03 мая 2018

Хранилище данных как основа создания корпоративной системы бизнес-аналитики

Что нужно держать в фокусе внимания в первую очередь, когда в компании планируется построение такой сложной и недешевой системы, как хранилище, рассказывает изданию Global CIO архитектор хранилищ данных департамента BI ГК «КОРУС Консалтинг» Илья Ситник.

С одной стороны, системы бизнес-аналитики разрабатываются и внедряются довольно давно, существуют у каждого предприятия в том или ином виде, с другой стороны, не хватает согласованной терминологии. Также на существующих аналитических системах мы часто сталкиваемся с тем, что в них отсутствует изначально заложенный фундамент, обеспечивающий управляемое функционирование и органическое развитие таких систем.

Что такое аналитические системы

Прежде всего хотелось бы рассказать про категоризацию ИТ-систем в целом. Они делятся на два больших класса — оперативные системы (OLTP) и аналитические системы (BI).

OLTP-системы автоматизируют бизнес-процессы компании, заставляют «вращаться колеса бизнеса» и могут быть так же разнообразны, как и

различные виды бизнеса и функциональные направления внутри компании. Это системы класса ERP, CRM, системы документооборота и электронного обмена данными, решения для контроля доступа, электронные торговые площадки и другие, — многообразие этих продуктов поистине безгранично. Одним словом, такие системы помогают осуществлять ежедневные операции компании и являются неотъемлемой частью ведения бизнеса. При поломке такой системы могут нарушиться ключевые бизнес-процессы — легко представить себе масштабы последствий при выходе из строя системы обработки заказов интернет-магазина либо системы регистрации пассажиров на авиарейс.

BI-системы, с другой стороны, позволяют наблюдать и анализировать результаты бизнес- процессов - следят «за вращением колес». Они не столь критичны для проведения ежедневных операций, но их значение проявляется на уровнях выше — тактическом и стратегическом. Менеджмент, вооруженный качественной и эффективной аналитической системой, может видеть как текущее состояние компании и проводить исторический анализ, так и заглядывать в будущее, делать прогнозы по развитию организации, принимать обоснованные управленческие решения.

Классификация BI-систем

Чтобы лучше разобраться в типах BI-систем давайте обратимся к истории развития аналитических систем.

Вначале автоматизировалась регламентная (стандартизованная) отчетность путем оптимизации процессов ее подготовки с помощью ИТ-системы. Такое

применение аналитической системы упрощает, ускоряет подготовку отчетности и делает ее более надежной и качественной, но принципиально не изменяет подходы к анализу данных.

С развитием технологий появилась возможность создания решений для динамического интерактивного анализа. Такие решения позволяют проводить в онлайн-режиме различные виды анализа, конструировать аналитику по произвольным атрибутам, настраивать различные фильтры, оперативно конфигурировать любой требуемый табличный и графический вид представления отчетности. Наконец, прогнозная (предиктивная) аналитика позволяет с помощью методов машинного обучения и искусственного интеллекта создавать новые данные, находить неявные закономерности, делать прогнозы на будущее, проводить what-if анализ.

В соответствии с этой классификацией можно говорить о функциональном разделении BI-систем, причем конкретная система может выполнять одну или несколько функций:

- Анализ регламентной (стандартизованной) отчетности. Подразумеваются, как правило, достаточно детальные отчеты, созданные по заданным шаблонам. Эти отчеты используют менеджеры оперативного уровня для различных целей, например, для менеджера по продажам это могут быть «план оплат на неделю» или «факт оплат на сегодня». В этом смысле аналитические системы пересекаются с оперативными системами - современные CRM-, ERP-системы также обладают встроенной системой отчетности. Различие же заключается в том, что BI-система может собирать данные из разных источников и предоставлять результаты в едином

интерфейсе.

- Динамическая, также называемая OLAP, отчетность. Такой вид анализа позволяет выявлять закономерности, тренды, оценивать влияние бизнес-факторов на основе фактических данных.
- Прогнозная и продвинутая аналитика.

Как связаны аналитические системы и хранилище данных

Хранилище данных (ХД) является основой и ядром аналитической системы. Стоит отметить, что это не просто некая база данных, хранящая данные на постоянной или временной основе и используемая в процессе подготовки аналитических материалов (что часто соответствует интуитивному и несколько вульгарному представлению о сущности ХД), а информационная система, обладающая определенными свойствами.

ХД — важнейшая часть процесса принятия управленческих решений — и это еще раз говорит о том, что аналитическая система нацелена на высокоуровневый обзор состояния компании.

ХД является предметно-ориентированной системой: данные организуются в объектную модель, отвечающую предметной области конкретной компании.

ХД интегрирует данные, что означает, что данные собираются из разных источников, но в ХД они очищаются и приводятся к «единому общему знаменателю» сущностей объектной модели, поэтому говорят о том, что ХД - это интегрированная система.

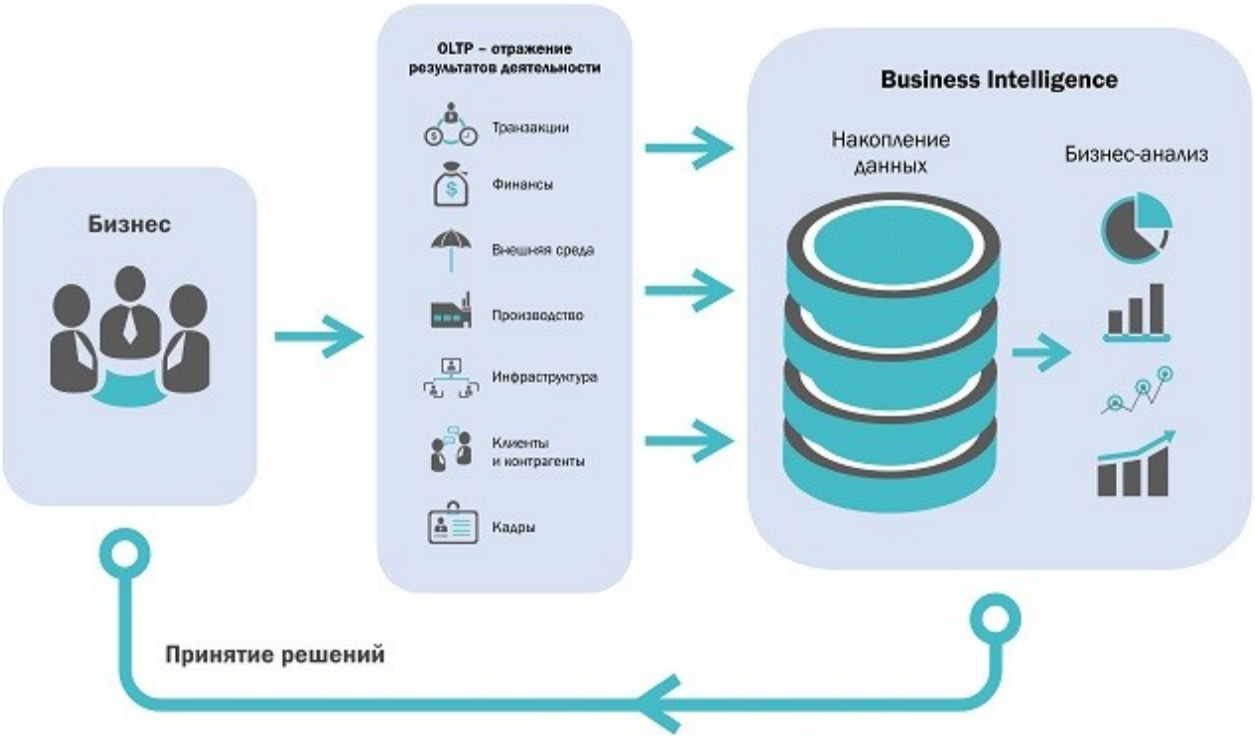
ХД неволатильно, то есть всегда достаточно статично и организовано таким образом, что обновление данных происходит за счет отслеживания изменений, произошедших в информационных системах-источниках. Например, данный принцип запрещает полную перезаливку данных при обновлении, что порой встречается на хранилищах небольшого объема. Такой подход хоть и быстрее в реализации, но приводит к проблемам в долгосрочной перспективе. В ХД должны быть предусмотрены механизмы инкрементальной загрузки данных.

ХД хранит всю истории деятельности компании. Это означает, что хранятся все данные, загруженные из любых источников. Переход с одной производственной системы на другую, архивация данных и изменение горизонта хранения данных в исходной системе не должны влиять на базовый принцип хранения данных в ХД за все время.

Важным моментом являются надежность и адаптивность ХД. Оно должно быть построено таким образом, чтобы гибко реагировать на структурные изменения в системах-источниках. ХД должно быть робастным, что означает, что изменения входящих данных определенного масштаба должны приводить к изменениям в хранилище такого же или меньшего масштаба, например, изменение или удаление какого-то поля одной таблицы не должно приводить к остановке обновления всего хранилища. Появление нового источника данных должно укладываться в существующую архитектуру, а не приводить к запуску нового проекта по переделке хранилища, а то и построению новой системы рядом со старой.

В силу того, что ХД объединяет данные всей компании, инициативы по ее внедрению должны координироваться между всеми отделами — потенциальными пользователями системы, то есть принято бизнес-средой. Каждое из подразделений, поставляющих данные в ХД и планирующее использовать результаты внедрения системы, должно быть активно вовлечено в проект развития ХД, проверять качество данных, принимать результаты на всех этапах проекта. В противном случае, легко может произойти ситуация, когда после завершения проекта результаты не используются в должной мере из-за их несогласованности с потребностями конкретных бизнес-пользователей. Здесь мы видим еще одно принципиальное различие между оперативными и аналитическими системами. Развитие первых – в первую очередь зона ответственности ИТ-отдела, вторые инициируются и развиваются как мероприятие стратегического уровня с зоной курирования топ менеджментом компании.

Таким образом, ХД решает задачу предоставления консолидированного набора данных (собранных из разных источников) о состоянии компании, дает согласованную «картину мира» в заданной временной отрезок. Это устойчивый фундамент, на основании которого можно строить и развивать аналитические системы компании, «взращивая» их в ходе эволюции управленческих процессов.



Подходы и методики по построению ХД

Наиболее известными методиками к построению ХД являются подходы Инмона и Кимбола. Метод проектирования Кимбола работает от частного к общему (Bottom-up дизайн) и означает соединение разрозненных витрин данных, построенных для решения определенных задач предметной области, в единое ХД. При этом такие витрины данных (денормализованные по определению) являются одновременно и пользовательскими базами для построения отчетности, и местом хранения данных.

Несмотря на некоторые плюсы, к которым относится относительная быстрота разработки, следование этому подходу на сложных комплексных проектах приводит к существенным проблемам, среди которых дублирование

алгоритмов и процессов загрузки, избыточность и противоречивость данных, сильная зависимость модели ХД от бизнес-требований, что приводит к волатильности модели, сложности поддержки инкрементальной загрузки и проблемам при поддержке клиентов, которым требуется не многомерная витрина данных (например, при использовании ХД как источника данных для других ИС, в том числе оперативных).

В целом можно сказать, что подход Кимбола больше подходит для пилотных проектов и небольших проектов с простой структурой источников и витрин данных.

Другой подход — методология Инмона, состоящая в концепции построения единого централизованного места хранения данных и порождении пользовательских витрин данных уже из него (Top-down дизайн). Иногда

считают, что этот подход противоположен методике Кимбола, но скорее можно говорить о том, что подход Инмона включает в себя концепцию витрин данных, глубоко проработанную Кимболом, нежели противоречит ей. Водораздел же между двумя концепциями заключается именно в наличии единой базы данных, которая является централизованным местом хранения данных и служит источником для витрин данных. Это положении крайне важно и поэтому заслуживает повторения и акцентирования внимания.

Таким образом, при проектировании ХД мы начинаем с построения объектной модели предметной области бизнеса и проекции этой модели на информационный уровень. При этом не обязательно строить всю модель и только потом приступать к построению витрин, вполне допустим итерационной подход. Несмотря на некоторые минусы, к которым относится более сложная структура ХД и ETL, дополнительные затраты на первый запуск, этот подход обладает важными преимуществами, которые в долговременной перспективе могут оказать принципиальное влияние на исход проекта в целом. К ним относятся создание объектной модели бизнеса, единый язык аналитиков, разработчиков и пользователей, возможность разделения задач между разработчиками ХД и разработчиками витрин данных и отчетности, повторное использование алгоритмов и данных, непротиворечивость данных (single version of truth), устойчивость модели данных к изменчивым требованиям к отчетности и поддержка историчности мастер-данных.

Существует еще третий, менее известный подход — data vault, который заключается в развитии идеи о централизованном хранении данных, но с большим акцентом на разнородность и изменчивость данных, то есть на обеспечение таких свойств ХД, как интегрированность и адаптивность.

В наших проектах мы стараемся учитывать специфику каждого проекта для выбора оптимальной архитектуры, но магистральный подход, которого мы придерживаемся – развитие по возможности единой интегрированной универсальной базы (в силу вышеописанных преимуществ методики Инмона), которая является ядром хранилища, и на базе которого строятся пользовательские витрины данных.

Отдельным вопросом стоят формирования требования к проекту и оценка подобных проектов, но это тема для отдельной статьи.

Источник: Global CIO