, которые идентичны первичным ключам, найденным в таблицах измерений. На рисунке 1 мерой является Sales (Продажи).

Аналитики часто должны группировать, агрегировать и объединять данные. Эти операции в реляционных базах данных являются ресурсоемкими. Данные OLAP могут быть предварительно рассчитаны и предварительно агрегированы, что ускоряет анализ.

Все инструменты OLAP построены на 4 основных аналитических операциях:

- Консолидация (свертка) осуществляется объединение данных, которое можно вычислить во многих измерениях.
- Развертывание – это противоположная методика консолидации, которая позволяет пользователям посмотреть детализированные данные, выполненные при консолидации.
- Срез – это метод, при котором пользователи извлекают (нарезают) нужный набор данных.
- Вращение – это трансформация положения измерений на отображаемом отчете.

Интеллектуальный анализ данных состоит из извлечения, преобразования и загрузки данных из различных источников. Источники, хранящие и управляющие данными в многомерной системе баз данных, обеспечивают доступ к данным бизнес-аналитиков и специалистов в области информационных технологий, анализируют данные прикладным программным обеспечением, представляют данные в полезном формате, таком как график или таблица.

Обзор аналогов

Вместе с дальнейшим развитием OLAP появляются и развиваются другие подходы к хранению данных и операциям. Помимо быстро развивающегося в памяти OLAP, есть еще аналоги традиционному хранилищу данных.

1. Озера данных

По сравнению с традиционным хранилищем данных озеро данных (data lake) гораздо менее структурировано. Традиционные хранилища данных

основаны на технологиях СУБД. В отличие от таблиц, используемых в реляционных базах данных, многомерные кубы данных, найденные в этих СУБД, являются нереляционными хранилищами данных, поскольку данные загружаются в кубический массив. Однако хранилище данных обычно содержит больше данных, чем сам куб данных, что позволяет аналитикам выполнять запросы ниже самых гранулярных измерений в кубе. Как правило, отношения между таблицами в хранилище данных определяются с использованием звездообразной схемы, которая организует данные в мерах и измерениях для облегчения объединения таблиц с помощью запросов. Связывая внешние ключи с первичными ключами, схемы типа «звезда» связывают одни и те же факты с несколькими измерениями, что позволяет выполнять многомерный анализ в кубе данных.

Напротив, данные в озере данных гораздо менее структурированы и схематизированы. Во многих случаях данные все еще предварительно обрабатываются перед хранением, но без использования жестких структур (таких как схема типа «звезда»), обычно используемых для хранилищ данных.

В озере данных все данные хранятся независимо от источника и его структуры. Данные хранятся в необработанном виде. Он преобразуется только тогда, когда он готов к использованию. Хранилище данных будет состоять из данных, извлеченных из транзакционных систем, или данных, которые состоят из количественных метрик с их атрибутами. Данные очищены и преобразованы.

В хранилище данных информация структурируется с использованием схем при загрузке в хранилище - подход, известный как схема записи (поскольку данные схематизируются при записи в память).

В случае озера данных схема определяется, когда данные запрашиваются с помощью API или SQL - подхода, известного как схема чтения, которая требует гораздо менее жесткого моделирования данных.

Таким образом, данные озера требуют гораздо меньше предварительных данных моделирования. Схема на подходе чтения позволяет хранить неоднородные и многочисленные наборы данных. Данные озера также идеально

подходят для случаев, когда вы не уверены, как анализировать свои наборы данных или храните информацию из социальных сетей вместе с записями о продажах и данными о транзакциях. Основная проблема озер данных заключается в том, что ими обычно трудно управлять. Часто требуется участие ИТ-отдела, поскольку бизнес-аналитики могут не иметь дополнительных навыков. Озера данных используют компании СИБУР и MyGames[6].

Кроме того, озера данных достаточно глубоки, чтобы обрабатывать наборы данных в терабайтном и петабайтном масштабе, генерируемые массовыми развертываниями датчиков в промышленных или ориентированных на потребителя проектах.

2. Анализ в стиле OLAP с помощью инструментов самообслуживания

Инструменты самообслуживания BI используют технологию, отличную от традиционных инструментов OLAP, поддерживаемых хранилищами данных. В частности, инструменты самообслуживания используют кэши данных хранилища столбцов, а не кубы данных OLAP. Кэши данных не требуют записи или чтения на диск. Вместо этого они доступны через память, поэтому процесс запросов становится намного быстрее.

Существует такая аналитическая модель Kyubit, которая является аналитической функцией самообслуживания BI. Она быстро создается с использованием данных из результатов запросов SQL и файлов CSV. Набор значений из результатов запросов SQL или файлов CSV преобразуется в аналитические модели, и инструменты самообслуживания BI. С помощью «Аналитических моделей» конечный пользователь может создавать аналитику, отчеты, визуализации и информационные панели, используя показатели, измерения, срезы и многие функции, аналогичные анализу OLAP.

После импорта данных из файлов CSV или результатов SQL-запроса конечному пользователю необходимо определить типы данных (мер и измерений) и затем можно настраивать аналитические модели, поведение которых в многомерном анализе очень похоже на анализ куба OLAP.

Единственной технологической предпосылкой является MS SQL Server, который в любом случае является обязательным условием для всего продукта.

Анализ самообслуживания предлагает тот же вид многомерного взаимодействия с наборами данных, для которого изначально требовалось использование кубов OLAP, хотя пользователи определяют свои собственные пути детализации, а не следуют тем, которые были структурированы в куб данных во время моделирования данных.

Таким образом, выбор аналогов сводится в зависимости от потребностей. Для организаций, которые в основном хранят данные транзакций для исторической аналитики, хранилище данных по-прежнему имеет смысл. Однако, при сопоставлении транзакционных данных с масштабами веб-данных или большими наборами данных, озеро данных будет больше подходить.

Постановка задачи

На основе проведённого анализа можно сделать вывод, что OLAP-куб является очень хорошей методикой анализа больших объемов данных. Поэтому целью моей работы будет проектирование и формирование OLAP куба для ветеринарных дистрибьюторов с расчетом плана и факта продаж, а также расчет KPI по каждому менеджеру по продажам в РФ.

Для создания куба я буду использовать средства SQL Server 2012 BI Edition, а именно Microsoft SQL Server Analysis Services (SSAS), SQL Data Tools for Visual Studio (BI) и Excel.

Данные, необходимые для отчетов будут выгружены из OLAB-куба в Excel, т.к. этот инструмент является доступным всем пользователям, и все умеют с ним взаимодействовать.

На рисунке 3 представлена графическая нотация гиперкуба, предназначенного для формирования отчетности по плану и фактам продаж. Гиперкуб Vet_Sales обозначается символом куба, из него выходят линии к принадлежащим ему измерениям: Линия, Продукт, Сотрудник, Клиент, Филиал, Дата. Измерения обозначаются прямоугольником со скощенными краями, а от

измерений выходят линии, которые принадлежат иерархиям этого измерения и обозначаются овалом. От куба также есть разветвления: Plan и Sales, которые являются мерами данного куба и обозначаются прямоугольником.

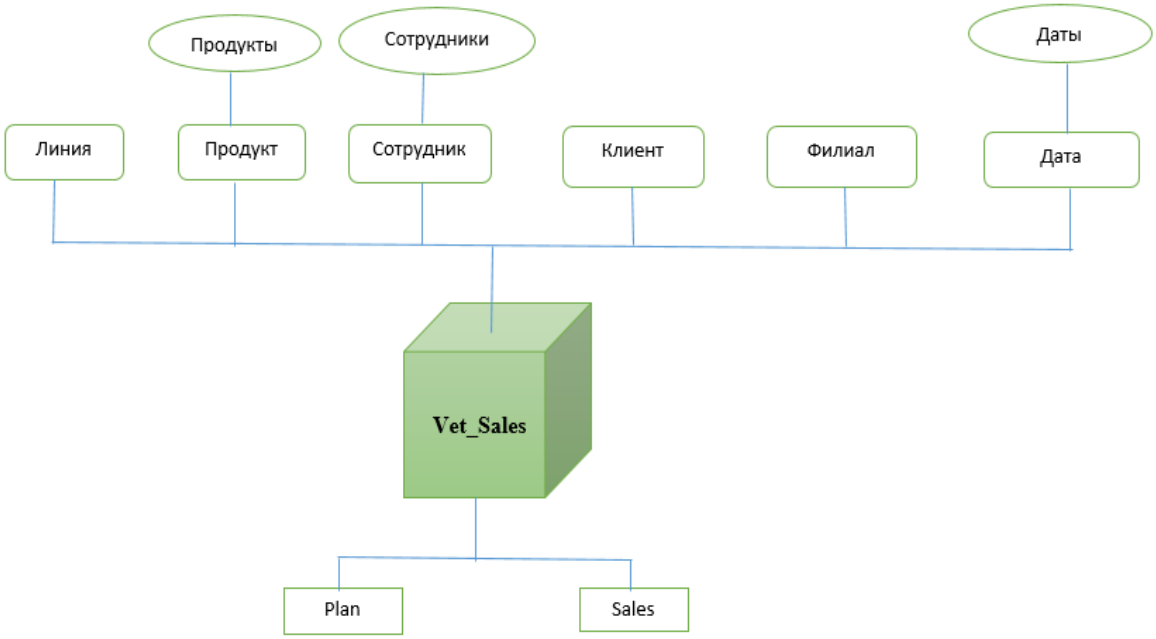

Рисунок 3 – Графическая нотация

В таблице 1 даны представления, которые необходимы для проектирования куба, представление календаря создадим автоматически, используя встроенный календарь.

Таблица 1

Представления

Employee_View	Сотрудники
Product_vet	Продукты
UCClients	Уникальные клиенты
Lines	Линии
Supplier	Поставщики (дистрибьюторы)

В таблице 2 представлена сетка измерений, которая необходима для просмотра и редактирования связей между измерениями куба и группами мер. Каждая связь измерений представляется ячейкой в сетке, в которой группы мер

представлены столбцами, а измерения представлены строками. Данную информацию можно увидеть на вкладке Dimension Usage (Использование измерений) Visual Studio.

Таблица 2

Сетка измерений

	Продажи (уп, руб)	План (уп, руб)
продукты	Product_id	Product_id
сотрудники	Emp_id	Emp_id
клиенты	Uc_id	
дистрибьюторы	Filial_id	
календарь	date	date
линия	Line_id	

Пользователями данного куба будут менеджеры по продажам, которые будут видеть свои плановые и фактические показатели, также для удобства будет рассчитан KPI каждого сотрудника. Так как менеджеру необходимо видеть только себя и своих коллег, куб будет иметь несколько ролей доступа.

Показатели KPI представляют собой бизнес-метрики, создаваемые для наблюдения за развитием в сторону определенных заданных целей. У показателя KPI имеется фактическое и плановое значение, представляющее собой количественную цель, достижение которой важно для успеха организации. Показатели KPI обычно отображаются в виде групп в системе показателей, демонстрируя общую работоспособность бизнеса.

Таким образом, благодаря внедрению OLAP куба с расчетом плана и факта продаж позволит быстро и удобно перейти к качественному анализу результатов деятельности организации по технологии OLAP, повысит эффективность информационно-аналитической и управленческой деятельности руководящего персонала, позволит быстрее и более обоснованно принимать оперативные и стратегические решения.

Библиографический список

1. Erik Thomsen. OLAP Solutions: Building Multidimensional Information Systems Second Edition. Wiley Computer Publishing John Wiley & Sons, Inc., 2002.

2. SQL Server 2012 Overview, data platform, store data [Электронный ресурс]. Режим доступа: http://www.microsoft.com/sql/default.mspx, свободный (дата обращения: 01.12.2019).

3. Алексей Федоров, Наталия Елманова Введение в OLAP [Электронный ресурс]. Режим доступа: http://kek.ksu.ru/EOS/DW/OLAP_Microsoft.pdf, свободный (дата обращения: 01.12.2019).

4. Фисун Н.Т., Горбань Г.В., Модели и методы построения системы OLAP для объектно-ориентированных баз данных // Объектные системы. – 2013.– №1 (7). Режим доступа: https://cyberleninka.ru/article/n/modeli-i-metody-postroeniya-sistemy-olap-dlya-obektno-orientirovannyh-baz-dannyh, свободный (дата обращения: 09.12.2019).

5. Swathi R Kasireddy. Thesis: Olap reporting application using office web components, – 2008.

6. Habr, Озеро данных для маркетинга — от монструозных таблиц до отчётов и визуализации, 2019 [Электронный ресурс]. Режим доступа: https://habr.com/ru/company/sibur_official/blog/461029/, свободный (дата обращения: 01.12.2019).

7. Бергер, А.Б. Microsoft® SQL Server 2005 Analysis Services. OLAP и многомерный анализ данных / Бергер А.Б., Горбач И.В., Меломед Э.Л., /Под общ. ред. А.Б. Бергера, И.В. Горбач. – СПб.: БХВ-Петербург, – 2007. – 928 с

8. Царькова Н.И., Суворов С.В., Жиляева И.А., Шебанова К.В. Множественное восстановление пропущенных данных с помощью глубинных нейробайесовских моделей – Российский экономический интернет-журнал, 2019г – № 4 – С. 45

9. Суворов С.В., Царькова Н.И., Спиридонова А.К. Анализ больших данных компании UBER TECHNOLOGIES INC с помощью технологии DATA

MINING – Управление экономическими системами: электронный научный журнал, 2019 г – № 7 (125) – С. 21

References

1. Erik Thomsen. OLAP Solutions: Building Multidimensional Information Systems Second Edition. Wiley Computer Publishing John Wiley & Sons, Inc., 2002.

2. SQL Server 2012 Overview, data platform, store data [Electronic resource]. Access mode: http://www.microsoft.com/sql/default.mspx, free (accessed: 12/01/2019).

3. Alexei Fedorov, Natalia Elmanova Introduction to OLAP [Electronic resource]. Access mode: http://kek.ksu.ru/EOS/DW/OLAP_Microsoft.pdf, free (access date: 12/01/2019).

4. Fisun N.T., Gorban G.V., Models and methods for constructing an OLAP system for object-oriented databases // Object systems. – 2013. – №1 (7). Access mode: https://cyberleninka.ru/article/n/modeli-i-metody-postroeniya-sistemy-olap-dlya-obektno-orientirovannyh-baz-dannyh, free (accessed: 12/9/2019).

5. Swathi R Kasireddy. Thesis: Olap reporting application using office web components, – 2008.

6. Habr, A lake of data for marketing - from monstrous tables to reports and visualizations, 2019 [Electronic resource]. Access mode: https://habr.com/en/company/sibur_official/blog/461029/, free (access date: 12/01/2019).

7. Berger, A. B. Microsoft® SQL Server 2005 Analysis Services. OLAP and multivariate data analysis / Berger A.B., Gorbach I.V., Melomed E.L., / Under the general. ed. A.B. Berger, I.V. Gorbach. – SPb.: BHV – Petersburg, – 2007. – 928 s.

8. Tsarkova N.I., Suvorov S.V., Zhilyaeva I.A., Shebanova K.V. Multiple recovery of missing data using deep neuro-Bayesian models - Russian Economic Internet Journal, 2019 – №4 – P. 45

9. Suvorov S.V., Tsarkova N.I., Spiridonova A.K. UBER TECHNOLOGIES INC Big Data Analysis Using DATA MINING Technology – Economic Systems Management: Electronic Scientific Journal, 2019 – №7 (125) – P. 21