Оперативная аналитическая обработка (OLAP) — это технология, которая упорядочивает большие коммерческие базы данных и поддерживает сложный анализ. Ее можно использовать для выполнения сложных аналитических запросов без негативного воздействия на системы транзакций.

Базы данных, в которых компании хранят свои транзакции и записи, называются базами данных оперативной обработки транзакций (OLAP). Такие базы данных обычно содержат записи, которые вводятся поочередно. Часто они содержат много ценных для организации сведений. Но базы данных, используемые для OLAP, не предназначены для анализа. Поэтому извлечение ответов из этих баз данных требует много времени и усилий. Системы OLAP предназначены для извлечения этих сведений бизнес-аналитики из данных максимально оптимальным способом. Это достигается благодаря тому, что базы данных OLAP оптимизированы для рабочих нагрузок с большим числом операций чтения и малым числом операций записи.

В основе концепции OLAP лежит принцип многомерного представления данных. В 1993 году в статье 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

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

Требования к средствам оперативной аналитической обработки

Кодд определил 12 правил, которым должен удовлетворять программный продукт 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.

Набор этих требований, послуживших фактическим определением OLAP, следует рассматривать как рекомендательный, а конкретные продукты оценивать по степени приближения к идеально полному соответствию всем требованиям.

Классификация продуктов OLAP по способу представления данных

В настоящее время на рынке присутствует большое количество продуктов, которые в той или иной степени обеспечивают функциональность OLAP. Обеспечивая многомерное концептуальное представление со стороны пользовательского интерфейса к исходной базе данных, все продукты OLAP делятся на три класса по типу исходной БД.

1. Самые первые системы оперативной аналитической обработки (например, Essbase компании Arbor Software, Oracle Express Server компании Oracle) относились к классу MOLAP, то есть могли работать только со своими

собственными многомерными базами данных. Они основываются на патентованных технологиях для многомерных СУБД и являются наиболее дорогими. Эти системы обеспечивают полный цикл OLAP-обработки. Они либо включают в себя, помимо серверного компонента, собственный интегрированный клиентский интерфейс, либо используют для связи с пользователем внешние программы работы с электронными таблицами. Для обслуживания таких систем требуется специальный штат сотрудников, занимающихся установкой, сопровождением системы, формированием представлений данных для конечных пользователей.

2. Системы оперативной аналитической обработки реляционных данных (ROLAP) позволяют представлять данные, хранимые в реляционной базе, в многомерной форме, обеспечивая преобразование информации в многомерную модель через промежуточный слой метаданных. К этому классу относятся DSS Suite компании MicroStrategy, MetaCube компании Informix, DecisionSuite компании Information Advantage и другие. Программный комплекс ИнфоВизор, разработанный в России, в Ивановском государственном энергетическом университете, также является системой этого класса. ROLAP-системы хорошо приспособлены для работы с крупными хранилищами. Подобно системам MOLAP, они требуют значительных затрат на обслуживание специалистами по информационным технологиям и предусматривают многопользовательский режим работы.
3. Наконец, гибридные системы (Hybrid OLAP, HOLAP) разработаны с целью совмещения достоинств и минимизации недостатков, присущих предыдущим классам. К этому классу относится Media/MR компании Speedware. По утверждению разработчиков, он объединяет аналитическую гибкость и скорость ответа MOLAP с постоянным доступом к реальным данным, свойственным ROLAP.

Помимо перечисленных средств существует еще один класс - инструменты генерации запросов и отчетов для настольных ПК, дополненные функциями OLAP или интегрированные с внешними средствами, выполняющими такие функции. Эти хорошо развитые системы осуществляют выборку данных из исходных источников, преобразуют их и помещают в динамическую многомерную БД, функционирующую на клиентской станции конечного пользователя. Основными представителями этого класса являются BusinessObjects одноименной компании, BrioQuery компании Brio Technology и PowerPlay компании Cognos.

Многомерный OLAP (***MOLAP***)

В специализированных СУБД, основанных на многомерном представлении данных, данные организованы не в форме реляционных таблиц, а в виде упорядоченных многомерных массивов:

1) гиперкубов (все хранимые в БД ячейки должны иметь одинаковую мерность, то есть находиться в максимально полном базисе измерений) или

2) поликубов (каждая переменная хранится с собственным набором измерений, и все связанные с этим сложности обработки перекладываются на внутренние механизмы системы).

Использование многомерных БД в системах оперативной аналитической обработки имеет следующие достоинства.

1. В случае использования многомерных СУБД поиск и выборка данных осуществляется значительно быстрее, чем при многомерном концептуальном взгляде на реляционную базу данных, так как многомерная база данных денормализована, содержит заранее агрегированные показатели и обеспечивает оптимизированный доступ к запрашиваемым ячейкам.
2. Многомерные СУБД 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 СУБД достаточно сложным, а иногда и невозможным.

С другой стороны, имеются существенные ограничения.

1. Многомерные СУБД не позволяют работать с большими базами данных. К тому же за счет денормализации и предварительно выполненной агрегации объем данных в многомерной базе, как правило, соответствует (по оценке Кодда) в 2.5-100 раз меньшему объему исходных детализированных данных.
2. Многомерные СУБД по сравнению с реляционными очень неэффективно используют внешнюю память. В подавляющем большинстве случаев информационный гиперкуб является сильно разреженным, а поскольку данные хранятся в упорядоченном виде, неопределенные значения удаётся 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 всего

используется в запросах. Поэтому в реальных системах приходится искать компромисс между быстродействием и избыточностью дискового пространства, занятого базой данных.

Следовательно, использование многомерных СУБД оправдано только при следующих условиях.

1. Объем исходных данных для анализа не слишком велик (не более нескольких гигабайт), то есть уровень агрегации данных достаточно высок.
2. Набор информационных измерений стабилен (поскольку любое изменение в их структуре почти всегда требует полной перестройки гиперкуба).
3. Время ответа системы на нерегламентированные запросы является наиболее критичным параметром.
4. Требуется широкое использование сложных встроенных функций для выполнения кроссмерных вычислений над ячейками гиперкуба, в том числе возможность написания пользовательских функций.

Реляционный OLAP (***ROLAP***)

Непосредственное использование реляционных БД в системах оперативной аналитической обработки имеет следующие достоинства.

1. 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.
2. В случае переменной размерности задачи, когда изменения в структуру измерений приходится вносить достаточно часто, ROLAP системы с динамическим представлением размерности являются оптимальным решением, так как в них такие модификации не требуют физической реорганизации БД.
3. Реляционные СУБД обеспечивают значительно более высокий уровень защиты данных и хорошие возможности разграничения прав доступа.

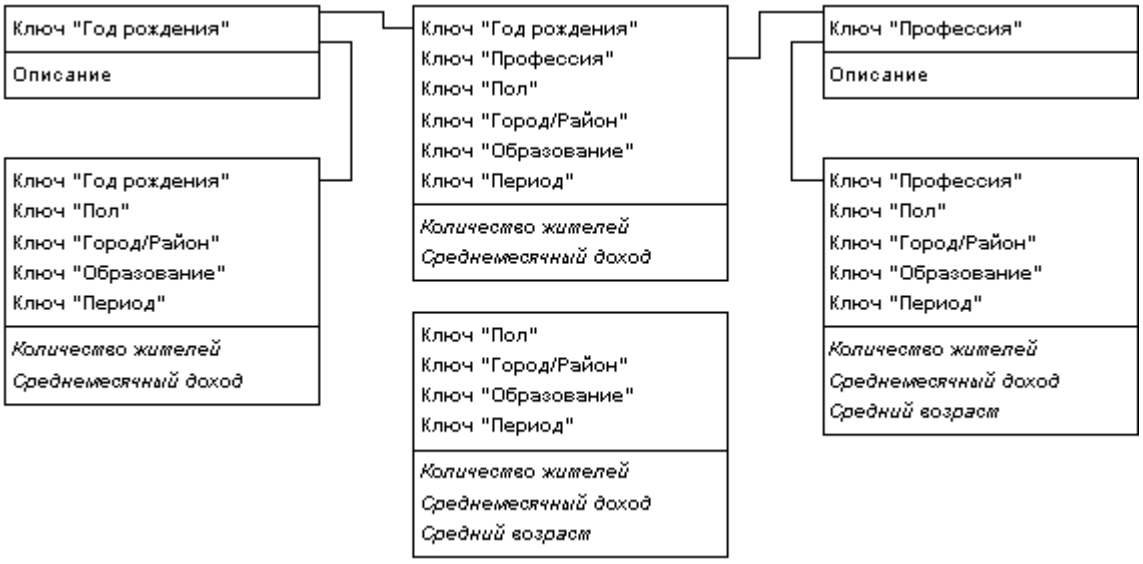
Главный недостаток ROLAP по сравнению с многомерными СУБД - меньшая производительность. Для обеспечения производительности, сравнимой с MOLAP, реляционные системы требуют тщательной проработки схемы базы данных и настройки индексов, то есть больших усилий со стороны администраторов БД. Только при использовании звездообразных схем производительность хорошо настроенных реляционных систем может быть приближена к производительности

систем на основе многомерных баз данных.

Описанию схемы звезды (star schema) и рекомендациям по ее применению полностью посвящены работы. Ее идея заключается в том, что имеются таблицы для каждого измерения, а все факты помещаются в одну таблицу, индексируемую множественным ключом, составленным из ключей отдельных измерений. Каждый луч схемы звезды задает, в терминологии Кодда, направление консолидации данных по соответствующему измерению.

В сложных задачах с многоуровневыми измерениями имеет смысл обратиться к расширениям схемы звезды - схеме созвездия (fact constellation schema) и схеме снежинки (snowflake schema). В этих случаях отдельные таблицы фактов создаются для возможных сочетаний уровней обобщения различных измерений. Это позволяет добиться лучшей производительности, но часто приводит к избыточности данных и к значительным усложнениям в структуре базы данных, в которой оказывается огромное количество таблиц фактов.

Увеличение числа таблиц фактов в базе данных может проистекать не только из множественности уровней различных измерений, но и из того обстоятельства, что в общем случае факты имеют разные множества измерений. При абстрагировании от отдельных измерений пользователь должен получать проекцию максимально полного гиперкуба, причем далеко не всегда значения показателей в ней должны являться результатом элементарного суммирования. Таким образом, при большом числе независимых измерений необходимо поддерживать множество таблиц фактов, соответствующих каждому возможному сочетанию выбранных в запросе измерений, что также приводит к неэкономному использованию внешней памяти, увеличению времени загрузки данных в БД схемы звезды из внешних источников и сложностям администрирования. Частично решают эту проблему расширения языка SQL (операторы "GROUP BY CUBE", "GROUP BY ROLLUP" и "GROUP BY GROUPING SETS"); кроме того, авторы статей предлагают механизм поиска компромисса между избыточностью и быстродействием, рекомендуя создавать таблицы фактов не для всех возможных сочетаний измерений, а только для тех, значения ячеек которых не могут быть получены с помощью последующей агрегации более полных таблиц фактов.



В любом случае, если многомерная модель реализуется в виде реляционной базы данных, следует создавать длинные и "узкие" таблицы фактов и сравнительно небольшие и "широкие" таблицы измерений. Таблицы фактов содержат численные значения ячеек гиперкуба, а остальные таблицы определяют содержащий их многомерный базис измерений. Часть информации можно получать с помощью динамической агрегации данных, распределенных по незвездообразным нормализованным структурам, хотя при этом следует помнить, что включающие агрегацию запросы при высоконормализованной структуре БД могут выполняться

довольно медленно.

Ориентация на представление многомерной информации с помощью звездообразных реляционных моделей позволяет избавиться от проблемы оптимизации хранения разреженных матриц, остро стоящей перед многомерными СУБД (где проблема разреженности решается специальным выбором схемы). Хотя для хранения каждой ячейки используется целая запись, которая помимо самих значений включает вторичные ключи - ссылки на таблицы измерений, несуществующие значения просто не включаются в таблицу фактов.

Интеллектуальный анализ данных

ИАД (Data Mining) - это процесс поддержки принятия решений, основанный на поиске в данных скрытых закономерностей (шаблонов информации). При этом накопленные сведения автоматически обобщаются до информации, которая может быть охарактеризована как знания.

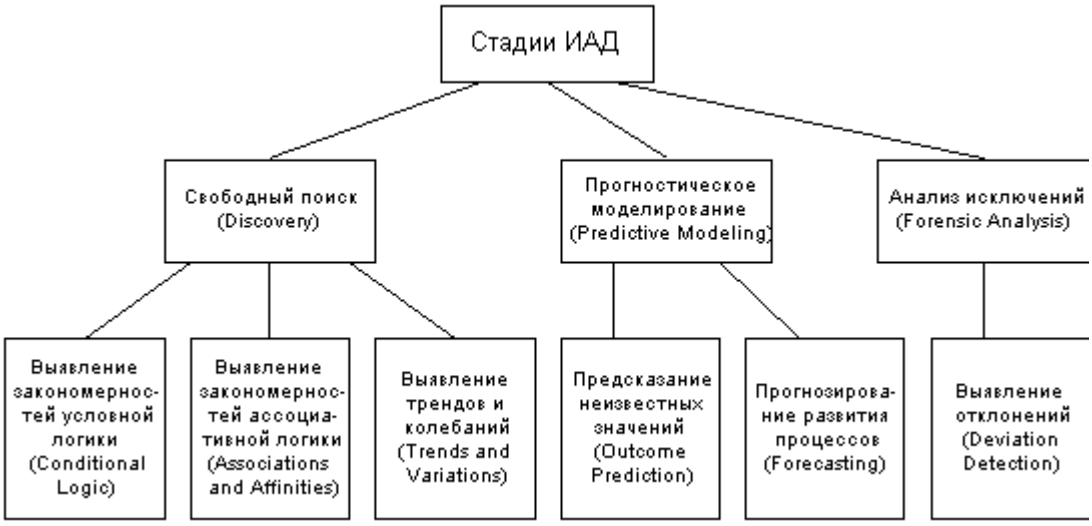
В общем случае процесс ИАД состоит из трёх стадий:

1) выявление закономерностей (свободный поиск);

2) использование выявленных закономерностей для предсказания неизвестных значений (прогностическое моделирование);

3) анализ исключений, предназначенный для выявления и толкования аномалий в найденных закономерностях.

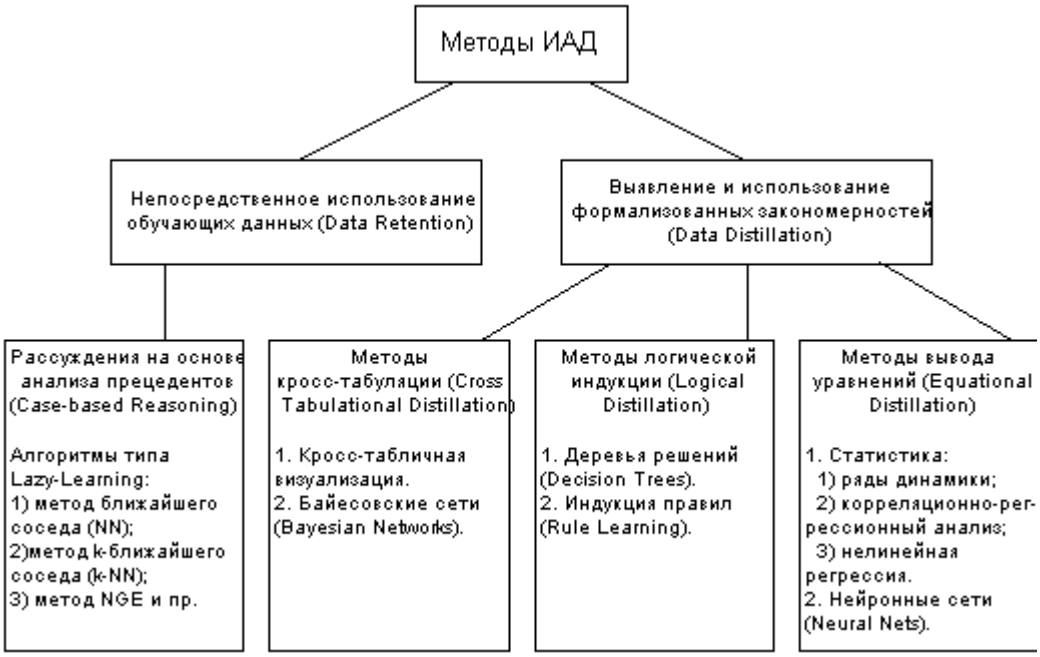
Иногда в явном виде выделяют промежуточную стадию проверки достоверности найденных закономерностей между их нахождением и использованием (стадия валидации).



Все методы ИАД подразделяются на две большие группы по принципу работы с исходными обучающими данными.

1. В первом случае исходные данные могут храниться в явном детализированном виде и непосредственно использоваться для прогностического моделирования и/или анализа исключений; это так называемые методы рассуждений на основе анализа прецедентов. Главной проблемой этой группы методов является затрудненность их использования на больших объемах данных, хотя именно при анализе больших хранилищ данных методы ИАД приносят наибольшую пользу.
2. Во втором случае информация вначале извлекается из первичных данных и преобразуется в некоторые формальные конструкции (их вид зависит от конкретного метода). Согласно предыдущей классификации, этот этап выполняется на стадии свободного поиска, которая у методов первой группы в принципе отсутствует. Таким образом, для прогностического моделирования и анализа исключений используются результаты этой стадии, которые гораздо более компактны, чем сами массивы исходных данных. При этом полученные конструкции могут быть либо "прозрачными" (интерпретируемыми), либо "черными ящиками" (нетрактуемыми).

Две эти группы и примеры входящих в них методов представлены на рисунке.



Интеграция OLAP и ИАД

Оперативная аналитическая обработка и интеллектуальный анализ данных - две составные части процесса поддержки принятия решений. Но сегодня большинство систем OLAP заостряет внимание только на обеспечении доступа к многомерным данным, а большинство средств ИАД, работающих в сфере закономерностей, имеют дело с одномерными перспективами данных. Эти два вида анализа должны быть тесно объединены, то есть системы OLAP должны фокусироваться не только на доступе, но и на поиске закономерностей. Как заметил N. Raden, "многие компании создали ... прекрасные хранилища данных, идеально разложив по полочкам горы неиспользуемой информации, которая сама по себе не обеспечивает ни быстрой, ни достаточно грамотной реакции на рыночные события".

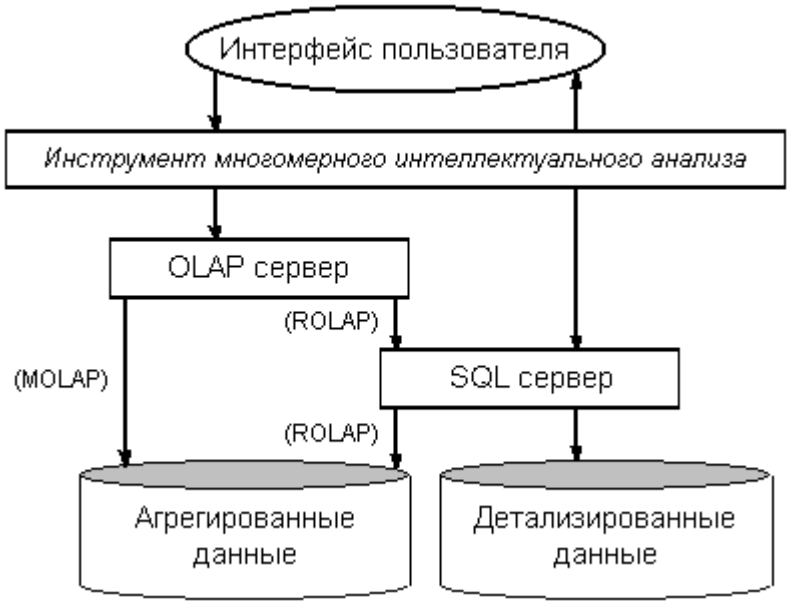
K. Parsaye вводит составной термин "OLAP Data Mining" (многомерный интеллектуальный анализ) для обозначения такого объединения. J. Han предлагает еще более простое название - "OLAP Mining", и предлагает несколько вариантов интеграции двух технологий.

1. "Cubing then mining". Возможность выполнения интеллектуального анализа должна обеспечиваться над любым результатом запроса к многомерному концептуальному представлению, то есть над любым фрагментом любой проекции гиперкуба показателей.
2. "Mining then cubing". Подобно данным, извлечённым из хранилища, результаты интеллектуального анализа должны представляться в гиперкубической форме

для последующего многомерного анализа.

3. "Cubing while mining". Этот гибкий способ интеграции позволяет автоматически активизировать однотипные механизмы интеллектуальной обработки над результатом каждого шага многомерного анализа (перехода между уровнями обобщения, извлечения нового фрагмента гиперкуба и т. д.).

К сожалению, очень немногие производители предоставляют сегодня достаточно мощные средства интеллектуального анализа многомерных данных в рамках систем OLAP. Проблема также заключается в том, что некоторые методы ИАД (байесовские сети, метод k-ближайшего соседа) неприменимы для задач многомерного интеллектуального анализа, так как основаны на определении сходства детализированных примеров и не способны работать с агрегированными данными [20].



Критерии оценки существующих продуктов

Как и в любой другой области, в сфере OLAP не может существовать однозначных рекомендаций по выбору инструментальных средств. Можно только заострить внимание на ряде ключевых моментов и сопоставить предлагаемые возможности программного обеспечения с потребностями организации.

1. Удобство и богатство возможностей средств администрирования. Работа администратора является самой важной и самой сложной частью эксплуатации OLAP-системы. Поэтому следует обращать внимание на удобство интерфейса администрирования, а более того - на спектр его функциональных

возможностей. Как формируются новые измерения? Как модифицируется существующая модель? Требуется ли создание базы данных жестко заданной структуры, или можно анализировать данные, собранные в ранее созданных базах (в случае ROLAP)? На все эти вопросы необходимо получить ясный и четкий ответ.

2. Гибкость настройки и наглядность форм демонстрации результатов. Интуитивность представления информации - главная изюминка OLAP. Насколько качественно и удобно формируются отчёты? Наглядны ли графические возможности, существует ли связь с ГИС-технологиями? Налажены ли механизмы экспорта результатов в стандартные форматы?
3. Спектр методов постобработки данных, доступность средств интеллектуального анализа. Богаты ли аналитические возможности инструмента? Есть ли в нём элементы Data Mining, и если есть, какие преимущества они могут обеспечить при использовании?
4. Возможность обработки больших хранилищ данных с приемлемой производительностью. Если необходим планомерный непрерывный анализ большого хранилища данных организации, требуется выяснить объективные ограничения продукта с точки зрения предельных размеров исходных баз данных.
5. Возможность увязки OLAP-инструментария со всеми СУБД, используемыми в организации. Как показывает практика, интеграция разнородных продуктов в устойчиво работающую систему - один из наиболее важных вопросов, и его решение в ряде случаев может быть связано с большими проблемами. Необходимо разобраться, насколько просто и надёжно можно интегрировать средства OLAP с существующими в организации СУБД.

Источники

1. https://docs.microsoft.com/ru-ru/azure/architecture/data-guide/relational-data/online-analytical-processing
2. https://infopedia.su/17xedca.html
3. http://www.e-biblio.ru/book/bib/01_informatika/inform_analit_systemy/posob/332.2.5.html
4. https://studopedia.net/1_50319_trebovaniya-k-sredstvam-operativnoy-analiticheskoy-obrabotki-dannih.html
5. Петров В.Ю. Информационные технологии в менеджменте. Учебное пособие - Санкт-Петербург: СПб: Университет ИТМО, 2015