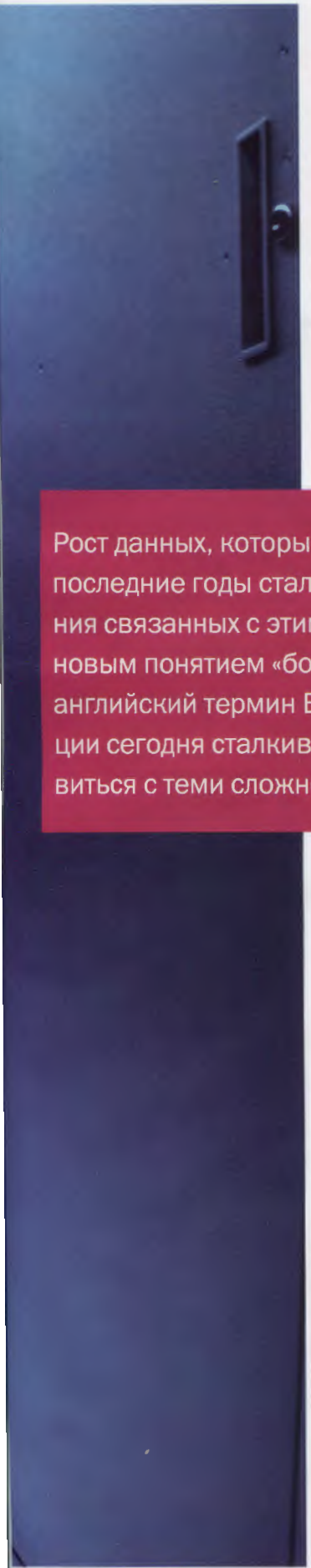
Наталья Жилкина/
nzhilkina@computerra.ru/

BIG DATA: изменение ландшафта



Прежде чем говорить о проблемах больших массивов данных, надо определиться с самим понятием. Популярный термин BIG DATA многие используют для обозначения экспоненциального роста, проблемы доступности и использования информации в ИТ-ландшафте завтрашнего дня, которому сопутствует накопление огромных массивов данных.

Однако в ИТ-индустрии нет пока общей договоренности о том, что понимать под термином. Многие эксперты считают, что это понятие надо соотносить не с объемом, а со скоростью роста данных, при которой предприятие не успевает реагировать на возникающие проблемы.

Рост данных, которыми приходится управлять организациям, в последние годы стал настолько впечатляющим, что для описания связанных с этим проблем специалисты стали оперировать новым понятием «большие данные». Более точно его отражает английский термин BIG DATA. Это явление, с которым организации сегодня сталкиваются все чаще, не будучи в состоянии справиться с теми сложностями, которые ему сопутствуют.

Термин BIG DATA традиционно используется для описания массивных объемов данных, которые анализируются сверхкрупными организациями вроде Google либо коллективами, работающими над грандиозными научными проектами, – такими как NASA.

– Однако для большинства корпоративных структур значение термина BIG относительно: все зависит от размера организации, – поясняет Тобиаш Ратай, руководитель российского офиса компании Teradata. – В большей степени речь идет о поиске новой ценности как внутри традиционных источников данных, так и за их пределами. Две трети фирм, опрошенных компанией Tech Target, обеспечивают онлайн-доступ к данным, сохраняющим свою ценность на протяжении более года. А 43% опрошенных поддерживают такой уровень доступа к ценным данным более трех лет.

Большие объемы данных сегодня поступают из различных источников. Крупным «каналом поставки» BIG DATA являются медицинские учреждения, где хранятся карты пациентов, результаты томографических исследований, рентгенограммы и проч.

Технологические прорывы в сфере медиаданных (видео высокого разрешения и 3D-технологии) способствуют резкому увеличению их объема.

«Быстрый рост данных наблюдается в сфере телевещания и киноиндустрии: с развитием цифровых технологий и переходом от пленки к компьютерной анимации, визуализации и рендерингу на специализированных рабочих станциях, объемы цифрового контента многократно возросли, – рассказывает Иван Ермаков, руководитель отдела поддержки приоритетных заказчиков HP в России. – При этом, некоторые технологии требуют просто наличия большого количества свободного дискового пространства, а некоторые еще и увеличения скорости доступа и возможности одновременной обработки миллионов объектов».

Огромные объемы данных обрабатываются в области фармацевтики, генной инженерии, нефтегазодобычи. Еще одна область работы с BIG DATA – тепловыделяющие элементы ядерных реакторов (ТВЭЛы).

Теория относительности

Понятие BIG DATA сегодня во многом зависит от того, с чего начинает компания. 23% респондентов, опрошенных Tech Target, уже управляют более чем 10 терабайтами данных для получения аналитических срезов, а у одной трети организаций, как ожидается, будет параллельно работать сто и более пользователей. Таким образом, данные начинают работать в связке BIG DATA и BIG USAGE (крупномасштабное использование), принося новую ценность.

По информации TDWI Research, во многих компаниях сегодня стремительно расширяются объемы хранилищ данных (Data Warehouse). В 2009 году 62% организаций имели менее 3 ТБ данных в своих хранилищах. К 2012-му 59% этих же организаций оценивают свои хранилища более чем в 3 ТБ, а 34% заявляют, что будут иметь свыше 10 ТБ.

Тобиаш Ратай убежден, что BIG DATA – это принципиально новое направление в индустрии, и практический интерес к этой теме чрезвычайно высок. «У людей зачастую существует путаница в том, что же на самом деле означает понятие BIG DATA», – говорит он. Вот некоторые характеристики понятия: это другая аналитика, это разные структуры данных, это многообразие аналитических отчетов, которые требуются в работе. BIG DATA означает различные обязательства, которые прежде никогда не существовали.

«Однозначно сказать, что такое „большой объем данных", можно лишь с учетом масштаба конкретного потребителя, с привязкой к определенному времени, поскольку даже для одной компании, но в разные временные промежутки этот термин будет характеризовать различные величины», – считает Алексей Пилипчук, технический директор компании «Техносерв».

Владимир Кирсанов, технический директор департамента корпоративных систем компании «Астерос», отмечает, что возникновение понятия BIG DATA связано именно со стремительным разрастанием информации в компаниях. «В настоящее время ее объемы достигают

не просто терабайтов, а петабайтов и даже экзабайтов, – делится он наблюдениями. – Как и „облака", этот термин трактуется пока весьма неоднозначно. Зачастую он используется в любом контексте при работе с данными большого объема».

В концепции специалистов «Астероса» понятие BIG DATA применимо не просто к колоссальному объему данных, но и к массивам информации, обработка и анализ которых требует нетривиального подхода, нестандартных технических и программных решений. Например, оптимизированных ХД и приложений для анализа как структурированных, так и неструктурированных данных.

– Под BIG DATA понимаются объемы данных, превосходящие на один или несколько порядков объемы, с которыми работают современные информационные системы, – отмечает Валерий Юринский, директор отделения технологического консалтинга компании «ФОРС». – Проблемы BIG DATA связаны с необходимостью обработки постоянно и резко увеличивающихся объемов информации, поиска и классификации данных в условиях роста сложности и числа их отдельных элементов. Данная тенденция просматривается уже в том, как сегодня строятся корпоративные хранилища информации.

97% специалистов, работающих с большими объемами данных, утверждают, что технологии хранения и обработки BIG DATA нуждаются в совершенствовании. Статистические исследования показывают, что стремление к изменению ситуации в этой сфере продолжает расти.

Технологические цунами

Причины появления больших объемов неструктурированного контента – технологические волны, считает Галина Аристова, директор по маркетингу компании RadiusGroup. «За короткий промежуток времени прошли три больших технологических волны, связанных с эволюцией Интернета, – отмечает она. – Первая волна Web 1.0 принесла возможность оцифровки текстов: компании создали себе сайты и разместили там свои тексты. Спустя несколько лет пришла технологическая волна Web 2.0, которая позволила создавать социальные сети; при этом значительная часть контента перекочевала в «облако». Что касается объемов неструктурированного контента, которым обмениваются пользователи, – они резко возросли за счет аудиовидеофайлов, а также за счет выхода корпоративного контента за пределы Firewall в «облако» – в Wiki, подкасты, блоги, социальные сети и т. д. Следующая технологическая волна – Web 3.0 – подняла объемы неструктурированного контента, в том числе корпоративного, на новую высоту – за счет выхода контента за пределы компьютера, на мобильные устройства. Теперь человек носит на ладони телевизор, диктофон, плеер и так далее – а значит, имеет при себе в любое время в любом месте всю необходимую ему информацию; без преувеличения можно сказать, что весь мир (весь контент мира) имеется в оцифрованном виде. Причем значительные его массивы являются неотъемлемой частью контента корпоративного, поскольку используются в коммуникации с существующими и потенциальными клиентами, партнерами, а также для связи сотрудников внутри компании.

■ **Иван Ермаков: ИТ-стратегия организаций меняется из-за требований бизнеса, которому просто необходимо, чтобы оставаться конкурентным, работать с большими объемами данных.**

Точки роста

Тобиаш Ратай приводит выводы исследования Gartner (Gartner CEO Advisory: Big Data Equals Big Opportunity by Stephen Prentice, March 31, 2011): BIG DATA предоставляет огромные возможности, которые ни государственный сектор экономики, ни бизнес не могут игнорировать.

– Ландшафт BIG DATA требует изменения бизнес-стратегии, – говорит Ратай. – В этом нуждается каждая компания. Организации должны быть готовы к взрывному росту данных, планировать его и искать пути для внутренних преобразований своего бизнеса. В использовании BIG DATA кроется громадный потенциал. Выиграют те компании, которые будут наживать капитал от стратегического использования данных и аналитики и использовать их повсеместно в своей организации. Это станет чрезвычайно важно, так как рынки консолидируются, конкуренция ужесточается и обостряется, а бизнес и регулирующая среда – более сложными.

■ BIG DATA: большая перемена

■ **Из интервью Стивена Бробста, технического директора компании Teradata, опубликованного в BBC News Business 27 июля 2011 (Stephen Brobst, chief technology officer (CTO) of Teradata).**

Компании встали перед дилеммой «больших данных». Следующий большой технологический прорыв, очевидно, будет концентрироваться вокруг идеи BIG DATA. Причем «большие» эти данные не только в смысле величины их объема, но также и в смысле разнообразия типов.

По мере того как все бóльшую роль начинают играть социальные медиаданные, сенсорные данные, природа информации поразительно меняется. Традиционно в тех записях, которые хранились как результат бизнес-деятельности, самым существенным были строки и столбцы. А сейчас данные становятся базисом для создания новых направлений в бизнесе, и эта информация приходит в организации в самых разных видах. Мы считаем, что разнообразие типов данных, как и рост их объемов, ведет к полному изменению сущности аналитических систем. В будущем каждый объект будет «знать» о месте своего расположения, о своей температуре, влажности… Обо всех симптомах, жизненно важных для него.

Кстати, в подобной роли может выступать и человек. Например, как объект медицинских исследований. Или это может быть объект в логистической цепочке, о местонахождении которого на производственном предприятии всегда известно. О любом объекте, который поставляется с одного завода на другой, сенсорные устройства позволят собрать огромное количество данных, которые могут быть проанализированы для выработки наилучшего решения.

Поэтому массивы данных, относящиеся к категории BIG DATA, которые являются большими не только по своему объему, но также и по разнообразию типов, со всей очевидностью, станут следующим крупным явлением в нашей индустрии. <

Источник: BBC News Business

ИТ-стратегия организаций меняется из-за требований бизнеса, которому просто необходимо, чтобы оставаться конкурентным, работать с большими объемами данных. Кстати, технологии виртуализации при неумелом использовании, добавляют много проблем, связанных с ростом данных — бесконтрольное размножение виртуальных сред способно поглощать как объемы хранения, так и доступную производительность систем хранения, — обращает внимание на проблему Иван Ермаков.

В компаниях, столкнувшихся с необходимостью обработки больших объемов данных, важно обеспечить возможность прозрачного масштабирования без прерывания сервисов, — как производительности, так и объема ИТ инфраструтур, обрабатывающих эти данные. В такой среде необходимо обеспечить возможность платить по мере роста, а так же обеспечить единое и простое управление всеми элементами таких инфраструктур, — добавляет он.

■ **Валерий Юринский: «Укрупнение бизнеса при его территориальном распределении существенно влияет на концентрацию информации».**

■ **Галина Аристова: «За короткий промежуток времени прошли три больших технологических волны, связанных с эволюцией Интернета».**

Крупнейший мировой вендор в области интеграции данных, корпорация Informatica в России и странах СНГ, полгода тому назад даже выделила специальное направление бизнеса по работе с BIG DATA. «Сегодня многие активно стремятся к централизации, — рассказывает Олег Гиацинтов, технический директор компании Data Integration Software (авторизованный дистрибьютор Informatica в России и странах СНГ). — До определенного времени крупные финансовые или телекоммуникационные компании в России имели децентрализованную распределенную структуру, и каждый филиал работал большей частью на себя, автономно занимаясь развитием своей инфраструктуры; при этом данные в центр передавались частично. В настоящее время мы видим, как российские компании активно централизуют свои системы, что приводит к росту данных в одной точке. С другой стороны, рынок развивается, компании предлагают все больше услуг, а это дополнительные данные, которые нужно обрабатывать.

Специалисты Informatica выделяют три категории BIG DATA. Во-первых, это транзакционные, структурированные данные, которые генерируются основными корпоративными системами. Во-вторых, неструктурированная информация, объем которой растет гораздо быстрее: данные взаимодействия (BIG interaction data), приходящие от различных дополнительных систем, атрибутивная информация, социальные медиаданные, записи почтовых систем, голосовых вызовов. В-третьих, единые средства обработки укрупненных данных (big data processing), которые предназначены для обработки результата объединения транзакционных и данных взаимодействия. Следовательно, нужны принципиально новые решения, которые позволяют обработать эти данные.

— Нужно говорить если не о принципиально новых решениях, то хотя бы о конвергенции всех отдельных элементов ИТ-инфраструктуры, с тем, чтобы обеспечить единое надежное функционирование и быстрое простое управление не только отдельными хранилищами данных, но всей инфраструктурой в целом с тем, чтобы отдельные ее элементы органично дополняя друг друга, решали задачи бизнеса, — подводит итог Александр Грубин, технический консультант департамента систем хранения данных, HP в России.

Суть — в деталях

Валерий Юринский отмечает, что укрупнение бизнеса при его территориальном распределении существенно влияет на концентрацию информации. Сотрудники главного офиса собирают и производят тщательную обработку больших массивов сведений, поступающих из различных региональных подразделений. Штаб-квартира требует от филиалов все более и более детализированных данных, которые затем хранятся, обрабатываются и анализируются с целью принятия верных и своевременных управленческих решений.

■ **Тобиаш Ратай: «Парадигма BIG DATA требует изменения бизнес-стратегии».**

В использовании BIG DATA кроется громадный потенциал. Выиграют те компании, кто будет наживать капитал от стратегического использования данных и аналитики и применять их повсеместно в своей организации.

«В предельном случае это может привести к удвоению объема хранимой фактографической информации, — подчеркивает Юринский. — К тому же в оперативной работе все шире используются исторические данные. Этому способствует как сложившаяся структура информационного обеспечения бизнеса, так и собственно природа корпоративной информации, используемой для управления и учета бизнес-процессов».

Олег Гиацинтов также обращает внимание на новые источники информации:

— Каждая компания хочет выжить на рынке и заработать как можно больше денег. А чтобы больше заработать, нужно больше информации о клиенте, которую лучше всего брать из не принадлежащих самой компании источников. Сегодня таким активно используемым источником стали социальные сети. Постепенно компании начинают пользоваться этим ресурсом, чтобы извлечь информацию. Это делается и при приеме новых сотрудников на работу, и в маркетинговых целях. Многие компании хотели бы иметь эти сведения у себя в постоянном режиме или получать их как минимум по запросу.

Денис Соловьев, руководитель практики SAP компании «СИТРОНИКС Информационные Технологии», констатирует, что укрупнение бизнеса

влечет за собой увеличение клиентской базы компании. А появление нового клиента – не просто очередная запись в базе данных, а история работы с клиентом в компании, история общения с ним account-менеджеров, различные измерения (такие как масштаб бизнеса, принадлежность определенным группам) и иные показатели. «Всю эту информацию нужно хранить и анализировать, – подчеркивает Соловьев. – И если компания заинтересована в оптимизации своего бизнеса, повышении качества предоставляемых услуг, то полнота информации о клиенте для ее дальнейшего развития будет играть одну из ключевых ролей».

■ **Денис Соловьев: «Укрупнение бизнеса влечет за собой увеличение клиентской базы компании».**

■ **Александр Грубин: Нужно говорить если не о принципиально новых решениях, то хотя бы о конвергенции всех отдельных элементов ИТ инфраструктур.**

Несмотря на то что технологии активно развивались, в подавляющем большинстве организаций данные хранились по старинке – на карточках, в тетрадках и пр.

«И только сейчас, – говорит Алексей Пилипчук, – когда правительство страны стало внимательнее к электронным услугам, проблема обработки большого объема данных приобрела особую остроту. Это и архивы различных государственных учреждений, предоставляющих услуги населению, и картотека больниц, поликлиник и прочих организаций».

Одной из причин значительного роста объемов данных Валерий Юринский считает их более подробную детализацию. «В Москве раньше собирались показания счетчиков воды только по дому в целом, израсходованное количество делилось пропорционально числу проживающих в квартире, и на этом основании выставлялись счета на оплату, – поясняет он. – Теперь требуется собирать сведения, накапливаемые индивидуальными квартирными счетчиками. Пока данные поступают от жильцов ежемесячно, но многие уже установленные современные счетчики могут отправлять их автоматически, с получением сведений ежесуточно, ежечасно или ежеминутно. Другая причина увеличения объема корпоративных данных – это использование новых источников получения цифровой информации: подписанные документы сканируются; предметы и люди фотографируются; общение с клиентами по телефону записывается и так далее».

Причины возникновения BIG DATA – в технологическом развитии общества, убежден Владимир Кирсанов: «За последнее десятилетие почти все российские организации перешли на автоматизированные системы управления, у населения появилась возможность использовать электронные системы оплаты услуг (коммунальные платежи, билетные системы, системы оплаты проезда и т. д.). Чем крупнее компания, чем больше рынков, на которых она присутствует, тем большим объемом корпоративных данных она обладает».

■ **Алексей Пилипчук: «Однозначно сказать, что такое „большой объем данных", можно лишь с учетом масштаба конкретного потребителя, с привязкой к определенному времени».**

В качестве примера Кирсанов приводит крупную страховую компанию с филиалами по всей России. В ее базе находятся персональные данные клиентов, информация о транспортных средствах, о каждом страховом случае. В базу попадают все обращения за медицинской помощью, миллионы страховок для выезда за рубеж, перелеты, проезд транспортом.

– Несмотря на территориальное распределение, все данные находятся в одной базе, где и обрабатываются, – подчеркивает специалист. – При этом наличие офисов по всей стране не влияет на расширение базы.

Стратегические цели

Денис Соловьев рекомендует рассматривать большие объемы данных только на основе измерения хранимой информации в терабайтах, петабайтах или объемах бэкапирования. Зачастую встречаются ситуации, когда в компании функционирует большое ХД, при этом таблицы или партиции, где располагаются данные, используются довольно редко. Проблема больших объемов информации возникает для каждой компании, когда уменьшается производительность аналитической системы или снижается скорость выполнения запросов хранилищем данных из-за того, что его архитектура и поставленные перед ним задачи не могут удовлетворять тем объемам информации, которыми должны оперировать пользователи в своей профессиональной деятельности. При этом неважно, каков реальный объем – два или 100 терабайт. В такой ситуации архитектура хранилища и принципы ее работы должны меняться. Поиск и предсказание этих проблем – одна из задач архитектора хранилища. Если оно имеет качественно проработанную структуру и гибкую модель, то даже не самые мощные БД могут успешно функционировать на больших объемах.

Причиной того, что объемы данных резко возрастают, Тобиаш Ратай называет такой фактор, как быстрое увеличение производительности систем управления данными в расчете на единицу стоимости. Это позволяет организациям собирать больше детализированных транзакционных данных, таких как веб-клики, патентные или поступающие с точек продаж данные. «Например, телекоммуникационные компании сегодня могут хранить детализированные записи звонков в течение месяцев и даже лет – вместо того чтобы подводить итоги и архивировать данные в конце каждого операционного дня или недели, – отмечает Тобиаш Ратай. – Аналитики могут использовать эти исторические подробности, например, для лучшего понимания модели и поведения заказчика».

В дополнение к этому существует несколько источников данных, которые организации хотели бы включить в аналитическую орбиту. В их числе – социальные медиаданные (то есть блоги, онлайн-дискуссии, краткие обсуждения), сенсорные данные (от RFID-чипов), GPS-данные, различные устройства и специализированные системы вроде SmartMeters либо традиционные неструктурированные данные, такие как аудио, изображения, видео и текст (почта, веб-сайты, документы).

– Анализируя эту детализированную информацию, – резюмирует Тобиаш Ратай, – организации могут извлечь большую ценность, и это должно стимулировать их к тому, чтобы собирать еще больше данных. Получается, что чем больше данных собирает организация, тем больше моделей и возможностей проникновения в суть проблемных ситуаций она извлечет из их обработки. <