Big Data как отрасль:
идеи, перспективы, основные тренды

Ольга СВИТНЕВА,
менеджер продукта, VK Cloud

Понятие Big Data, или большие данные, обрело большую популярность и широкое применение. Появился термин еще в 2008 г., и с тех пор его значение и трактовка значительно расширились. Сегодня можно говорить о Big Data как о классе комплексных платформенных продуктов и даже самостоятельной отрасли программного обеспечения. Как удалось Big Data пройти путь от буквального понимания («большой объем данных») до комплексного определения, объединяющего широкий спектр программных продуктов и технологий?

Большие данные в современном контексте

Дословный перевод термина говорит об объемах данных, но это не единственная и не основная черта отрасли Big Data. Большие данные предполагают нечто большее, чем просто хранение или анализ значительных объемов информации. Закономерный вопрос: что считать большим, значительным в применении к величине данных в мире, где поток информации разнообразен и с трудом поддается подсчетам?

В сущности, понятие больших данных подразумевает работу с информацией не просто огромного объема, но и разнообразного состава, в том числе находящуюся в разных источниках, обновляемую и зачастую плохо структурированную, например с видео- и аудиозаписями, текстовыми документами, веб-информацией и потоками сообщений из соцсетей, метеорологическими данными и координатами геолокации.

Значительный объем данных генерируется за пределами компаний, потребляющих их. Поэтому есть потребность в создании связей между данными в их корректной интерпретации. Наиболее популярные источники больших данных:

- корпоративные системы: транзакции в базах данных и файловых хранилищах;
- интернет-сайты, соцсети, СМИ, IoT;
- показания приборов реального мира (датчиков, сенсоров, регистраторов).

Все это значительно расширяет горизонт больших данных. Согласно определению, предложенному консалтинговой компанией Forrester, *«большие данные объединяют техники и технологии, которые извлекают смысл из данных на экстремальном пределе практичности»*.

Как распознать большие данные

Поскольку объем не является точным критерием того, что такое большие данные, как же распознать большие данные?

Правило 8V

Наиболее популярный подход – правило 8V, согласно которому в определении Big Data должны фигурировать восемь характеристик: объем (Volume), скорость (Velocity), разнообразие (Variety), достоверность (Veracity), изменчивость (Variability), жизнеспособность (Viability), визуализация (Visualization), ценность (Value). При этом каждая V важна в совокупности с другими.

Первые три V наиболее очевидны. Объем информации увеличивается по экспоненте, данные постоянно обновляются и темп их обновления приобретает все больший вес в Big Data-проектах. То есть нужно успевать за скоростью, с которой данные создаются, и анализировать их в режиме реального времени. Разнообразие данных означает, что в проектах Big Data информация разных форматов, как структурированная, так и неструктурированная: текстовая, графическая, аудио и видео. Каждый тип данных требует соответствующего анализа и стека инструментов для их обработки.

По мере развития технологий расширился набор характеристик для отрасли Big Data и в определение были добавлены **Veracity, Variability, Viability, Visualization, Value**. Остановимся на них подробнее.

Veracity – достоверность. Это одна из важнейших характеристик в цепочке 8V, поскольку Big Data определяют не столько объемы,

сколько качество и пригодность данных для последующего принятия решения, что невозможно без их достоверности. Крайне важно заранее убедиться, что данные корректны, для этого существует отдельный класс решений – Data Quality и Data Management. Это еще один аргумент в пользу того, что Big Data – самостоятельная отрасль программных решений.

Variability – изменчивость. Значение одних и тех же данных может различаться в зависимости от контекста или области их применения. Изменчивость подразумевает наличие в отрасли алгоритмов, которые заточены под понимание контекста, и в состоянии расшифровать значения данных в конкретном смысловом приложении.

Второй аспект Variability – переменчивость во времени и под воздействием обстоятельств. Это значимый признак данных при проведении пролонгированных исследований и прогнозов.

Viability – жизнеспособность данных, причем в плане возможности их хранения, а также актуальности. Моментальное устаревание данных диктует потребность интенсивного развития стека стриминговых технологий, заточенных на анализ данных в режиме реального времени.

Visualization – визуализация (на дашбордах, панелях мониторинга, графиках и т. п.). Цель визуализации – сделать их читабельными и доступными для восприятия, для чего предназначен отдельный класс ПО, обозначаемого термином Business Intelligence.

Value – ценность. Имеется в виду извлечение максимальной пользы из результатов анализа больших данных. Это аспект применения данных – как трактовать массивы информации для извлечения коммерческой ценности. Важно, как использовать данные, удастся ли компании опираться на идеи, полученные из аналитики.

Задачи больших данных

Характеристики отрасли Big Data дают понимание, что речь идет о решении сложных задач с множеством переменных и часто нетривиальными условиями, в частности, такими как:

- очистка данных, их категоризация и обогащение (краудсорсинг);
- смешение и интеграция разнородных данных (цифровая обработка сигналов и обработка естественного языка);
- выявление закономерностей, обучение ассоциативным правилам, кластерный и регрессионный анализ (Data Mining);
- прогнозная аналитика;
- имитационное моделирование;
- машинное обучение, искусственные нейронные сети, сетевой анализ, методы оптимизации и генетические алгоритмы;
- распознавание образов.

Принципы работы с большими данными

Особенности отрасли формируют принципы и подходы к работе с данными. Рассмотрим основные из этих принципов.

Принцип горизонтальной масштабируемости. Увеличение количества физических или виртуальных вычислительных узлов, чтобы ускорить обработку информации. Чем больше поток, тем больше мощности задействуется.

Принцип отказоустойчивости. Позволяет ИТ-инфраструктуре продолжать работу, предотвращая сбои, вызванные точкой отказа. Чем больше объем данных, тем больше требуется технических мощностей на их обработку и тем выше риск сбоя.

Принцип локализации. Как правило, хранятся данные на одних серверах, а обрабатываются на других. По мере увеличения объема информации растут затраты на ее передачу. Потому одна из проблем отрасли Big Data – оптимизация доставки.

На пути к демократизации данных

Big Data сегодня – это уже не конкретная технология, а самостоятельная отрасль ПО. В том числе это технологии хранения больших объемов структурированных и неструктурированных данных, технологии загрузки, обработки и моделирования данных, отслеживания потока данных, управления их качеством, предоставления данных потребителю. И это далеко не полный перечень, но базовый из стека отрасли Big Data, предполагающий взаимодействие множества специалистов разной квалификации. Основной фокус отрасли – организовать непрерывный и беспрепятственный доступ к данным всем специалистам, чтобы извлечь полезную информацию.

Популярный сейчас тренд – Data Operations (или DataOps). Речь идет об интеграции аналитики, разработки и эксплуатации при работе с большими данными, или «DevOps для Big Data».

DataOps – это концепция, набор практик непрерывной интеграции данных между процессами, командами и системами.

В приоритете непрерывная доставка аналитических знаний и ориентация на удовлетворенность клиента. Характерная черта DataOps – кросс-функциональные группы, включающие специалистов по эксплуатации, программной инженерии, разработке архитектуры, планированию и управлению продуктами, а также специалистов по подготовке и исследованию данных. Таким образом, синергия специалистов в одном жизненном цикле работы с данными обеспечивает наибольшую эффективность.

Инфраструктура DataOps насчитывает пять основных элементов:

1. Технологии работы с данными.
2. Адаптивная архитектура, позволяющая непрерывно совершенствовать технологии и процессы.
3. Процессы обогащения данных.
4. Методологии для построения аналитики и развертывания конвейеров данных.
5. Культура и люди.

Данные не статичны, одних технологий их обработки недостаточно. От компаний требуются маневренность и приверженность культуре работы с данными.

Из этого вытекают следующие принципы DataOps:

- думайте о сервисах, а не о серверах;
- инфраструктура работы с данными – это код;
- автоматизируйте все;
- не забывайте про DevOps: компания, управляемая данными, – это DataOps и DevOps, реализуемые по принципам Agile.

Реализация DataOps

На практике реализация такого подхода требует платформенного продукта, объединяющего тех, кто создает данные, и тех, кто их использует. Платформа по работе с Big Data представляет собой комплексную экосистему интегрированных между собой масштабируемых сервисов по работе с данными, в том числе для их хранения, обработки, анализа, визуализации, моделирования, управления качеством. При этом платформа как продукт нацелена на решение таких задач бизнеса, как:

- упрощение и повышение точности планирования на основе данных, в том числе за счет реализации подходов Data first и Data driven;
- увеличение скорости запуска новых проектов, поскольку платформенные сервисы интегрированы между собой и охватывают весь жизненный цикл работы с данными;
- повышение качества клиентских сервисов, эффективности работы с поставщиками и контрагентами благодаря быстрой и точной аналитике, прогнозированию.

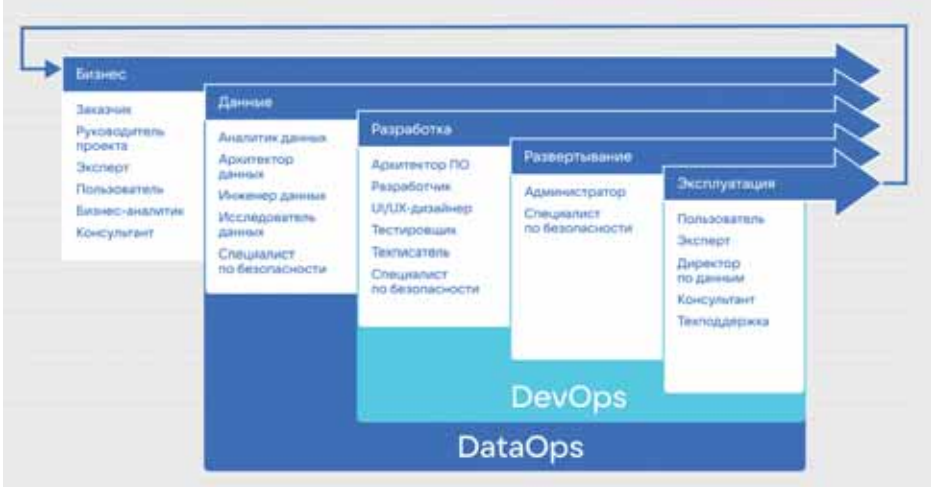
Помимо этого платформа должна соответствовать регуляторным принципам: импортозамещение, соблюдение законодательства по обработке персональных данных, обеспечение информационной безопасности.

Не менее важно наличие в составе платформы инструментов самообслуживания для демократизации данных. Это необходимая составляющая для реализации DataOps-подхода, а также технологий и процессов, обеспечивающих быстрый Time-to-market решений для бизнеса. Таким образом, Big Data – не только самостоятельная отрасль ПО, технологий, процессов, но и комплексная платформа управления большими данными.

Именно логическая и технологическая связанность всех модулей платформы позволяет создать среду для организации взаимодействия участников бизнеса на всем жизненном цикле создания и использования данных. Следующая важная составляющая отрасли – цепочка создания ценности данных.

Цепочка создания ценности

В упрощенном виде цепочку создания ценности данных можно представить в виде схемы.



Цепочка создания ценности данных – это механизм, который определяет набор повторяющихся процессов для извлечения ценности данных шаг за шагом на протяжении их жизненного цикла, от необработанных данных до аналитических и пригодных для принятия управленческих решений. Цепочка состоит из нескольких этапов:

- генерация данных;
- сбор, обработка, проверка и хранение;
- моделирование, анализ и визуализация;
- обмен: предоставление выходных данных для использования внутри компании и, возможно, за ее пределами с партнерами и клиентами.

Эффективность цепочки обеспечивают процессы агрегирования и эксплуатации данных, что достигается благодаря платформенности решений. Достижения в области цепочки создания ценности позволяют монетизировать данные, т. е. получать прибыль, причем это может быть как продажа данных напрямую, так и косвенное использование для создания ценности. Эта идея согласуется с экосистемой больших данных, которая позволяет на практике извлекать ценность и прибыль из информации.

Модульная архитектура платформы

Платформа больших данных сложна в проектировании, и, хотя сегодня ведется много разговоров на тему «как стать Data-driven-company», результатами могут похвастаться немногие. Платформа должна позволять прозрачно управлять движением потоков данных. Для этого она должна быть построена по модульному принципу, где каждый модуль обеспечивает решение конкретных задач и нативно интегрируется с другими модулями. Среди таких модулей стоит отметить следующие.

1. **Модуль захвата и обработки данных**. ETL/ELT-инструменты, позволяющие собирать, доставлять, реплицировать, очищать, преобразовывать, дедуплицировать, сортировать и выполнять другие функции обработки данных. Оркестраторы для планирования заданий, организации конвейера данных, определения зависимостей, мониторинга, а также готовые инструменты интеграции данных. Загрузка и обработка данных должна предусматривать различные сценарии: пакетный режим, микробатчи, стриминг, потоковая загрузка данных на основе захвата изменений и т. п.
2. **Модуль хранения**. Технологии реализации хранилищ, озер и витрин данных, в том числе реляционные, документоориентированные, графовые, колоночные, поисковые, In-memory, Time-series базы данных. Обработка любого типа нагрузки на чтение и запись (OLAP, OLTP, Streaming) должна быть решена инструментарием данного модуля. Он, в свою очередь, должен обеспечивать реализацию необходимого подхода построения платформы данных: традиционного, Advanced architecture, Modern Data architecture, Lambda architecture, Data Mesh architecture.
3. **Модуль управления данными**. Инструменты отслеживания потока данных, управления их качеством, дата-каталоги для организации доступа к качественным и интегрированным метаданным, а также бизнес-глоссарии и централизованные унифицированные справочники. Значимость модуля подчеркивается его архитектурной особенностью – надстройкой над прочими модулями платформы, по сути, весь поток данных контролируется этим модулем.
4. **Модуль аналитики**. Средства анализа и визуализации данных, инструменты отслеживания метрик и формирования прогнозов, ситуационные центры.
5. **Модуль машинного обучения**. Среда для работы специалистов по исследованию данных, инструменты создания статистических моделей для нахождения закономерностей на основе массивов данных, инструменты подготовки и обучения моделей, а также широкий спектр средств и методов применения искусственного интеллекта.

Компонент	Назначение	Возможная реализующая технология
Загрузчик данных	Средство доставки данных до хранилища	Apache NiFi, Apache Flink
Шина обмена данными	Система, через которую будет проходить обмен данными, в том числе в режиме реального времени	Apache Kafka – брокер сообщений с горизонтальным масштабированием и высокой пропускной способностью
Слой хранения данных	Целевая система, в которую загружаем данные	Apache HDFS + Hive, S3, ClickHouse, Greenplum, Tarantool, Apache Kudu, Apache Impala
Вычислительный движок	Позволяет делать различные фильтрации, сортировки и прочие операции	Apache Flink, Apache Spark
Оркестратор	Связывает весь процесс воедино, организуя многоэтапную обработку данных	Dagster, Apache Airflow
Каталог данных	Интегрированные качественные метаданные	CKAN DataHub Magda
Data Lineage (сервис отслеживания потока данных)	Показывает подробные сведения о потоке данных от системы-источника к системе-приемнику и позволяет отследить преобразования и взаимосвязи как технических, так и бизнес-метаданных	Open Lineage
Data Quality (сервис качества данных)	Выполнение задач управления качеством данных, включая исправления, дополнения, стандартизацию и устранение дубликатов данных, исправление ошибок и настройку бизнес-правил. Ключевая задача сервиса – очистка данных	Great Expectations
Аналитика и BI-системы	Средства подготовки, комбинации и визуализации данных	Apache Superset, Dremio, Apache Drill ML- и DS-сервисы (машинное обучение и среда для специалистов Data Science)

Безусловно, модулей может быть больше и они могут быть узко заточенными под сценарии конкретного бизнеса. Однако такой базовый набор модулей дает представление о платформе как экосистеме интегрированных сервисов, цель которой – превращение данных в стратегический ресурс компании и извлечение из них прибыли.

Рассмотрим примеры технологий Open Source-стека для ряда задач из отрасли больших данных.

Большие данные – логическое следствие значения, которое цифровые технологии приобрели в нашей жизни, где данные множатся с беспрецедентной скоростью, представляют огромную ценность для тех, кто может справиться с их масштабом и раскрыть заложенные в них потенциал и знания. Поэтому Big Data как самостоятельная отрасль не только предлагает более эффективные методы работы с большими данными, но и выступает в роли комплексного платформенного продукта. ■