

Открытые системы

Открыты для вас. 20 лет

№06
2013

ISSN 1028-7493

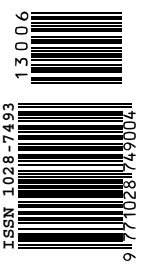
ИТ для бизнеса —
архитекторам
информационных систем

www.osmag.ru

АНАЛИЗ НЕСТРУКТУРИРОВАННЫХ ДАННЫХ

Виртуализация Больших Данных

- Инструменты исследования масс-медиа
- Детские болезни SDN
- Визуальный анализ данных
- Надежные высокопроизводительные облака



Открытые системы

ЖУРНАЛ ДЛЯ АРХИТЕКТОРОВ ИНФОРМАЦИОННЫХ СИСТЕМ

www.osmag.ru



Реклама 16+

1580 руб.*

2100 руб.*

В РЕДАКЦИИ

www.osp.ru/subscribe/os/

+7 495 725-47-85

e-mail: xpress@osp.ru

ПО ПОЧТЕ

«Пресса России» 72773

«Роспечать НТИ» 59869

«Почта России» 99482

ОПЛАТИТЬ

VISA

МИР

деньги

ROBO

QIWI

WebMoney

*Стоимость подписки указана за 12 месяцев, с учетом доставки «Почта России» на территории РФ

Большой Разрыв

Компаниям и организациям, по наблюдениям аналитиков Aberdeen Group, все труднее доставлять информацию нужным людям в нужное время. Мало того, 65% менеджеров сегодня сталкиваются с сокращением «окна» от момента события до реагирования на него — времени, отведенного для принятия взвешенных решений на основе достоверных данных, становится все меньше. Такая ситуация наблюдается как в отдельных отраслях (например, по данным SITA, лишь 9% авиакомпаний считают, что сегодня качество данных, которыми они оперируют, удовлетворяет всем их требованиям, при этом только 7% достигли необходимого уровня интеграции различных источников данных в пределах своей компании), так и во всей Цифровой Вселенной.

С 2005 по 2020 год объем всех генерируемых на Земле данных вырастет со 130 до 40 000 экзабайт, удваиваясь каждые два года, однако инвестиции в инфраструктуру их обработки, включая оборудование, ПО, телекоммуникации, сервисы и кадры, будут расти не более чем на 40% в год. Получается, что инвестиции в пересчете на один гигабайт сократятся с двух долларов до двадцати центов. Конечно, это всего лишь усредненные показатели, и темпы инвестиций в такие «горячие» области ИТ, как управление хранением, безопасность, Большие Данные и облака, будут существенно выше. Важно другое — 80% всех данных, причем в большинстве своем неструктурированных, создают предприятия, однако они способны извлечь пользу лишь из тонкого слоя этого объема. В 2012 году, как отмечают в IDC, 23% всего цифрового пространства потенциально было доступно для конструктивного анализа, но технологии, особенно работы с неструктурированными данными, пока не позволяют сделать и этого — сегодня не более 3% доступных для обработки данных может хоть как-то учитываться, а реально аналитиками для извлечения скрытых тенденций и получения новых знаний используется лишь 0,5%. Такую диспропорцию впору назвать Большим Разрывом между объемами данных, потенциально несущих полезные сведения и реально используемых. Потребуется еще много времени, работы и значительных инвестиций для совершенствования технологий.

Это была плохая новость, а хорошая такова: по мере расширения Цифровой Вселенной в ней растет и количество данных, полезных для анализа. По прогнозам экспертов IDC, к 2020 году таковых будет уже 33% от накопленного объема. Иначе говоря, как отмечают авторы этого номера, посвященного технологиям сбора, анализа и визуализации неструктурированных данных, при обеспечении учета всех этих данных «точность принятия решения будет стремиться к абсолютной».

Безусловно, не все данные обязательно полезны, тем не менее некоторые их типы, например записи с камер наблюдения, могут представлять интерес и открывать дополнительные возможности, если кадры будут захвачены и проанализированы в реальном времени. Обработка неструктурированных данных этого типа может ускорить расследование преступлений, повысить эффективность аналитики розничных продаж и борьбы с терроризмом. То же можно сказать и про данные, поступающие от различных встроенных датчиков и медицинского оборудования, — при наличии средств их обработки в реальном времени получаемые сведения могут стать жизненно важными как для отдельного человека, так и для мониторинга потенциальных вспышек вирусов, распространения эпидемий и т. д.

Отдельной строкой в потоке неструктурированных данных стоят потребительские изображения — сегодня принято много говорить о себе, размещать свои фотографии, фото своих близких и друзей. Такие изображения, при умелом использовании, стоят тысячи слов для продавцов, производителей, маркетологов, политиков и представителей других заинтересованных профессий. Среди статей этого номера есть описание конкретных решений, позволяющих оценивать настроения в обществе, прогнозировать выборы, анализировать СМИ, для того чтобы из врага сделать союзника, вовремя отловить отрицательный отзыв и отреагировать на него, посулив покупателю скидку или извинившись за промах продавца. Потенциал традиционных подходов к развитию средств аналитики и визуализации ее результатов на базе лишь структурированных данных сегодня оказался практически исчерпан — эти подходы недостаточно динамичны и не способны адаптироваться к новым источникам дан-



ных, поэтому в этой области мы и наблюдаем сейчас заметные изменения.

Авторы этого номера затронули и другой аспект Большого Разрыва: доля сведений, не предназначенных для широкой огласки в Цифровой Вселенной, растет быстрее ее самой — с трети в 2010 году до 40% в 2020-м, однако защищается только половина информации, которую следовало бы уберечь от посторонних глаз. Неожиданно выяснилось, что набирающие популярность технологии вроде SDN при определенных условиях более уязвимы, чем традиционные. Как отмечает Валерий Коржов, объектов для целенаправленной атаки в случае программно-конфигурируемых сетей стало больше, чем, например, для классической IP-сети. В SDN за состоянием сети следит один контроллер, который может стать самой уязвимой точкой, и тогда данные пользователей окажутся незащищенными.

В аббревиатуре ИТ на «И», как известно, приходится 90% нагрузки, а на технологии — лишь 10%, следовательно, ни продукты, ни решения не заменят профессионалов в области работы с данными. Однако пока, как отмечает Артем Гришковский, большинство руководителей компаний не могут оценить объемы и качество скрытой информации, которую содержат неструктурированные данные, и не понимают, из каких источников их можно брать, как эти данные коррелируют со структурированной корпоративной информацией и какое преимущество получит компания от интегрированного анализа всех своих данных. ■

Дмитрий Волков