Business Intelligence (BI)

«Вряд ли найдётся человек, которому ни разу не приходило в голову проследить забавы ради шаг за шагом все, что привело его к известному выводу. Это — преувлекательное подчас занятие, и кто впервые к нему обратится, будет поражён, какое неизмеримое на первый взгляд расстояние отделяет исходный пункт от конечного вывода и как мало они друг другу соответствуют». Эдгар Алан По, «Убийство на улице Морг», 1841 г.

Создавая первый в мире детективный рассказ, Эдгар Алан По вряд ли думал о том, что через 100 лет представители бизнеса настолько озаботятся вопросом, как обеспечить обоснованность выводов и решений, что появится отдельное направление в сфере ИТ – Business Intelligence. Термин BI появился впервые в 1958 году в статье Ханса Питера Луна и трактовался как возможность понимания связей между представленными фактами. Сейчас, когда объемы данных, которые приходится анализировать для принятия решений, существенно выросли, и их обработка стала невозможной без специальных технических средств (как аппаратного, так и программного обеспечения), этот термин стал в первую очередь использоваться в контексте ИТ-систем:

Business Intelligence (BI) – широкий класс информационных систем предприятия, направленных на поддержку принятия решений, включающий в себя следующие функциональные блоки:

- ***Загрузка и преобразование данных (ETL, Extract-Transform-Load)***
- ***Корпоративное хранилище данных (DWH, Data Warehouse)***
- ***Продвинутая аналитика (Advanced Analytics)**, включая предиктивную аналитику (Predictive Analytics), текстовую аналитику (Text Mining) и прочий инструментарий инженеров по данным (Data Scientists)*
- ***Автоматизация отчетности (Reporting) и самостоятельное исследование данных (Self-service BI)**, включая инструменты OLAP (On-Line Analytical Processing) и визуальные конструкторы интерактивных дашборддашбордов и строгой отчетности*
- ***Ручной ввод данных** – встроенные в BI платформу инструменты автоматизации ручного ввода данных, в том числе формы ввода (Data Entry), обратная запись в системы-источники (Write Back) и решения на основе форм, адаптированные для задач консолидации отчетности, бюджетирования и планирования (Integrated Planning).*

Рисунок 1. Принцип работы BI-системы

Также класс BI зачастую пересекается с решениями класса CPM (Corporate Performance Management) с тем различием, что если BI – это как правило инструментарий, не привязанный изначально к какой-либо предметной области, то CPM уже обычно содержит методологические наработки, позволяющие ускорить внедрение.

Часть 1. Зачем нужна бизнес-аналитика?

Все просто. Мир в целом и бизнес как его составная часть давно уже стали data-центричными. Именно накопленные за время работы данные являются самым важным активом любого бизнеса независимо от его размера и отраслевой принадлежности. И с каждым годом данных становится все больше: аналитики IDC оценивали глобальный объем хранимых в 2017 году данных в 33 зетабайта, а на 2025 год прогнозировали его увеличение до 175 зетабайт.

И хотя весомую долю в этом объеме составляет развлекательный контент, с каждым технологическим витком все тяжелеющий, бизнес-данные не отстают. По данным исследования Global Data Protection Index Study компании Dell Technologies, в 2018 году организации в среднем обрабатывали 9,7 ПБ данных, причем за два года этот показатель вырос более чем в 5 раз.

Однако данные недостаточно накопить: их необходимо анализировать, управлять ими, заставить работать – именно в этом самый значимый вызов для современных компаний, особенно если они видят себя частью цифровой экономики. С учетом все увеличивающегося темпа бизнеса у руководителей компаний остается все меньше времени на принятие решений, тогда как решения все так же должны быть взвешенными и обоснованными. Так что классические методы управления, анализа и оценки становятся неэффективными. BI-

платформы позволяют решить эту проблему, создавая основу для правильного управления бизнесом и повышения его эффективности.

Предпосылки внедрения BI

- Рост объемов поступающей информации
- Сокращение времени на принятие решений
- Сложность процессов управления
- Неэффективность традиционных методов управления, анализа и оценки

Excel или BI?

Естественным «предшественником» BI-систем практически в любой компании является Excel (или другие табличные процессоры), на которых решаются задачи сбора, анализа и визуализации данных.

Периодически среди экспертов возникает спор о том, что лучше и удобнее для аналитики – Excel или BI. Копий сломано много, однако однозначного победителя нет, да и, вероятнее всего, не будет. Так что целесообразно признать: Excel и BI дополняют друг друга, а вовсе не конкурируют.

Excel и BI, хоть и совпадают частично по функционалу, но решают разные задачи – и в этом их принципиально отличие. Опыт показывает: выбор в пользу табличного процессора или системы бизнес-анализа определяется не масштабом организации или проекта, а именно решаемой задачей. Если в конкретной ситуации крайне важна именно скорость внесения изменений, а методическая устойчивость общего процесса вторична, то Excel окажется более удобен. Именно так происходит, например, когда требуется загрузить новые данные, связать их с уже существующими, изменить методику подсчета лишь одного из показателей и удалить неактуальные данные. Это не значит, что BI-система сделать этого не сможет – вовсе нет, просто стоимость внесения изменений будет излишне высокой.

С другой стороны, при агрегировании данных из всех подразделений для формирования единой отчетности практически невозможно корректно свести в одну таблицу большое количество Excel-файлов. Велика вероятность появления ошибок, источника которых будет не найти. BI-система для выполнения такой задачи лучше подойдет.

Именно по этой причине в любой системе бизнес-аналитики обязательно имеются модули интеграции с табличными процессорами. На рисунке Рисунок 2 показано, как обеспечивается интеграция средств бизнес-аналитики и Excel в зрелых аналитических платформах.

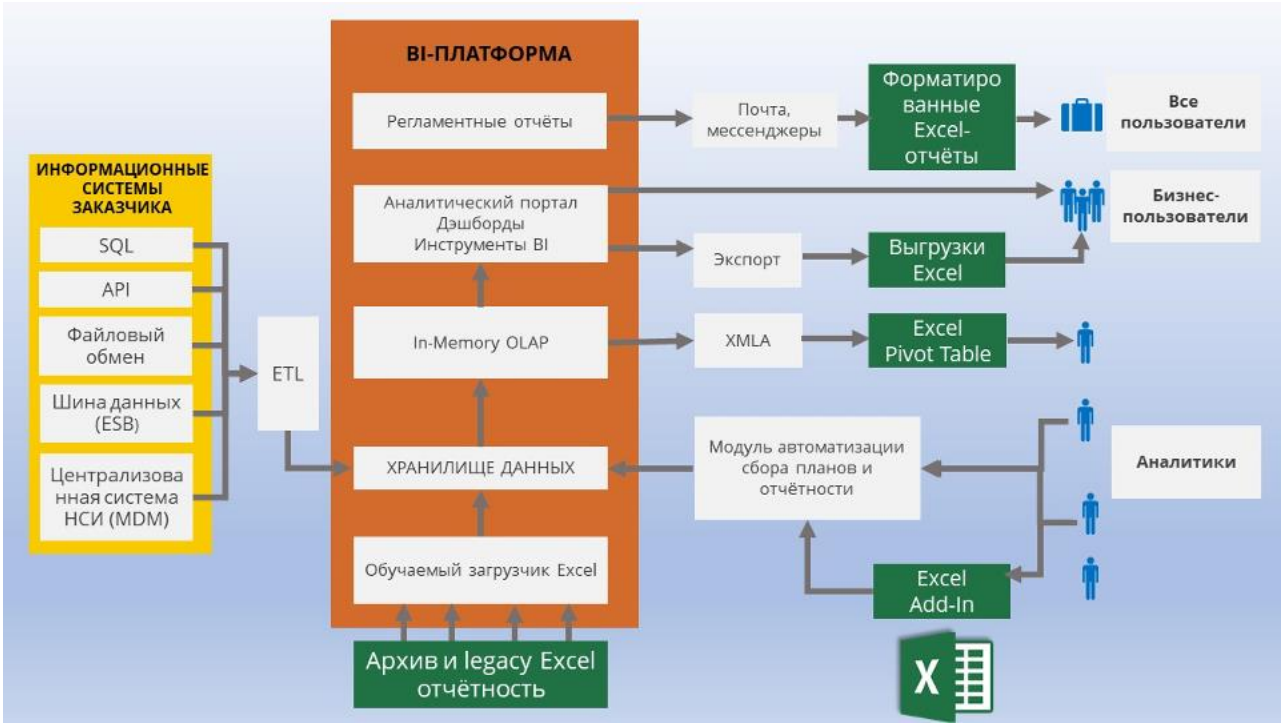

Рисунок 2. Взаимодействие средств бизнес-аналитики и табличных процессоров

КАЧЕСТВЕННЫЕ ВЫГОДЫ ОТ ВНЕДРЕНИЯ BI

1. В компании или организации появляется ЕДИНАЯ «ТОЧКА ПРАВДЫ». Типичная картина на многих совещаниях – расхождение информации из разных департаментов или филиалов. И понять, где ошибка, далеко не всегда просто. BI дает возможность создать единый для всех источник информации, доступный в любой момент времени и позволяющий делать выводы, строить планы для любого подразделения компании без разногласий и споров.

2. Скорость получения информации – одна из важнейших составляющих принятия управленческих решений. BI обеспечивает МГНОВЕННЫЙ ДОСТУП К ИНФОРМАЦИИ в режиме реального времени, поэтому топ-менеджмент сразу получает ответы на важные бизнес-вопросы и управленческие решения принимаются не на основе интуиции, а на основе фактов и цифр.

3. Одной из важных характеристик современного мира является высокий темп жизни и бизнес-процессов, поэтому руководителям важно, чтобы ДОСТУПНОСТЬ ИНСТРУМЕНТОВ ПОДДЕРЖКИ ПРИНЯТИЯ РЕШЕНИЙ была обеспечена не только в офисе. BI-платформы позволяют наладить вывод информации на мобильные устройства руководителя, а кроме того использовать предсказательные сценарии и обучаемые математические модели, т.е. могут подстраиваться под потребности конкретного руководителя и предприятия и прогнозировать будущее.

4. Чем больше менеджеров и рядовых сотрудников компании получают доступ к «единой правде», тем более верные решения они принимают и тем более эффективными будут все процессы. Добиться этого, если для каждого аналитического запроса потребуется привлекать ИТ-специалиста, невозможно. BI-платформа – это УНИВЕРСАЛЬНОЕ РЕШЕНИЕ, т.е. сотрудники могут создавать отчеты без привлечения ИТ-специалистов с учетом разграничения доступа к информации по ролевому признаку.

КОЛИЧЕСТВЕННЫЕ ВЫГОДЫ ОТ ВНЕДРЕНИЯ

В зависимости от исходного состояния процессов, отраслевой специфики и многих других параметров количественные выгоды от внедрения BI будут различаться. Однако еще в 2018 году компания IDC провела исследование Time to Value and ROI From BI, в котором выяснила окупаемость и рентабельность инвестиций в одну из известных BI-платформ. Как показал опрос, после внедрения время, затрачиваемое на поиск и доступ к информации, сокращается на 51%, а на анализ информации тратится на 48% времени меньше. При этом среднее время реализации и окупаемости проекта по внедрению информационно-аналитической системы составляет 12 недель, а средний срок окупаемости – 28 недель при среднем ROI в 186%.

Дополнительные финансовые показатели:

- 31% – снижение расходов, связанных с поддержкой отчетности;
- 30% – снижение расходов на разработку новой отчетности;
- 20% – снижение операционных расходов;
- 16% – увеличение выручки.

Важно понимать, что технологии не стоят на месте: с тех пор BI-платформы получили новый функционал, повысили производительность и т.д. Это значит, что если бы подобное исследование проводилось сейчас, то результаты могли бы быть иными – возможно, более впечатляющими.

Часть 2. Принцип работы BI-системы

BI-платформы могут отличаться друг от друга по большому числу параметров – производительности, функциональности, наличию облачных версий и многим другим. Но если посмотреть на схему их работы, то она у всех принципиально схожа. Есть источники информации, с которыми интегрируется BI-платформа, которая предоставляет различный инструментарий, а на выходе пользователи получают регламентированные и прогнозные отчеты, дашборды и графики на ПК, мобильных устройствах, видеостенах и т.д. Для обеспечения этих функций предусмотрены технические подсистемы, выполняющие различные функции обработки данных.

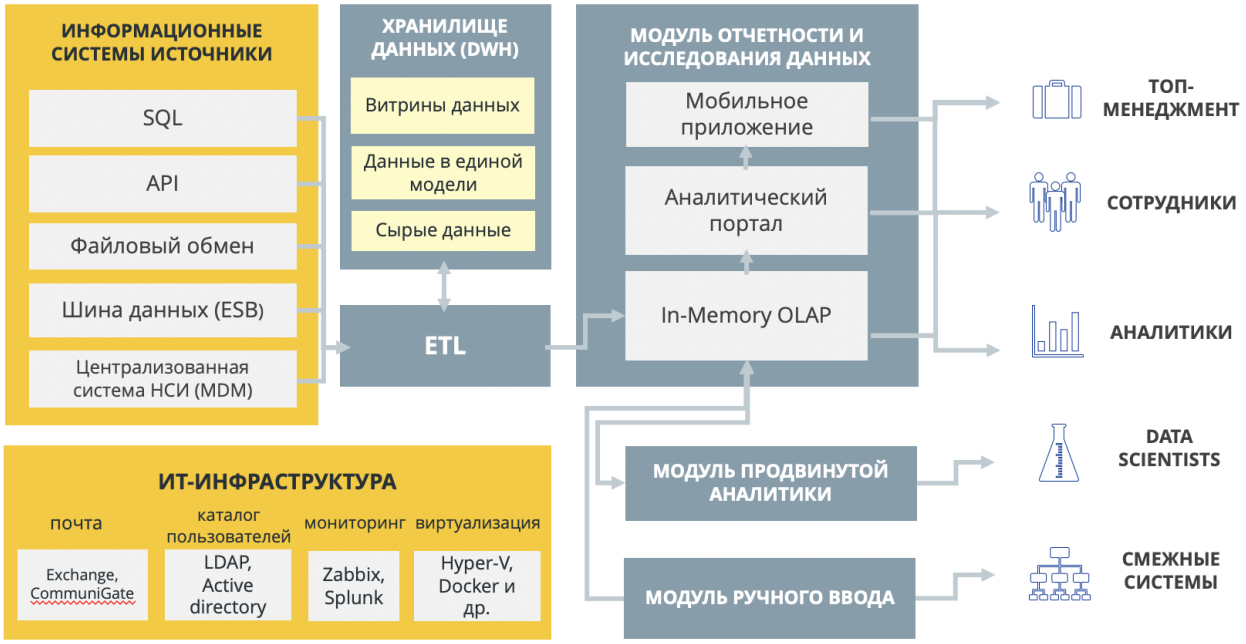

Рисунок 3. Типовая структура BI-решения

КОМПОНЕНТЫ ТИПОВОГО BI-РЕШЕНИЯ

1. Загрузка и преобразование данных (ETL)

Данные, необходимые для работы BI-платформы, извлекаются из различных ИТ-систем с помощью ETL-инструментов и поступают в специальные хранилища данных. В отличие от баз данных в этих хранилищах информация структурируется специальным образом для более эффективного анализа и обработки запросов. Преобразования данных на уровне хранилища данных также, как правило, оркестрируются ETL инструментом. Конечным результатом преобразования данных на уровне DWH являются так называемые «витрины данных» (Data Marts) – очищенные наборы данных, к которым пользователи получают прямой доступ с помощью OLAP инструментов.

Все ETL-решения можно распределить на два класса, отличающихся по подходу к работе.

Визуальный ETL (он же visual ETL) позволяет создавать схемы из готовых блоков. Script ETL – в противовес визуальному – предлагает использовать определенный скрипт-язык, оптимизированный под обработку данных.

Преимуществом visual ETL, безусловно, является простота и наглядность, но есть и минусы, которые весьма существенны. Как только задачи усложняются, схемы и блоков становятся слишком сложными и для понимания, и для составления, поэтому возникает потребность писать код.

Скриптовые ETL позволяют избежать таких сложностей, однако их использование требует определенных навыков и знаний. К тому же script ETL дает больше возможностей для развития. Определенным преимуществом обладают ETL на Python, во-первых, потому что это простая среда программирования, доступная даже детям, а во-вторых, для Python есть готовая экосистема готовых технологий и библиотек – от готовых коннекторов до библиотек для data-scientist.

В поисках золотой середины сейчас наблюдается следующий тренд: разработчики ETL делают ставку на смешанные, гибридные варианты инструментов, хотя в ДНК этих решений всегда заложен лишь один из вариантов.

КАКИМ ДОЛЖЕН БЫТЬ ETL-ИНСТРУМЕНТ?

ETL- инструменты, как уже говорилось выше, необходимы для извлечения данных (E – extract), преобразования, приведения их к необходимому формату, обработки в соответствии с заданными правилами, комбинировании с другими данными (T – transform), а также для загрузки данных (L – load), записи данных в хранилище или в другую базу данных. Процессы это тонкие и очень значимые для работоспособности и эффективности BI-платформы, поэтому необходим правильный ETL-инструмент.

Практика показывает, что аналитики, да и в целом те, кто работает с данными, 20% своего времени уделяют анализу и визуализации, а 80 % тратят на преобразование, очистку и выгрузку данных. Понятно, что им необходим инструмент, который эту проблему решает. Проблема в том, что ETL на рынке много, но выбрать надо оптимальный.

Можно выделить *5 основных критериев для анализа ETL-инструментов*.

1. ETL должен быть **простым** в изучении. Во-первых, благодаря этому им сможет пользоваться даже неопытный сотрудник, что в конечном счете расширяет список тех, кому доступны данные. Во-вторых, сейчас прослеживается глобальный тренд к упрощению ИТ-инструментов: человек не должен тратить полжизни на изучение решения – при идеальных условиях нужно просто начать с ним работать.

2. В ETL должно быть **максимальное количество готовых коннекторов**. Никто не хочет писать решение с нуля при появлении нового источника данных. На рынке представлен ряд ИТ-систем, которыми так или иначе пользуется весомая часть компаний и организаций. ETL должен предоставлять коннекторы для интеграции BI-платформы с максимальным числом самых распространенных из подобный решений.

3. ETL должен быть **универсальным**, т.е. работать с большим количеством BI-систем. Как и другие специалисты, аналитики переходят из одной компании в другую, а следовательно,

могут работать с разными BI-системами. ETL должен поддерживать если не все, то практически все крупные платформы, представленные на рынке.

4. ETL **не должен ограничивать развитие**. На данный момент одной из важнейших проблем многих ETL-инструментов является ограниченность: простые задачи ETL-инструмент решает, но по мере усложнения задач возможностей не хватает. Аналитикам необходимы ETL-инструменты, которые способны развиваться и расти вместе с пользователем и его компетенциями и потребностями.

5. ETL-инструмент должен быть либо **недорогим, либо бесплатным**, причем не в виде trial-версии, а навсегда.

Что сейчас есть на рынке?

Все ETL-решения можно оценить по двум параметрам: стоимость и функциональность. Если расположить самые распространенные инструменты на схеме, где по горизонтали движение от простого к сложному, а по вертикали – от бесплатных или open-source решений до инструментов с высокой стоимостью, то получается следующая картина (авторская методика):

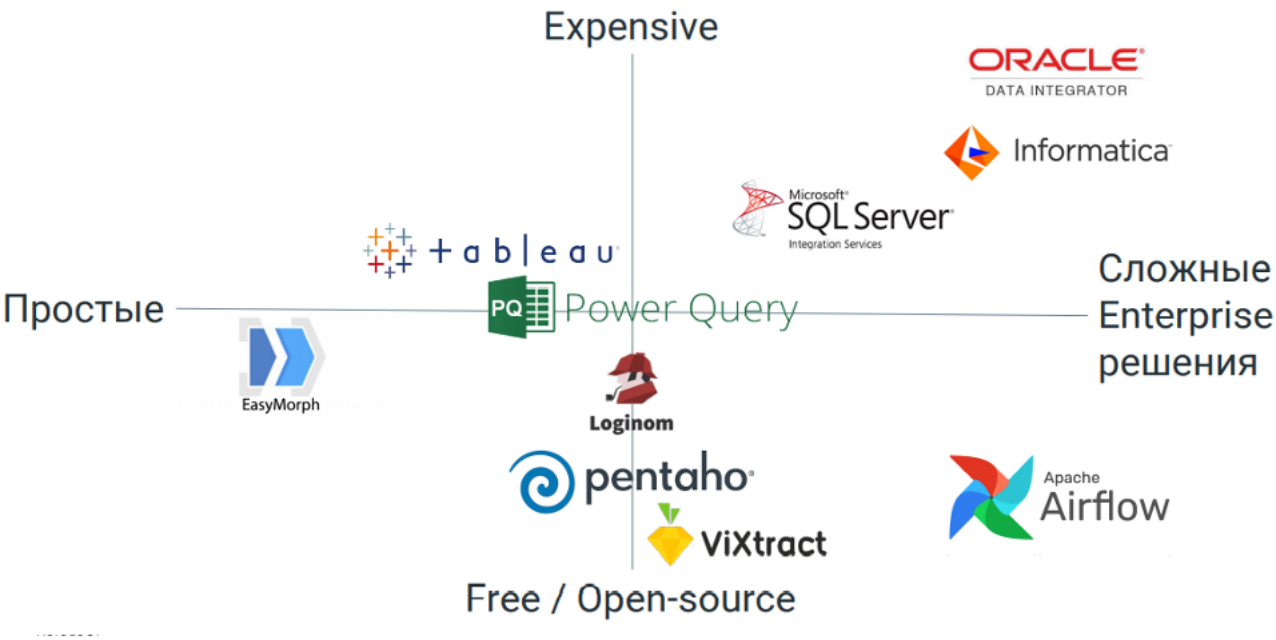

Рисунок 4. Распределение ETL-инструментов на российском рынке по стоимости и функциональности (авторская методика)

Решения, которые тяготеют к правому верхнему углу, это сложные, тяжелые, широко функциональные и при этом дорогостоящие решения. SQL Server – чуть более демократичное решение. В правом нижнем углу – Apache Airflow – open-source-решение, но также отличающееся сложностью и ориентированное на enterprise-сегмент.

В центре схемы расположились решения, средние по соотношению цена-функциональность. Однако если все решения проанализировать по пяти предложенным выше критериям, окажется, что ни одно из них им полностью не удовлетворяет. Те решения, которые тяготеют к краям схемы, выпадают по одному из показателей – либо слишком дороги, либо недостаточно функциональны. Все инструменты, которые расположены в центре схемы, не являются универсальными. Но этот рынок развивается, постоянно пополняясь новыми решениями, которые вполне могут добиться соответствия всем пяти критериям.

2. Корпоративное хранилище данных (Data Warehouse, DWH)

В контексте систем бизнес-аналитики под хранилищем данных подразумевается в первую очередь не система хранения данных, а именно программный и логический уровни организации информации. Именно правильная гармонизация данных из различных источников – это наиболее сложная и важная задача создания DWH, качественное решение позволит своевременно обеспечить всех пользователей качественной непротиворечивой информацией. С точки зрения ПО, большинство DWH работают на стандартных реляционных СУБД, таких как Oracle, MS SQL Server и PostgreSQL. В ряде случаев используются инструменты для автоматической генерации схемы данных в СУБД в соответствии с популярными моделями данных (см. ниже), но это не является обязательным требованием.

Модель Инмона (3NF Schema) – классическая модель организации данных в DWH, тяготеющая к максимальной нормализации данных, аналогично тому, как организованы данные в учетных системах. Плюсы – многие вопросы консистентности данных решаются на уровне схемы данных, во многих случаях можно перености схему из систем-источников. Минусы – высокая трудоемкость гармонизации данных, высокая стоимость внесения изменений в схему.

Модель Кимбалла (Star Schema) или многомерная модель – самая часто используемая модель, балансирующая между полностью денормализованными таблицами (датасетами) и нормализованной схемой. Сущности в Star Schema делятся на таблицы фактов (Fact Tables) и измерения (Dimensions). Это разделение всегда зависит от конкретных решаемых задач аналитики. Плюсы – модель отлично ложится на возможности OLAP инструментов и позволяет напрямую работать с данными из систем аналитики, есть отработанная годами теория моделирования. Минусы – в сложных предметных областях может потребоваться построение большого количества таблиц фактов (и соответствующих им многомерных кубов) под конкретные задачи аналитики.

Data Vault и Anchor modeling – относительно новые модели организации данных, призванные решить проблемы отслеживания историчности данных и изменения бизнес-сущностей со временем за счет разделения метаданных на структурные и описательные части. На сегодняшний день эти модели используются достаточно редко из-за более высокого порога входа для аналитиков и ETL-инженеров, а также из-за плохой совместимости с популярными инструментами анализа и визуализации данных.

Следует отметить, что в последнее время с ростом объемов данных в DWH начали активно применяться распределенные СУБД, такие как Greenplum, Vertica и ClickHouse. Для таких хранилищ данных применяется термин Data Lake («озеро данных»). Ключевыми отличиями концепции Data Lake от традиционных DWH являются:

- возможность хранения всех исходных (исторических, нетрансформированных) данных;
- поддержка Real-time загрузки, аналитики и выгрузки данных;
- хранение и обработка неструктурированных (тексты, изображения, аудио, видео) данных в едином контуре со структурированными.

С технической точки зрения основная особенность Data Lake – это применение распределенных вычислений (с использованием кластерной или массивно-параллельной архитектуры). При этом важно отметить, что несмотря на распространенное заблуждение, само по себе использование технических средств класса Data Lake не отменяет необходимости методической и аналитической проработки модели данных и логики трансформаций, поэтому Data Lake – это не «простая и удобная замена устаревшим DWH», а высокопроизводительная альтернатива классическим DWH, применяемая при повышенных требованиях к объему и частоте обновления данных, когда хранилище на основе реляционной СУБД не справляется.

3. Продвинутая аналитика (Advanced Analytics)

Если под аналитикой мы подразумеваем все ответы на вопросы "Что случилось?" (что у нас с прибылью, с расходами, что мы должны банку, что мы ожидаем от новой стратегии), то продвинутая аналитика – это ответы на вопросы "Почему так произошло?" и «Что мы можем сделать?» (почему прибыль такая маленькая или резко повысилась, почему мы не успели завершить задачу, почему запустился определённый процесс?) Хотя ответы на эти вопросы будут разными, и во втором случае потребуют более глубокого подхода, обе задачи являются сложными: аналитика и продвинутая аналитика требуют серьёзных инструментов – алгоритмических и инструментальных.

Споры о том, что важнее в прикладном плане – классическая автоматизация отчётности или продвинутая аналитика, не утихнут ещё долго, но роль последней год от года только растёт, и в современной BI-платформе инструменты продвинутой аналитики, основанные на науке о данных (Data Science) обязательно должны присутствовать. Конечно, продвинутая аналитика – это очень обширная область знаний сама по себе, и может применяться в компании вне контекста BI-систем. Однако, как правило, в BI-системах также встроены инструменты продвинутой аналитики (как некоторое подмножество всей палитры инструментов Advanced Analytics), которые позволяют BI-аналитикам иметь их под рукой и быстро применять к решению своих задач. При этом в одной организации могут эффективно сосуществовать подразделения BI и Advanced Analytics (часто называемое Data Science отделом), и это накладывает определенные требования на унификацию инструментария продвинутой аналитики.

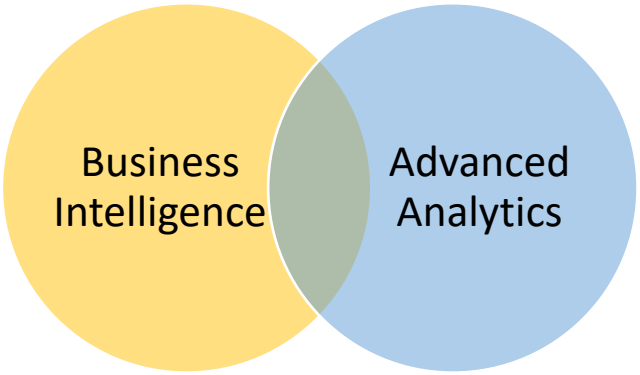

Рисунок 5. Взаимосвязь между BI и Data Science

Продвинутая аналитика во многом основана на методах математической статистики, которые с развитием вычислительных мощностей и ростом количества доступных прикладных алгоритмических реализаций, сделали качественный скачок и создали новую область знаний – машинное обучение. Что означает «продвинутая аналитика средствами машинного обучения»? Автоматизировать аналитические задачи можно разными способами: например, средствами дискретного программирования (когда мы в явном виде указываем компьютеру, что делать), полуавтоматически (взаимодействие человека и какой-то системы). Человек показывает машине много данных, правильный способ с ними работать, правильные решения задачи – а потом спрашивает её, как это сделать на новых примерах. Есть специальные подходы, которые без явных инструкций позволяют машине научиться решать такие задачи на новых примерах, отвечать на вопрос.

Этапы решения задачи продвинутой аналитики

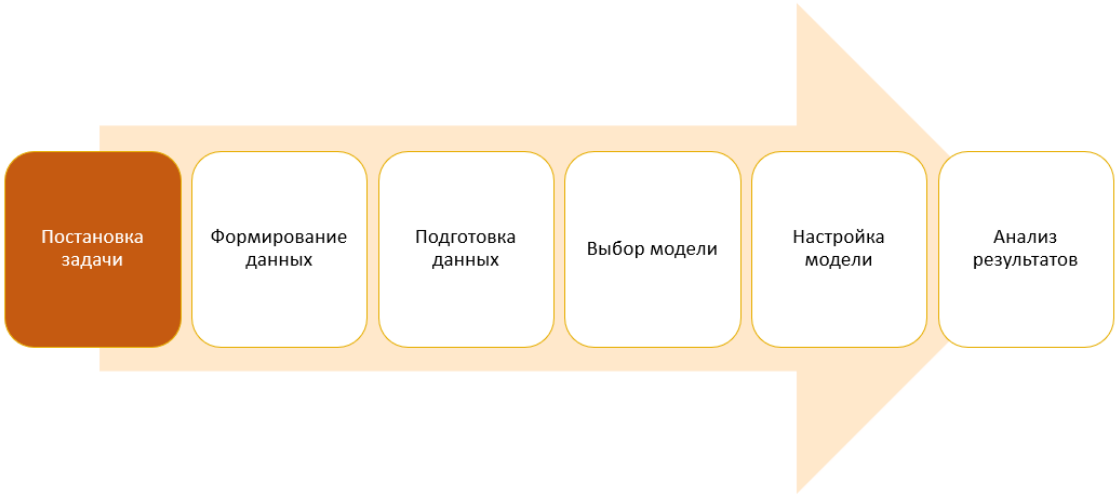

Рисунок 6. Этапы решения задачи продвинутой аналитики

Самый первый и главный этап – постановка задачи, затем происходит формирование данных, которые потом их нужно подготовить (в частности, очистить – это называется препроцессингом). Далее идёт выбор соответствующей математической модели, наиболее подходящей для решения поставленной задачи; и если предыдущим пунктам можно научиться на курсах за несколько дней, то вопрос выбора модели и её настройки – это не просто сложная область, а, по мнению многих, вопрос искусства: и именно этому люди учатся много лет. Важнейший пункт – анализ результатов, после которого последовательность, чаще всего, повторяется заново – с немного переформулированной задачей, скорректированными данными и другой моделью. Это происходит до тех пор, пока не будут получены удовлетворительные результаты. Разберём подробнее каждый этап:

1. **Постановка задачи.** Как учат на всех математических специальностях - это важнейший этап: без постановки не бывает решения. Обязательный пункт в постановке задачи: **когда мы будем считать задачу успешно решенной? В каких границах мы считаем результат приемлемым?**
   Важным этапом является определение, к какому типу относится задача, потому что это влияет на непосредственный выбор модели.

- Самый распространенный тип задач – задачи классификации, когда результат работы модели – один из нескольких известных вариантов. Давать ли кредит клиенту, который пришёл в банк?
- Следующий класс задач – задачи регрессии, когда мы заранее не знаем, сколько есть возможных вариантов, зачастую их бесконечное количество. Например, не просто "даём/не даём кредит", а "сколько даём"?
- Каков будет спрос завтра на наш продукт? Какова глубина грунтовых вод в данной местности? Обе этих задачи чаще всего решаются, когда есть примеры готовых решений. Это называется "задача с учителем".
- Второй важный класс задач – задача без учителя, когда нет данных вообще: мы ничего не знаем, банк только открылся. Полноценно такую задачу решить не получится, но есть, например, другие типы задач, которые успешно могут быть решены без правильных ответов – задачи кластеризации. Задача поиска аномалий – это типичная задача кластеризации: мы ищем области, где хорошо, и области, где плохо, то есть, резко отличающиеся друг от друга состояния.
- Ещё один класс задач: мы хотим понять, какие факторы влияют на наш целевой показатель – на то, что является определяющим фактором формирования, например, цены билета на самолёт.
- Задача ассоциации: например, что из одежды с чем сочетается или как расставлять товар на полках, чтобы он пользовался популярностью и т д.

2. **Формирование данных** – основной вопрос на данном этапе – это достаточность их количества и качества. Данные могут быть очень плохие, могут быть даже не оцифрованы (тогда есть задача их оцифровать), могут быть в разных источниках (тогда нужны решения для группировки данных из разных источников). К сожалению, очень мало по каким задачам известен конкретный статус, конкретная оценка, сколько данных нужно подавать: чаще всего это именно практическая область. Есть способы

автоматического дополнения данных – но, к сожалению, сильно расширить выборку ими не получается так, чтобы не потерять в точности.

3. **Подготовка данных** (pre-processing) состоит из 3 самых важных пунктов, их редко можно миновать. Русских терминологий для этих пунктов не создано:
    a. **data sanitizing – очистка данных**. Мы убираем все неполные данные (про кого-то из людей неизвестен возраст, когда мы пытаемся дать ему кредит). Можно их удалить, можно заполнить пропущенные значения средними – чтобы хотя бы использовать остальные данные.
    b. **feature selection** – **выделить признаки из данных**. Признаки – это параметры, по которым мы решаем нашу задачу. Чем меньше признаков, тем модели легче справиться. Есть такое понятие – проклятие размерности – необходимо избавляться от признаков по максимуму, особенно, если они связаны между собой.
    c. **Трансформация данных** – **feature transformation**. Любые категориальные данные, выражаемые не числом, а типом данных, нужно перевести в спец. формат, этот процесс называется **нормализацией**. Вместо 1 столбца заведём 3 и поставим 1 в том из них, который соответствует нашему типу. Это больше похоже на числа: с бинарными данными (0/1) компьютер работает хорошо. Это отдельный пункт – **работа с категориальными данными** – всегда учитывать и смотреть, не отдали ли вы системе такие данные.
4. **Выбор модели**. Для каждого класса задач продвинутой аналитики существует множество математических моделей и их прикладных реализаций (например, в виде библиотек на Python). Для успешного решения задачи аналитик должен подобрать одну или несколько наиболее подходящих моделей и собрать на их основе цепочку преобразования данных, позволяющие получить ответ на поставленный вопрос – например, рассчитать прогноз продаж по определенной номенклатурной группе. При выборе модели аналитик должен учитывать допущения, на основе которых создавалась модель (например, простая линейная регрессия требует, чтобы влияющие факторы были независимы друг от друга), проверять соответствие результатов работы модели фактическим данным, а также опираться на отраслевые лучшие практики и свой личный опыт решения аналогичных задач.
5. **Настройка модели.** После выбора и тестирования модели необходимо выполнить её реализацию и тонкую настройку. Одна из проблем, которая может возникнуть на этом этапе – это необходимость портирования кода модели с одного инструмента на другой при переходе в промышленную эксплуатацию. Одно из преимуществ использования Python – это универсальность, позволяющая избежать этой проблемы. В современных BI-системах, как правило, предусмотрены удобные инструменты, которые позволяют с минимальными трудозатратами создать на основе модели аналитическое приложение с простым пользовательским интерфейсом и запустить её в промышленную эксплуатацию.
6. **Анализ результатов.** На каждом этапе эксплуатации модели должна проводиться оценка соответствия модели критериям работоспособности, разработанными на этапе постановки задачи. Эта оценка может выполняться как вручную, так и в автоматизированном режиме. Последний вариант особенно полезен на практике именно в промышленной эксплуатации, поскольку исходные данные будут постоянно

изменяться и можно легко пропустить момент, когда исходные допущения модели перестанут выполняться, и модель начнет выдавать некорректные данные. Избежать этого позволяют автоматические проверки.

Инструментарий продвинутой аналитики

Существует большое количество различных инструментов продвинутой аналитики. Исторически это были, в основном, проприетарные инструменты – как в составе BI-платформ, так и самостоятельные продукты. Однако, сегодня значительная часть специалистов по данным переходя на использование инструментария с открытым кодом, т.к. открытый код позволяет множеству людей совершенствовать продукт, по нему есть много справочной информации, техподдержки, много вопросов и ответов. Один из вариантов – язык Python, для которого не только создано множество готовых аналитических решения, но и порог вхождения у него достаточно низкий.

При этом в любой современной BI платформе должен быть предусмотрен интерфейс интеграции с открытым инструментарием, так как модели, разработанные Data Scientist'ами, необходимо публиковать на аналитическом платформе, чтобы они имели простой и удобный интерфейс и были доступны бизнес-пользователям. Для различных платформ способы интеграции могут отличаться – от базовой возможности вызова функций R или Python в конструкторе дашбордов или ETL-инструменте, до полноценной интеграции с JupyterHub (популярной открытой среды разработки продвинутой аналитики на Python) в которой доступна отладка модели в режиме реального времени, API для работы с DWH и виджетами дашборда, автоматическая генерация кода и многое другое.

4. Автоматизация отчетности (Reporting) и самостоятельное исследование данных (OLAP, Self-service BI)

Непосредственно для всестороннего анализа данных в BI-решениях используются OLAP-инструменты (On-Line Analytical Processing, оперативная аналитическая обработка данных). Это ключевой элемент архитектуры современных BI-систем, так именно благодаря OLAP анализ данных становится доступен бизнес-пользователям, а не только разработчикам и специалистам по базам данных.

Концептуально OLAP инструменты построены на так называемой «многомерной» модели данных. **Многомерная модель данных** - это альтернатива обычной табличной или "реляционной" модели, которая лучше подходит для аналитической работы с данными. Сводные таблицы ("pivot tables") Excel построены именно на многомерной логике. Идея модели заключается в разделении данных на показатели (или "меры", Measures) и измерения (справочники, разрезности, Dimensions). По показателям мы агрегируем (например, считаем сумму или среднее), а по измерениям группируем или фильтруем.

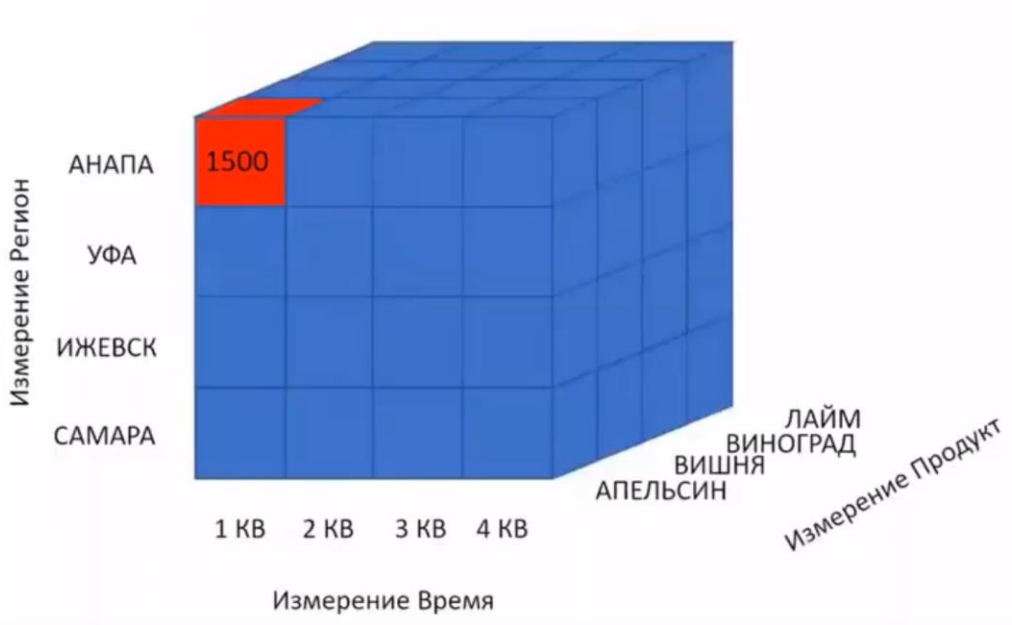

Рисунок 7. Пример многомерного OLAP куба

Именно в рамках многомерной модели появился термин «куб» применительно к анализу данных. Идея в том, что если изобразить измерения, как оси многомерного пространства, то сложную таблицу с большим количеством колонок удобно представить в виде многомерного куба (дискретного гиперкуба), в ячейках которого располагаются значения. Тогда любые отчетные таблицы можно получить из гиперкуба операциями проекции, когда по измерениям происходит либо фильтрация, либо агрегация.

		Самара	Ижевск	Уфа	Анапа	Итого
1 квартал	Апельсин	-	-	2 500 ₽	1 500 ₽	**4 000 ₽**
	Вишня	-	-	2 000 ₽	2 000 ₽	**4 000 ₽**
	Виноград	1 000 ₽	3 000 ₽	-	-	**4 000 ₽**
	Лайм	2 000 ₽	2 000 ₽	-	-	**4 000 ₽**
	Итого 1 кв	**3 000 ₽**	**5 000 ₽**	**4 500 ₽**	**3 500 ₽**	**16 000 ₽**
2 квартал	Апельсин	4 000 ₽	-	-	-	**4 000 ₽**
	Вишня	1 000 ₽	3 000 ₽	-	-	**4 000 ₽**
	Виноград	-	-	1 500 ₽	2 500 ₽	**4 000 ₽**
	Лайм	-	-	2 000 ₽	2 000 ₽	**4 000 ₽**
	Итого 2 кв	**5 000 ₽**	**3 000 ₽**	**3 500 ₽**	**4 500 ₽**	**16 000 ₽**
Итого		**8 000 ₽**	**8 000 ₽**	**8 000 ₽**	**8 000 ₽**	**32 000 ₽**

Рисунок 8. Пример сводной таблицы, полученной в результате проекции OLAP куба

Современные BI-системы включают в себя множество оптимизаций, которые позволяют достичь максимальной производительности на аналитической OLAP нагрузке. Выделим три основных подхода к архитектуре BI-систем: ROLAP, In-Memory OLAP и гибридный вариант. Разберем каждый из них подробнее.

Реляционный OLAP (Relational OLAP, ROLAP)

ROLAP (реляционный OLAP) — OLAP-системы, имеющие прямой доступ к существующим базам данных или использующие данные, выгруженные в собственные локальные таблицы. Аналитические запросы при использовании схемы ROLAP строятся над виртуальным многомерным представлением данных, а их выполнение происходит на уровне реляционной базы данных, то есть фактически выполняются SQL-запросы над реляционной системой.

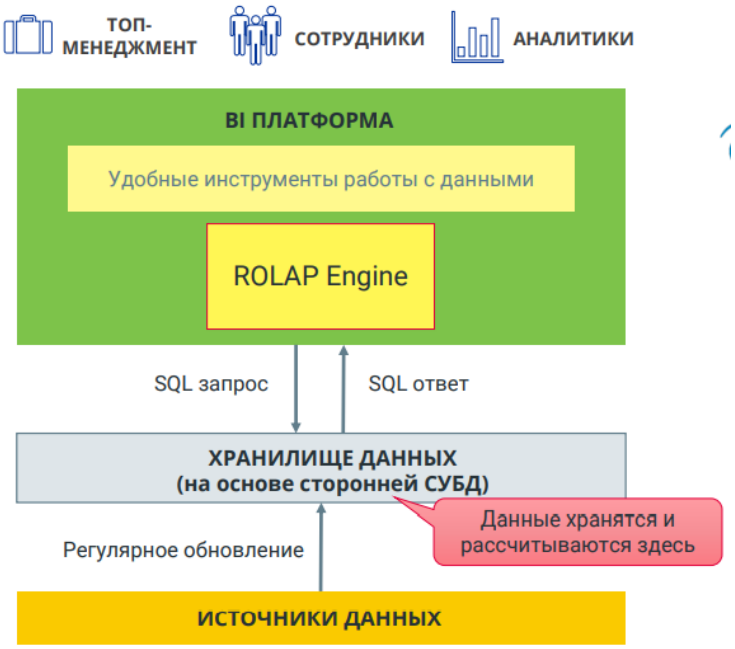

Рисунок 9. ROLAP архитектура

К самым значимым преимуществам схемы можно отнести тот факт, что ограничения объема данных в такой системе определяются только возможностями СУБД DWH, следовательно ROLAP могут работать с очень большими базами данных. Кроме того, реляционные ROLAP обеспечивают высокий уровень защиты данных и хорошие возможности разграничения прав доступа, масштабируемость и короткое время загрузки.

Однако ROLAP работают крайне медленно, если не используется аналитическая колоночная СУБД (например, ClickHouse или HP Vertica). Также к недостаткам схемы относятся низкая эффективность использования кэша, повышенные требования к инфраструктуре, ограниченная аналитическая функциональность (что, разумеется, крайне важно для BI-систем) и сложность развертывания и поддержки.

По данной схеме работают решения Pentaho, Metabase, SAP Business Objects, Alpha BI, Форсайт и ряд других.

In-Memory OLAP

Технология In-memory позволяет загружать информацию непосредственно в оперативную память, что позволяет отвечать на запросы пользователи без обращения к хранилищу данных. При этом преимущества In-Memory OLAP технологии заключаются не столько в работе в оперативной памяти (современные реляционные СУБД тоже активно используют кэширование в ОЗУ), сколько в специальных оптимизациях под аналитическую нагрузку, которая сильно отличается от транзакционной. Среди таких оптимизаций можно отметить:

- Колоночное хранение и выполнение запросов. В In-Memory OLAP системах каждая колонка таблицы хранится в последовательных блоках памяти, что существенно повышает эффективность работы кэша центрального процессора, снижая количество «промахов» и перезагрузок страниц данных из ОЗУ.
- Сжатие данных. Низкая кардинальность (большое количество повторяющихся значений) многих столбцов в OLAP таблицах позволяют применять сверхбыстрые алгоритмы, такие как Run-Length Encoding.
- Активное использование векторных инструкций и сверширoких регистров современных процессоров. Расчет функций агрегации на большом количестве значений очень хорошо ложится на возможности современных процессоров, такие как сверхширокие регистры (AVX-256, AVX-512) и векторные инструкции (SIMD).
- Поздняя материализация. Планы запросов, оптимизированные под большое количество установленных фильтров на данные и, напротив, небольшое количество используемых столбцов к каждом запросе.
- Прочие эвристики. Каждая реализация In-Memory OLAP использует собственный набор эвристик для оптимизации производительности. Из наиболее часто используемых можно отметить:
    - Автоматическая нормализация-денормализация
    - Сортировка по календарному измерению
    - Разделяемый пользовательский кэш
    - Разделяемый кэш подзапросов
    - Обратные индексы для строковых полей

Таким образом, системы, основанные на In-Memory OLAP технологии, демонстрируют наивысшую производительность, если допущения по характеру аналитической нагрузки выполняются.

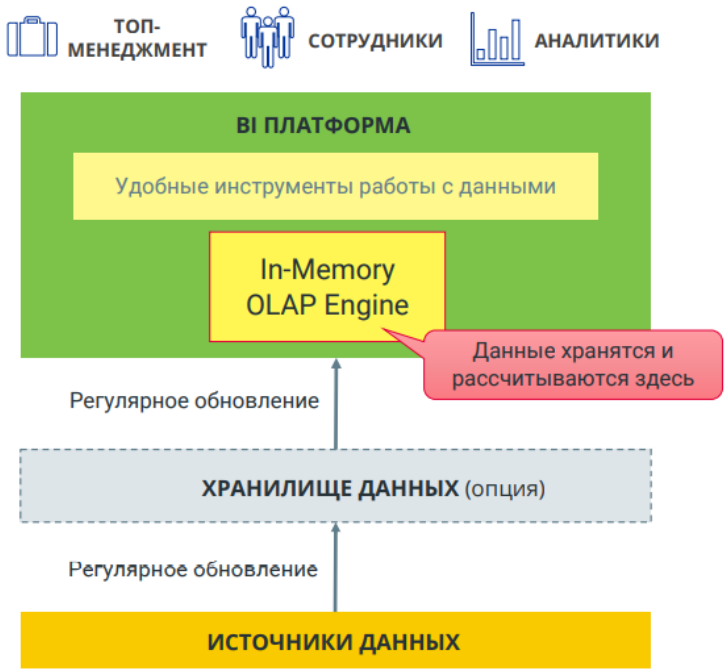

Рисунок 10. Архитектура In-Memory OLAP

Также среди преимуществ технологии – простая установка и использование (не нужен DBA), развитые функции для аналитиков и возможность реализовать «умный кэш» - эффективное использование ресурсов на большом количестве пользователей.

Недостатки также есть. Например, использовать данную технологию невыгодно, если данных в сжатом виде в RAM больше 200-300 Гбайт. К тому же не всегда просто минимизировать задержку отображения обновлений данных на дашбордах, так как обновления из источников носят регулярный характер и вносят задержку актуализации.

По данной схеме работают платформы Qlik, Visiology, Полиматика.

Гибридная схема OLAP

Так как у обеих схем есть как плюсы, так и минусы, вполне объяснимо, что разработчики BI-платформ попытались найти золотую середину – гибридные решения, которые объединяют возможности сразу двух схем. Гибридный OLAP обеспечивает быструю обработку запросов в сочетании с высокой масштабируемостью. При использовании этой технологии многомерные данные хранятся как в оперативной памяти, так и в реляционном представлении. Доступ к данным осуществляется с помощью API для генерации многомерных запросов, тогда как к реляционным БД доступ производится посредством запросов на SQL.

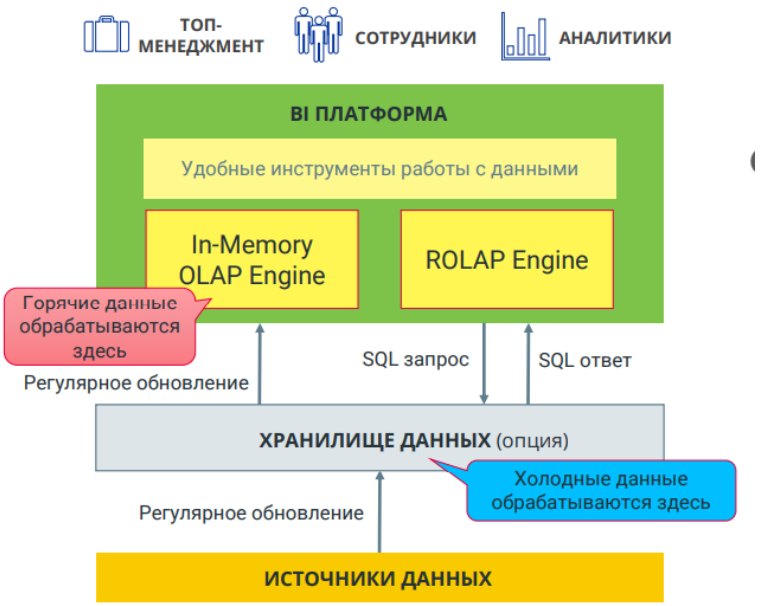

Рисунок 11. Гибридная OLAP архитектура

Но и тут не обошлось без нюансов. Прежде всего, для онлайн работы на всем объеме данных нужно использовать аналитическую колоночную MPP СУБД, а разделять данные на горячие и холодные нужно вручную на уровне ETL-процедур – это как минимум добавляет работы. К тому же совместимость с каждой СУБД должна быть реализована и поддерживаться вендором BI отдельно. Также важно, что при такой схеме работы системы не все аналитические функции будут доступны на всем объеме существующих данных.

Технология поддерживается такими платформами, как Qlik, Tableau, Visiology, Power BI.

Визуализация и самостоятельное исследование данных

Для представления данных по запросу пользователей предусмотрены различные графические средства – отчеты, графики, диаграммы, дашборды, которые настраиваются за счет выбранных параметров. В современных BI-решениях пользователи могут в динамическом режиме менять графическое представление, масштабировать или объединять данные, изменять цвета визуализации и «проваливаться» в данные для их детализации или поиска причин отклонений. Визуальные представления делают сложные массивы данных интуитивно понятными даже неспециалистам.

Бизнес-пользователи получают доступ к аналитике без программирования и привлечения ИТ-специалистов. Для этого в BI-платформах предусмотрен настраиваемый **аналитический веб-портал**. Через него можно получить доступ ко всем инструментам платформы – разумеется, в соответствии с предоставленными конкретному пользователю правами доступа. Также портал позволяет работать со всеми опубликованными дашбордами и обеспечивает совместную работу с данными с ПК, ноутбука или планшета.

В зависимости от платформы он может содержать разный набор стандартных шаблонов визуальных элементов, нестандартные виджеты для визуализации информации, а также предлагать различные компоненты для персонализации дизайна.

Рисунок 12. Примеры управленческих дашбордов

***Дашборды** – это интерактивные отчеты, создаваемые платформой на основе объединенных данных из различных источников – учетных систем, различных СУБД, систем обработки Big Data, а также данных, собранных вручную через Excel или модуль сбора данных. Преимущество дашбордов – удобная и понятная визуализация, для создания которой не требуется привлечение ИТ-специалиста, а настройка дашборда под запросы пользователя может быть очень точной.*

5. Ручной ввод данных.

Зачастую для составления отчетности и планирования в качестве средства автоматизации во многих организациях используется Microsoft Excel. И это, повторимся, действительно эффективный инструмент, если задача достаточно проста, а количество источников данных невелико. Однако по мере усложнения процесса трудозатраты на консолидацию данных и поддержку решения резко возрастают. Весьма значимым риском является и потенциальное

появление большого количества ошибок в данных и отчетах, связанных с человеческим фактором.

Решить проблему позволяют инструменты ввода данных, характерные для многих BI-систем. Эти инструменты позволяют сохранить простоту конструирования форм ввода бизнес-пользователями (например, финансистами/экономистами), при этом предоставляя дополнительные возможности, а именно:

- форматно-логический контроль (ФЛК, автоматические проверки вводимых данных);
- бизнес-процессы согласования;
- бизнес-правила и формулы расчета (применяется, например, для трансформаций и аллокаций);
- уведомления и контроль сроков заполнения форм;
- сохранение истории изменений данных и аудиторский след;
- разграничение прав доступа;

Когда в процессе ввода данных участвуют десятки и сотни людей, управлять этим процессом в ручном режиме уже нецелесообразно. Именно поэтому BI активно используется для задач распределенного сбора данных, управленческой отчетности, планирования и бюджетирования.

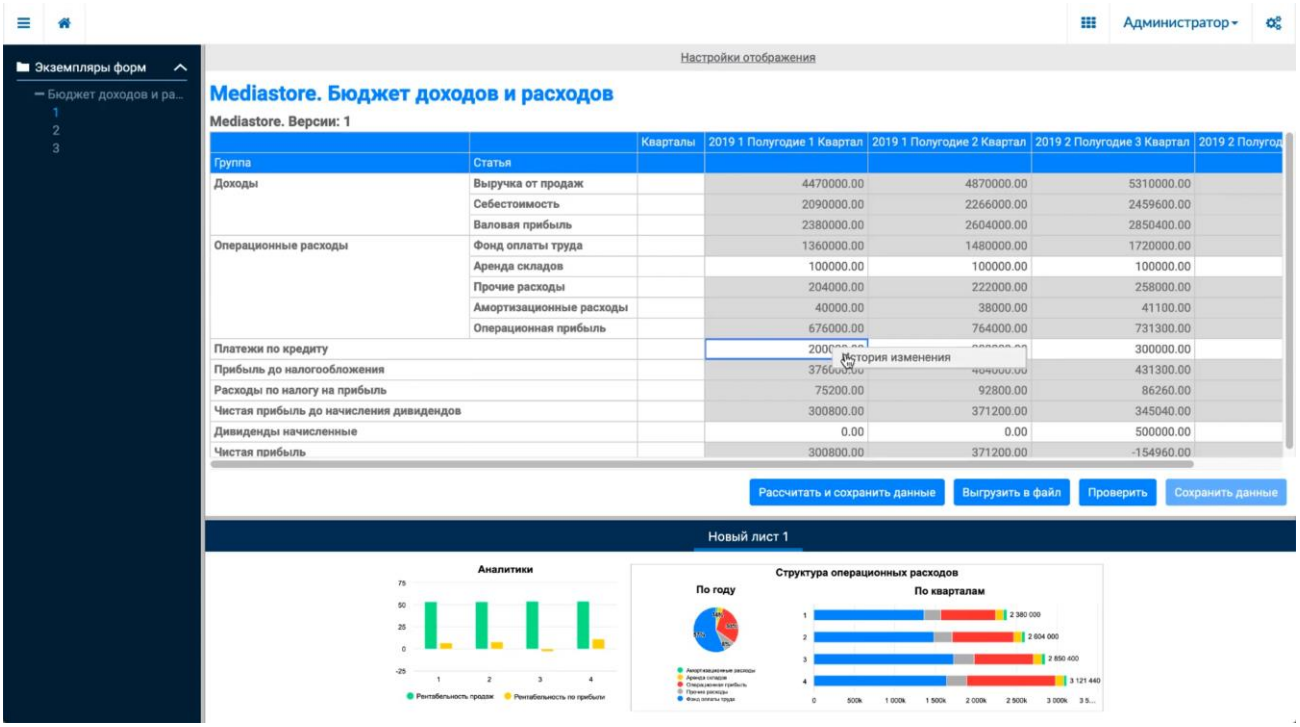

Рисунок 13. Пример формы ввода в BI-системе

Аналитические модули планирования и бюджетирования позволяют обеспечить операционное и финансовое планирование на корпоративном уровне. Так как это необходимые компоненты реализации стратегии для компаний, то и функционал для планирования и бюджетирования

интегрирован практически во все современные BI-платформы. В отличие от модулей планирования ERP-систем, интегрированные в BI охватывают всю компанию и ведутся в агрегированных показателях на достаточно долгосрочные плановые периоды, в том числе с разбивкой по подпериодам. ERP же обеспечивают текущее планирование в рамках отдельных подразделений. Таким образом планирование и бюджетирование в BI и ERP-системах дополняют друг друга, охватывая широкий спектр планов - от стратегических до оперативных.

Часть 3. Особенности реализации BI-проектов

BI-платформы могут применяться в любых отраслях и направлениях бизнеса – от ретейла и логистики до спорта и госорганизаций. Один из базовых кейсов использования основан на потребности руководства компании видеть показатели эффективности в режиме онлайн и оперативно на них реагировать.

Так, BI позволяет анализировать такие финансовые показатели, как EBITDA, дебиторская и кредиторская задолженности, остатки на счетах и быстро принимать соответствующие решения, если зафиксированы отклонения от нормы или плана.

Распространенный вариант – анализ продаж. Это больная точка ретейлеров и дистрибьюторов, которым необходимо быстро подстраивать закупки и маркетинговую активность в соответствии с трендами быстро меняющегося спроса. Поэтому анализ продаж, заказов, товарных остатков и прочих показателей для этой отрасли крайне актуален.

Еще один типичный кейс для BI – аналитика маржинальности по направлениям. Он позволяет определить, какое из направлений требует большего внимания руководства, выявить среди них высокомаржинальные и развивать их более активно, повышая таким образом эффективность всего бизнеса.

Для производственных компаний одно из актуальных направлений использования BI – контроль эффективности производственных рабочих и оборудования. Например, постоянный контроль и оценка результативности работы персонала позволят отыскать и устранить ошибки в планировании всего бизнеса, распределить программы повышения квалификации для повышения эффективности, оптимизировать закупки и продажи.

Анализ административных и общепроизводственных расходов также относится к распространенным BI-кейсам. Руководители предприятий и организация хотят знать, какие расходы можно минимизировать – для результативности бизнеса это всегда важно.

Так что BI применима везде, где руководство хочет понимать, что влияет на различные показатели и как их можно улучшить.

Например, типичные болевые точки производственных предприятий – это отсутствие гибкости в производственных процессах, необходимость обеспечивать ликвидность складских запасов

и своевременные поставки комплектующих и материалов, а также сбои в балансе процесса производства. В этом случае внедрение BI-платформы поможет в режиме онлайн проводить анализ структуры, качества и возраста дебиторской и кредиторской задолженностей, а также структуры себестоимости – все это позволит построить систему бюджетирования и бизнес-планирования, оптимизировать и улучшить контроль прямых и косвенных затрат. Анализ трендов спроса на продукцию, загрузки склада и оборачиваемости продукции дает информацию для оптимизации производственного процесса, маркетинговой активности, что в конечном итоге может стать точками роста прибыли.

Типовые ошибки при внедрении BI-систем

Как и любой другой проект, внедрение BI-системы может оказаться неудачным. Разберем наиболее часто встречающиеся ошибки, которые приводят к провалу. Большее количество провалов BI-проектов связано с типовыми проблемами внедрения ИТ систем, такими как недостаточное вовлечение руководства/бизнеса, поставка результата только на финальном этапе проекта, отсутствие компетенции по системе на стороне заказчика. Однако, можно выделить и специфичные для BI-проектов ошибки, которые могут вызвать сложности с внедрением.

1. *Не устанавливать измеримых целей внедрения BI*

Эффективность внедрения BI-системы объективно гораздо сложнее измерить, чем эффективность внедрения большинства других ИТ-систем, потому что основной эффект от внедрения – это повышение качества управленческих решений, которое сложно обоснованно выразить в деньгах. Это часто приводит к тому, что внедрение начинается без установки измеримых критериев успешности. Тем не менее, такие критерии все же должны быть установлены, например, «срок предоставления актуальной информации в отчете А сократился до 1 дня» или «руководитель А удовлетворен качеством информации в отчете Б (нет возражений по качеству данных) и пользуется системой не реже 3 раз в неделю»

2. *«Отчет из PPT перенесли в дашборд»*

Одна из значимых ошибок – перенос отчетов пользователей в систему в том формате, в котором они были созданы. Весьма вероятно, пользователи просто не знали, что есть другие более удобные и эффективные варианты визуализации данных. Если руководствоваться исключительно «привычками» пользователей, наработанными в компании схемами, действительно заметных результатов от проекта добиться не удастся.

3. *Строить BI на загрузке Excel-файлов*

Повторимся: Excel – действенный и полезный инструмент. При этом во многих компаниях данные из ИТ-систем выгружаются в Excel в полу-ручном режиме для получения разных отчетов. Ошибкой будет интегрировать BI с такими Excel-файлами, пренебрегая данными,

которые есть в самих ИТ-системах компании – именно там скрыта наиболее точная и полная информация, на основе которой и должны приниматься управленческие решения. Конечно, для такой интеграции придется преодолеть как технические, так и организационные проблемы (и загрузка Excel может быть эффективна на промежуточных этапах), но только в этом случае возможно достичь достоверности данных.

Часть 4. Тренды BI

Несмотря на то, что Business Intelligence – это зрелая технология, многие годы находящаяся на «плато продуктивности», она продолжает развиваться. На сегодняшний день можно выделить 5 основных и самых значимых трендов в области BI.

Augmented Analytics

Augmented Analytics или дополненная (расширенная) аналитика считается большинством аналитиков определяющим или даже прорывным трендом ближайшего будущего. Augmented Analytics изменяет сами сценарии разработки, потребления и совместного использования аналитического контента.

На практике это означает следующее. Если раньше с данными работали в режиме pull (делали запрос к системе и получали ответ), то сейчас парадигма меняется и данные благодаря Augmented Analytics приходят к пользователю сами, когда это ему нужно. Например, можно выбрать показатели, по которым система сама будет генерировать инсайты – например, будет предлагает обратить внимание на какие-то аномалии. При этом пользователь сможет даже не заходить в BI-систему – она будет сама присылать сообщения, причем не по расписанию, а по мере возникновения аномалии.

Это быстрорастущая область, в которой уже сейчас экспериментирует большинство разработчиков, так что это драйвер рынка на годы вперед.

Natural Language Query

На самом деле этот тренд каждому из нас уже хорошо знаком благодаря взаимодействиям с голосовыми помощниками. Речь про запросы на естественном языке: если раньше BI-системе нужно было отправлять запрос в соответствии с определенными правилами, то вскоре систему можно будет просто спросить, например, какой тренд был самым заметным в прошлом месяце или как растут продажи определенного товара. Такие разработки уже есть, в том числе в платформах Power BI, Qlik, Visiology, уже сейчас их можно опробовать.

Возможно, это покажется не столь важным трендом на первый взгляд, но это обманчивое ощущение. Дело в том, что даже в самых крупных и продвинутых компаниях проникновение

BI среди персонала не превышает 25%, т.е. лишь четверть сотрудников пользуются возможностями бизнес-аналитики. Причина проста: для оставшихся 75% сотрудников это слишком сложно, нужно обучаться, что затратно и требует времени и сил. Если бы им предоставили возможность просто говорить с системой, то количество пользователей, которые могли бы извлекать нужную информацию из данных компании, принося тем самым ей пользу, выросло бы.

AutoML

AutoML – это автоматизированное машинное обучение. Алгоритмов машинного обучения много, но пока до сих пор нужен специалист, чтобы подобрать правильный вариант и эффективно развивать систему. Сейчас появляется класс технологий, которые позволяют на простых кейсах автоматически перебирать большое количество разных моделей и по техническим характеристикам подбирать лучшие. Это как минимум снижает затраты дата-сайентистов, потенциально же в будущем может появиться продвинутая аналитика, которой смогут пользовать бизнес-пользователи без глубоких знаний математики, программирования и т.д.

Embedded Analytics

Embedded Analytics или встроенная аналитика. Во многих отраслях почувствовали, что использование голой BI-платформы трудоемко – надо настраивать, проводить интеграцию и т.д. Поэтому появился новый тренд: в отраслевой софт (например, ERP-систему) встраивается аналитическая система, которая для пользователя выглядит как часть уже существующего решения. Но при этом пользователь может глубоко проваливаться в данные, как в BI-системе, а не просто смотреть отчеты, как в классических отраслевых решениях.

Data Governance + Data Ethics

Развитие последнего тренда обусловлено технологическим развитием и увеличение роли данных в мире и бизнесе. Как только все поняли, какую силу дают данные, стали задумываться об этике их использования. Это в первую очередь касается больших данных, но и корпоративных тоже. Отсюда растущее внимание к защите данных, к ограничению использования персональных данных, к разработке соответствующих политик. BI также затронет этот тренд – он будет развиваться все активнее.