РОМАН ДАВЫДОВ

В ОМУТ С ГОЛОВОЙ?

ЧТО МОЖЕТ ДАТЬ «ОЗЕРО ДАННЫХ» БИЗНЕСУ

СОБЕСЕДНИК

РОМАН ДАВЫДОВ,
архитектор Центра внедрения бизнес-систем компании «Инфосистемы Джет»

Роман Давыдов

МЕТОДОЛОГИЯ DATA LAKE ПРИ ПРАВИЛЬНОМ ИСПОЛЬЗОВАНИИ ПОЗВОЛЯЕТ СПРАВИТЬСЯ С ОБРАБОТКОЙ И ХРАНЕНИЕМ УВЕЛИЧИВАЮЩИХСЯ ОБЪЕМОВ ДАННЫХ И СОКРАТИТЬ ОБЪЕМ ИНВЕСТИЦИЙ В КЛАССИЧЕСКОЕ ХРАНИЛИЩЕ.

Существующие в ряде компаний стандартные решения для хранения *(корпоративные хранилища данных, ХД)* и анализа информации часто не справляются с новыми объемами данных или становятся дорогостоящими в том случае, когда данные все же удается прогрузить. Высокая стоимость хранения и работ по загрузке данных в хранилища — вот насущная проблема для ИТ-менеджмента. Если прибавить к этому прогнозы, согласно которым объемы ежегодно генерируемых данных растут по экспоненциальному закону, приходится искать новые методы их хранения и обработки. Дополнительные требования к организации хранения данных предъявляют и новые методы их анализа, основанные на алгоритмах машинного обучения. Если просто брать данные из хранилища, обычно требуются промежуточные шаги в виде выгрузок в файлы, что накладно по ресурсам и по времени. Но означает ли это, что в новых условиях классические ХД будут неактуальны? Сдадут ли они все свои позиции новой методологии организации работы с данными — Data Lake?

Сначала немного теории. «Озеро данных» — это централизованное хранилище, которое позволяет хранить большие объемы необработанных данных в первоначальном формате *(файлы журналов, интернет-клики, объекты JSON, изображения, сообщения в социальных сетях)*. Это позволяет масштабировать данные, экономя время на определение их структуры и преобразований.

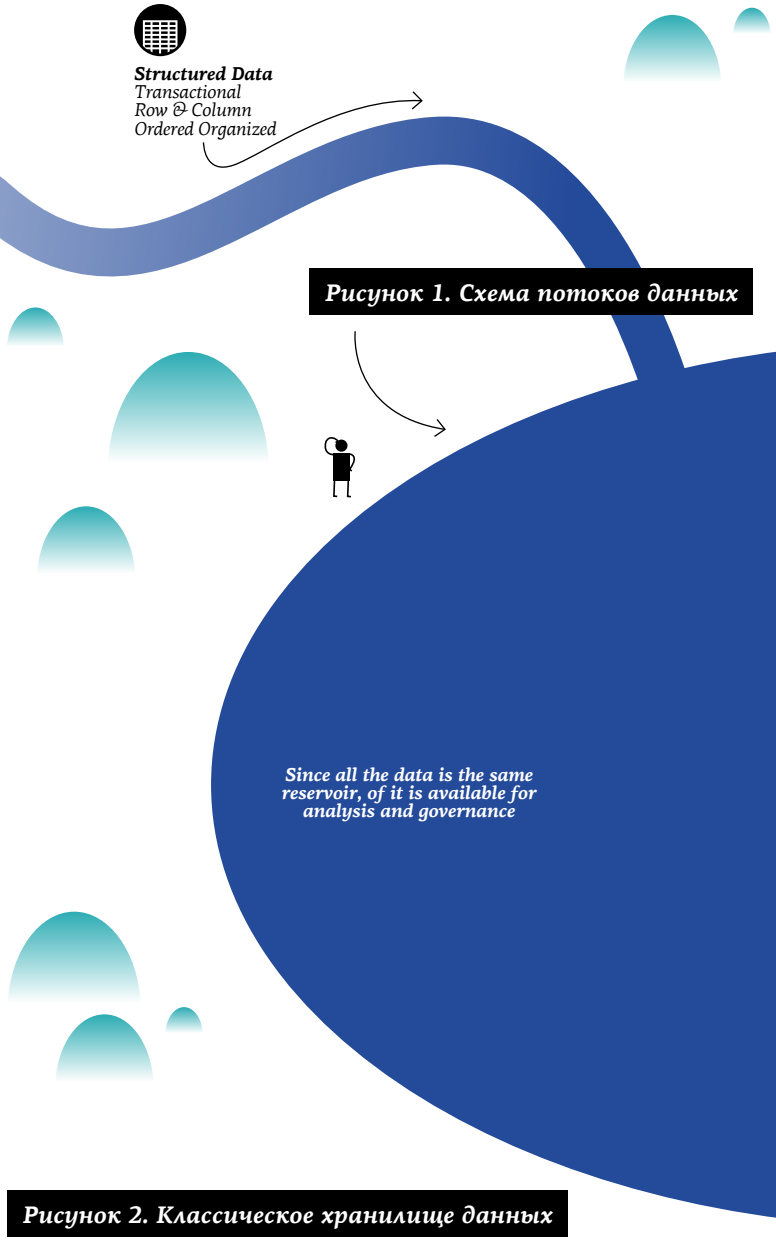


Рисунок 1. Схема потоков данных

При этом стоит отметить, что модернизация среды ХД является нетривиальной задачей, и очень немногие компании сегодня действительно готовы полностью заменить свое хранилище на Data Lake. Дело в том, что:

- Эти технологии сложны, соответственно, специалистов, обладающих нужными знаниями и способных оптимально настроить систему, очень мало.
- Средства управления Data Lake не всегда имеют удобный интерфейс.
- Требуются дополнительное обучение и повышение квалификации конечных пользователей, которые будут работать с новым форматом хранилища. Знания SQL здесь явно не достаточно.

Поэтому первым шагом на пути модернизации ХД зачастую становится создание гибридной архитектуры: дополнение существующего хранилища «озером данных». Data Lake обеспечивает большую гибкость и скорость при обработке и сборе неструктурированных, полуструктури-

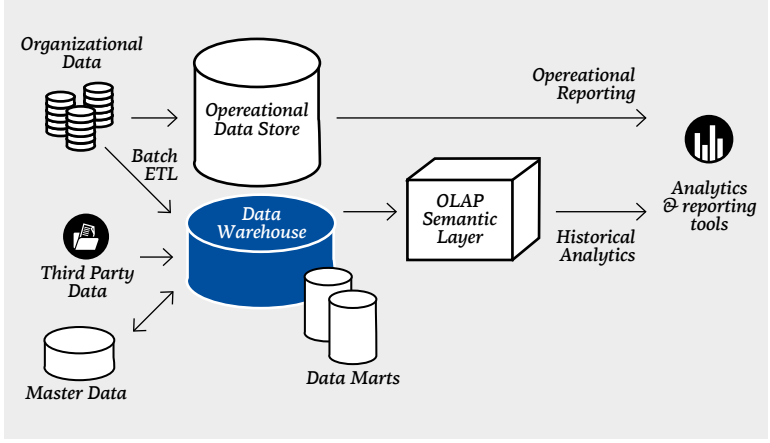


Рисунок 2. Классическое хранилище данных

рованных и потоковых данных, а также сохраняет уже реализованные потоки данных в хранилище для отчетности и бизнес-аналитики.

Подобное расширение хранилища дает ряд преимуществ. Так, обеспечение большей емкости в меньшем объеме экономит деньги: масштабируемые архитектуры могут хранить необработанные данные в любом формате за меньшую стоимость по сравнению со стандартным вариантом. Гибкая архитектура «озера данных» также обеспечивает более быструю загрузку данных и их параллельную обработку.

Не последний момент — оптимизация вычислительных ресурсов. Дорогостоящие вычислительные ресурсы сервера хранилища не должны уходить только на преобразование данных. При этом, по статистике, около 70–80% мощностей сервера используется для ETL-процессов, и на предрасчетную аналитику остается

Рисунок 3. Схема гибридного решения

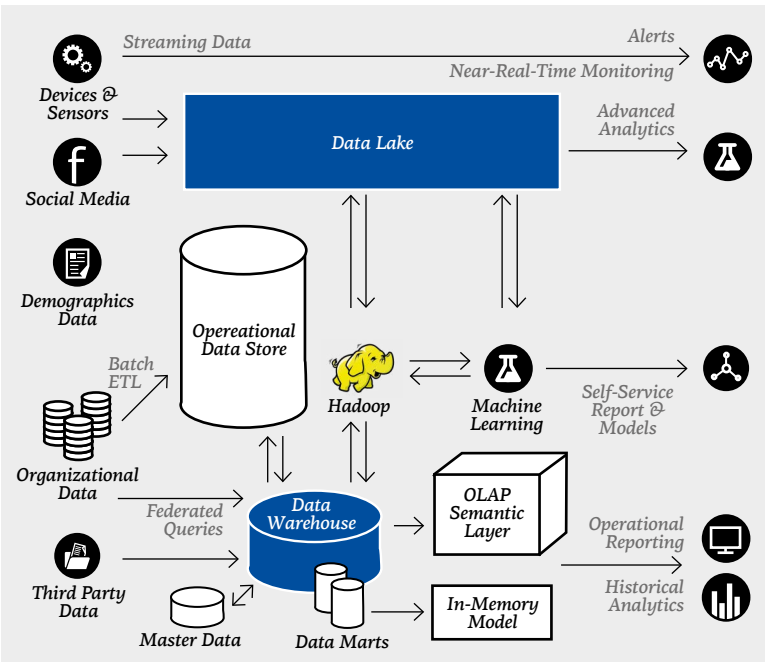


Рисунок 4. Data Lake без взаимодействия с хранилищем данных

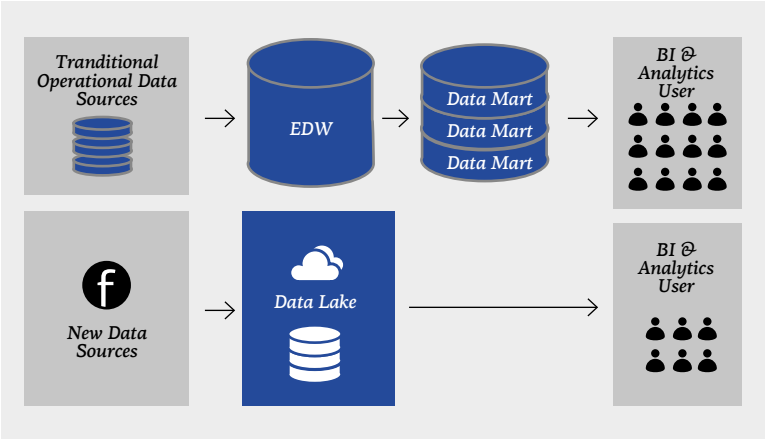


Рисунок 5. Data Lake со взаимодействием с хранилищем данных

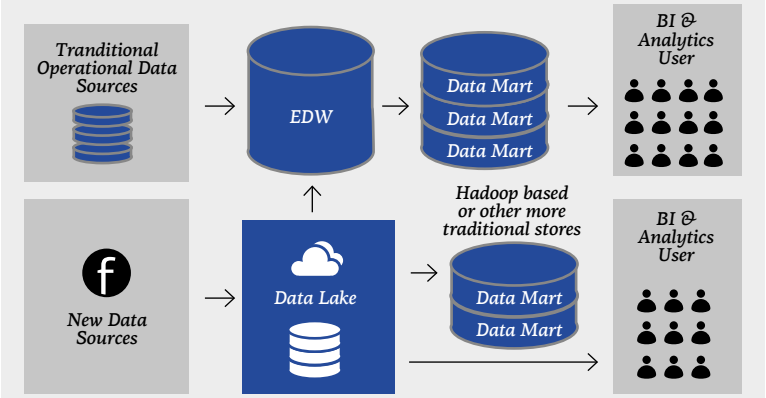
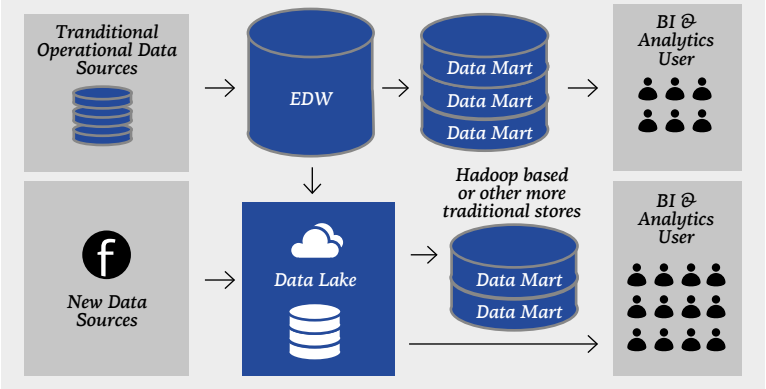


Рисунок 6. Data Lake используется как основное хранилище данных



мало вычислительной мощности. Расширение ХД позволяет более рационально расходовать вычислительные ресурсы.

Кроме того, встречаются задачи, когда нужно *максимально оперативно загрузить данные и проанализировать их*. Если пойти по пути их загрузки в ODS-область *(Operational Data Storage)* ХД, о быстром выполнении анализа можно забыть в связи с большими организационными издержками. При использовании «озер данных» эта задача решается за меньшее время.

ВАРИАНТЫ ГИБРИДНОЙ АРХИТЕКТУРЫ

Ниже представлены различные варианты вовлечения «озера данных» в экосистему BI. Data Lake может выступать в качестве отдельного нового источника данных *(вариант 1, рис. 4)*.

В этом случае «озеро данных» рассматривается как источник для построения отчетности BI, отличный от ХД. Вариант реализуют, когда нужно быстро и малыми ресурсами подключить новые источники с неизвестным качеством данных, которые будут использоваться небольшой группой пользователей *(они обычно являются заказчиками этих данных)*. Доступа к данным у остальных пользователей ХД на данном этапе нет.

«Озеро данных» может быть дополнительным источником для загрузки данных в ХД *(вариант 2, рис. 5)*.

В этом случае Data Lake используется как недорогой источник ХД *(как и в предыдущем варианте)*. После проведения анализа *(после реализации схемы 1)* выявляют значимость данных и необходимость их загрузки в ХД. При этом «озеро данных» используется как ODS для хранилища. Основная идея заключается в том, чтобы не загружать новые данные в дорогостоящее по ресурсам ХД, а проделывать эту процедуру только с реально ценными данными. При таком подходе Data Lake является также источником, из которого черпают информацию для анализа с помощью алгоритмов машинного обучения.

Возможны варианты, когда на данных Data Lake формируются свои витрины для анализа. Храниться они могут как в самом «озере данных», так и в реляционной системе. Центральным источником данных для отчетности остается ХД.

И наконец, «озеро данных» стоит рассмотреть как основной источник *(основное хранилище)* для бизнес-аналитики *(вариант 3, рис. 6)*.

Потоки данных из ХД перенаправляются в «озеро данных» и там дополняются. Такой подход позволяет уменьшить размер ХД и снизить темпы его роста. Хранилище продолжает поддерживать важные задачи, такие как нормативная отчетность для ЦБ. Но большинство задач управленческой отчетности и аналитики переносятся в «озеро данных». Преимущество такого подхода заключается в том, что Data Lake содержит больше данных, чем ХД, причем с гораздо большим объемом истории. Кроме того, оно может содержать данные более детализированного транзакционного уровня, в отличие от высоко агрегированного представления, которое мы часто видим в ХД.

В итоге можно запускать процессы отчетности на огромных объемах данных с большей историей и разрабатывать аналитические модели на более высоком уровне детализации данных. В этом случае «озеро данных» становится основным источником для большинства пользователей.

ПРИМЕРЫ РАСШИРЕНИЯ ХРАНИЛИЩА С ПОМОЩЬЮ «ОЗЕРА ДАННЫХ»

Ниже представлены реальные кейсы использования Data Lake в качестве расширения стандартного хранилища.

Первый относится к сфере ретейла. Топ-менеджмент одной компании поставил перед ИТ-подразделением задачу по использованию методов машинного обучения для более эффективного проведения маркетинговых кампаний. Но мощностей на это не хватало: большую часть рабочего времени и ресурсов сервера хранилища занимали пресловутые процессы ETL. Новая гибридная среда — «озеро данных» *(2-й вариант)* на Hadoop от Hortonworks и хранилище на Oracle — увеличила емкость хранилища в несколько раз. Кроме того, возросла скорость обработки данных. Решение сократило затраты на аппаратные ресурсы на 30%.

Второй пример относится к компании, работающей на рынке финансовых услуг. Для контроля транзакций клиентов с целью борьбы с мошенническими операциями айтишникам нужно было подключить несколько потоков транзакционных данных из разных систем-источников. Требовалось хранить данные с историей более 5 лет, что было очень дорого в случае использования ресурсов ХД. Для первичного анализа данные загрузили в Data Lake согласно варианту 1. Затем провели анализ качества данных и частично загрузили их в хранилище *(2-й вариант)* для общего пользования. В результате был обеспечен более быстрый доступ к анализу данных и построению отчетов, а также удалось сэкономить ресурсы по разработке ETL-процедур при последующей загрузке части данных в ХД.

ЗАКЛЮЧЕНИЕ

Методология Data Lake при правильном использовании позволяет справиться с обработкой и хранением увеличивающихся объемов данных и сократить объем инвестиций в классическое хранилище. Возможность использовать методы машинного обучения на данных Data Lake будет дополнительным драйвером их внедрения.

В свою очередь, ХД не потеряют своей актуальности. Они будут постепенно отдавать данные для анализа в «озера», но все-таки оставят себе самую важную отчетность. По мере улучшения средств реализации Data Lake и накопления опыта специалисты компании будут плавно переходить от варианта 1 к варианту 2. Вариант 3 в текущих условиях для большинства отечественных компаний недостижим, при этом он является той целью, к которой, по нашему мнению, все будут стремиться при развитии своих BI-стратегий.