УДК 004.421

ОСНОВНЫЕ ТЕНДЕНЦИИ ИСПОЛЬЗОВАНИЯ БОЛЬШИХ ДАННЫХ

Диденко А.А., аспирант гр. ИВа-191, III курс
Научный руководитель: Пимонов А.Г., д.т.н., профессор
Кузбасский государственный технический университет имени Т.Ф. Горбачева
г. Кемерово

Большие данные доказывают свою ценность для организаций всех типов и размеров в самых разных отраслях. Предприятия, которые активно используют большие данные, получают ощутимые преимущества при ведении бизнеса от повышения эффективности операций до оптимизации продуктов и услуг для клиентов. В результате этого организации находят применение таким большим хранилищам данных, а также развиваются технологии, методы анализа и подходы к работе с большими данными. Кроме того, продолжают появляться новые методы и архитектуры для сбора, обработки, управления и анализа всех видов данных в организации.

Работа с большими данными – это больше, чем просто работа с большими объемами хранимой информации. И она имеет свои специфические проблемы. Объем (V) – это лишь одна из многих «V» (velocity, volume, value, variety and veracity) проблем больших данных [1], которые организациям необходимо решить. Обычно также существует значительное разнообразие данных – от структурированной информации, хранящейся в базах данных, распределенных по всей организации, до огромных объемов неструктурированных и полуструктурированных данных, находящихся в файлах, изображениях, видео, датчиках, системных журналах, тексте и документах, включая бумажные документы, которые ждут оцифровки. Кроме того, эта информация часто создается и изменяется с большой скоростью (velocity) и имеет разный уровень качества данных (veracity), что создает дополнительные проблемы при управлении данными, их обработке и анализе. Четыре основные тенденции в области больших данных помогают организациям решать эти проблемы. Все они представлены на рис. 1.

Больше данных, увеличение разнообразия данных способствуют прогрессу в обработке и развитию граничных вычислений. Неудивительно, что скорость генерации данных продолжает увеличиваться. Большая часть этих данных генерируется не из транзакций, происходящих в базах данных, а из других источников, включая облачные системы, интеллектуальные устройства, такие как смартфоны и голосовые помощники, а также потоковое видео. Эти данные в значительной степени не структурированы и в прошлом в основном оставались необработанными и неиспользованными организациями, превращая их в так называемые темные данные.

Рисунок 1 – Основные тенденции в области больших данных

Это подводит нас к самой большой тенденции в области больших данных: источники, не являющиеся базами данных, будут по-прежнему оставаться доминирующими генераторами данных, что, в свою очередь, заставит организации пересмотреть свои потребности в обработке данных. Голосовые помощники и устройства IoT (internet of things) [2], в частности, способствуют быстрому росту потребностей в управлении большими данными в таких разных отраслях, как розничная торговля, здравоохранение, финансы, страхование, производство и энергетика, а также на самых разных рынках государственного сектора. Этот взрыв разнообразия данных заставляет организации думать не только о традиционных хранилищах данных, но и о средствах обработки всей этой информации.

Кроме того, необходимость обработки генерируемых данных перемещается на сами устройства, поскольку отраслевые прорывы в вычислительной мощности привели к разработке все более совершенных устройств, способных самостоятельно собирать и хранить данные, не нагружая сеть, хранилище и вычислительную инфраструктуру. Например, мобильные банковские приложения могут выполнять множество задач по удаленному внесению и обработке чеков без необходимости отправлять изображения туда и обратно в центральные банковские системы для обработки.

В знак того, что все это становится актуальным, исследование планов расходов на ИТ на 2022 год, проведенное подразделением Tech Target Enterprise Strategy Group, показало, что главными приоритетами организаций для поддержки своих инициатив в области данных являются продвижение использования технологий следующего поколения, а также перенос данных из устаревших систем в современные и расширение возможности обработки данных там, где они были созданы.

Использование устройств для распределенной обработки воплощено в концепции граничных вычислений (Edge computing) [3], которая перекладывает вычислительную нагрузку на сами устройства до отправки данных на серверы. Граничные вычисления оптимизируют производительность и хранилище, уменьшая потребность в передаче данных по сетям, снижая затраты на вычисления и обработку, особенно на облачное хранилище, пропускную

способность и расходы на обработку. Граничные вычисления помогают ускорить анализ данных и быстрее реагируют на запросы пользователя.

Например, в секторе здравоохранения быстро растет рынок мобильных устройств, таких как Fitbit, Apple Watch и устройства Google Android, что стимулирует рост телемедицины и позволяет поставщикам медицинских услуг собирать важные данные о пациентах в режиме реального времени. В дальнейшем эти данные используются для анализа через приложения для обработки больших данных и аналитики, предназначенных для улучшения результатов лечения пациентов.

Потребность в хранении больших данных стимулирует инновации в облачных и гибридных облачных платформах, рост озер данных (data lakes). Чтобы справиться с неумолимым увеличением объема генерируемых данных, организации тратят все больше своих ресурсов на хранение этих данных в ряде облачных и гибридных облачных систем, оптимизированных для всех V больших данных [1]. В предыдущие десятилетия организации управляли собственной инфраструктурой хранения, что приводило к созданию огромных центров обработки данных, которыми предприятия должны были управлять: защищать и эксплуатировать. Переход к облачным вычислениям изменил эту динамику. Перекладывая ответственность на поставщиков облачной инфраструктуры, таких как AWS, Google, Microsoft и IBM, организации могут работать с почти неограниченными объемами новых данных и платить за хранение и вычислительные мощности по требованию без необходимости поддерживать свои собственные дата-центры.

Некоторые отрасли сталкиваются с трудностями при использовании облачной инфраструктуры из-за нормативных или технических ограничений. Например, строго регулируемые отрасли, такие как здравоохранение, финансовые услуги и правительство, имеют ограничения, препятствующие использованию общедоступной облачной инфраструктуры. Поэтому, за последнее десятилетие поставщики облачных услуг разработали способы предоставления более удобной для регулирования инфраструктуры, а также гибридные подходы, которые сочетают аспекты сторонних облачных систем с локальными вычислениями и хранилищем для удовлетворения критически важных потребностей инфраструктуры. Эволюция как общедоступных, так и гибридных облачных инфраструктур, несомненно, будет продолжаться по мере того, как организации ищут экономические и технические преимущества облачных вычислений.

Помимо инноваций в области облачного хранения и обработки, предприятия переходят на новые подходы к архитектуре данных, которые позволяют им справляться с проблемами разнообразия, достоверности и объема больших данных. Вместо того, чтобы пытаться централизовать хранение данных в хранилище данных, которое требует сложного и трудоемкого извлечения, преобразования и загрузки данных, предприятия развивают концепцию озера данных (data lakes) [4]. Озера данных хранят структурированные и неструктурированные наборы данных в их собственном формате. Этот подход переносит

ответственность за преобразование и обработку на конечные точки, которые имеют разные потребности в данных. Озеро данных также может предоставлять общие сервисы для анализа и обработки данных.

Внедрение расширенной аналитики, машинного обучения и других технологий искусственного интеллекта резко возрастает. При огромном количестве генерируемых данных традиционные подходы к аналитике сталкиваются с трудностями, поскольку их нелегко автоматизировать для масштабного анализа данных. Технологии распределенной обработки, особенно продвигаемые платформами с открытым исходным кодом, такими как Hadoop и Spark, позволяют организациям обрабатывать петабайты информации с высокой скоростью. Системы машинного обучения и искусственного интеллекта позволяют им легче выявлять закономерности и аномалии, а также делать прогнозы быстрее, чем раньше. Предприятия используют технологии аналитики больших данных для оптимизации своих инициатив в области бизнес-аналитики, переходя от медленных инструментов отчетности, зависящих от технологии хранилища данных, к более интеллектуальным, быстро реагирующим приложениям, которые обеспечивают большую прозрачность поведения клиентов, бизнес-процессов и операций в целом.

Ни одна технология не была столь революционной для аналитики больших данных, как системы машинного обучения и искусственного интеллекта (ИИ). ИИ используется организациями любого размера для оптимизации и улучшения своих бизнес-процессов. Машинное обучение позволяет им легче выявлять шаблоны и обнаруживать аномалии в больших наборах данных, чтобы обеспечить прогнозную аналитику и другие расширенные возможности анализа данных. Сюда входят системы распознавания изображений, видео и текстовых данных, автоматизированная классификация информации, возможности обработки естественного языка для чат-ботов и анализа голоса и текста, машинная автоматизация бизнес-процессов, системы с высокой степенью персонализации и рекомендаций, и системы поиска оптимальных решений в море данных (sea of data) [5].

Действительно, с помощью искусственного интеллекта и машинного обучения компании используют свои среды больших данных для обеспечения более глубокой поддержки клиентов с помощью интеллектуальных чат-ботов и более персонализированного взаимодействия, и при этом не требуется значительного увеличения штата службы поддержки клиентов. Эти системы с поддержкой ИИ способны собирать и анализировать огромные объемы информации о клиентах и пользователях, особенно в сочетании со стратегией озера данных, которая может собирать широкий спектр информации из многих источников.

Предприятия также видят инновации в области визуализации данных. Люди понимают значение данных, когда они представлены в наглядной форме, например, в виде диаграмм и графиков. Появляющиеся формы визуализации данных предоставляют возможности аналитики с поддержкой ИИ даже обычным бизнес-пользователям. Это помогает организациям выявлять ключевые

идеи, которые могут улучшить процесс принятия решений. Компании открывают для себя ценность принятия решений на основе больших данных всей организации. Расширенные формы инструментов визуализации и аналитики позволяют пользователям даже задавать вопросы на естественном языке, при этом система автоматически определяет правильный запрос и отображает результаты в зависимости от контекста.

DataOps и управление данными выходят на первый план. Многие аспекты обработки, хранения и управления большими данными будут развиваться еще долгие годы. Большая часть этих инноваций обусловлена технологическими потребностями, а также частично изменениями в том, как мы думаем о данных и относимся к ним.

Одной из областей инноваций является появление DataOps, методологии и практики, которые фокусируются на гибких, итерационных подходах к работе с полным жизненным циклом данных, когда они проходят через организацию. Вместо того, чтобы думать о данных по частям с отдельными людьми, занимающимися созданием, хранением, транспортировкой, обработкой и управлением данными, процессы и платформы DataOps удовлетворяют потребности организации на протяжении всего жизненного цикла данных от создания до архивирования.

Точно так же организации все чаще сталкиваются с вопросами управления данными, конфиденциальности и безопасности. В прошлом предприятия часто не слишком заботились о конфиденциальности данных и управлении ими, но новые правила делают их гораздо более ответственными за то, что происходит с личной информацией в их системах. Из-за широко распространенных нарушений безопасности, подрыва доверия клиентов к корпоративным методам обмена данными и проблем с управлением данными на протяжении их жизненного цикла организации все больше вовлекаются в управление данными и прилагают больше усилий для надлежащей защиты данных и управления ими, особенно когда они пересекают международные границы. Появляются новые инструменты, чтобы гарантировать, что данные остаются там, где они должны оставаться, и будут защищены и надлежащим образом отслежены на протяжении всего их жизненного цикла.

В совокупности эти тенденции в области больших данных делают работу в пространстве больших данных интересной в 2022 году и, без сомнения, в обозримом будущем.

Список литературы:

1. What are the 5 V's of Big Data? [Электронный ресурс]. – URL: https://www.teradata.com/Glossary/What-are-the-5-V-s-of-Big-Data (дата обращения: 29.03.2022).

2. Что такое интернет вещей? [Электронный ресурс]. – URL: https://www.kaspersky.ru/resource-center/definitions/what-is-iot (дата обращения: 29.03.2022).

3. What is edge computing? [Электронный ресурс]. – URL: https://www.ibm.com/cloud/what-is-edge-computing (дата обращения: 29.03.2022).

4. Data Lake [Электронный ресурс]. – URL: https://www.bigdataschool.ru/wiki/data-lake (дата обращения: 29.03.2022).

5. A Sea of Data: Pattern Recognition and Corporate Animism (Forked Version) [Электронный ресурс]. – URL: https://mediarep.org/bitstream/handle/doc/13257/Pattern_Discrimination_1I-22_Steyerl__Sea_of_Data_.pdf?sequence=5 (дата обращения: 29.03.2022).