УДК 528.065/.067:004.6

ТЕХНОЛОГИИ BIG DATE И НЕКОТОРЫЕ АСПЕКТЫ ИХ ИСПОЛЬЗОВАНИЯ

К.С. АЛЕКСЕЕВА, А.В. КИРИЛЛОВА
(Представлено: П.Ф. Парадня)

В статье рассматривается трактовка термина «Big Date» или по-другому «большие данные» и различные аспекты этой терминологии. Приведены практические примеры использования Big Data и риски, связанные с этим. Уделено внимание применению больших данных в науках о Земле.

Определение Big Data впервые появилось в конце 2000-х г. в английском словаре The Oxford English dictionary, которое можно перевести как: «Различные инструменты, а также подходы и методы к обработке как структурированных, так и неструктурированных больших данных для выполнения каких-либо задач и целей» [1].

Большие данные часто называют наборам большого количества информации, которые имеют сложную гетерогенную или неопределенную структуру. Иногда большие данные упоминаются как неструктурированная информация, но это неверно – большие данные всегда имеют структуру, они могут быть сложными, поскольку поступают из разных источников и содержат совершенно разную информацию или полностью неизвестны. То есть, как правило, невозможно свести этот беспорядок условно в одну таблицу.

Большие данные, хотя и существуют уже несколько лет, раньше не были очень востребованными. Их было сложно обрабатывать и анализировать – для этого требовались значительные вычислительные мощности, длительное время и финансовые затраты. Все изменилось, когда появилась технология обработки многогигабайтных массивов информации в быстрой оперативной памяти. Прорывы в этой области связаны с выпуском свободно распространяемой платформы Hadoop, которая включает библиотеки, утилиты и фреймворки для работы с Big Data. Компоненты Hadoop сегодня используются в большинстве коммерческих платформ и систем таких компаний, как SAP, Oracle, IBM.

В 2001 г. вышло важнейшее исследование Дуга Ланей, которое указало три главные характеристики больших данных: объем, скорость, разнообразие (так называемые три «V»: Volume, Velocity, Variety) [1]. Большие данные характеризуются большими размерами, большими скоростями их новой генерации и притока, неоднородностью и неупорядоченностью.

Если обратиться к сервису GoogleTrends, который пользуется большими данными для анализа (массивы поисковых запросов и анализ документов на наличие ключевых слов), то рост популярности поискового запроса и темы «большие данные» в мире начинается с 2011 г.

На данный момент цифровые технологии заполнили большую часть жизни человека. Например, размер данных, находящихся в хранилищах по всему миру, увеличивается с каждой минутой, и поэтому условия хранения должны меняться с одинаковой скоростью, открывая новые возможности для увеличения объема. В настоящее время термин «Big Data» все чаще используется для обзора не только массивов данных, но и инструментов, используемых для их обработки, а также потенциальных преимуществ, которые могут быть получены в результате трудоемкого анализа.

Все знают, что основные информационные потоки создают и обрабатывают не люди. Это вызвано внедрением роботизированных машин в жизнь людей, которые на протяжении всего времени связаны друг с другом. Например, системы наблюдения, смартфоны, механизмы для мониторинга, сенсоры и многое другое. Это повлекло за собой огромный скачок в росте больших данных, что вынуждает наращивать количество рабочих серверов, развивать и включать новые data-центры.

Подумав о практических аспектах использования больших данных, исследователи больших данных уверяют, что в будущем людей ожидает ситуация, когда мир адаптируется к каждому человеку. Эксперты преобразовали данные в последовательности цифр для всех человеческих обязательств и интересов – теперь остается только понять, как эти данные будут использоваться.

В настоящие время применение больших данных можно наблюдать в различных сферах деятельности. Например, в розничной торговле – это история о потребителях: что они покупают, подробная информация о чеках, скидках в настоящие время в различных торговых центрах и другое.

Банки и страховые фирмы так же собирают информацию о клиентах, их действиях, денежных переводах и т.д.

Большие данные также определяют формирование социального сектора. Возможность собирать и анализировать информацию со счетчиков воды, газа и электроэнергии является первым и наиболее важным шагом на пути к значимому потреблению ресурсов как на уровне домохозяйств, так и в масштабах жилищных компаний. Так, например, внедрение больших данных позволило эстонской распредели-

тельной компании Elektrilevi вместе с Egissson, которая запустила интеллектуальную систему учета электроэнергии, повысить эффективность производства на 20% только за первые два года реализации плана и избежать дорогостоящих потерь из-за своевременного обнаружения неисправностей [11].

В телекоммуникациях большие данные – это вся служебная информация, поступающая с сетевых устройств, история использования всех видов услуг, информация о местоположении и весь трафик, который можно анализировать, а также вся служебная информация, вплоть до текстовых сообщений. Операторы имеют доступ к такой информации, но в соответствии с «Законом о защите персональных данных» они не имеют права использовать эту информацию без разрешения владельца устройства. Однако, они могут анализировать обширный трафик, который не содержит личной информации, могут выполнять сортировку клиентов по типам в соответствии с поведением и предпочтениями потребителей и т.п. Кроме того большие данные могут применяться для выявления и предотвращения краж, мошенничества (действий киберпреступников, направленных на кражу денежных средств).

По сути, фирмы обращаются к большим данным, чтобы повысить эффективность принимаемых решений и снизить риски неправильных решений. Но считается, что сами большие данные также сопряжены с рисками:

– Риск конфиденциальности. Если произошла потеря данных, и они попали в руки конкурентов, для компании это достаточно серьёзный инцидент, так как будет нанесёт ущерб ее репутации.

– Риск потери данных. Это относится к потере данных в целом, например, в результате действий мошенников или чрезвычайных ситуаций. Чтобы предотвратить возникновение этих проблем, необходимо создать резервную копию данных.

– Риск переполнения хранилища. Обычно это происходит в результате неправильного хранения данных. Необходимо правильное формирование хранилища и тщательный отбор данных.

– Риск снижения эффективности больших данных. Отбор действительно важных данных должен быть выполнен достаточно обоснованно. Из-за накопления нежелательной информации полезность содержимого данных снижается.

– Риск ошибок больших данных. Даже незначительные упущения могут привести к значительным проблемам. И ошибки не исключены в случае работы с большим объемом данных. В итоге необходимо периодически пересматривать данные и анализировать эффективность инструментов.

– Риск экономической неблагоприятности. Не всегда аналитики находят необходимую для них информацию в нужном объёме, и избавиться от этого риска невозможно. Однако, правильно распоряжаясь ресурсами, есть шанс его минимизировать.

Ключом к появлению больших данных можно считать неявную информацию. Неявные знания - это тип знаний, которые нелегко передать другим. Сама по себе неявная информация характеризуется слабой структурой, которая считается признаком больших данных [2]. Например, особенность получения информации при цифровой аэрофотосъемке и дистанционном зондировании заключается в том, что данные сначала накапливаются, а затем начинают обрабатываться через некоторое время. В то же время на аэрокосмических изображениях могут быть изображения сложных комбинаций различных объектов. В аэрокосмической фотографии изображение создается как совокупность наложенных друг на друга объектов. Например, облака могут закрывать часть области, при фотографировании водной поверхности она становится прозрачной до определенной глубины. И все объекты на снимке отображаются так, как если бы они были на одном слое.

Сложность как источник больших данных. Индивидуальность термина «сложность» заключается в том, что это объект (или атрибут), связанный с другим объектом. Это приводит ко всевозможным сложностям. Например, выделяют виды сложности по связи с объектом: сложность организационно-технической системы, сложность процесса (действия), сложность явления, условная колмогоровская сложность, простая колмогоровская сложность, префиксная сложность, сложность ситуации, сложность теории и т. д. Таким образом, термин «сложность» требует указания связанного объекта, для которого оценивается сложность. В противном случае степень сложности станет недостаточной [3; 4]. Сложности качественно разных сущностей или разных атрибутов могут быть несопоставимыми.

Различают разные сложности одного и того же объекта:

– структурная сложность объекта [5];

– сложность процессов, в которых участвует объект;

– сложность получения решения в разрешенное время - временная сложность;

– сложность, возникающая из-за ограниченного объема памяти вычислительной системы, обрабатываемой в больших объемах - емкостная сложность;

– сложность определения местоположения в пространстве - пространственная сложность позиционирования;

– сложность формы объекта - морфологическая сложность;

– сложность ситуации, в которой находится объект - ситуационная сложность;
– сложность местоположения объекта - позиционная сложность;
– сложность декодирования объектов - криптографическая сложность [5];
– сложность определения явления, с которым объект связан;
– сложность теории, описывающей поведение объекта и т. д.

Следовательно, для полноты исследования можно упомянуть «обобщенную сложность» объекта и «качественную сложность».

Для характеристики больших данных обычно используются критерии «три V»: объем (volume – V1), скорость (velocity – V2), разнообразие (variety – V3), и следует добавить сложность (covplex – C1).

Например, обработка серии аэрофотоснимков и космических снимков приводит к получению файлов большого размера. Эта ситуация усугубляется появлением сканеров с высоким разрешением, которые значительно увеличивают информационную емкость изображений и создают проблемы для их обработки. Критерий V1 проявляется в науках о Земле, например, при хранении гигабайтных и терабайтных файлов при работе с многомасштабными картами и банками пространственных данных [4]. Критерий V2 – при уравнивании больших систем уравнений [6]. Это также происходит при оперативном управлении движущимися объектами. Критерий V3 характеризует моделирование сложных систем большого регионального охвата и семантический анализ информационных объектов [7]. Критерий C1 применяется при топологическом анализе сложных транспортных и других сетей [8; 9].

В заключение следует отметить, что появление больших данных можно рассматривать как отражение процессов глобализации. Их анализ требует применения высокопроизводительных вычислительных технологий и инструментов. Их основная проблема – это, во-первых, сложность, а во-вторых, большой физической объем информации. Большие объемы данных создают проблемы при создании источников информации из этих данных. По сути, большие данные рассматриваются как новая форма информационного барьера. Они качественно отличаются от обычных данных тем, что при их обработке и анализе происходит семантический разрыв. С одной стороны, большие данные приводят к созданию и решению новых задач, а с другой - к созданию интегрированных и инклюзивных систем и технологий.

ЛИТЕРАТУРА

1. Как зарождалась эра Big Data [Электронный ресурс]. – Режим доступа: https://rb.ru/story/era-big-data/. – Дата доступа: 30.03.2022.
2. Сигов, А.С. Неявное знание: оппозиционный логический анализ и типологизация / А.С. Сигов, В.Я. Цветков // Вестник Российской Академии Наук. – 2015. – С. 800–804.
3. Вьюгин, В.В. Колмогоровская сложность и алгоритмическая случайность / В.В. Вьюгин. – М. : ИППИ РАН, 2012. – 131 с.
4. Тихонов, А.Н. Основы управления сложной организационно-технической системой. Информационный аспект. / А.Д. Иванников, И.В. Соловьёв, В.Я. Цветков. – М. : МаксПресс, 2010. – 228 с.
5. Tsvetkov V.Ya. Complexity Index // European Journal of Technology and Design. – 2013. – Vol. (1). №1. – P. 64–69.
6. Михайлович, К. Геодезия (уравнительные вычисления) / К. Михайлович ; пер. С.В. Лебедева ; под ред. В.Д. Большакова. – М. : Недра, 1984.
7. Цветков В.Я., Железняков В.А. Мультимасштабная электронная карта как основа системы учёта земель // Государственный советник. – 2014. – № 1. – С. 28–37.
8. Железняков, В.А. Интеллектуальное обновление информации в банке геоданных // Инженерные изыскания. – 2012. – № 5. – С. 58–61.
9. Свами, М. Графы, сети и алгоритмы. / М. Свами, К. Тхуласираман. – М. : Мир, 1984. – Т. 198.
10. Большие данные (Big Data) – одна из ключевых технологий будущего [Электронный ресурс]. – Режим доступа: https://www.kommersant.ru/doc/2614791. – Дата доступа: 15.04.2022.