# Big Data: новые возможности для растущего бизнеса

**Сергей Артемов, эксперт группы перспективных технологий компании «Инфосистемы Джет»**

Среди наиболее обсуждаемых тем в ИТ-изданиях в последнее время выделяется феномен Big Data, или проблема «Больших данных». Стоит отметить, что проблема хранения и обработки большого объема данных стояла всегда, но с развитием ИТ она стала беспокоить не только ряд крупнейших корпораций, но и гораздо более широкий круг компаний. Сегодня Big Data не просто модный термин – для многих организаций это стало насущной проблемой, требующей немедленного решения.

Говоря «Большие данные», надо понимать, что, как правило, под этим подразумевается большой объем плохо структурированных данных, обработка которых привычными методами невозможна или экономически нецелесообразна. Типичный пример – это записи о транзакциях (например, данные Call Data Records у сотовых операторов или данные платежей в процессинговых центрах), данные с телеметрических датчиков, журналы активности пользователей в крупных интернет-проектах или социальных сетях.

**Разложить по полочкам**

Методику и инструменты работы со структурированными данными ИТ-индустрия создала давно – это реляционная модель данных и системы управления БД. Но современной тенденцией является потребность обработки большого объема неструктурированных данных, и это та область, где прежние подходы работают плохо. Именно эта потребность требует новой методики обращения с данными, и сейчас все более популярной становится модель работы с Big Data, реализованная в проекте Apache Software Foundation Apache Hadoop (http://hadoop.apache.org).

Не вдаваясь глубоко в технические подробности, можно сказать, что Apache Hadoop состоит из двух компонентов: это распределенная кластерная система Hadoop Distributed File System (HDFS) и программный интерфейс Map Reduce.

В основе модели работы Apache Hadoop лежат три основных принципа. Во-первых, данные равномерно распределяются на внутренних дисках множества серверов, объединенных HDFS. Во-вторых, не данные передаются программе обработки, а программа – к данным. Третий принцип – данные обрабатываются параллельно, причем эта возможность заложена архитектурно в программном интерфейсе Map Reduce.

Таким образом, вместо привычной концепции «база данных + сервер» у нас имеется кластер из множества недорогих узлов (так называемое «commodity hardware»), каждый из которых является и хранилищем, и обработчиком данных, а само понятие «база данных» отсутствует.

В этой системе запрос на обработку данных представляет собой небольшую программу. По умолчанию она – на языке Java, но фактически можно использовать любой язык

программирования. Apache Hadoop framework передает и одновременно запускает эту программу на узлах кластера, хранящих обрабатываемые данные. Так как HDFS равномерно распределяет данные по всем узлам, скорее всего, будут задействованы все доступные серверы кластера. В свою очередь, результаты работы от всех узлов агрегируются и передаются пользователю.

Стоит отметить, что подобная система обладает двумя важными характеристиками. Во-первых, любой сколь угодно сложный анализ большого объема данных сводится к их обработке на локальных дисках сервера, поэтому максимально возможное время реакции известно заранее. Во-вторых, система масштабируется симметрично и линейно – при добавлении новых узлов возрастает и вычислительная мощность, и дисковая емкость – поэтому время обработки данных не зависит от их объема.

Очевидно, что подобный подход к обработке данных имеет хорошие перспективы и со временем будет находить все более широкое распространение. Сфера применения данной технологии достаточно широка: это и поисковые системы, и платформа для систем бизнес-аналитики, и преобразование больших объемов данных (например, ETL-задачи или конвертация аудио- и видеоинформации), и универсальный framework для параллельного запуска различных задач.

Так, газета New York Times использовала Apache Hadoop для преобразования четырех терабайт TIFF-изображений (включая TIFF-изображения размером в 405 КБ, SGML-статьи – в 3,3 МБ и XML-файлы – в 405 КБ) в изображения PNG-формата размером в 800 КБ, пригодные для публикации через HTTP-сервер. Весь процесс преобразования занял 36 часов.

Компания Factual (http://www.factual.com) использовала Apache Hadoop для осуществления стрессового тестирования своего продукта. Организаторам тестирования понадобилось лишь написать небольшой код, принимающий на вход список URL и выполняющий HTTP-запрос по каждому из них. Apache Hadoop и Amazon Elastic Cloud справились со всем остальным (http://www.cloudera.com/blog/2011/02/gratuitous-hadoop-stress-testing-on-the-cheap-with-hadoop-streaming-and-ec2/).

**Актуально в России**

Для России тема Big Data актуальна ничуть не меньше, чем для остального мира. Особенно если относиться к этому явлению не как к неожиданно возникшей досадной проблеме, а как к новой возможности сделать свой бизнес более эффективным, «на полную катушку» используя информацию, которая раньше не анализировалась или анализировалась частично.

За примерами далеко ходить не надо: сотовые операторы и сейчас хранят и анализируют данные о звонках абонентов. Речь идет, разумеется, не о содержании разговоров, а о характеристиках звонка: длительность, номер вызываемого абонента и т.д. На основе этой информации работает ряд информационных систем сотовых операторов: системы противодействия мошенничеству (Fraud Control), бизнес-аналитики и биллинговые системы.

Проблема состоит в том, что, как правило, каждая из подобных систем загружает огромные объемы данных CDR в локальную БД и только потом работает с ними. При этом одни и те же данные оказываются размноженными в нескольких экземплярах, помещенными на High-End дисковые массивы, для них приобретаются и поддерживаются

реляционная система управления БД и высокопроизводительные High-End серверы, копии данных от каждой системы хранятся в системе резервного копирования и, возможно, реплицируются на удаленную площадку.

В результате инфраструктура для анализа подобных данных получается чрезвычайно затратной, и приходится идти на компромисс – сокращать «глубину» анализируемой информации. Гораздо более выгодно и эффективно сделать общее хранилище подобных данных на основе Hadoop, а различные системы аналитики «научить» обращаться за требуемой информацией к централизованному хранилищу. Такое решение менее затратно с точки зрения инфраструктурных компонентов и позволяет не ограничивать объем данных для анализа.

В решениях Big Data могут быть заинтересованы крупные online-магазины. Согласно исследованиям, у продавца есть 10 дней после покупки товара, в течение которых можно предложить что-то из сопутствующих товаров и услуг покупателю, которые он с большой вероятностью согласится приобрести. После этого предлагать что-либо бесполезно – человек уже купил то, что ему было необходимо, или «наигрался» с новой вещью. С помощью аналитики систем на базе Hadoop можно быстро формировать персональные рекомендации для каждого клиента, основанные на информации о совершенных ранее покупках, а не просто предлагать сопутствующие товары. Также возможно рекомендовать другие варианты, которые почти наверняка заинтересуют покупателя, например, новые книги авторов, которых ранее читал клиент, новые диски музыкальных групп, фильмы и т.д.

Государственные заказчики также не являются исключением: на основе данных системы наблюдения за дорожным трафиком можно оперативно получать информацию о том, где камеры последний раз зафиксировали требуемую машину, ее маршрут, характеристики машинного потока, среднюю скорость движения, прогноз плотности трафика.

**Современный подход к решению проблем Big Data**

Занимаясь проблемами Big Data, необходимо иметь ввиду, что пользователям требуется решение, способное легко интегрироваться в существующую инфраструктуру ЦОД и обеспечить все три этапа обработки информации: сбор, ее организацию и анализ. Таким образом, современное решение для Big Data – это не изолированная система, а комплекс систем, каждая из которых выполняет собственные задачи и легко интегрируется с другими.

При этом важно правильное позиционирование новых систем, подобных вышеописанному Apache Hadoop. Их, разумеется, нельзя продвигать как универсальную замену больших баз данных, они конкурентоспособны в своей области приложения (начальная обработка или преобразование больших объемов разнородных данных). В связи с этим попытка возложить на Hadoop OLTP-подобную нагрузку только из-за большого объема информации обречена на провал – классические базы данных справляются с этой работой значительно успешнее.

То есть новые решения, позволяющие эффективно работать с неструктурированными данными, при работе со структурированной информацией находятся в заведомо проигрышном положении по сравнению с традиционными реляционными БД. Поэтому эффективное решение проблем Big Data будет представлять собой совместную работу Hadoop, реляционных БД и аналитического ПО.

В этом свете логичным выглядит подход к решению проблем Big Data от компании Oracle (http://www.oracle.com/us/products/database/big-data-appliance/overview/index.html). Oracle разбивает жизненный цикл обработки информации на три этапа и использует для каждого из них собственное решение:

1. **Сбор, обработка и структурирование данных**. В качестве решения используется Oracle Big Data Appliance – это предустановленный Hadoop-кластер, Oracle NoSQL Database и средства интеграции с другими хранилищами данных. Задача Oracle Big Data Appliance состоит в хранении и первичной обработке неструктурированной или частично структурированной информации, то есть как раз в том, что у систем на базе Hadoop получается лучше всего.

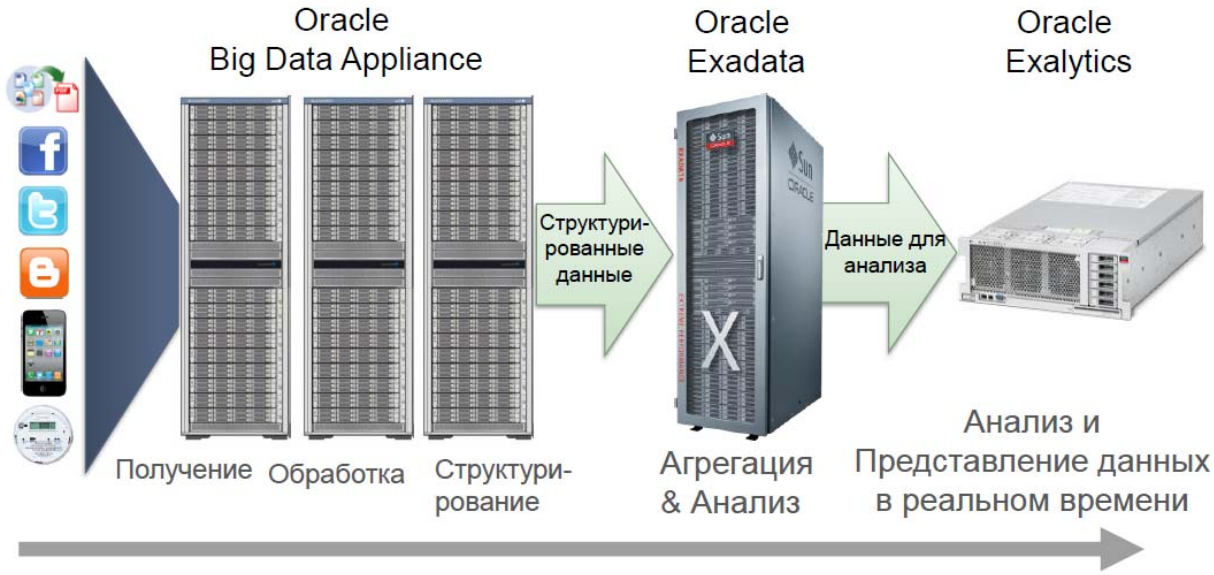


**Рис. 1.** Инфраструктура поддержки жизненного цикла обработки Big Data

2. **Агрегация и анализ данных.** Так как задача структурирования и начальной обработки данных решена, для работы со структурированными данными используется комплекс Oracle Exadata. Модули интеграции Oracle Big Data Appliance позволяют как оперативно загружать данные в Oracle Exadata, так и получать доступ к данным «на лету» из Oracle Exadata.

3. **Аналитика данных в реальном времени**. Ну, и для максимально оперативного анализа полученных данных используется Oracle Exalytics Database Machine, которая позволяет решать аналитические задачи фактически в режиме «online».

В заключение хочется отметить, что тема «Больших данных» предполагает не только развитие технологий и продуктов, способных обработать огромные объемы данных. Также требуются аналитики – специалисты, которые способны из моря первичных данных получать уникальные знания. Без этого все технологии обработки не имеют ценности и теряют смысл.