УДК [611.018.51+615.47]:612.086.2

ИСПОЛЬЗОВАНИЕ ОБЛАЧНЫХ ВЫЧИСЛЕНИЙ ПРИ РАБОТЕ С BIG DATA: ОСОБЕННОСТИ И ТРУДНОСТИ

Е.А. Гриз
Студент 4 курса, кафедра ЭВМ, БГУИР

С.Н. Нестеренков
Кандидат технических наук, доцент, декан факультета компьютерных систем и сетей

А.Н. Марков
Старший преподаватель, магистр технических наук, заместитель начальника Центра информатизации и инновационных разработок БГУИР

Центр информатизации и инновационных разработок Белорусского государственного университета информатики и радиоэлектроники, Республика Беларусь
Белорусский государственный университет информатики и радиоэлектроники, Республика Беларусь.
E-mail: evgeniy.hryz@gmail.com, s.nesterenkov@bsuir.by, a.n.markov@bsuir.by

Е. А. Гриз
Студент 4 курса специальности "Вычислительные машины, системы и сети" БГУИР.

С. Н. Нестеренков
Кандидат технических наук, доцент, декан факультета компьютерных систем и сетей Белорусского государственного университета информатики и радиоэлектроники, доцент кафедры программного обеспечения информационных технологий. Автор публикаций на тему машинного обучения, алгоритмов принятия решений, искусственных нейронных сетей и автоматизации.

А. Н. Марков
Магистр технических наук, старший преподаватель кафедры ПИКС, заместитель начальника Центра информатизации и инновационных разработок Белорусского государственного университета информатики и радиоэлектроники.

Аннотация. Big Data относится к огромным разнородным объемам данных, которые могут поступать из разных источников с меняющимися скоростями. Обработка и анализ этих данных могут помочь компаниям лучше узнать потребности своих клиентов, их модели поведения, оптимизировать использование своих ресурсов и увеличить прибыль. Облачные вычисления подходят для хранения и обработки Big Data, поскольку предоставляют on-demand сервисы, способные быстро масштабироваться в зависимости от нагрузки и требующие оплаты по факту использования. В этой статье описаны особенности использования облачных вычислений с Big Data, а также возникающие при этом проблемы и трудности.

Ключевые слова: Big Data, Cloud, Облачные вычисления.

Введение.

Объем информации, собранной с мобильных устройств, смарт-часах, публикаций в социальных сетях, фотографий и видео увеличивается с каждой минутой и почти удвабивается каждый год. Этот объем данных включает структурированные и неструктурированные данные, а также данные из разных источников, поэтому они не могут храниться в обычной реляционной базе данных. Компании в сфере здравоохранения,

финансов, инженерии используют эти данные для анализа и принятия решений. Так они могут лучше понять потребности своих клиентов и предсказывать их желания, а вследствие оптимизировать использование своих ресурсов.

Облачные вычисления – это модель обеспечения доступа по требованию к вычислительным ресурсам, таким как устройства хранения данных, серверам, приложениям. При этом ресурсы, доступные клиенту могут легко масштабироваться в зависимости от потребностей клиента, а клиент будет платить только за то, что он использовал. Облачные вычисления позволяют перенести затраты и обязанности от клиента к поставщику облачных услуг, что способствует развитию малых компаний, для которых начало работы в IT-бизнесе требует больших усилий и затрат. Иными словами, компании передают управление частью своего бизнеса поставщику облачных услуг.

Эластичность, оплата по факту использования, низкие первоначальные инвестиции, короткое время выхода на рынок и передача рисков провайдеру облачных услуг делают облачные вычисления универсальной парадигмой для развертывания новых приложений, которые были бы экономически нецелесообразны в традиционных условиях.

Основная цель этой статьи – предоставить обзор возможностей и трудностей, связанных с Big Data и показать реальные примеры компаний, использующих Big Data и облачные вычисления.

Big Data: общая концепция.

Огромные объемы данных, для хранения, управления и анализа которых не подходят традиционные базы данных, получили название Big Data. Для их хранения, манипулирования и анализа требуется масштабируемая архитектура. Хотя большие данные в основном связаны с хранением огромных объемов данных, они также касаются способов их обработки и создания на их основе некоторых выводов [1]. Big Data обладает следующими характеристиками:

1) Объем (Volume). При обработке и хранении больших объемов данных нужно учитывать масштабируемость, чтобы система могла расти; доступность, гарантирующая доступ к данным и способам выполнения над ними операций; пропускную способность и производительность.

2) Разнообразие (Variety) означает различные типы данных из различных источников.

3) Скорость (Velocity) относится к различным скоростям, с которыми потоки данных могут поступать или выходить из системы. В условиях постоянно возрастающего темпа генерации данных скорость передачи и доступа к данным должна оставаться высокой, чтобы обеспечить доступ в реальном времени различным приложениям, зависящим от этих данных.

4) Ценность (Value) означает реальную ценность данных, насколько полезными эти данные могут быть. Огромные объемы данных бесполезны, если они не представляют ценности.

5) Достоверность (Veracity) означает степень надежности данных, то есть конфиденциальность, целостность и доступность.

Big Data и облачные вычисления.

Облачные вычисления предоставляют вычислительные услуги, такие как серверы, хранилища, базы данных, сети, программное обеспечение, аналитика через интернет для более быстрого внедрения инноваций, гибких ресурсов, трудоемких вычислений, параллельной обработка данных [2, 3]. Существуют различные категории, по которым могут быть сгруппированы системы облачных вычислений. Одним из наиболее часто используемым критерием для классификации этих систем является уровень абстракции, предлагаемый для пользователя системы. Выделяют три разных уровня: инфраструктура как услуга (IaaS), платформа как услуга (PaaS) и программное обеспечение как услуга (SaaS), как показано на рисунке 1.

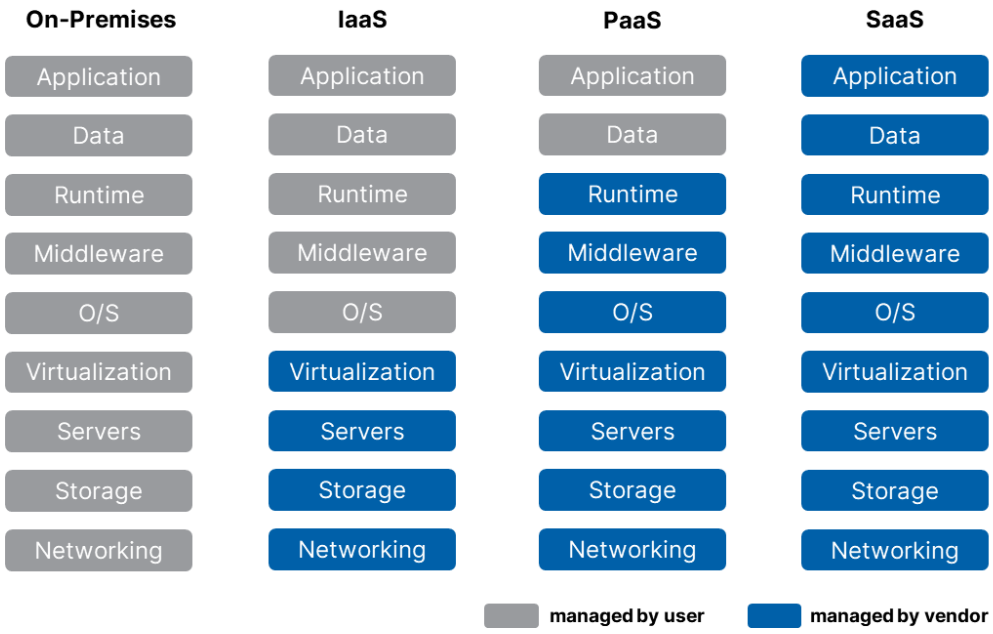

Рисунок 1 – Различия между моделями облачных вычислений

SaaS: программное обеспечение как услуга

Программное обеспечение как услуга представляет собой крупнейший рынок облачных вычислений и наиболее часто используемый вариант среди облачных сервисов. SaaS доставляет приложения пользователям через международную сеть, доступ осуществляется клиентом через браузер, что устраняет необходимость для клиента загружать или устанавливать какое-либо ПО. В SaaS производитель управляет средой выполнения приложений, данными, промежуточным ПО, ОС, виртуализацией, серверами, хранилищем и сетью.

PaaS: платформа как услуга

Модель «Платформа как услуга» предоставляет аппаратные и программные инструменты через Интернет, которые используются разработчиками для создания специализированных приложений. PaaS делает разработку, тестирование и развертывание приложений быстрым, простым и экономичным процессом. Эта модель позволяет бизнесам проектировать и создавать приложения, интегрированные в программные компоненты PaaS, в то время как провайдеры облачных услуг управляют ОС, виртуализацией, серверами, хранилищами, сетью и самим программным обеспечением PaaS. Такие приложения являются масштабируемыми, так как они обладают облачными характеристиками.

IaaS: инфраструктура как услуга

Модель облачных вычислений «Инфраструктура как услуга» обеспечивает платформу для доступа, мониторинга и управления инфраструктурами удаленного центра обработки данных, такими как сервисы хранения, вычисления и сетевые сервисы, используя технологию виртуализации. Пользователи IaaS отвечает за управление приложениями, данными, средой выполнения, промежуточным программным обеспечением и ОС, в то время как поставщики все еще управляют виртуализацией, серверами, жесткими дисками, хранилищем и сетью. IaaS обеспечивает те же возможности, как и центры обработки данных, но без необходимости их физической поддержки.

Большие данные и облачные вычисления тесно связаны друг с другом. Большие данные больше связаны с извлечением ценности из данных, в то время как облачные

вычисления ориентированы на масштабируемое, гибкое обслуживание по запросу и с оплатой по факту использования. Большие данные требуют огромных вычислительных мощностей по требованию и распределенного хранилища, в то время как облачные вычисления беспрепятственно предоставляют компьютерные ресурсы по запросу, необходимые для анализа больших данных.

Примеры использования Big Data в облаке.

Netflix использует AWS (Amazon Web Services) практически для всех своих потребностей в области вычислений и хранения данных, включая базы данных, аналитику, механизмы рекомендаций, транскодирование видео и многое другое. Netflix собирает данные о 151 миллионе пользователей и внедряет анализ данных для выявления моделей поведения клиентов. Затем, использует эту информацию для рекомендации фильмов и сериалов на основе предпочтений пользователей. Например, Netflix знает дату и время просмотра сериала, какое устройство было использовано, был ли сериал поставлен на паузу и был ли просмотр возобновлен после паузы и т. д. Используя все эти данные Netflix создает рекомендации на основе пользовательских предпочтений и может предсказывать успех новых сериалов. На ноябрь 2020 г. показатель удержания клиентов Netflix составил 74% [4].

Мобильные телефоны Nokia используются многими людьми для связи. Чтобы улучшить взаимодействие пользователей с их телефонами, Nokia собирает большие объемы данных с мобильных телефонов. Компания реализовала Hadoop data warehouse как инфраструктуру, которая могла хранить этот огромный объем неструктурированных данных, собираемых с мобильных телефонов, сервисов, лог-файлов и других источников. Nokia также способна выполнять сложную обработку на своих данных и получать аналитические сведения об взаимодействиях пользователей с телефонами [5].

Проблемы и трудности.

Поскольку объем данных растет быстрыми темпами, хранение всех данных физически неэффективно. Поэтому корпорации должны создавать политики для определения жизненного цикла и даты истечения срока действия данных. Кроме того, они должны определить, кто и с какой целью может получать доступ к данным клиентов. По мере того как данные перемещаются в облако, также появляются такие проблемы как безопасность и конфиденциальность.

Big Data СУБД обычно имеют дело с большим количеством данных из нескольких источников, и поэтому неоднородность также является проблемой, которая в настоящее время изучается. К другим проблемам относятся восстановление данных после катастрофы и быстрая загрузка данных в облако.

Безопасность.

Проблема безопасности становится важной для корпораций при рассмотрении возможности загрузки данных в облако. Возникают такие вопросы, как кто является реальным владельцем данных, где находятся данные, кто имеет к ним доступ и какие разрешения у них есть.

а) Кто является реальным владельцем данных, и кто имеет к ним доступ?

Клиенты облачного провайдера платят за услугу и загружают свои данные в облако. Однако кому именно действительно принадлежат данные? Более того, может ли провайдер использовать данные клиента? Какой уровень доступа провайдер имеет к данным и с какими целями может их использовать? Может ли поставщик облачных услуг извлечь выгоду из этих данных?

Таким образом, в интересах клиента предоставить ограниченный доступ к данным, чтобы свести к минимуму доступ к ним и гарантировать, что только уполномоченный персонал обращается к данным по уважительной причине.

Большинство проблем с безопасностью обычно возникают внутри организаций, поэтому имеет смысл, чтобы компании анализировали все политики доступа к данным, прежде чем заключать контракт с поставщиком облачных услуг.

б) Где находятся данные?

Конфиденциальные данные, которые являются законными в одной стране, могут быть незаконными в другой стране, поэтому нужно договориться с клиентом о местонахождении данных, так как его данные могут считаться незаконными в некоторых странах и привести к судебному преследованию.

Конфиденциальность.

Сбор данных и использование аналитических инструментов вызывает ряд проблем с конфиденциальностью. Обеспечение безопасности и конфиденциальности данных стало чрезвычайно сложным, поскольку информация распространена и по всему миру, часто не в одном экземпляре для защиты от потери. Аналитики часто собирают конфиденциальную информацию пользователей, такую как их медицинские записи, потребление энергии, онлайн-активность, покупки в супермаркетах и т. д. Для решения этой проблемы традиционно организации использовали различные методы деидентификации (анонимизация или шифрование данных), чтобы отделить данные от реальных личностей. Хотя в последние годы было доказано, что даже при анонимизации данных их все же можно повторно идентифицировать и отнести к конкретным лицам [6 – 8]. Способ решить эту проблему заключался в том, чтобы рассматривать все данные как персональные данные, подпадающие под действие нормативно-правовой базы.

Законы о конфиденциальности и защите данных основаны на индивидуальном контроле над информацией и на таких принципах, как минимизация и ограничение данных. Тем не менее, неясно, всегда ли сведение к минимуму сбора информации является практическим подходом к конфиденциальности. В настоящее время подходы к конфиденциальности основаны на согласии пользователя.

Разнородность данных.

Большие данные относятся к большим объемам данных, но также и к разным скоростям (т. е. данные поступают с разной скоростью в зависимости от скорости источника и задержки сети) и большому разнообразию. Данные поступают в Big Data СУБД с разной скоростью и в разных форматах из разных источников. Это связано с тем, что разные сборщики информации предпочитают свои собственные форматы и протоколы для записи данных, а характер различных приложений также приводит к различным представлениям данных. Работа с таким большим разнообразием данных и различными скоростями – сложная задача, которую должны решать Big Data системы. Эта задача усугубляется тем, что постоянно создаются новые типы файлов без какой-либо стандартизации.

Управление данными.

Цены на аппаратное обеспечение снижаются и продолжат снижаться. Однако Big Data СУБД включает и другие затраты: обслуживание инфраструктуры, энергия, лицензии на ПО. Все эти затраты входят в совокупную стоимость владения (Total cost of ownership).

Совокупная стоимость владения увеличивается пропорционально росту объема хранимых данных, поэтому этот рост должен контролироваться. Это означает, что компании должны создавать политики, определяющие жизненный цикл данных и то, как управлять данными в течение жизненного цикла.

Восстановление данных после катастрофы.

Потеря данных для бизнеса означает потенциальную потерю прибыли. В случае чрезвычайных ситуаций или опасных происшествий, таких как землетрясения, наводнения и пожары, потери данных должны быть минимальными. Для выполнения этого требования в случае какого-либо инцидента данные должны быть быстро доступны с минимальной задержкой и потерями. Это реализуется созданием резервных копий данных и процедурами восстановления, используя резервные копии. Поскольку периодические резервные копии петабайт данных отнимает много времени и денег, необходимо определить подмножество данных ценных для организации для резервного копирования.

Заключение.

Объем данных, генерируемых в настоящее время различными видами деятельности общества никогда не был таким большим и продолжает расти. Эта тенденция рассматривается компаниями как способ получения преимущества над своими конкурентами: если одна компания способна извлечь ценную информацию из данных быстрее чем другие, она сможет получить больше клиентов, увеличить доход из одного клиента, оптимизировать свою работу, и сократить расходы. Тем не менее, аналитика больших данных по-прежнему остается сложной задачей, требующей дорогого ПО и больших вычислительных инфраструктур.

Облачные вычисления помогают решить эти проблемы, предоставляя ресурсы по требованию с затратами, пропорциональными фактическому использованию. Более того, облачные вычисления позволяет быстро масштабировать инфраструктуру, адаптируя системы к фактическим нуждам.

Хотя Big Data системы являются мощными инструментами, которые позволяют как предприятиям, так и науке получать представление о данных, есть некоторые проблемы, требующие дальнейшего изучения. Необходимо приложить дополнительные усилия для разработки механизмов безопасности и стандартизации типов данных. Восстановление данных при их потере и определение жизненного цикла данных также являются проблемами.

В этом материале сделан обзор Big Data в облачных средах, подчеркнуты их преимущества и показано, что обе технологии очень хорошо работают вместе, а также представлены проблемы, с которыми сталкиваются эти две технологии.

Список литературы

[1] Big Data: Concepts, Technology and Architecture / B. Balusamy [и др.]. – Hoboken : John Wiley & Sons, Inc., 2021. – 368 с.

[2] Kavis, M. J. Architecting the Cloud: Design Desicions for Cloud Computing Service Models / M. J. Kavis. – Hoboken : John Wiley & Sons, Inc., 2014. – 224 с.

[3] Big-Data Analytics and Cloud Computing: Theory, Algorithms and Applications / M. Trovati [и др.]. – Berlin : Springer International Publishing, 2015. – 185 с.

[4] Disney+ takes customer retention crown from Netflix [Электронный ресурс]. – Режим доступа: https://secondmeasure.com/datapoints/disney-takes-customer-retention-crown-from-netflix/. – Дата доступа: 25.03.2022.

[5] Nokia: Using Big Data to Bridge the Virtual & Physical Worlds [Электронный ресурс]. – Режим доступа: https://hadoopilluminated.com/hadoop_illuminated/cached_reports/Cloudera_Nokia_Case_Study_Hadoop.pdf. – Дата доступа: 25.03.2022.

[6] Taran Singh Bharati, Challenges, Issues, Security And Privacy Of Big Data / Taran Singh Bharati // International Journal of Scientific & Technology Research. – 2020. – Т. 9, № 2. – С. 5.

[7] Omer, T. Privacy in the Age of Big Data: A Time for Big Decisions / T. Omer, P. Jules // Stanford Law Review. – 2012. – . – Т. 64, № 63. – С. 7.

[8] Security and Privacy for Big Data, Cloud Computing and Applications / W. Ren [и др.]. – London : The Institution of Engineering and Technology, 2019. – 328 с.

[9] Нестеренков, С.Н. Применение больших данных в электронном образовании / С.Н. Нестеренков, М.И. Макаров, Н.В. Ющенко, А.Д. Радкевич // BIG DATA and Advanced Analytics = BIG DATA и анализ высокого уровня : сб. материалов V Междунар. науч.-практ. конф. (Республика Беларусь, Минск, 13-14 марта 2019 года). В 2 ч. Ч. 2 / редкол. : В. А. Богуш [и др.]. - Минск : БГУИР, 2019. - С. 242-245.

[10] Беляк, А. А. Анализ производительности технологии Hadoop / А. А. Беляк, С. Н. Нестеренков // BIG DATA and Advanced Analytics = BI DATA и анализ высокого уровня: сб. научных статей VII Междунар. науч.-практ. конф. (Республика Беларусь, Минск, 19-20 мая 2021 года): / редкол. : В. А. Богуш [и др.]. – Минск : Бестпринт, 2021. – С. 343–346.

USING CLOUD COMPUTING WHEN WORKING WITH BIG DATA: FEATURES AND CHALLENGES

E.A. HRYZ
Pregraduate student of the BSUIR

S.N. NESTERENKOV,
PhD, Associate Professor, Dean of the Faculty of Computer Systems and Networks

A.N. MARKOV
Senior lecturer of the department, Deputy head of the Center for Informatization and Innovative Developments

Center for Informatization and Development of the Belarusian University of State Informatics and Radioelectronics, Republic of Belarus
E-mail: evgeniy.hryz@gmail.com, s.nesterenkov@bsuir.by, a.n.markov@bsuir.by

Abstract. Big Data refers to huge heterogeneous amounts of data that can come from different sources at varying speeds. Processing and analyzing this data can help companies better understand the needs of their customers, their behavior patterns, optimize the use of their resources and increase profits. Cloud computing is suitable for storing and processing Big Data, as it provides on-demand pay-as-you-go services that can quickly scale depending on the workload. This article describes the features of using cloud computing with Big Data, as well as the problems and challenges that arise.

Key words: Big Data, Cloud, Cloud computing.