ПРЕДИЗВИКАТЕЛСТВА НА ИНФОРМАЦИОННАТА СИГУРНОСТ ПРИ ГОЛЕМИТЕ ДАННИ

Красимир Трайков, Стефка Толева-Стоименова
Университет по библиотекознание и информационни технологии

Резюме: Бързият растеж на големите данни промени коренно много стопански сектори, предоставяйки нови възможности за иновации и повишаване на ефективността. В същото време огромният обем, скорост и разнообразие на данните доведоха до сериозни предизвикателства, свързани с информационната сигурност. Настоящият доклад разлежда явлението големи данни, техните свойства и приложение в различни стопански сектори. Основната цел на изследването е свързана с анализ на актуалното състояние на проучванията в областта на сигурността на големите данни, както и съществуващите методи и техники за подобряване на информационната сигурност и поверителност.

Ключови думи: големи данни, информационна сигурност, поверителност.

Въведение

Експоненциалното нарастване на обема на данните, предизвикано от технологичния прогрес, доведе до появата на концепцията за големите данни (Big data). Тези данни се отличават с огромен обем, висока скорост на генериране и разнообразие, поради многобройните им източници като социални медии, интернет на нещата (IoT) и корпоративни системи. Анализът на големите данни играе ключова роля за организациите в различни сектори като им позволява да извличат ценни прозрения и да вземат информирани решения.

С увеличаване на обема на данните расте и необходимостта от тяхната защита. Информационната сигурност в контекста на големите данни е от критично значение поради чувствителната природа на данните и потенциала за значителни вреди при компрометиране. Рисковете включват пробиви в данните, неоторизиран достъп и манипулиране на данни, което може да има сериозни последици, както за индивидите, така и за организациите. Затова е от изключителна важност да се гарантира цялостта, поверителността и достъпността на данните.

Целта на този доклад е да представи актуалните предизвикателства и решения в областта на сигурността на големите данни. Докладът е структуриран в две основни секции. Първата разглежда ключови аспекти на големите данни и тяхното приложение в различни отрасли. Втората секция представлява преглед на текущото състояние на сигурността на големите данни и различни методи и техники за подобряване на информационната

сигурност и защитата на личната информация. В заключение са очертани бъдещи направления за изследвания в тази област.

Методология на изследването

Методологията на изследването се базира на задълбочен научната литература, свързана с информационната сигурност на големите данни, както и систематизирането и обобщаването на събраната информация, с оглед идентифициране на ключовите предизвикателства в тази област.

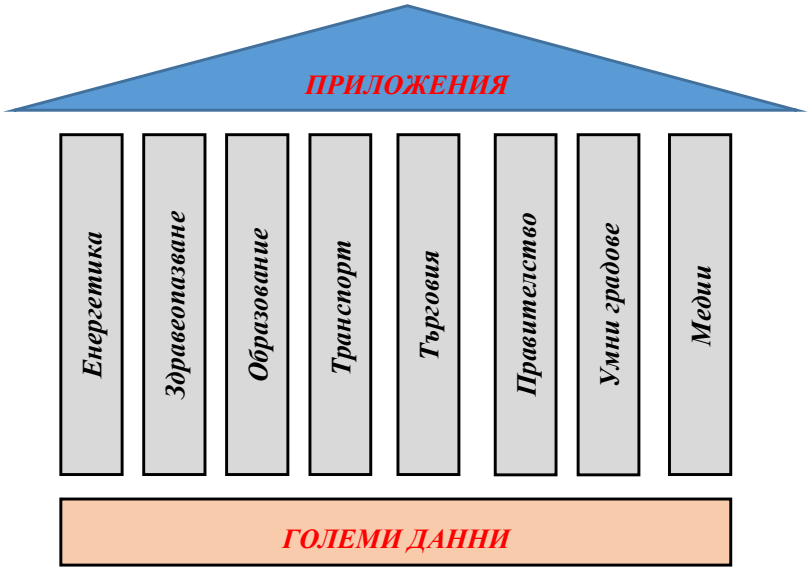
Големи данни и техни приложения

През 1997 г. за първи път се споменава терминът „големи данни“. Обемът на данните започва да се превръща в предизвикателство, тъй като надхвърля възможностите на компютърните памет и дискове, изисквайки нови ресурси и технологии [1]. През 2001 г. Laney въвежда концепцията за „трите V-та“ (Volume, Variety, Velocity), които определят обема, многообразието и скоростта на големите данни [2]. С времето дефиницията на големите данни се разширява, добавяйки четвърта характеристика – **истинност** (Veracity, Enterprice Big Data Framework), или **променливост** (Variability, NIST Big Data Interoperability Framework), а по-късно и **стойност** (Value, Yuri Demchenko). Броят на характеристиките продължава да нараства като през 2017 г. Tom Shafer изготвя списък с 42 V-та, включващ нови свойства като **виралност** (Virality), **местоположение** (Venue), **речник** (Vocabulary), **неопределеност** (Vagueness) и др. [3]. Обикновено редът на тези характеристики не отразява тяхната значимост, тъй като при различни приложения тя е различна.

Големите данни намират приложение в множество сектори, както е показано на фиг. 1. В здравеопазването се използват за анализ на клинични резултати, прогнозиране на епидемии, в банковия сектор – за предотвратяване на измами, оптимизация на на инвестиционните стратегии и анализ на риска, в транспорта – за проследяване и оптимизация на операциите, на дейностите в „умни“ градове. В търговията и маркетинга те помагат за оптимизиране на веригата на доставките и събиране на данни за клиентските предпочитания, а в медиите, включително социалните медии, се използват за анализ на потребителските предпочитания и поведение, и т.н.

Генерирането, обработката, съхранението, споделянето, преносът, анализът и визуализацията на големи данни създават нови технологични предизвикателства. Те изискват увеличена изчислителна мощност, нови методи за съхранение и високоскоростни канали за пренос на данни. Събирането на данни в реално време предоставя нови възможности за по-добро разбиране на информацията, като например комбинирането на административни данни с големи обеми от данни от различни източници – мобилни устройства, сензори, социални медии и други обществени източници. Вследствие на това се

появява необходимостта както от нови инструменти и методи, така и от нови умения и компетенции за работа с тези данни.



Фиг. 1. Приложение на големите данни

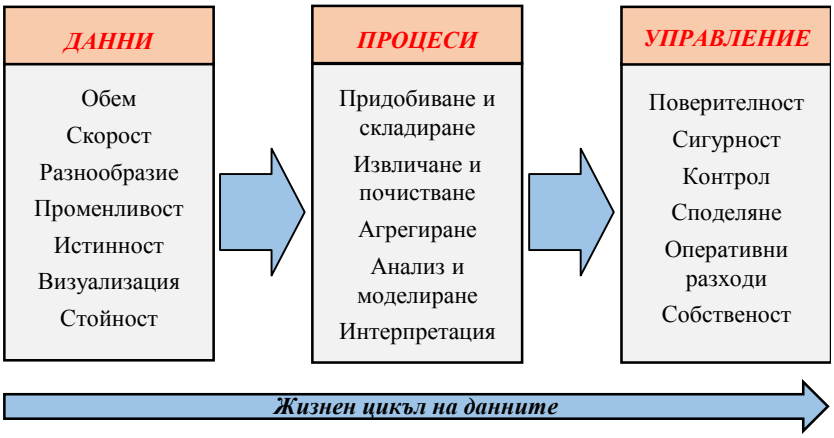
Информационна сигурност при големи данни

Информационната сигурност обхваща мерки за защита на информацията от неоторизиран достъп, неволно или преднамерено разкриване, повреда, унищожаване, модификация или загуба (ISO 2013). За решаването на проблемите, свързани с информационната сигурност, се прилагат комплексни решения, които осигуряват конфиденциалност, цялостност и достъпност на данните през целия им жизнен цикъл.

При работа с големи данни се появяват редица предизвикателства, свързани с техните характеристи и управлението им, както е показано на фиг. 2 [4]. Въпреки напредъка в проучванията, свързани със сигурността на големите данни, все още съществуват значителни трудности в различни етапи, като генериране, събиране, обработка, съхранение и споделяне на данни.

Огромният обем данни затруднява прилагането на ефективни мерки за сигурност. Освен това, хетерогенността на източниците и форматите добавя допълнителна сложност към защитата на данните. Това означава, че традиционните методи за защита, използвани при по-малки и хомогенни масиви от данни, често не са достатъчни и не могат да се прилагат. Бързото и непрекъснато генериране на данни също води до трудно осигуряване на сигурността в реално време [5]. Сред основните предизвикателства са

необходимостта от по-ефективни методи за запазване на поверителността, по-добри техники за откриване на вътрешни заплахи, както и цялостни рамки за интегриране на мерките за сигурност в различни платформи на големи данни [6]. Липсата на стандартизирани протоколи и насоки в различни сектори и региони допълнително ограничава възможността за прилагане на последователни и ефективни мерки за сигурност на глобално ниво [7].



Фиг. 2. Предизвикателства при работа с големи данни (адаптирана от Sivarajah, U. *et al., 2016)*

Управлението на данните е тясно свързано със сигурността, и ключови проблеми като пробиви на данни, вътрешни заплахи и трудност при осигуряването на поверителността на данните. Наличните решения включват шифроване (криптиране) на данни, контрол на достъпа и различни методи за запазване на поверителността.

Шифроване на данни: Шифроването на данни, както в покой (data at rest), така и в движение (data in motion) е основна мярка за защита срещу неразрешен достъп и пробиви. Криптографските алгоритми намират приложение при идентификация и автентикация, контрол на достъпа и осигуряване на поверителността и цялостта на данните. С развитието на изчислителната мощност на компютърните системи, се развиват и усъвършенстват алгоритмите като хомоморфно шифроване, сигурно многостранно изчисление (multi-party computation) и диференциална поверителност, които позволяват изчисленията върху шифрованите данни без тяхното разшифроване [8]. Хомоморфното шифроване все още се сблъсква с проблеми по отношение на производителността при работа с големи обеми от данни [9, 10]. Сериозен напредък в проучванията представлява появата на схеми за напълно хомоморфно шифроване (FHE) [9, 11]. Сигурното многостранно

изчисление позволява комбиниране на данни от множество източници, без да се разкриват основните чувствителни данни на субектите. По време на процеса на изчисление данните остават шифровани и разпределени между различни страни. Съществуващите алгоритми за сигурно многостранно изчисление имат недобра мащабируемост по отношение на размера на данните, което ги прави прекалено бавни при работа с големи данни. В [12] е предложено решение, което използва хибридни протоколи, което включва допълнителни стъпки извън стандартното многостранно изчисление, с цел ускоряване на процеса.

Контрол на достъпа: Контролът на достъпа се осъществява с помощта на надеждни технологии и политики като защитен хардуер, софтуер и биометрични мерки и услуги. Съществуват различни модели за контрол на достъпа, като дискреционен контрол на достъпа (DAC), задължителен (MAC), базиран на роли (RBAC) и базиран на атрибути (ABAC). Тези методи са от съществено значение за запазване на поверителността при извършване на анализ на големи данни, позволяват подробна настройка на правата за достъп, но управлението им става сложно в големи и разпределени системи [13].

Методи за запазване на поверителността: Поверителността често се разглежда като част от сигурността, като при нея фокусът е върху скриването или премахването на връзката между данните и на това за кого се отнасят данните, или върху шифроването на данните, за да се предпазят поверителни данни. Като следствие, това означава, че защитата на поверителността се отнася до:

1) анонимизация или псевдонимизация на лични данни чрез изтриване или разделяне на идентификатори и

2) скриване на съдържанието чрез шифроване или други мерки за сигурност.

Анонимизацията на данните представлява необратимо унищожаване на връзката между данните и техния произход. Получените данни не трябва да могат да бъдат използвани за идентифициране на дадена самоличност или да се свържат с други данни за същия индивид. Съществуват два основни подхода за извършване на процеса на анонимизирането: първият се базира на рандомизация, а вторият – на генерализация. Рандомизацията включва различни технически методи (включително добавяне на шум, пермутация, диференциална поверителност), които изменят точността на данните с цел да се премахне връзката между тях и субекта на данни [14, 15]. Генерализацията, от своя страна, се изразява в генерализиране, или разводняване, на атрибутите на субектите на данни чрез промяна на мащаба или поредността на величините (например регион вместо град, месец

вместо седмица). Използват се техники като k-анонимност, също l-многообразие, t-близост.

От друга страна, псевдонимизацията е процес, при който данните, идентифициращи директно даден индивид, се заменят с данни, които го идентифицират индиректно (чрез изкуствени идентификатори). За разлика от анонимизацията, псевдонимизацията не премахва напълно цялата идентифицираща информация от данните и по този начин възможността за идентификация на субектите. Тя само ограничава връзката между данните и оригиналната идентичност на индивида (например чрез криптографски методи). Псевдонимизираните данни са защитени срещу идентификация, но те все още са лични и позволяват повторно възстановяване на идентичността, докато анонимизираните данни не подлежат на повторно идентифициране.

И двата процеса – анонимизация и псевдонимизация на данните – могат да доведат до загуба на точност и полезност на данните [15]. В [16] е направен анализ на подходи за защита на чувствителни данни в Интернет на нещата, организирани в осем категории: техники за анонимизация (включително k-анонимност, l-разнообразие, t-близост), методи за обърквane, многослойно машинно обучение, децентрализирано машинно обучение, хомоморфно шифроване, модели за защита на потоци от данни с различни нива на поверителност, създаване на сбити версии на данните и създаване на сигурни индивидуални хранилища на данни. В [17] е предложен модел за защита на чувствителни данни, събирани от социални мрежи, който включва три стъпки: клъстериране на данните, осигуряване на k-анонимност и прилагане на l-разнообразие и t-близост. В [18] е разработен концептуален модел, представящ съвременни подходи, методи и техники за защита на лични и чувствителни данни.

Изводи/Дискусия

Вследствие на извършения сравнителен анализ на различните техники за сигурност и техните предимства и недостатъци става ясно, че за постигане на цялостна и ефективна защита на данните е необходим холистичен подход, който адресира едновременно различни аспекти на сигурността. Наред с това съществуват редица етични и правни съображения, свързани с гарантирането на неприкосновеността на личните данни, с постигане на баланс между полезност от събираните данни и поверителност, особено в области, в които анализът на данни може да донесе значителни ползи за обществото. Общият регламент за защита на данните (General Data Protection Regulation, GDPR) въведе редица значителни промени спрямо действащата преди това правна рамка и постави по-високи изисквания към експертите относно защитата на личните данни на европейските граждани.

Подобни правни рамки съществуват и в други региони, например Законът за поверителност на потребителите в Калифорния (California Consumer Privacy Act, CCPA). Етичното използване на големите данни изисква постоянно адаптиране на законодателството към развитието на технологиите и тясно сътрудничество между политици, специалисти и общество.

Заключение

Днес потребителите и организациите са изправени пред предизвикателството на големите данни. Анализът на големи данни се използва успешно в редица сфери за подпомагане на процесите на вземане на решение чрез извличане на надеждна релевантна информация. В същото време се появява необходимостта от обезпечаването на информационната сигурност и разработването на по-ефективни техники за запазване на поверителността, подобряване на мащабируемостта на съществуващите решения и интеграция на мерките за сигурност в различни платформи за големи данни. За повишаване на сигурността на данните се използват иновативни технологии като блокчейн и квантово изчисление. Също така се изследват възможности относно разработване на потребителски ориентирани решения за сигурност, които да бъдат внедрявани и управлявани от различни по големина организации. Сътрудничеството между академичната общност, индустрията и правителствата ще бъде от ключово значение при справяне с тези нововъзникващи предизвикателства.

References/Литература

1. **Cox**, M., D. **Ellsworth**. Application-controlled demand paging for out-of-core visualization, Proceedings of the IEEE 8th conference on Visualization, 1997, pp. 235 – 244.
2. **Laney** D. 3D data management: Controlling data volume, velocity and variety. META Group Research Note, 6 February 2001.
3. **Shafer** T. The 42 V's of Big Data and Data Science, 2017. https://www.elderresearch.com/blog/42-v-of-big-data.
4. **Sivarajah**, U., M. M. **Kamal**, Z. **Irani**, V. **Weerakkody**. Critical analysis of Big Data challenges and analytical methods. – In: *Journal of Business Research*, 2017, Vol. 70, pp. 263 – 286. https://doi.org/10.1016/j.jbusres.2016.08.001.
5. **Hashem**, I. A. T., I. **Yaqoob**, N. B. **Anuar**, S. **Mokhtar**, A. **Gani**, S. U. **Khan**. The rise of "big data" on cloud computing: Review and open research issues. – In: *Information Systems*, Vol. 47, 2015, pp. 98 – 115. https://doi.org/10.1016/j.is.2014.07.006.
6. **Santosh**, A., N. **Ranganathan**. A System Architecture for the Detection of Insider Attacks in Big Data Systems. – In: *IEEE Transactions on Dependable and Secure Computing*, 2016, Vol. 15, pp. 974 – 987. Doi: 10.1109/TDSC.2017.2768533.

7. **Wang**, Y., L. **Kung**, T. **Byrd**. Big data analytics: Understanding its capabilities and potential benefits for healthcare organizations. – In: *Technological Forecasting and Social Change*, 2018, Vol. 126, pp. 3 – 13.
8. **Hervais**, S. F. Big Data: Opportunities and Privacy Challenges, 2015. https://doi.org/10.48550/arXiv.1502.00823.
9. **Gentry**, C. Fully homomorphic encryption using ideal lattices. – In: *STOC '09:* Proceedings of the forty-first annual ACM symposium on Theory of computing, 2009, Vol. 9, pp. 169 – 178. https://doi.org/10.1145/1536414.1536440.
10. **Acar**, A., H. **Aksu**, A. **Selcuk Uluagac**, M. **Conti**. A Survey on Homomorphic Encryption Schemes: Theory and Implementation. – In: *ACM Comput. Surv*., July 2019, 51, 4, Article 79. https://doi.org/10.1145/3214303.
11. **Alabdulatif**, A., I. **Khalil**, X. **Yi**. Towards secure big data analytic for cloud-enabled applications with fully homomorphic encryption. – In: *Journal of Parallel and Distributed Computing*, 2020, Vol. 137, pp. 192 – 204. https://doi.org/10.1016/j.jpdc.2019.10.008.
12. **Volgushev**, N., M. **Schwarzkopf**, B. **Getchell**, M. **Varia**, A. **Lapets**, A. **Bestavros**. Conclave: secure multi-party computation on big data. – In: *Proceedings* of the Fourteenth EuroSys Conference 2019 (EuroSys '19). Association for Computing Machinery, New York, NY, USA, Article 3, pp. 1 – 18, 2019. https://doi.org/10.1145/3302424.3303982.
13. **Brown** L. Access Control Mechanisms in Distributed Systems. – In: *Computer Security Journal*, 2019, Vol. 40, pp. 98 – 115.
14. **Zhao**, Y., J. **Chen**. A Survey on Differential Privacy for Unstructured Data Content. – In: *ACM Computing Surveys (CSUR)*, Vol. 54, Issue 10, Article No 207, pp. 1 – 28. https://doi.org/10.1145/3490237.
15. **Majeed** A., S. **Lee**. Anonymization Techniques for Privacy Preserving Data Publishing: A Comprehensive Survey. – In: *IEEE Access*, 2021, Vol. 9, pp. 8512 – 8545. Doi: 10.1109/ACCESS.2020.3045700.
16. **Torre**, D., A. **Chennamaneni**, A. **Rodriguez**. Privacy-Preservation Techniques for IoT Devices: A Systematic Mapping Study. – In: *IEEE Access*, 2023, Vol. 11, pp. 16323 – 16345. https://doi.org/10.1109/ACCESS.2023.3245524.
17. **Gangarde**, R., A. **Sharma**, A. **Pawar**. Enhanced Clustering Based OSN Privacy Preservation to Ensure k-Anonymity, t-Closeness, l-Diversity, and Balanced Privacy Utility. – In: *Computers, Materials & Continua*, 2023, 75 (1), pp. 2171 – 2190. https://doi.org/10.32604/cmc.2023.035559.
18. **Ivanova**, M., I. **Trifonova**. Privacy Preserving Techniques and Their Applications in Elearning. – In: *Science* Series "Innovative STEM Education", Vol. 05, Institute of Mathematics and Informatics – Bulgarian Academy of Sciences, pp. 93 – 102. https://doi.org/10.55630/STEM.2023.0512

За авторите

Красимир Трайков е докторант в катедра „Компютърни науки" на УниБИТ. Има две магистърски степени: по телекомуникационни технологии от Колежа по телекомуникации и пощи и по корпоративна сигурност от Военната академия „Г.

С. Раковски. Работи като Senior Information Security Compliance Officer. Интересите му са в сферата на информационната и корпоративната сигурност, киберсигурността, защитата и етичните аспекти на данните.

За контакт с автора: 4623495-2@unibit.bg

Стефка Толева-Стоименова е доцент в катедра „Компютърни науки“ на УниБИТ. Завършила е Техническия университет – София, а през 2011 г. получава докторска степен в УниБИТ. Нейните публикации и изследователски интереси са в научните области информатика, информационни системи, наука за информирането и наука за данните.

За контакт с автора: s.toleva@unibit.bg

INFORMATION SECURITY CHALLENGES IN BIG DATA CONTEXT

Krasimir Traykov, Stefka Toleva-Stoimenova
University of Library Studies and Information Technologies

Abstract: Various sectors has been revolutionized by the rapid growth of big data, which offering unprecedented opportunities for innovation and efficiency. At the same time, the massive volume, velocity, and variety of data have led to significant issues and challenges in information security. This report examines issues related to the “Big data” phenomenon, its main features and applications in some strategic sectors. The aim of the research is to examine the current state of big data security, to analyze different methods and techniques for improving information security and privacy.

Keywords: big data, information security, privacy.

About the Authors

Krasimir Traykov is a PhD student in Computer Science the Department at the ULSIT. He has completed two MSc degrees: in Telecommunication Technologies from the College of Telecommunications and Posts in Sofia and Corporate Security from the G. S. Rakovski Military Academy. He works as a Senior Information Security Compliance Officer. His interests are in the fields of information and corporate security, cybersecurity, data protection and data ethics.

To contact the Author: 4623495-2@unibit.bg

Stefka Toleva-Stoimenova is an Associate Professor in Computer Science Department at the ULSIT. She has obtained her MSc degree from the Faculty of Automation and System Design, Technical University – Sofia. In 2011, she received a PhD degree from ULSIT. Her publications and main research interests are in the fields of Informatics, Information Sistems, Informing Science, and Data Science.

To contact the Author: s.toleva@unibit.bg