МИНОБРНАУКИ РОССИИ

Федеральное государственное бюджетное образовательное учреждение
высшего образования
«МИРЭА – Российский технологический университет»
РТУ МИРЭА

Институт Информационных технологий
Кафедра Инструментального и прикладного программного обеспечения

САМОСТОЯТЕЛЬНАЯ РАБОТА СТУДЕНТА

по дисциплине

«Хранилище данных и OLAP-технологии»

по теме

«Концепции аналитической обработки данных (OLAP)»

Студенты групп ИКБО-00-15 Апальков П.Ю.

Зачтено Богорадникова А.В.

Москва 2018

Введение

Современный уровень развития аппаратных и программных средств с некоторых пор сделал возможным повсеместное ведение баз данных оперативной информации на разных уровнях управления. В процессе своей деятельности промышленные предприятия, корпорации, ведомственные структуры, органы государственной власти и управления накопили большие объемы данных. Они хранят в себе большие потенциальные возможности по извлечению полезной аналитической информации, на основе которой можно выявлять скрытые тенденции, строить стратегию развития, находить новые решения.

В последние годы в мире оформился ряд новых концепций хранения и анализа корпоративных данных:

1) Хранилища данных, или Склады данных (Data Warehouse)
2) Оперативная аналитическая обработка (On-Line Analytical Processing, OLAP)
3) Интеллектуальный анализ данных - ИАД (Data Mining)

Технологии OLAP тесно связаны с технологиями построения Data Warehouse и методами интеллектуальной обработки - Data Mining. Поэтому наилучшим вариантом является комплексный подход к их внедрению.

Способы аналитической обработки данных

Для того чтобы существующие хранилища данных способствовали принятию управленческих решений, информация должна быть представлена аналитику в нужной форме, то есть он должен иметь развитые инструменты доступа к данным хранилища и их обработки.

Очень часто информационно-аналитические системы, создаваемые в расчете на непосредственное использование лицами, принимающими решения, оказываются чрезвычайно просты в применении, но жестко ограничены в функциональности. Такие статические системы называются в литературе Информационными системами руководителя (ИСР), или Executive Information Systems (EIS) [3]. Они содержат в себе предопределенные множества запросов и, будучи достаточными для повседневного обзора, неспособны ответить на все вопросы к имеющимся данным, которые могут возникнуть при принятии решений. Результатом работы такой системы, как правило, являются многостраничные отчеты, после тщательного изучения которых у аналитика появляется новая серия вопросов. Однако каждый новый запрос, непредусмотренный при проектировании такой системы, должен быть сначала формально описан, закодирован программистом и только затем выполнен. Время ожидания в таком случае может составлять часы и дни, что не всегда приемлемо. Таким образом, внешняя простота статических СППР, за которую активно борется большинство заказчиков информационно-аналитических систем, оборачивается катастрофической потерей гибкости.

Динамические СППР, напротив, ориентированы на обработку нерегламентированных (ad hoc) запросов аналитиков к данным. Наиболее глубоко требования к таким системам рассмотрел E. F. Codd в статье [11], положившей начало концепции OLAP. Работа аналитиков с этими системами заключается в интерактивной последовательности формирования запросов и изучения их результатов.

Но динамические СППР (системы поддержки принятия решений) могут действовать не только в области оперативной аналитической обработки (OLAP); поддержка принятия управленческих решений на основе накопленных данных может выполняться в трех базовых сферах [21].

1. Сфера детализированных данных. Это область действия большинства систем, нацеленных на поиск информации. В большинстве случаев реляционные СУБД отлично справляются с возникающими здесь задачами. Общепризнанным стандартом языка манипулирования реляционными данными является SQL. Информационно-поисковые системы, обеспечивающие интерфейс конечного пользователя в задачах поиска детализированной информации, могут использоваться в качестве надстроек как над отдельными базами данных транзакционных систем, так и над общим хранилищем данных.
2. Сфера агрегированных показателей. Комплексный взгляд на собранную в хранилище данных информацию, ее обобщение и агрегация, гиперкубическое представление и многомерный анализ являются задачами систем оперативной аналитической обработки данных (OLAP) [11, 10, 6]. Здесь можно или ориентироваться на специальные многомерные СУБД [6], или оставаться в рамках реляционных технологий. Во втором случае заранее агрегированные данные могут собираться в БД звездообразного вида, либо агрегация информации может производиться на лету в процессе сканирования детализированных таблиц реляционной БД.
3. Сфера закономерностей. Интеллектуальная обработка производится методами интеллектуального анализа данных (ИАД, Data Mining) [19, 25], главными задачами которых являются поиск функциональных и логических закономерностей в накопленной информации, построение моделей и правил, которые объясняют найденные аномалии и/или прогнозируют развитие некоторых процессов.

Некоторые авторы выделяют в отдельную область анализ отклонений (например, в целях отслеживания колебаний биржевых курсов). В качестве примера может быть приведен статистический анализ рядов динамики. Чаще, однако, этот тип анализа относят к области закономерностей.

Полная структура информационно-аналитической системы, построенной на основе хранилища данных, показана на рис. 1. В конкретных реализациях отдельные компоненты этой схемы часто отсутствуют.

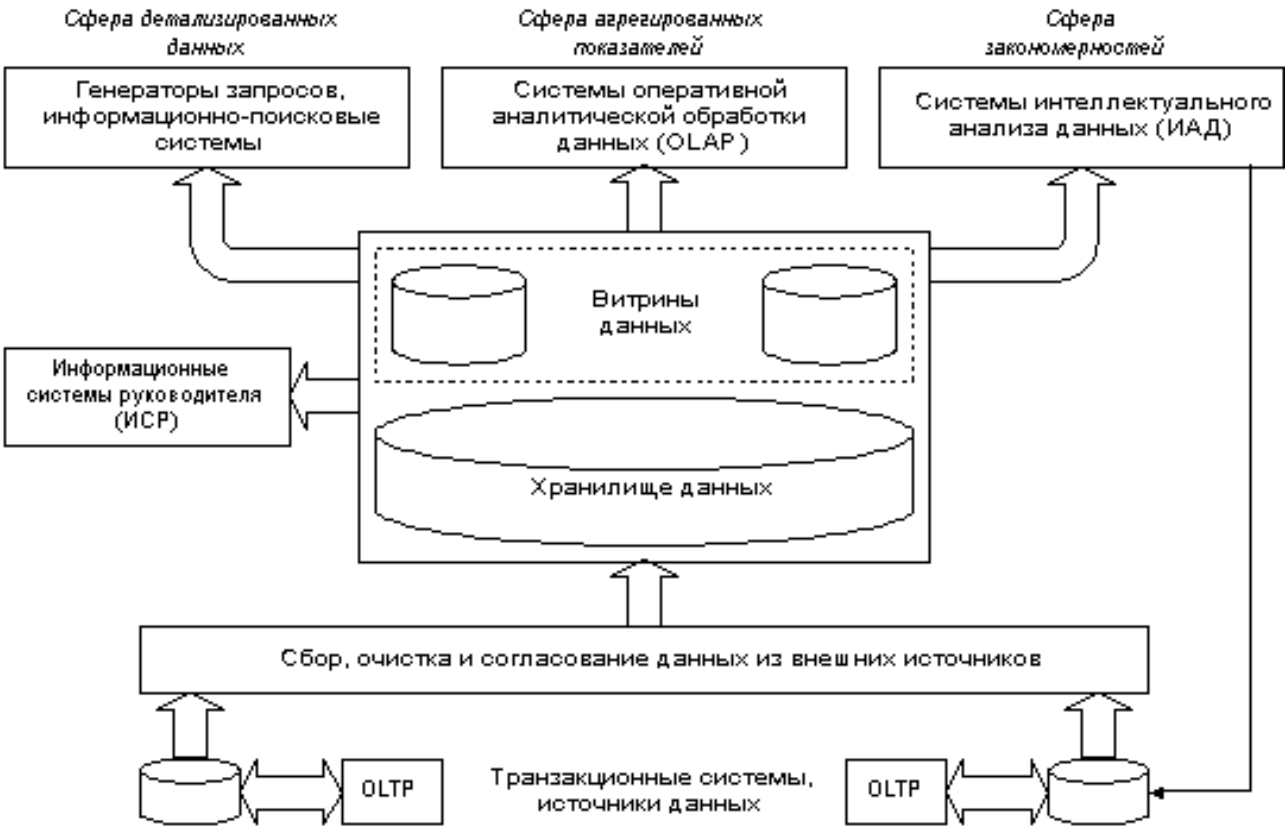

Рис. 1. Полная структура корпоративной информационно-аналитической системы (ИАС)

Оперативная аналитическая обработка данных

В основе концепции OLAP лежит принцип многомерного представления данных. В 1993 году в статье [11] E. F. Codd рассмотрел недостатки реляционной модели, в первую очередь указав на невозможность "объединять, просматривать и анализировать данные с точки зрения множественности измерений, то есть самым понятным для корпоративных аналитиков способом", и определил общие требования к системам OLAP, расширяющим функциональность реляционных СУБД и включающим многомерный анализ как одну из своих характеристик.

В большом числе публикаций аббревиатурой OLAP обозначается не только многомерный взгляд на данные, но и хранение самих данных в многомерной БД [6]. Вообще говоря, это неверно, поскольку сам Кодд отмечает, что "Реляционные БД были, есть и будут наиболее подходящей технологией для хранения корпоративных 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".

Такая путаница приводит к противопоставлениям наподобие "OLAP или ROLAP", что не совсем корректно, поскольку ROLAP (реляционный OLAP) на концептуальном уровне поддерживает всю определенную термином OLAP функциональность. Более предпочтительным кажется использование для OLAP

на основе многомерных СУБД специального термина MOLAP, как это и сделано в [4, 9].

По Кодду, многомерное концептуальное представление (multi-dimensional conceptual view) представляет собой множественную перспективу, состоящую из нескольких независимых измерений, вдоль которых могут быть проанализированы определенные совокупности данных. Одновременный анализ по нескольким измерениям определяется как многомерный анализ. Каждое измерение включает направления консолидации данных, состоящие из серии последовательных уровней обобщения, где каждый вышестоящий уровень соответствует большей степени агрегации данных по соответствующему измерению. Так, измерение Исполнитель может определяться направлением консолидации, состоящим из уровней обобщения "предприятие - подразделение - отдел - служащий". Измерение Время может даже включать два направления консолидации - "год - квартал - месяц - день" и "неделя - день", поскольку счет времени по месяцам и по неделям несовместим. В этом случае становится возможным произвольный выбор желаемого уровня детализации информации по каждому из измерений. Операция спуска (drilling down) соответствует движению от высших ступеней консолидации к низшим; напротив, операция подъема (rolling up) означает движение от низших уровней к высшим (рис. 2).

Рис. 2. Измерения и направления консолидации данных

Требования к средствам оперативной аналитической обработки

Кодд определил 12 правил, которым должен удовлетворять программный продукт класса OLAP (табл. 1).
Таблица 1 Правила оценки программных продуктов класса OLAP

1	Многомерное концептуальное представление данных (Multi-Dimensional Conceptual View)	Концептуальное представление модели данных в продукте OLAP должно быть многомерным по своей природе, то есть позволять аналитикам выполнять интуитивные операции "анализа вдоль и поперек" ("slice and dice"), вращения (rotate) и размещения (pivot) направлений консолидации.
2	Прозрачность (Transparency)	Пользователь не должен знать о том, какие конкретные средства используются для хранения и обработки данных, как данные организованы и откуда берутся.
3	Доступность (Accessibility)	Аналитик должен иметь возможность выполнять анализ в рамках общей концептуальной схемы, но при этом данные могут оставаться под управлением оставшихся от старого наследства СУБД, будучи при этом привязанными к общей аналитической модели. То есть инструментарий OLAP должен накладывать свою логическую схему на физические массивы данных, выполняя все преобразования, требующиеся для обеспечения единого, согласованного и целостного взгляда пользователя на информацию.
4	Устойчивая производительность (Consistent Reporting Performance)	С увеличением числа измерений и размеров базы данных аналитики не должны столкнуться с каким бы то ни было уменьшением производительности. Устойчивая производительность необходима для поддержания простоты использования и свободы от усложнений, которые требуются для доведения OLAP до конечного пользователя.
5	Клиент - серверная архитектура (Client-Server Architecture)	Большая часть данных, требующих оперативной аналитической обработки, хранится в мэйнфреймовых системах, а извлекается с персональных компьютеров. Поэтому одним из требований является способность продуктов OLAP работать в среде клиент-сервер. Главной идеей здесь является то, что серверный компонент инструмента OLAP должен быть достаточно интеллектуальным и обладать способностью строить общую концептуальную схему на основе обобщения и консолидации различных логических и физических схем корпоративных баз данных для обеспечения эффекта прозрачности.

6	Равноправие измерений (Generic Dimensionality)	Все измерения данных должны быть равноправны. Дополнительные характеристики могут быть предоставлены отдельным измерениям, но поскольку все они симметричны, данная дополнительная функциональность может быть предоставлена любому измерению. Базовая структура данных, формулы и форматы отчетов не должны опираться на какое-то одно измерение.
7	Динамическая обработка разреженных матриц (Dynamic Sparse Matrix Handling)	Инструмент OLAP должен 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 разреженность данных.
8	Поддержка многопользовательского режима (Multi-User Support)	Зачастую несколько аналитиков имеют необходимость работать одновременно с одной аналитической моделью или создавать различные модели на основе одних корпоративных данных. Инструмент OLAP должен предоставлять им конкурентный доступ, обеспечивать целостность и защиту данных.
9	Неограниченная поддержка кроссмерных операций (Unrestricted Cross-dimensional Operations)	Вычисления и манипуляция данными по любому числу измерений не должны запрещать или ограничивать любые отношения между ячейками данных. Преобразования, требующие произвольного определения, должны задаваться на функционально полном формульном языке.
10	Интуитивное манипулирование данными (Intuitive Data Manipulation)	Переориентация направлений консолидации, детализация данных в колонках и строках, агрегация и другие манипуляции, свойственные структуре иерархии направлений консолидации, должны выполняться в максимально удобном, естественном и комфортном пользовательском интерфейсе.
11	Гибкий механизм генерации отчетов (Flexible Reporting)	Должны поддерживаться различные способы визуализации данных, то есть отчеты должны представляться в любой возможной ориентации.
12	Неограниченное количество измерений и уровней агрегации (Unlimited Dimensions and Aggregation Levels)	Настоятельно рекомендуется допущение в каждом серьезном OLAP инструменте как минимум пятнадцати, а лучше двадцати, измерений в аналитической модели. Более того, каждое из этих измерений должно допускать практически неограниченное количество определенных пользователем уровней агрегации по любому направлению консолидации.