## Глава 3.1. Business Intelligence (BI)

* Иван Вахмянин, сооснователь Visiology*

*«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.*

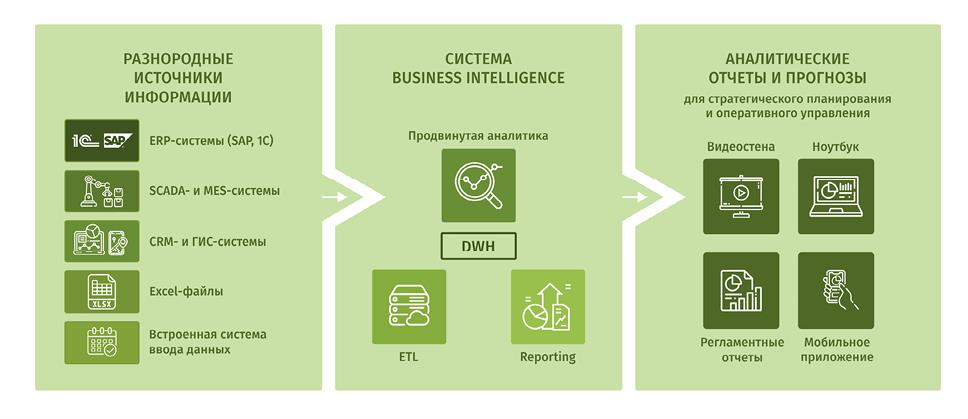
***Эдгар Алан По, «Убийство на улице Морг», 1841 г.***

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)*** *– технологическое решение, которое позволяет собирать, интегрировать, анализировать и представлять бизнес-информацию для поддержки принятия решений. BI-система позволяет использовать данные, собранные из различных источников, обрабатывать их и представлять в удобном для анализа виде, например, в виде отчетов, дашбордов, графиков и таблиц. Основные функциональные блоки BI - это:*

* *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 дашбордов и строгой отчетности.*
* *Ввод данных – встроенные в BI платформу инструменты автоматизации 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).*

Стоит отметить, что для эффективного использования BI-системы необходимы соответствующие специалисты – в первую очередь, аналитики BI. Несмотря на то, что в современных BI-системах начинают появляться инструменты, основанные на ИИ (подробнее в заключительной части главы), основную часть работы по сбору, анализу и визуализации данных пока выполняют люди – а BI-системы позволяют им автоматизировать рутинные задачи и сосредоточиться на интерпретации и анализе.

**



***Рис. 3.1.1.*** *Принцип работы BI-системы.*

### 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 одним из самых важных активов любого бизнеса, вне зависимости от его размера и отраслевой принадлежности. И с каждым годом их становится все больше.

Однако данные недостаточно накопить: их необходимо анализировать, управлять ими, заставить их работать – именно в этом состоит самый значимый вызов для современных компаний, особенно если они видят себя частью цифровой экономики. С учетом растущего темпа бизнеса у руководителей компаний остается все меньше времени на принятие решений, тогда как сами решения по-прежнему 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 (далеко за пределы объема, доступного для анализа с помощью Excel).
* Сокращение времени на принятие решений.
* Сложность процессов управления.
* Неэффективность традиционных методов управления, анализа и оценки.

#### Управление на основе данных

В открытых обсуждениях до сих пор можно встретить мнение, что BI — это не более чем графики и диаграммы, которые можно построить при помощи любых инструментов — даже карандашом нарисовать. Но современные организации стремятся внедрить BI-платформы, а далее — распространить практику BI на широкий спектр отделов потому, что системы бизнес-аналитики позволяют решить более сложную задачу и реализовать практику управления на основе данных.

В современном, быстро меняющемся мире, уже недостаточно просто накопленного опыта и экспертизы, чтобы принять верное решение. Внезапное изменение спроса, трансформация логистических цепочек, снижение производительности труда отдела, преждевременный износ деталей станков, увольнение сотрудников группы разработки — все это события, которые могут повлиять на результаты компании и требуют моментальной реакции со стороны руководства. Но если в компании не налажен процесс сбора и анализа информации с возможностью ее изучения и визуализации, об управлении на основе данных можно забыть. Руководитель узнает о том, что произошли какие-то важные изменения только в конце месяца или даже квартала, и их последствия уже успеют оказать серьезное влияние на показатели компании в целом.

BI-платформа — это в первую очередь средство решения задачи управления на основе данных. Правильно настроенные механизмы сбора информации, эффективное размещение данных в хранилище, а также удобные инструменты визуализации и изучения закономерностей (data discovery) позволяют модернизировать процессы принятия решений на всех уровнях — от стратегических задач до ежедневных рутинных операций.

#### Качественные выгоды от внедрения BI

1. Ключевая выгода от внедрения – это **высокое качество и точность данных** в необходимых бизнесу аналитических отчетах, значительное повышение доверия руководителей к аналитике. Минимизация человеческого фактора в процессе подготовки и обновления отчетности, документирование и автоматизация обработки информации позволяют радикально снизить количество ошибок и неточностей.
2. 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.
3. Скорость получения информации – одна из важнейших составляющих принятия управленческих решений. BI обеспечивает **ускорение доступа к информации,** вплоть до получения данных в режиме реального времени, поэтому топ-менеджмент вовремя получает ответы на важные бизнес-вопросы, и управленческие решения чаще принимаются не на основе интуиции, а на основе фактов и цифр.
4. Чем больше менеджеров и рядовых сотрудников компании получают доступ к «единой правде», тем более верные решения они принимают и тем эффективнее будут процессы в компании. Если для каждого аналитического запроса потребуется привлекать ИТ-специалиста, добиться нужной доступности будет невозможно. BI-платформа позволяет масштабировать доступность данных в формате **Self-service аналитики**, 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 информации и доступ к ней, сокращается на 51%, а на анализ информации тратится на 48% меньше времени. 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%.

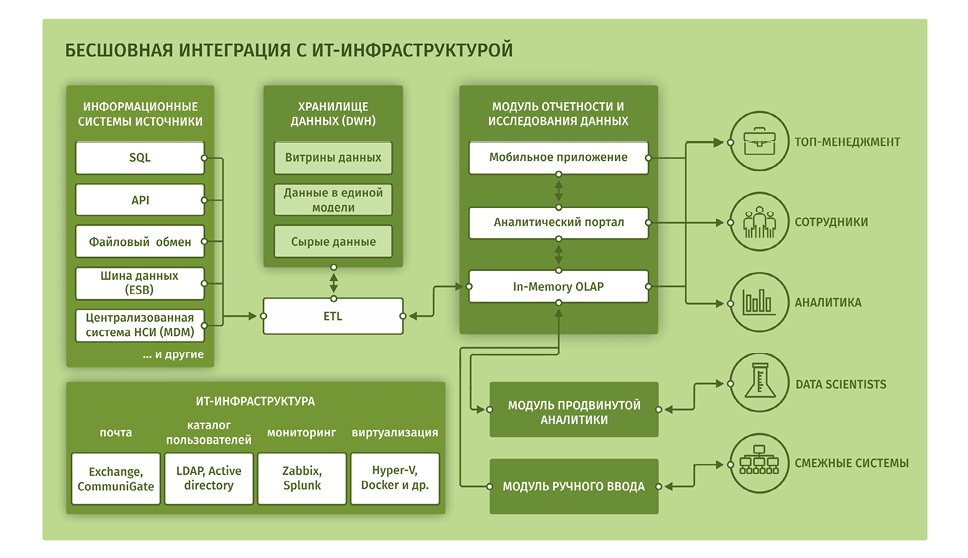
Дополнительные финансовые показатели:

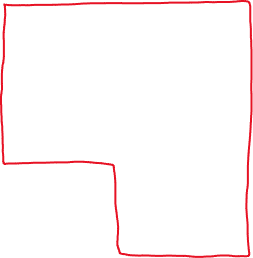
* 31% – снижение расходов, связанных с поддержкой отчетности;
* 30% – снижение расходов на разработку новой отчетности;
* 20% – снижение операционных расходов;
* 16% – увеличение выручки.

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

### Принцип работы BI-системы

BI-платформы могут отличаться друг от друга по многим параметрам – производительности, функциональности, наличию облачных версий и многим другим. Но если посмотреть на схему их работы, то принципиально она схожа у всех. Есть источники информации, с которыми интегрируется BI-платформа, предоставляющая 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.

**



BI-система организации



ROLAP / In-memory OLAP

***Рис. 3.1.3.*** *Типовая структура BI-решения.*

#### 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. **Скриптовый ETL** (Script ETL) – в противовес визуальному – предлагает использовать определенный скрипт-язык, оптимизированный для обработки данных.

Преимуществами Visual ETL, безусловно, являются простота и наглядность, но есть и минусы, которые весьма существенны. Как только задачи усложняются, схемы и блоки 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 начинающим, а во-вторых, для Python есть готовая экосистема готовых модулей и технологий – от коннекторов до библиотек для продвинутых инженеров данных.

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, как и все, кто работает с данными, уделяют анализу и визуализации 20% своего времени, а 80% тратят на преобразование, очистку и выгрузку данных. Понятно, что им необходим инструмент, который решает эту проблему. Загвоздка в том, что ETL на рынке много, а выбрать нужно 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 – в идеальном сценарии 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 подобных решений, хотя бы самых распространённых.
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, работают 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: инструмент решает простые задачи, но по мере усложнения задач возможностей не хватает. Аналитикам необходимы ETL-инструменты, которые способны развиваться вместе с пользователем и его компетенциями и потребностями.
5. ETL-инструмент желательно должен быть либо **недорогим**, либо **бесплатным**, причем не в виде trial-версии, а навсегда.

В любом проекте комплексного внедрения BI реализация процессов ETL является отдельной непростой задачей. Именно поэтому на начало 2024 года на российском рынке доминируют ETL-решения с открытым исходным кодом. К ним относятся в первую очередь Apache AirFlow и Apache NiFi, а также простой в использовании инструмент Vixtract. Используя open-source ETL, команды разработки и системные интеграторы получают возможность создавать сложные схемы сбора данных в аналитическое хранилище, подключать широкий спектр источников и кастомизировать практически любые параметры таким образом, чтобы загрузка и трансформация данных происходила быстро и эффективно, не мешая работе основных бизнес-систем.

В качестве альтернативы, получившей широкое распространение в России, можно отметить платформу Loginom. Это развитый инструмент ETL с визуальным интерфейсом и высоким уровнем Self-Service. Большое количество уже готовых коннекторов позволяет настроить загрузку данных в режиме Drag&Drop. Поэтому Loginom часть используется в связках с различными BI-платформами как более удобное для заказчика решение ETL.

В некоторых BI-платформах также имеются встроенные системы ETL. Но так как они не являются специализированными инструментами, в подобных движках обычно реализован лишь базовый функционал, который подходит только для простейших задач или ознакомления с BI-платформой. На практике при переходе к более сложным моделям и подключению новых источников, возникает потребность в более гибких и функциональных решениях, что приводит либо к переходу на специализированное проприетарное ПО, либо к адаптации Open-Source инструментов.

Наиболее детальное рассмотрение различных ETL-инструментов приводится в исследовании “ETL Круг Громова”. Авторская методика Сергея Громова, руководителя BI Consult, позволяет оценить возможности и сопоставить стоимость всех доступных сегодня на российском рынке решений для загрузки и трансформации данных.

#### Корпоративное хранилище данных (Data Warehouse, DWH)

В контексте систем бизнес-аналитики под хранилищем данных подразумевается в первую очередь не СХД, а именно программный и логический уровни организации информации. Именно правильная гармонизация данных из различных источников – наиболее сложная и важная задача создания DWH. Практика показывает, что на формирование хранилища уходит до 80% ресурсов при реализации проекта корпоративного BI. Грамотное и профессиональное решение этой задачи позволят своевременно обеспечить всех пользователей качественной непротиворечивой информацией.

Если рассмотреть ПО для создания хранилищ, еще недавно большинство DWH работали на стандартных реляционных СУБД, таких как, например, MS SQL Server и PostgreSQL. Однако уверенный курс на Big Data и рост требований к производительности хранилища для выполнения все более сложных запросов со стороны аналитических платформ, приводит к стремительному росту популярности распределенной СУБД GreenPlum, колоночной СУБД ClickHouse, а также построенные на их основе российские СУБД Arenadata DB и Arenadata QuickMarts.

Вопрос оптимизации и автоматической подготовки хранилища при этом стоит все острее, а зарплаты архитекторов СУБД стремительно растут. Поэтому спрос на решения, которые позволят “из коробки”, обеспечить эффективное хранение и работу с данными, становится все выше. Таких решений пока немного, но уже сейчас можно найти инструменты для автоматической генерации схемы данных в СУБД в соответствии с популярными моделями данных (см. ниже). Под моделью данных в контексте DWH понимается методология организации схемы таблиц СУБД в зависимости от исходных данных и аналитических задач компании.

**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 перенести 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 (дата-сетами) 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 и проверенная 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 редко используются 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, например, Greenplum или 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 загрузки, аналитики и выгрузки данных в реальном времени;
* хранение и обработка неструктурированных (тексты, изображения, аудио, видео) данных в едином контуре со структурированными.

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.

#### 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, чего 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 обязательно должны присутствовать инструменты продвинутой аналитики, основанные на науке о данных (Data Science). Конечно, продвинутая аналитика – это очень обширная область знаний, и она может применяться в компании вне контекста BI-систем. Однако, как правило, в BI-системы также встроены инструменты продвинутой аналитики (как некоторое подмножество всей палитры инструментов Advanced Analytics), которые позволяют BI-аналитикам иметь их под рукой и быстро применять в решении своих задач. При этом в одной организации могут эффективно сосуществовать подразделения BI и Advanced Analytics (часто называемое отделом Data Science), и это накладывает определенные требования на унификацию инструментария продвинутой аналитики.

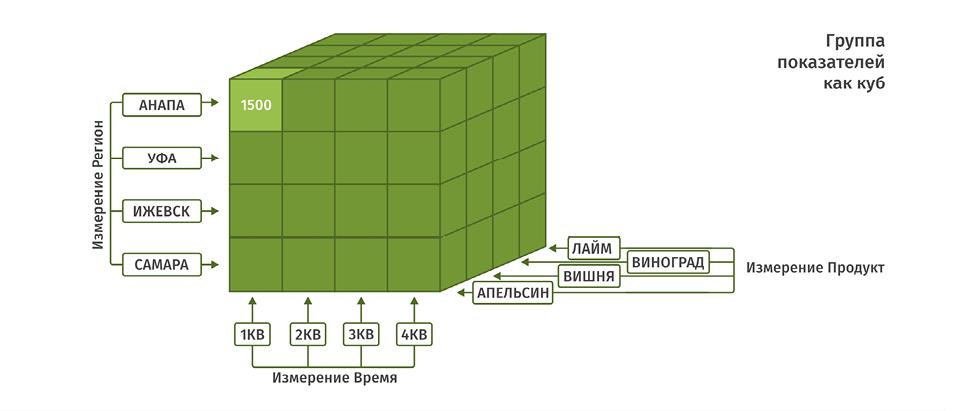
##### Инструментарий продвинутой аналитики

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 переходит 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 — модели, разработанные математиками и специалистами по продвинутой аналитике, необходимо публиковать на аналитической 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.

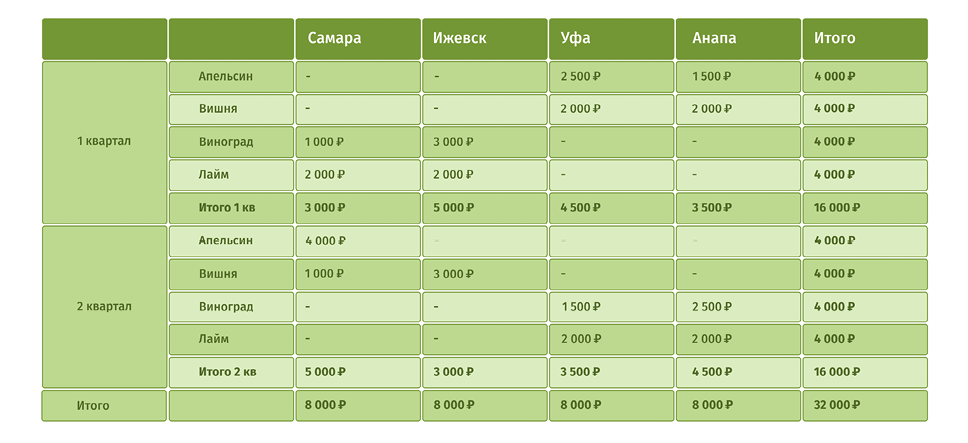
#### Автоматизация отчетности (Reporting) и самостоятельное исследование данных (OLAP, Self-service BI)

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 как именно благодаря OLAP анализ данных становится доступен бизнес-пользователям, а не только разработчикам и специалистам по базам данных.

**

***Рис. 3.1.7.*** *Пример многомерного OLAP куба.*

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.

**

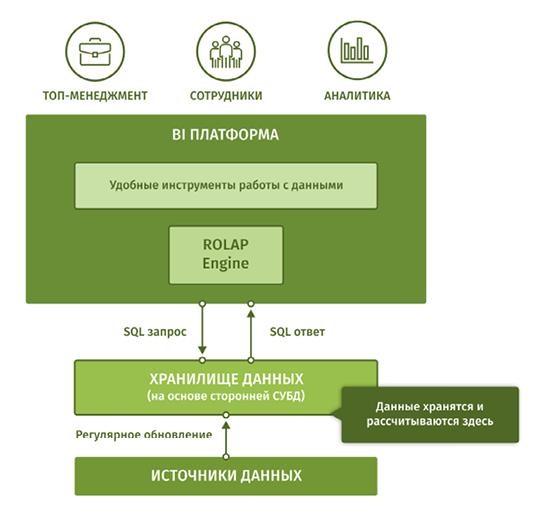
***Рис. 3.1.8.*** *Пример сводной таблицы, полученной в результате проекции OLAP-куба.*

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.

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)

Наиболее популярная схема работы с данными — это 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 над реляционной системой выполняются SQL-запросы.

**

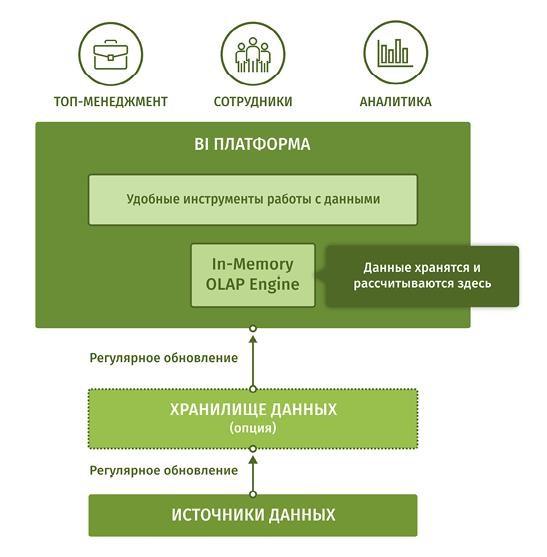
***Рис. 3.1.9.*** *ROLAP-архитектура.*

К 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.

Еще недавно к недостаткам ROLAP относили медленную работу по сравнению с другими моделями, но повышение скорости работы высокопроизводительных СУБД, а также наработки по их оптимизации привели фактически ко второму рождению ROLAP. Наработанные лучшие практики использования механизмов кеширования и индексирования, адаптация высокопроизводительных СУБД, таких как ClickHouse или Vertica, привели к тому, что на базе ROLAP удается одновременно решить задачу высокой скорости решения аналитических задач и обработки действительно больших объемов данных. Однако, стоит отметить, что высокая производительность OLAP требует серьезной оптимизации логической схемы СУБД и глубокой интеграции между DWH и BI. Одной из ключевых характеристик BI-системы является то, насколько эффективно она может выполниться эти оптимизации автоматически, разгружая дорогостоящего архитектора СУБД/DWH. Подробнее про требуемые для BI оптимизациях DWH можно прочитать в статье “7 направлений оптимизации ClickHouse, которые помогают в BI” (см. ссылку в конце главы).

##### In-Memory OLAP

Технология In-memory подразумевает загрузку информации непосредственно в оперативную память, что позволяет отвечать на запросы пользователей без обращения к хранилищу данных. Учитывая, что в числе преимуществ In-Memory OLAP — не только работа в оперативной памяти, но и в специальных оптимизациях под аналитическую нагрузку, подобная схема хорошо подходит для тех случаев, когда объем данных не слишком велик, а потребность в быстрой реакции и минимальных задержках — в приоритете.

**

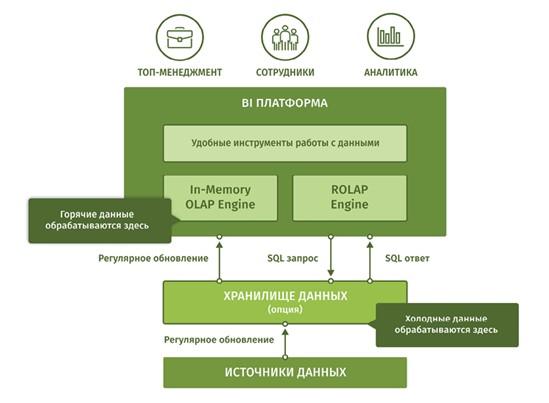
***Рис. 3.1.10.*** *Архитектура In-Memory OLAP.*

Технологии In-Memory OLAP демонстрируют наивысшую производительность (если выполняются допущения по характеру аналитической нагрузки). Также среди преимуществ подобного подхода – простая установка и использование (не нужен DBA), и возможность реализовать «умный кэш» — эффективное использование ресурсов на большом количестве пользователей.

Но у этой модели есть и недостатки. 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.

##### 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.

**

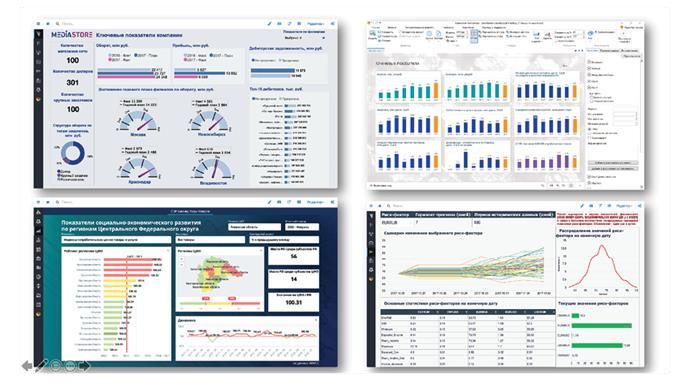
***Рис. 3.1.11.*** *Гибридная OLAP архитектура.*

Но и тут не обошлось без нюансов. Прежде всего, для работы на всем объеме данных онлайн 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.

##### 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 не специалистам.

***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.*

**

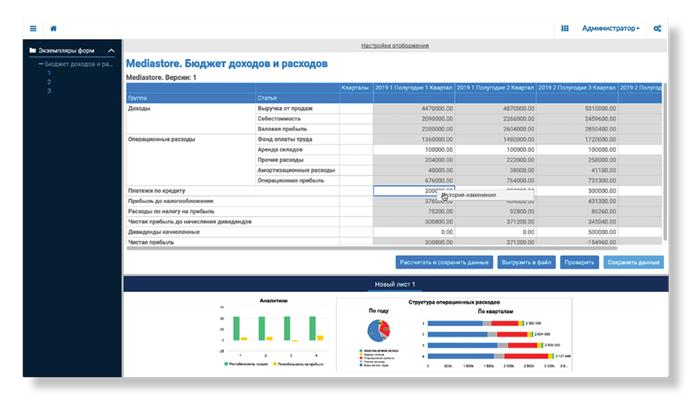
***Рис. 3.1.12.*** *Примеры управленческих дашбордов.*

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.

#### 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 в данных и отчетах ошибок, связанных с человеческим фактором.

**

***Рис. 3.1.13.*** *Пример формы ввода в BI-системе.*

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 (применяются, 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.

Применение модулей сбора данных BI платформ для решения задач планирования и бюджетирования позволяют обеспечить операционное и финансовое планирование на корпоративном уровне. Так как это необходимые компоненты реализации стратегии для компаний, то и функциональность для планирования и бюджетирования интегрирована практически во все современные BI-платформы корпоративного уровня. В отличие от специализированных модулей планирования учетных систем (например, построенных на базе 1С), интегрированные в BI модули, как правило, охватывают всю компанию и ведутся в агрегированных показателях на достаточно долгосрочные плановые периоды. Модули, построенные на базе учетных систем, лучше обеспечивают текущее планирование в рамках отдельных подразделений за счет документно-ориентированного подхода и жесткой привязке к нормативно-справочной информации (НСИ). Таким образом, планирование и бюджетирование в BI и учетных системах дополняют друг друга, охватывая широкий спектр планов — от стратегических до оперативных.

### Особенности реализации BI-проектов

BI-платформы могут применяться в любых отраслях и направлениях бизнеса – от ритейла 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. В торговле одной из самых распространенных задач для BI является анализ продаж. Практически для любых компаний хорошие результаты дает аналитика маржинальности по направлениям и маркетинговая аналитика.

Кстати, в последнее время отслеживание эффективности маркетинга стало настолько популярной практикой, что на рынке появились даже специализированные аналитические системы, предназначенные именно для решения задач так называемой «сквозной аналитики», когда оценка эффективности маркетинговых мероприятий основывается на отслеживании пути каждого клиента от контакта с рекламой до продаж, сервиса и прекращения взаимодействия с компанией. Такие системы отличаются тем, что имеют в комплекте наборы коннекторов к системам веб-аналитики, IP телефонии, системам управления интернет-рекламой – ко всем инструментам, откуда можно получить данные о пользовательском опыте клиента.

Для производственных компаний одно из актуальных направлений использования BI сегодня становится основным инструментом в достижении большей гибкости производственных процессов. Вместе с BI появляется возможность повысить ликвидность складских запасов, организовать своевременные поставки комплектующих и материалов, а также оперативно выявлять и устранять сбои в балансе процесса производства. А 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.

#### Подходы к использованию BI-систем

Стремительное развитие BI-платформ, как необходимых инструментов для принятия решений, привело к расширению функциональных возможностей подобных решений. Как следствие, в настоящее время сформировалось два подхода к развитию и использованию BI-платформы в организации.

1. **Централизованная разработка**

Наиболее простой для пользователей и традиционный вариант подразумевает подготовку отчетов и дашбордов специализированными разработчиками и аналитиками. Эта схема подразумевает постановку задачи на визуализацию тех или иных параметров функциональным заказчиком. Например, руководитель отдела кадров может запросить еженедельный отчет по прохождению программ обучения сотрудниками. Разработчики готовят формулы и представления и выдают заказчику готовый дашборд или отчет (смотря что он запрашивал), который будет доступен на аналитическом портале или в качестве рассылки.

1. **Self-Service и самостоятельная работа с данными**

Улучшение интерфейса BI-платформ, внедрение элементов визуального программирования и снижение порога входа для многих продуктов, доступных на рынке, создали условия для “самообслуживания” в сфере BI. Такой подход подразумевает, что заинтересованный в определенных аналитических выкладках сотрудник может самостоятельно подготовить себе дашборд, опираясь на уже имеющуюся модель данных. Причем это может быть как профессиональный аналитик, который ищет новые закономерности и исследует данные, так и просто бизнес-пользователь. Интересно, что в последнее время широта применения концепции Self-Service значительно расширяется, и пользователи без бэкграунда в программировании сами создают все более сложные отчеты. Интересно, что в этом немало способствует развитие “виртуальных помощников” аналитика, которые помогают писать запросы к данным на популярных языках. Например, бесплатный телеграм-бот ViTalk GPT позволяет на основе запросов на естественном языке готовить формулы на языке DAX и скрипты загрузки на языке Python, чтобы делать расчеты и обработку данных без дополнительной подготовки и обучения.

1. **Гибридный подход**

Однако на практике практически в любой компании наблюдаются запросы как на Self-Service, так и на централизованную разработку дашбордов и отчетов. И сегодня ожидания клиентов от BI-платформ сводятся одновременно к развитию средств самообслуживания и самостоятельной работы с данными и к расширению спектра возможных разработок, которые требуют навыков программирования. Благодаря этому достигается как более широкое проникновение практики BI, так и более глубокая кастомизация визуальных и расчетных элементов под нужды компании.

### Тренды BI

Еще недавно, когда речь шла про перспективы аналитических платформ, мы рассуждали про будущее Augmented Analytics (дополненная или расширенная аналитика), NLQ (обработка запросов на естественном языке), AutoML (автоматическое составление прогнозов). Однако на практике все эти направления оказались не настолько востребованными и не оправдали возложенных на них ожиданий. Тем временем, все изменилось в тот момент, когда на рынке появились технологии класса LLM/GPT.

#### Генеративный ИИ

Использование возможностей генеративного ИИ (ChatGPT и его аналогов), как и во всех других сферах, где эта технология применяется сегодня, привело к качественному росту эффективности работы специалистов. Например, после небольшого обучения LLM (Large Language Models) могут переводить “с человеческого” языка на машинный и помогать аналитикам с подготовкой запросов к платформе.

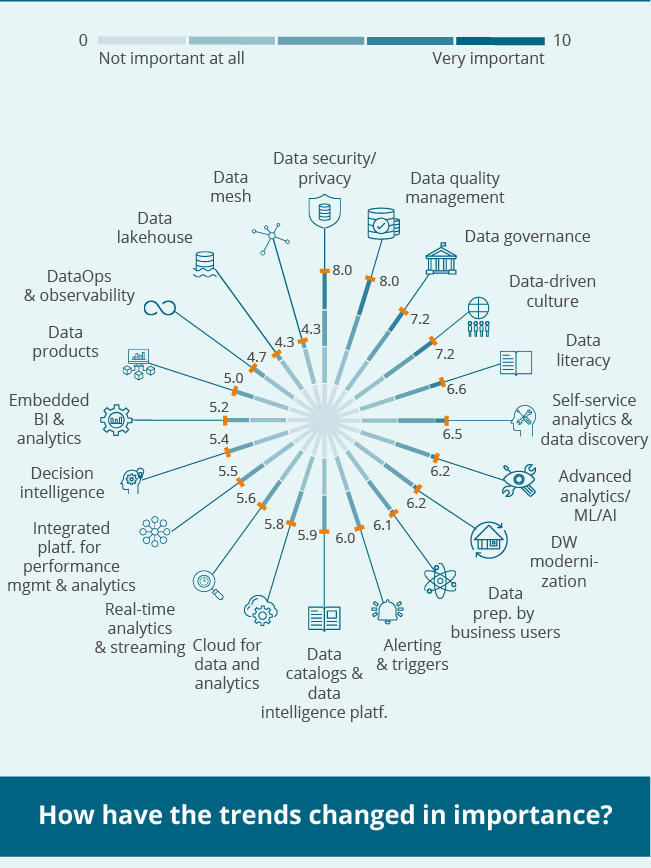
Основные варианты использования генеративного ИИ в рамках функциональности BI-систем:

1. Консультация пользователя по функциям и возможностям системы (“умная справка”). Современная BI-система имеет объем справочной системы в десятки тысяч страниц, которые еще и изменяются от версии к версии. Найти нужную информацию даже с использованием поиска бывает непросто. “Умная справка” же позволяет сгенерировать рекомендацию пользователю мгновенно и с учетом нюансов его запроса.
2. Помощь аналитикам в написании аналитических запросов и ETL-скриптов. Преимущества такого подхода заключаются в том, что ИИ может не только подсказать нужную формулу (хотя в режиме “умной справки” GPT показывается себя превосходно), но при правильной постановке задачи выдает готовый запрос, который можно просто взять и скопировать целиком. Подобная практика уже сейчас помогает экономить до 20% времени аналитиков. А учитывая, что любой отчет или дашборд — это фактически код разметки для браузера, готовить нужные представления скоро также можно будет по запросу к GPT-модулю.
3. Умный поиск по данным и представление результатов в виде нарратива. С использованием LLM очень просто добиться генерации связного ответа на бизнес-запросы вида “Как изменилась наша дебиторская задолженность за последний месяц и почему?”. К сожалению, пока ограниченность контекста LLM не дает реализовать этот вариант на полную мощность, и его продуктивное использование остается редкостью. Однако, учитывая скорость развития LLM технологий можно ожидать, что это изменится в ближайшем будущем.

Также LLM могут помочь анализировать неструктурированную информацию — например, задачи в Jira или другом таск-менеджере, или отзывы в Интернете. С использованием LLM API и простых ETL скриптов к подобным текстовым блокам легко добавлять структурированную разметку, которую далее уже можно анализировать с помощью BI.

#### Информационная безопасность и защита данных

За последнее время данные стали безусловно самым важным активом любой организации. И помимо методов работы с ними на первый план вышли вопросы безопасности и защиты этого ценного актива. Так, по оценкам аналитиков BARC, вопрос защиты корпоративных данных является задачей №1 наравне с поддержанием высокого качества и чистоты данных.



***Рис. 3.1.14.*** *Динамика трендов в BI по версии BARC BI Trend Monitor (см. ссылку в конце главы).*

Для мира аналитики и BI — это означает серьезную переработку процессов хранения и обработки данных. От самих BI-платформ начинают требовать все больше возможностей авторизации и разделения доступа, а от хранилищ данных и инструментов ETL — защищенности и надежности.

Компании, использующие ПО из категории Open Source сталкиваются с необходимостью выстраивания дополнительного эшелона защиты. Так, например, Apache SuperSet — одна из наиболее популярных открытых систем визуализации только за 2023 год была взломана дважды: в апреле тысячи серверов по всему миру пострадали от атак RCE, а в сентябре две уязвимости позволили злоумышленникам легко захватывать управление серверами Apache SuperSet.

Тем временем, продолжающаяся автоматизация критически важных секторов экономики диктует спрос на гарантированную и стабильную работу BI-платформ поверх защищенных операционных систем, таких как, например, Astra Linux SE и сертификацию ФСТЭК.

### Локализация российского рынка BI

Начиная с 2022 года российский рынок BI испытал кардинальный поворот, когда занимавшие по различным оценкам 80-90% западные компании (Tableau, Qlik и Microsoft), а также предоставляющие функционал BI в рамках комплексных корпоративных систем (Oracle и SAP), покинули рынок и оставили своих пользователей без возможности продления лицензий, поддержки и обновления платформ.

В результате компании оказались перед выбором — запускать миграцию на одну из российских BI-платформ, внедрять своими силами или с привлечением системного интегратора решение с открытым исходным кодом или остановить выбор на одной пока еще доступных зарубежных BI-платформ, принимая риски возможного ухода вендора с рынка в дальнейшем.

Интерес заказчиков вызвал целую волну пилотных проектов, в результате которых был накоплен первичный опыт миграций. Он обнаружил следующие тенденции:

* Перенос экспертизы происходит приблизительно на 40% быстрее внедрения “с нуля”
* Миграция остается полностью в компетенции ИТ, так как бизнес-задачи уже были поставлены
* В ходе подобных проектов требуется перестройка практически всего аналитического стека, включая хранилище данных и ETL
* Использование незрелых решений создает невероятную нагрузку на ИТ-отдел и лишает аналитиков и бизнес-пользователей ожидаемой гибкости

При этом сама задача выбора платформы для миграции оказалась намного более сложной, чем ожидалось. Учитывая значительную емкость рынка, свои предложения на начало 2024 года сформулировали более 60 разработчиков ПО, многие из которых начали развивать свои продукты только в 2022 году. Были подготовлены многочисленные рейтинги, которые помогают заказчикам сформировать хотя бы первичное представление о соответствии BI-платформ их ожиданиям — в их числе TAdviser, “Круг Громова”, Cnews, It-World и ряд других (см. ссылки в конце главы).

Сквозной анализ всех представленных рейтингов говорит о том, что на рынке присутствует порядка 15 BI-платформ, с которыми уже работают различные интеграторы, и которые могут показать определенный опыт внедрения. Однако, рынок российских BI-систем в реальности все более консолидируется и по данным исследования TAdviser 2024-го года ТОП-3 вендоров занимают более 60% рынка.

Агрессивный маркетинг ряда новых игроков часто ставит заказчиков в тупик, предлагая “лучшую замену” той или иной западной платформы, основываясь, например, только на схожести интерфейса. Однако практика уже показала, что цена ошибки при выборе очень высока как для организации в целом, так и для ответственных за принятие решений лиц. Поэтому основными ориентирами для перехода с недоступных более на рынке платформ являются:

* Изучение отраслевого опыта с уже представленными кейсами
* Пилотные проекты с обязательным нагрузочным и функциональным тестированием
* Оценка проекта внедрения с учетом всех затрат, включая создание и оптимизацию DWH, обеспечение безопасности, поддержку и сопровождение системы.
* Временные затраты, необходимые для внедрения и ввода эксплуатацию определенной платформы

### Полезные ссылки

1. **R** — язык программирования, изначально разработанный для статистической обработки данных. На сегодняшний день (2020 год) – один из двух наиболее популярных языков для работы с продвинутой аналитикой, наряду с Python. Подробнее см. [Ссылку 3.1.1.](https://ru.wikipedia.org/wiki/R_(%D1%8F%D0%B7%D1%8B%D0%BA_%D0%BF%D1%80%D0%BE%D0%B3%D1%80%D0%B0%D0%BC%D0%BC%D0%B8%D1%80%D0%BE%D0%B2%D0%B0%D0%BD%D0%B8%D1%8F)
2. **Python** — язык программирования общего назначения, являющийся в том числе стандартом де-факто для продвинутой аналитики и анализа данных, благодаря развитой экосистеме библиотек для научного анализа данных. Подробнее об использовании Python для Data Scince — см. [Ссылку 3.1.2.](https://www.coursera.org/learn/python-data-analysis)
3. **«Полная карта аналитики»** — регулярно обновляемая экспертами Visiology визуализация актуальных технологий аналитической обработки данных — см. [Ссылку 3.1.3.](https://ru.visiology.su/map)
4. **Поздняя материализация** — один из способов исполнения запросов к СУБД, во многих случаях имеющий преимущество по скорости на аналитической OLAP нагрузке. Подробнее об этом и других нюансах In-Memory OLAP – см. [Ссылку 3.1.4.](https://www.youtube.com/watch?v=rzdMWVC4mjA)
5. Статья о необходимых оптимизациях для создания высокопроизводительного DWH ‘7 направлений оптимизации ClickHouse, которые помогают в BI” - см ссылку <https://habr.com/ru/companies/visiology/articles/794456/>
6. Исследование трендов BI “BARC BI Trend Monitor 2024” - см. <https://barc.com/data-bi-and-analytics-trend-monitor-2024/>
7. TAdviser - анализ рынка BI в РФ в части лицензий и услуг - см. <https://www.tadviser.ru/index.php/%D0%A1%D1%82%D0%B0%D1%82%D1%8C%D1%8F:Business_Intelligence_(%D1%80%D1%8B%D0%BD%D0%BE%D0%BA_%D0%A0%D0%BE%D1%81%D1%81%D0%B8%D0%B8)>
8. Обзор IT-World. Business intelligence 2024: выбираем российскую платформу. См. <https://www.it-world.ru/tech/choice/213355.html>
9. Исследования российских BI и ETL продуктов - Круги Громова. См. <https://russianbi.ru/>