УДК 681.31(031)

05.00.00 Технические науки

АНАЛИЗ ВОЗМОЖНОСТИ ИСПОЛЬЗОВАНИЯ ТЕХНОЛОГИИ ОБРАБОТКИ БОЛЬШИХ ДАННЫХ В СИСТЕМАХ ДЛЯ ТЕРРИТОРИАЛЬНО-РАСПРЕДЕЛЕННЫХ КОМПЛЕКСОВ[1]

Видовский Леонид Адольфович
д.т.н., профессор

Янаева Марина Викторовна
к.т.н., доцент

Мурлин Алексей Георгиевич
к.т.н., доцент

Мурлина Владислава Анатольевна
к.т.н., доцент

Гвозденко Анастасия Алексеевна
студент
Кубанский государственный технологический университет, г. Краснодар, Россия

Статья посвящена анализу использования технологии обработки больших данных в информационных системах территориально-распределенных комплексов

Ключевые слова: ТЕРРИТОРИАЛЬНО-РАСПРЕДЕЛЕННЫЙ КОМПЛЕКС, БОЛЬШИЕ ДАННЫЕ, ПРОПУСКНОЙ КОНТРОЛЬ, ВИДЕОНАБЛЮДЕНИЕ, ИНФОРМАЦИОННАЯ СИСТЕМА

Doi: 10.21515/1990-4665-132-086

UDC 681.31(031)

Technical sciences

ANALYSIS OF THE OPPORTUNITY OF USING THE TECHNOLOGY OF PROCESSING LARGE DATA IN SYSTEMS FOR TERRITORIAL-DISTRIBUTED COMPLEXES

Vidovsky Leonid Adolfovich
Dr.Sci.Tech., Professor

Yanaeva Marina Viktorovna
Cand.Tech.Sci, associate professor

Murlin Aleksey Georgievich
Cand.Tech.Sci, associate professor

Murlina Vladislava Anatolevna
Cand.Tech.Sci, associate professor

Gvozdenko Anastasia Alekseyevna
student
Kuban State Technological University, Krasnodar, Russia

The article is devoted to the analysis of the use of large data processing technology in information systems of territorially distributed complexes

Keywords: TERRITORIAL-DISTRIBUTED COMPLEX, GREAT DATA, PASSENGER CONTROL, VIDEO-OBSERVATION, INFORMATION SYSTEM

При разработке систем контроля и наблюдения в территориально-распределенных комплексах возникает необходимость обработки и хранения больших объемов информации, что зачастую проблематично с учетом удаленности объектов комплекса. Территориально распределенный комплекс – это совокупность объектов, имеющих связи в рамках одной организации, расположенных на значительном удалении друг от друга, что характерно для крупных компаний. Технология обработки больших объемов данных занимается проблемами, связанными с их хранением и

[1] Работа выполнена при финансовой поддержке РФФИ, № 17-02-00475а.

обработкой. В информационных технологиях большие данные (англ. Big Data) представляют собой серию подходов, методов и инструментов обработки неструктурированных и структурированных данных огромных объемов и значительного многообразия для получения человеко-читаемых результатов, которые будут полезны в условиях быстрого прироста и распределения по многочисленным узлам вычислительной сети.

Системы, основанные на Big Data, обладают следующими общими принципами построения:

– данные системы состоят из огромного количества узлов, состоящих из дешевого оборудования;

– каждый узел представляет из себя сервер хранения и обработки данных;

– обработка данных ведется в режиме MapReduce;

– сохранение данных в нескольких копиях, как правило в трех, поэтому отказ узла или двух узлов не критичен для системы;

– неограниченное масштабирование объемов информации. [4]

Существует несколько современных технологий обработки Big Data:

– NoSQL DB (СУБД, построенные по принципу «ключ-значение», обеспечивают быструю запись и выборку по ключу);

– MapReduce (фреймворк для распределенных вычислений и обработки данных на тысячах узлах);

– Hadoop (это лидирующая реализация MapReduce (проект Apache); является масштабируемой пакетной обработкой (имеет большое количество наработок);

– HDFS (Hadoop Distributed File System, используется для построения дешевых, распределенных, масштабируемых хранилищ).

Само понятие Big Data означает огромнейшие хранимые и обрабатываемые массивы из сотен гигабайт, и даже петабайт данных. Те

данные, которые можно обработать и извлечь из них некоторое количество полезной информации. Приведем в таблице сравнение обычной базы данных с базой больших данных.

Таблица – Сравнительный анализ

Характеристика	Традиционная БД	База больших данных
Объем информации	от гигабайт (10^9 байт) до терабайт (10^{12} байт)	от петабайт (10^{15} байт) до эксабайт (10^{18} байт)
Способ хранения	централизованный	децентрализованный
Структурированность данных	структурирована	полуструктурирована и неструктурирована
Модель хранения и обработки данных	вертикальная модель	горизонтальная модель
Взаимосвязь данных	сильная	слабая

Для Больших Данных определяющими характеристиками являются «три V»:

а) volume – величина физического объема;

б) velocity – скорость скорости прироста данных и необходимости высокоскоростной обработки и получения результатов);

в) variety – возможности одновременной обработки разнообразных типов, полуструктурированных и структурированных данных.

Основные проблемы системы BigData сводятся к этим трем характеристикам. Требуются специальные условия для хранения огромных объемов информации, и это вопрос пространства и возможностей. Старые методы обработки оказывают значительное влияние на скорость из-за чего возможны замедления и «торможения», это еще один вопрос интенсивности: чем быстрее процесс, тем продуктивнее результат (больше отдача). По причине разрозненности качества, форматов и источников возникает проблема неоднородности и неструктурированности.

В Big Data еще существует проблема отсутствия предела «величины» данных. Её достаточно трудно предугадать, поэтому неизвестно какие технологии и сколько финансовых затрат потребуется на её устранение. Поскольку ресурсы не бесконечны, то становится нецелесообразно хранить все возможные данные. Зачастую возникает вопрос отказа от части данных. Данный вопрос является главной причиной отсрочки внедрения в компании проектов BigData.[2]

Не меньшей проблемой становится подбор данных, требуемых к обработке, и алгоритма анализа. Зачастую отсутствует понимание, какие данные следует собирать и хранить, а какие можно игнорировать. Исходя из этого становится очевидной проблема нехватки в данной отрасли профессиональных специалистов, которые способны на глубинный анализ, создание отчетов для решения бизнес-задач и как следствие извлечение прибыли из BigData.

При разработки поисковых систем возникают вопросы, имеющие этический характер – это необходимость без согласия пользователя накопления о нем всей доступной информации: IP-адреса, геолокации, личных данных, интересов, онлайн режимов работы. Данная информация позволяет демонстрировать контекстную рекламу в соответствии с интересами пользователя в интернете.

Исходя из того, что BigData накапливает различную информацию о пользователях возникает проблема обеспечения безопасности хранения и использования данных. Некоторые данные используются для решения многих бизнес-задач. Но никто не может гарантировать безопасность аналитической платформы, которая получает данные о посетителях в автоматическом режиме (просто за посещение данного сайта).

Также в современной информационной среде наблюдается вирусная активность и хакерские атаки, которые не всегда могут выдержать даже отлично защищенные серверы правительственных

спецслужб, тем более распределенные информационные системы удаленных объектов предприятия.

В этом году члены Cloud Security Alliance[2] провели анализ и выявили главные проблемы современных компаний, возникающие при хранении больших объемов данных:

– безопасность вычисления в распределенных программных системах;

– безопасность нереляционных баз данных;

– безопасность хранения данных;

– проверка достоверности;

– мониторинг безопасности в режиме реального времени;

– data mining и аналитика, которые сохраняют конфиденциальность;

– шифрование управления доступом и обеспечение безопасности соединения;

– фрагментарный контроль доступа, т.е. возможность сегментировать данные по степени их конфиденциальности;

– происхождение данных. [5]

Таким образом, основные принципы работы с использованием технологии Big Data могут быть сформулированы следующим образом:

1) Горизонтальная масштабируемость. Данных очень много, поэтому любая система, в которой возможна обработка больших данных,

[2] Cloud Security Alliance – некоммерческая организация, являющаяся лидером в области разработки стандартов, рекомендаций и инициатив, направленных на повышение безопасности и защищенности использования облачных вычислений. Деятельностью Cloud Security Alliance руководит обширная коалиция ведущих мировых экспертов в отрасли, передовых корпораций в IT-индустрии, известных профессиональных ассоциаций и крупнейших провайдеров облачных услуг (Google, eBay, SalesForce.com, RackSpace и др.). Российское отделение Cloud Security Alliance Russian Chapter создано на базе ассоциации профессионалов в области информационной безопасности (Russian Information Security Systems Professional Association, RISSPA), которая действует с 2006 года и объединяет специалистов и руководителей в области информационной безопасности России.

должна быть расширяемой. Следовательно, рост объема данных должен быть прямо пропорционален увеличению количества железа в кластере.

2) Отказоустойчивость. Принцип из первого пункта подразумевает, что в одном кластере может быть много машин. Ярким примером является Hadoop-кластер Yahoo, который имеет более 42000 машин. Очевидно, что часть включенных в кластер машин будет гарантированно выходить из строя. Данные сбои должны учитываться методами, применимыми в работе с большими данными.

3) Локальность данных. Принцип локальности данных, заключающийся в обработке данных на машине, на которой хранятся эти данные. Иначе при работе с большими системами, где данные распределены по большому количеству машин, в случае нахождения физически данных на одном сервере, а обработки на другом – расходы на передачу данных могут превысить расходы на саму обработку.

Большинство современных методов работы с большими данными в большей или меньшей степени соответствуют этим трем принципам. Для того, чтобы им следовать – необходимо придумывать какие-то методы, способы и парадигмы разработки средств разработки данных.

Методы и техники анализа, применимые к большим данным, выделены в отчете McKinsey[3]:

– методы класса Data Mining: обучение ассоциативным правилам, классификация, кластерный анализ, регрессионный анализ;

– краудсорсинг (с данными работает достаточно широкий, неопределенных круг лиц, обеспечивающий категоризацию и обогащение данных, люди находятся вне трудовых отношений и привлечённы на основании публичной оферты);

[3] McKinsey & Company – международная консалтинговая компания, специализирующаяся на решении задач, связанных со стратегическим управлением.

– data fusion and integration (набор техник, для интеграции разнородных массивов данных из различных источников для возможности глубинного анализа);

– машинное обучение, включая обучение с учителем и без учителя, а также Ensemble learning (англ.) — использование моделей, построенных на базе статистического анализа или машинного обучения для получения комплексных прогнозов на основе базовых моделей;

– искусственные нейронные сети, сетевой анализ, оптимизация, в том числе генетические алгоритмы;

– распознавание образов;

– прогнозная аналитика;

– имитационное моделирование;

– spatial analysis - класс методов, использующих топологическую, геометрическую и географическую информацию о данных;

– статический анализ;

– визуализация аналитических данных — представление информации в виде рисунков, диаграмм, с использованием интерактивных возможностей и анимации как для получения результатов, так и для использования в качестве исходных данных для дальнейшего анализа. [4]

Технология BigData используется во многих сферах деятельности. В первую очередь это выявление с помощью BigData предпочтений клиентов, оценка эффективности маркетинговых кампаний или проверка анализа рисков. На рисунке представлены результаты опроса IBM Institute, о направлениях использования BigData в компаниях. [1]

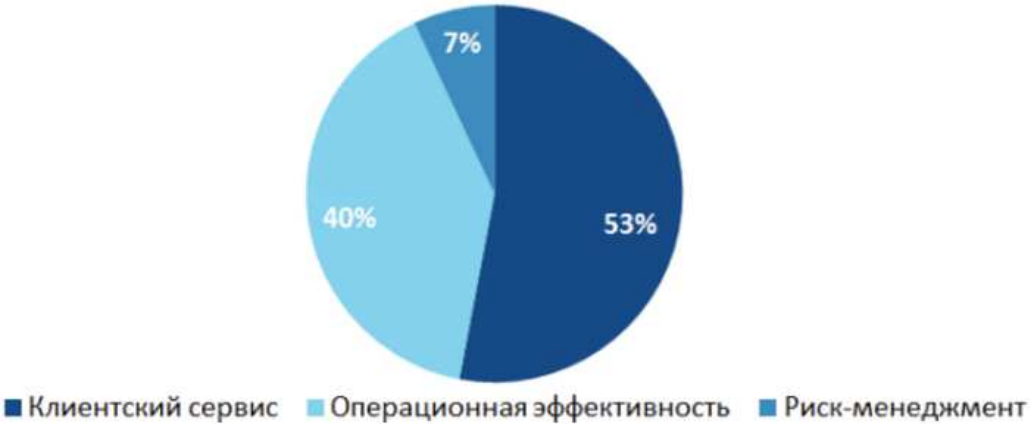

Рисунок – Результаты опроса IBM Institute

Из диаграммы видно, что большинство компаний используют большие данные в сфере клиентского сервиса, вторым по популярности направлением является операционная эффективность, в сфере управления рисками BigData на текущий момент наименее распространены.

В настоящее время Big Data является одной из самых быстрорастущих сфер информационных технологий, согласно статистике общий объем получаемых и хранимых данных удваивается каждые 1-2 года.

Стоит отметить, что Big Data обычно хранятся и организуются в распределенных файловых системах. Иными словами, информация хранится на нескольких жестких дисках стандартных персональных компьютеров. Так называемая «карта» (map) отслеживает, на каком компьютере и диске хранится конкретная часть информации. Каждую часть информации принято сохранять несколько раз, как правило – три, чтобы обеспечить надежность и отказоустойчивость.

Информация по территориально-распределенному комплексу, которую необходимо обрабатывать и хранить системе контроля и наблюдения на первый взгляд может показаться незначительной. Первой основной функцией программного продукта является распознавание

государственных регистрационных знаков. Поэтому следует хранить и обрабатывать следующую информацию:

– номерные знаки, используемые для обучения системы;

– характеристики камер видеофиксации;

– видеозаписи с камер наблюдения;

– кадры с изображением отдельных автомобилей до распознавания;

– кадры с распознанными номерами;

– информация об автомобилях, принадлежащих компании, владельцу комплекса;

– журнал фиксации движения автотранспорта.

Для анализа поведенческих ситуаций требуется хранить и обрабатывать следующую информацию:

– поведенческие образы сотрудников, разграниченные в соответствие с нормами поведения, используемые для обучения системы;

– видеопоток с камер;

– кадры для анализа ситуаций;

– кадры с зафиксированным аномальным поведением;

– информация о срабатывании систем безопасности.

Наименее затратным по объему информации является реализация третьей функции – системы контроля и управления доступом, которая потребует лишь место на диске для:

– хранения информации о посещении сотрудниками комплекса;

– хранения отчетов;

– информация о сотрудниках.

И это только основные объекты, требующие хранения и обработки информации.

Преимущество использования технологии Big Data заключается в хранении всего возможного объема информации, не беспокоясь о том, какая часть данных актуальна для последующего анализа и принятия

решения. Поэтому для распределенных систем контроля и наблюдения для предприятий с территориально-распределенной структурой, где объекты находятся на значительном удалении актуальны вопросы хранения всех объемов, имеющейся информации для последующего анализа и обработки. Недостатком является последующая трудоемкая обработка этих «всех данных». Поскольку физические накопители, используемые для хранения информации, могут достигать своего предела, то необходимо грамотно обрабатывать полученные данные и сохранять лишь ценную для компании информацию.

Список литературы

1. Веретенников А. В. BigData: анализ больших данных сегодня // Молодой ученый. — 2017. — №32. — С. 9-12.

2. Смирнов Д. Защита Big Data: проблемы и решения // Еженедельник IT Weekly. – 2017. – № 13. – С. 22-30.

3. Проектирование интеллектуальных систем контроля доступа на объекты территориально-распределенных комплексов / Л.А. Видовский, М.В. Янаева, А.Г. Мурлин и др. // Политематический сетевой электронный научный журнал Кубанского государственного аграрного университета (Научный журнал КубГАУ) [Электронный ресурс]. – Краснодар: КубГАУ, 2016. – №09(123). – IDA [article ID]: 1231609130. – Режим доступа: http://ej.kubagro.ru/2016/09/pdf/130.pdf.

4. Натан Марц, Джеймс Уоррен. Большие данные. Принципы и практика построения масштабируемых систем обработки данных в реальном времени. – Москва: Вильям, 2016. – 368с.

5. Cloud Security Alliance [Электронный ресурс]. – Москва, 2017. – Режим доступа: https://cloudsecurityalliance.org

References

1. Veretennikov A. V. BigData: analiz bol'shih dannyh segodnja // Molodoj uchenyj. — 2017. — №32. — S. 9-12.

2. Smirnov D. Zashhita Big Data: problemy i reshenija // Ezhenedel'nik IT Weekly. – 2017. – № 13. – S. 22-30.

3. Proektirovanie intellektual'nyh sistem kontrolja dostupa na ob#ekty territorial'no-raspredelennyh kompleksov / L.A. Vidovskij, M.V. Janaeva, A.G. Murlin i dr. // Politematicheskij setevoj jelektronnyj nauchnyj zhurnal Kubanskogo gosudarstvennogo agrarnogo universiteta (Nauchnyj zhurnal KubGAU) [Jelektronnyj resurs]. – Krasnodar: KubGAU, 2016. – №09(123). – IDA [article ID]: 1231609130. – Rezhim dostupa: http://ej.kubagro.ru/2016/09/pdf/130.pdf.

4. Natan Marc, Dzhejms Uorren. Bol'shie dannye. Principy i praktika postroenija masshtabiruemyh sistem obrabotki dannyh v real'nom vremeni. – Moskva: Vil'jam, 2016. – 368s.

5. Cloud Security Alliance [Jelektronnyj resurs]. – Moskva, 2017. – Rezhim dostupa: https://cloudsecurityalliance.org