ПРОБЛЕМАТИКА BIG DATA В ИНФРАСТРУКТУРЕ УРОВНЯ ПРЕДПРИЯТИЯ

Сисюков А. Н. Email: Sisyukov1134@scientifictext.ru

Сисюков Артем Николаевич – кандидат технических наук, доцент,
кафедра технологии приборостроения, факультет систем управления и робототехники,
Санкт-Петербургский национальный исследовательский университет
информационных технологий, механики и оптики, г. Санкт-Петербург

Аннотация: *в статье затронута проблематика использования технологии больших данных (Big Data) в инфраструктуре предприятия. Предложен подход к интеграции с существующими решениями предприятия с традиционной схемой хранения и анализа данных. Рассмотрены основные аспекты традиционных хранилищ данных и особенности не реляционных структур Big Data. Обоснована необходимость применения Big Data на предприятии. Приводятся ключевые компоненты в построении платформы больших данных, а также требования к инфраструктуре, сбору, хранению и анализу Big Data.*

Ключевые слова: *Big Data, базы данных, киберфизические системы, анализ данных.*

BIG DATA CHALLENGES FOR ENTERPRISE LEVEL INFRASTRUCTURE

Sisyukov A.N.

Sisyukov Artem Nikolaevich – PhD in Technics, Associate Professor,
INSTRUMENTATION TECHNOLOGIES DEPARTMENT,
FACULTY OF CONTROL SYSTEMS AND ROBOTICS,
ITMO UNIVERSITY
SAINT PETERSBURG NATIONAL RESEARCH UNIVERSITY OF INFORMATION TECHNOLOGIES,
MECHANICS AND OPTICS, ST. PETERSBURG

Abstract: *the article touches the problems of using Big Data technologies in enterprise infrastructure. The approach to integrate with existing enterprise solutions using the traditional approach to storage and data analysis is considered. The basic aspects of traditional data warehouses and the non-relational structures of Big Data are considered. The necessity of using Big Data at the enterprise is grounded. Key components in the construction of a large data platform, as well as infrastructure requirements, collection, storage and analysis of Big Data are presented.*

Keywords: *Big Data, databases, cyber-physical systems, data analysis.*

УДК 004.62

Массив больших данных (Big Data) на предприятии, как правило, формируются из совокупности источников, таких как традиционные данные предприятия, данные, регистрируемые программно и извлекаемые с сенсоров (машинные данные), социальные данные.

К первым относятся транзакционные данные ERP и CRM систем, часть транзакций бухгалтерских систем и онлайн магазинов [1]. К машинным данным можно отнести логи программных систем, информация с торговых площадок, потоки с промышленных датчиков, логи оборудования. К социальным - социальные сети, микроблоги и массивы отзывов клиентов.

В соответствии с оценками McKinsey Global Institute объем данных растет по 40% в год. Тем не менее объем не единственный имеющий значение атрибут, присущий большим данным. Можно выделить 4 основных характеристики, определяющие большие данные (4V) по версии IDC (International Data Corporation) - Объем (Volume), Скорость (Velocity), Разнообразие (Variety) и Ценность (Value).

Так, к примеру, тяжелое промышленное оборудование, буровые установки, нефтедобывающее оборудование может генерировать по 10 терабайт информации в пределах получаса, что отражает характеристику объема.

Социальные сети и микроблоги генерируют не столь большой объем информации как промышленные системы, но дают существенный прирост данных для оценок в сфере взаимоотношений с клиентами. К примеру сравнительно небольшой размер твита в 100-150 символов при высокой скорости формирования твитов дает существенное увеличение объема данных порядка по 8 терабайт в день.

Традиционные форматы данных достаточно определенны и слабо изменчивы. В контрасте нетрадиционные форматы данных показывают высокую подверженность изменениям, чему соответствует третья характеристика Big Data. Новые типы данных необходимы для сбора результирующей информации при добавлении новых сервисов, размещении новых датчиков или запуска новых маркетинговых компаний.

Экономическая значимость различных данных значительно отличается. Обычно полезная информация скрывается за большим объемом неструктурированных данных. Задачу опознать значимые

данные, извлечь и трансформировать их для анализа отражает характеристика Value.

Для наиболее эффективного использования больших данных предприятиям требуется вовлекать свою ИТ инфраструктуру в обработку этих высокоемких, быстрорастущих и разнообразных источников данных и интегрировать их с уже существующими данными предприятия для анализа.

Необходимость больших данных

При обработке и анализе больших данных в сочетании с традиционными корпоративными данными, предприятия могут получить более глубокое и четкое понимание своего бизнеса, что приводит к повышению производительности, усилению конкурентной позиции и появлению новых возможностей.

К примеру производственные предприятия внедряют сенсоры в свои продукты в целях приема потока телеметрии. В автомобильной промышленности такие системы, как OnStar® компании General Motors или RLink® от Renault, предоставляют услуги связи, безопасности и навигации. Достаточно важно то, что эта телеметрия предоставляет модели использования, показатели отказов и другие индикаторы для улучшения продукции, которые могут снизить затраты на разработку и сборку.

Использование киберфизических систем в совокупности с существующим промышленным оборудованием на производственном предприятии позволяет оптимизировать производство, прогнозируя расходы сырья и материалов, предотвращая отказы оборудования [2].

Розничные торговые площадки обычно знают, кто покупает их продукцию. Информация, недоступная им ранее определяется массивами больших данных. Обработка данных социальных сетей, журналов активности с сайтов электронной торговли помогает им понять, кто и почему покупку не совершил. Это позволяет проводить более эффективную сегментацию клиентов и строить целевые маркетинговые кампании, повысить эффективность логистических цепочек за счет более точного планирования спроса [3].

Построение платформы больших данных

Для хранилищ данных, веб-магазинов или любой ИТ-платформы, инфраструктура больших данных вносит свои требования [4]. При встраивании компонентов платформы больших данных важно помнить, что конечная цель - легкость их интеграции с корпоративными данными, позволяющая проводить глубокую аналитику по совокупному набору данных.

Требование к инфраструктуре

Требования к инфраструктуре больших данных охватывают сбор данных, хранение (организацию) данных и анализ данных.

Сбор больших данных

Фаза сбора является одним из основных требований к инфраструктуре еще до появления больших данных. Поскольку Big Data относятся к потокам данных с более высокой скоростью и большим разнообразием, инфраструктура, необходимая для поддержки сбора больших данных, должна обеспечивать низкую, предсказуемую задержку как при сборе, так и при выполнении коротких простых запросов. Среда должна обеспечивать обработку больших объемов транзакций, часто в распределенной среде и поддерживать гибкие динамические структуры данных.

Базы данных NoSQL часто используются для сбора и хранения больших данных. Они хорошо подходят для динамических структур данных и обладают высокой масштабируемостью. Информация, хранимая в базе данных NoSQL, как правило, очень разнообразна, поскольку системы предназначены для простого захвата всех данных без категоризации и синтаксического анализа данных, присущих фиксированной схеме.

К примеру, БД NoSQL часто используются для сбора и хранения данных в социальных сетях. Хотя приложения, ориентированные на клиента, часто меняются, базовые структуры хранения остаются простыми. Вместо разработки схемы с отношениями между объектами эти простые структуры часто содержат только главный ключ для идентификации позиции данных, и контейнер основного содержимого, такого как идентификатор клиента и профиль клиента. Такая простая динамическая структура позволяет осуществлять изменения без дорогостоящих реорганизаций на уровне хранилища (например, добавлять новые поля в профиль клиента).

Хранение больших данных

В классических условиях хранение данных сводится к интеграции данных. Поскольку на входе огромный массив больших данных, существует тенденция к хранению данных в первичном месте после сбора, что экономит время и деньги, не перемещая большие объемы данных. Инфраструктура, необходимая для организации больших данных, должна уметь обрабатывать и манипулировать данными в исходном хранилище, поддерживать высокую пропускную способность (часто в пакетном режиме) для обработки больших наборов данных и взаимодействовать с большим разнообразием форматов данных, от неструктурированных до структурированных.

Hadoop - технология, которая позволяет организовывать и обрабатывать большие объемы данных, сохраняя при этом данные в исходном кластере хранения данных. Так распределенная файловая система Hadoop (HDFS) - это система долгосрочного хранения, например, для веб-журналов. Из веб-журналов

формируется последовательность действий в сеансе браузера, выполняя алгоритмы MapReduce в кластере и генерируя агрегированные результаты в нем. Полученные агрегированные результаты затем загружаются в систему реляционных СУБД.

Анализ больших данных

Поскольку данные не всегда перемещаются во время фазы хранения, анализ может также выполняться в распределенной среде, где некоторые данные остаются там, где они были первоначально сохранены с прозрачным доступом из хранилища данных. Инфраструктура, необходимая для анализа больших данных, должна поддерживать более глубокие аналитические методы, такие как статистический анализ и интеллектуальный анализ данных на более широком спектре типов данных, хранящихся в различных системах. Важно, что инфраструктура должна уметь интегрировать анализ по комбинации больших данных и традиционных данных предприятия. Свежее видение приходит не только от анализа новых данных, но и от анализа его в контексте старого.

Например, анализ данных по составу товаров в интеллектуальном торговом автомате в сочетании с планом обслуживания в месте установки оборудования будет определять оптимальный ассортимент и график пополнения торгового автомата.

Совместная интеграция

Системы NoSQL предназначены для сбора всех данных без категоризации и синтаксического анализа при входе в систему, поэтому данные очень разнообразны.

С другой стороны, системы SQL обычно помещают данные в четко определенные структуры и накладывают метаданные на данные, собранные для обеспечения согласованности и проверки типов данных.

В первом случае используются распределенные файловые системы и хранилища транзакций (ключ-значение).

Для интерпретации и выделения информации из данных в этих решениях используется парадигма программирования MapReduce [5].

Фактически хранилища «ключ-значение» или базы данных NoSQL - это базы данных OLTP

Системы больших данных оптимизированы для очень быстрого сбора данных в контексте простых шаблонов запросов. Базы данных NoSQL способны обеспечить очень высокую производительность, так как собранные данные быстро сохраняются по одному ключу, а не интерпретируются и преобразуются в схему. Таким образом, NoSQL база данных может быстро обрабатывать большое количество транзакций

.Однако из-за характера данных в БД NoSQL любая реорганизация данных требует программирования для интерпретации логики хранения.

Использование их в сочетании с отсутствием поддержки сложных шаблонов запросов, затрудняет конечным пользователям извлекать значение из БД NoSQL [6].

Чтобы получить максимальную отдачу от решений NoSQL и превратить их из специализированных, ориентированных на разработчиков решений в решения для предприятия, они в сочетании с решениями SQL должны быть включены в единую проверенную инфраструктуру, отвечающую требованиям к управляемости и безопасности современных предприятий.

Выводы

Анализ новых и разнообразных потоков цифровых данных позволяет выявить новые источники экономической значимости, предоставить свежие сведения о поведении клиентов и выявить тенденции рынка на раннем этапе. Но подобный приток новых данных порождает дополнительные проблемы перед ИТ отделами. Чтобы получить реальную пользу от больших данных необходимы правильные инструменты для сбора и организации широкого спектра типов данных из разных источников, а также для их легкого анализа в контексте всех корпоративных данных. Наряду с применением уже существующих реляционных инструментов продуктов компаний Oracle, IBM, Microsoft, Pentaho и их интеграцией с решением больших данных предприятия, компании могут приобретать, организовывать и анализировать все свои корпоративные данные (включая структурированные и неструктурированные), чтобы принимать наиболее обоснованные решения.

Список литературы / References

1. *Гобарева Я.Л., Городецкая О.Ю., Кочанова Е.Р.* Возможности технологии Big Data для повышения качества эксплуатации CRM-систем // Транспортное дело России, 2015. № 5. С. 62-63.
2. *Иващенко А.В., Двойнина О.В.* Анализ сверхбольших массивов данных (Big Data) в едином информационном пространстве научно-производственного предприятия // Наука и Мир, 2014. Т. 1. № 12 (16). С. 44-47.
3. *Сулейкин А.С.* Возможности применения технологии Big Data в крупном ритейле // European Research, 2015. № 10 (11). С. 77-79.
4. *Голов Н.И., Кравченко Т.К.* Проектирование хранилища данных для решения задач Big Data //

Информационные технологии в проектировании и производстве, 2014. № 1 (153). С. 56-61.