Big Data:
распознать и применить



Андрей ТАМБОВСКИЙ,
директор по технологиям, «ФОРС Дистрибуция»

Сложно представить современную компанию, которая не использует в своей работе информационные технологии. Системы класса ERP, CRM, HRM и многие другие – необходимые инструменты эффективного выстраивания и управления бизнес-процессами. Все большее внимание уделяется системам бизнес-аналитики, или Business Intelligence (BI), предназначенным для работы с данными. Зачастую от того, насколько быстро, точно и понятно BI-системы предоставляют информацию, зависит возможность принятия верных управленческих решений. В настоящее время динамику рынка BI определяют новые технологии, одна из них – Big Data – вызвала настоящий бум, открыв новые возможности в этом направлении.

Количество, переходящее в качество

Термин «большие данные» появился сравнительно недавно и довольно долго использовался в узких отраслях, в первую очередь в научных проектах, связанных с накоплением больших или сверхбольших объемов данных, появляющихся с высокой скоростью. Количество данных в мире стремительно растет и, по прогнозам аналитического агентства IDC, к 2018 г. достигнет 40 зеттабайт, что эквивалентно 40 трлн Гб и в 57 раз больше, чем количество песчинок на всех пляжах поверхности Земли. В ответ на растущую лавину информации возникла необходимость в решениях, позволяющих работать с большими объемами данных.

В эпоху информационных технологий, особенно после появления и активного освоения социальных сетей, накапливается значительное количество информации по каждому пользователю. Со стороны бизнеса возникла потребность эффективно использовать эти данные. Поставщики корпоративных ИТ-решений, отреагировав на новый запрос рынка, обратили внимание на технологии работы с большими объемами информации. Это положило начало широкому обсуждению новых технологий. Возникшие дискуссии были направлены на понимание определения больших данных и того, какое место они займут в устоявшейся архитектуре корпоративных информационных систем.

Сегодня классическими признаками больших данных принято считать:

- большой объем (Volume). Когда информации слишком много, процесс ее обработки и хранения традиционными способами становится трудоемким, что приводит к усовершенствованию технологий и инструментов;
- скорость появления (Velocity). Причем высокая скорость не только накопления, но и обработки данных, поскольку многим компаниям необходима информация в режиме реального времени;
- большая разнородность данных (Variety). Наличие как структурированной, так и неструктурированной информации разных форматов. Неструктурированная информация требует комплексного анализа перед обработкой;
- достоверность данных (Veracity). Особая значимость придается достоверности полученных данных. Так, например, необходимо разделять действия, выполняемые в сети людьми и роботами, что создает дополнительные сложности при анализе;
- целесообразность (Value). Многие компании стали задумываться о ценности накопленной информации, которая может помогать в усовершенствовании текущих процессов или оптимизации затрат.

Очевидно, что технологии Big Data позволяют не просто собирать и хранить данные – для этих задач давно используются другие решения. Например, архивное хранение на магнитной ленте – одна из старейших технологий в ИТ. Применение технологий Big Data предполагает возможность качественного анализа данных. Если классические системы аналитики связаны со сложностями в виде необходимости

передачи данных от систем хранения к серверам, то технологии больших данных используют для хранения серверы, что позволяет хранить и обрабатывать данные в одном месте.

Преимущество систем аналитики больших данных – используемый в них принцип массивно-параллельной архитектуры, согласно которому множество независимых серверов решают общую задачу. Подобная архитектура обеспечивает высокий уровень отказоустойчивости – при выходе из строя одного из узлов, на котором хранятся данные, не нарушается функциональность системы в целом. В этом случае количество переходит в качество, а производительность системы аналитики повышается путем простого добавления новых серверов. Причем использовать можно не самые дорогие серверы, что оптимизирует расходы по хранению и обработке данных.

Традиционные BI-системы: заменить нельзя использовать

Несмотря на то что многие характеристики систем на базе технологий Big Data совпадают с описаниями классических систем бизнес-аналитики, они существенно различаются.

В традиционных BI-системах большое внимание уделяется структуре данных и их качеству: для получения непротиворечивых и точных результатов строятся хранилища данных. В таких системах обрабатываются факты, результат обработки жестко детерминирован, а вид отчетов задается на этапе проектирования. Пользоваться традиционными BI-системами могут либо сотрудники компании, для которых создавался тот или иной отчет, либо профессиональные бизнес-аналитики, понимающие структуру источников данных и способ построения отчетов. Стоит также отметить взаимосвязь данных, поэтому потеря или отсутствие даже небольшой их части приводит к тому, что BI-система становится неактуальной для решения своих задач.

Если традиционные BI-системы «загоняют» пользователя в рамки линейного мышления и предопределенных вопросов, то системы, работающие на базе технологий больших данных, позволяют организовать поиск по принципу «не знаю, что ищу». Следовательно, поиск данных может осуществлять человек, не имеющий представления о том, как и в каких источниках хранятся данные, как они организованы и т. д. Фактически поиск может быть контекстным, знакомым сегодня практически всем по обычным поисковым системам в Интернете. Еще одно отличие систем Big Data – использование методов из теории математической статистики и моделирования. При обнаружении трендов, тенденций и других процессов, наличие которых можно выявить путем анализа больших данных, отсутствие или неточное, неполное указание данных не является критичным для получения объективного результата.

Однако за последние годы стало понятно, что традиционные BI-системы и системы Big Data – это не конкурирующие, а взаимодополняющие технологии. Появился даже термин «обогащенные данные». Иными словами, данные, полученные в результате работы систем Big Data, предоставляют дополнительные, расширенные характеристики, которых нет в корпоративных системах или хранилищах, поскольку они были получены из внешних источников.

По принципу целесообразности

Долгое время потенциальные пользователи технологий Big Data занимались в большей степени изучением самих инструментов, позволяющих получать интересную для бизнеса аналитику на основе тех данных, которые раньше не использовались или были задействованы в незначительной мере. Нередко проекты осуществлялись собственными силами ИТ-подразделений компаний, поскольку рынок не предоставлял широкого выбора готовых или настраиваемых решений такого класса от независимых разработчиков. В настоящее время можно утверждать, что пользователями и независимыми разработчиками накоплен опыт, позволяющий строить BI-системы с применением технологий Big Data.

> Традиционные BI-системы и системы Big Data – это не конкурирующие, а взаимодополняющие технологии.

Можно выделить два основных направления, в которых используются большие данные исходя из признака Value (целесообразности):

- решение задач по снижению рисков и возможных издержек для компании, выявление, прогнозирование и предупреждение ситуаций и факторов, которые могут повлечь их за собой. Системы, предназначенные для выявления попыток мошеннических действий (антифрод) или проведения сложного анализа событий, оправдывают затраты на их развертывание, поддержание и развитие за счет снижения прямых или возможных убытков;
- решение задач по повышению эффективности маркетинговых инструментов и активностей,

в частности, увеличение продаж за счет правильного таргетирования рекламных кампаний. В данном случае главный фактор, позволяющий оценить целесообразность использования Big Data, – возможность оценить эффективность маркетинговой деятельности, применив современные методы привлечения целевой аудитории к проводимым рекламным акциям (адресные рассылки, сегментация с геопозиционированием и др.).

Big Data: искусство применения

Важным элементом, характерным для сегодняшнего развития технологий Big Data, стало появление готовых программно-аппаратных комплексов, представленных на рынке ведущими поставщиками ИТ-решений. Преимущества таких комплексов по сравнению с самостоятельным развертыванием, сопровождением и развитием систем для поддержки проектов Big Data заключаются в скорости развертывания системы и появлении возможности получения от производителя, во-первых, технической поддержки аппаратных средств и программных компонентов, во-вторых, обновления программного обеспечения из единого источника и, в-третьих, дорожной карты развития системы аналитики по мере увеличения нагрузок на нее.

При переводе проектов из режима исследований в промышленную эксплуатацию возникают вопросы об обеспечении отказоустойчивости бизнес-критичных систем, о защите хранимых данных, а также о создании сред для разработки и тестирования. Использование готовых программно-аппаратных комплексов позволяет решать все эти вопросы в максимально короткие сроки и с минимальными рисками.

Среди решений корпорации Oracle, одного из признанных лидеров в области систем управления базами данных, имеется готовый программно-аппаратный комплекс (ПАК) для работы с большими данными – Oracle Big Data Appliance. Кроме вышеупомянутых преимуществ готовых решений ПАК Oracle Big Data Appliance открывает ряд дополнительных возможностей, в частности использование технологии Oracle Big Data SQL, которая делает общей точкой входа базу данных Oracle. Из нее можно обращаться к данным при помощи языка Oracle SQL независимо от того, где они находятся – в кластере Hadoop, реляционной или NoSQL-базе данных.

Немаловажное достоинство ПАК Oracle Big Data Appliance – появление приложений от независимых производителей ПО. Как правило, такие решения проходят процедуру проверки на совместимость с ресурсами ПАК. Кроме того, разработчики поддерживают модернизацию своих решений по мере обновления базовой системы. В результате заказчики могут выбирать готовые эффективные решения, а также использовать опыт и экспертизу компаний, специализирующихся на технологиях Oracle. Это позволяет заказчикам сокращать время внедрения решений, что особенно важно с учетом темпов развития современных технологий, в том числе для работы с большими данными.

> Согласно данным IDC, среднегодовой рост мирового рынка технологий Big Data составляет 31,7% и, по прогнозам, к 2018 г. достигнет 41,5 млрд долл.

Согласно данным IDC, среднегодовой рост мирового рынка технологий Big Data составляет 31,7% и, по прогнозам, к 2018 г. достигнет 41,5 млрд долл.
Что касается российских реалий, то интерес к новым технологиям повышается, однако на пути к практическому их применению очень часто возникает вопрос о том, какие результаты можно будет получить. При всех достоинствах систем Big Data достижение преимуществ от их использования зависит прежде всего от двух факторов: умения сформулировать бизнес-задачу, или так называемое datascience, и скорости внедрения решения. Применение крупным и средним бизнесом готовых решений ПАК позволяет достичь максимального эффекта от технологий Big Data. Компании, которые заинтересованы в использовании этих технологий, но не готовы инвестировать в создание собственных ресурсов, могут присмотреться к возможностям, предоставляемым облачными провайдерами.

В компании «ФОРС Дистрибуция» создан Центр компетенций по технологиям Big Data, на базе которого заказчики и партнеры – разработчики ПО совместно с экспертами «ФОРС Дистрибуция» тестируют различные продукты и решения. Базой Центра служат ПАК Oracle: Exadata Database Machine для создания хранилищ, использующих традиционную технологию СУБД Oracle, ExalyticsIn-Memory Machine – для работы приложений Oracle BI, Oracle Big Data Appliance – для обработки больших данных. Кроме того, в Центре представлен полный стек продуктов Oracle, которые можно использовать для создания практических решений, сооответствующих широкому спектру задач заказчика. ■