BLUE SKIES DEPARTMENT

Trustworthy Processing of Healthcare Big Data in Hybrid Clouds

Surya Nepal
Commonwealth Scientific and Industrial Research Organization

Rajiv Ranjan
Commonwealth Scientific and Industrial Research Organization

Kim-Kwang Raymond Choo
University of South Australia

As we delve deeper into the "Digital Age," we're witnessing an explosive growth in the volume, velocity, variety, veracity, and value (the 5Vs) of data produced over the Internet. According to recent Cisco[1] and IBM[2] reports, we now generate 2.5 quintillion bytes of data per day, and this is set to explode to 40 yottabyes by 2020[3]—that is, 5,200 gigabytes for every person on the earth. As noted in previous "Blue Skies" columns, data generated by Internet of Things (IoT) devices and sensors are part of the big data landscape.[4,5] IoT comprises billions of Internet-connected devices (ICDs) or "things," each of which can sense, communicate, compute, and potentially actuate, and can have intelligence, multimodal interfaces, physical/virtual identities, and attributes. ICDs can be mobile devices, sensors, medical imaging devices, individual archives, social networks, smart cameras, body sensors, automobile cosimulations, or software logs. In a nutshell, a large volume of veracity data is generated at high velocity from a variety of sources.

The amalgamation of ICDs with big data processing software frameworks and cloud-based hardware resources leads to the creation of novel big data applications in domains such as healthcare, traffic management, smart energy grids, and smart manufacturing. Managing large, heterogeneous, and rapidly increasing volumes of data, and extracting value out of such data, has long been a challenge. In the past, this was partially mitigated by fast processing technologies that exploited Moore's law. However, with a fundamental shift toward big data applications, data volumes are growing faster than they can be analyzed, regardless of increased CPU speeds or other performance improvements. Although the impetus for the remainder of our article comes from healthcare big data, the problems and solutions discussed are applicable to other application domains.

2325-6095/15/$31.00 © 2015 IEEE

Big Data in Healthcare

A 2015 Gartner report noted that data processing technologies haven't kept pace with the significant increase in the volume of digital healthcare data, and an integrated and trustworthy healthcare analytics solution can facilitate more effective decision making in patient care and risk management, improving quality of life, optimizing performance of services, and so on.[6] Medical professionals have made similar observations. For example, the chief information officer of Boston's Beth Israel Deaconess Medical Center explained that, "working with big data in hospital systems is hugely challenging but at the same time holds tremendous promise in providing more meaningful information to help clinicians treat patients across the continuum of care."[7]

Consider, for example, the problem of managing petabytes of multimedia content produced by advanced medical devices in the healthcare or medical domain as exemplified by the following inventions and reports.

- In conjunction with traditional x-rays, medical imaging can now delve deeper into the human body, discovering and analyzing smaller and smaller details. A research team from Williams College at Harvard University has developed a new type of optical medical imaging device that captures high-resolution live video of human cells and molecules.[8]
- A report from AT&T reveals that medical content (x-rays, computed tomography, genetic data, and other pathology test reports) archives are increasing by 20–40 percent each year.[9] In 2012, there were 1 billion of above mentioned content types in United States alone, accounting for one-third of global storage demand.
- According to another study, "In 2012, worldwide digital healthcare data was estimated to be equal to 500 petabyes and is expected to reach 25,000 petabyes in 2020."[10] Further, it's anticipated that in 2015, an average hospital will need to manage 665 terabytes of patient data, 80 percent of which will be unstructured medical imaging data.

The challenge is how to ensure data confidentiality and integrity when storing such data but still make it highly available, process it to extract actionable information for decision makers, including medical professionals, and share it with collaborators, while preserving the privacy of individual patients and giving them the full control of their data at all times. This challenge calls for a trustworthy big data processing platform.

> According to recent reports, we now generate 2.5 quintillion bytes of data per day, and this is set to explode to 40 yottabyes by 2020.

Private Clouds: What Are the Research Opportunities?

Existing technology deployments within a medical organization, including its internal, on-premise infrastructure (private clouds) for data storage and the image archiving and communication systems used by radiologists, radically limit efforts to harness the massive amount of medical imaging and other healthcare data. In other words, organizations face several limitations when using private clouds to process healthcare application data.

The first limitation is scalability. On-premise private cloud deployments might not consider future growth, resulting in limited scalability. This isn't surprising, as building highly scalable private clouds requires a large capital investment for procuring and installing computing and storage resources. However, the changing volume, velocity, and variety of data make it difficult to accurately plan private cloud capacity, and private clouds are often either under- or overprovisioned. To reduce capital investment, private clouds are always built with limited scalability

Analytics is another possible limitation. Analytics models and software frameworks required to manage heterogeneous data might not be available in the private cloud because of higher operational costs. In general, as Editor-in-Chief Mazin Yousif notes, public clouds support the most commonly used analytics models and software frameworks because of their commercial interests, and private clouds deploy analytics models and software frameworks not available from public cloud providers or analytics models and software frameworks developed in-house.

A third limitation is data sharing. Data must be shared with collaborators who don't have access to private clouds or who reside outside the perimeter defenses. For example, a medical practitioner from a hospital in a different jurisdiction might not be able to access the data stored in the private cloud because at present, healthcare providers are generally subject to exacting regulatory requirements to ensure the security and privacy of patient and other sensitive data.

Although private clouds are inherently trustworthy, these limitations hamper the use of private clouds for processing healthcare big data. We also note the evolution of externally hosted private clouds, which are managed by third parties but support strict security and privacy guarantees. For example, the Postgres Plus Cloud Database (PPCD) provides an externally hosted private cloud. This service includes strict security and auditing features that are in compliance with the Health Insurance Portability and Accountability Act (HIPPA).[11] PPCD's architecture ensures that database instances and data are hosted in complete isolation from other instances and data. However, this isn't possible with purely public clouds.

> The National Institute of Standards and Technology (NIST) defines the four stages of the big data lifecycle as collection, preparation, analysis, and action.

Further, an externally hosted private cloud model incurs higher leasing costs and offers fewer opportunities to optimize costs than public clouds (for example, leasing spot instances from the Amazon Elastic Compute Cloud is relatively cheaper than leasing standard instances). Although public cloud infrastructures offer the opportunity to optimize hosting costs, they're more prone to security and privacy attacks because of the multitenancy of virtual machines (VMs) and data.

On the other hand, public clouds support the scalability and easy sharing of data. Alan Sill, editor of the "Standards Now" column, also rightly pointed out that US-based cloud service providers must ensure that they meet HIPAA requirements and offer levels of service that provide privacy and are in compliance with various internationally recognized standards.

The National Institute of Standards and Technology (NIST) defines the four stages of the big data lifecycle as collection, preparation, analysis, and action.[12] At different stages of the data processing, however, the data could be targeted by an attacker. For example, big data processing frameworks don't allow application orchestrator such as Apache Yarn (http://hadoop.apache.org/docs/current/hadoop-yarn/hadoop-yarn-site/YARN.html) to control which physical server rack the mapper and reducer VMs are deployed on at runtime. Hence, these instances can be mapped to VMs from other applications because of multitenancy, exposing the data to different types of security and privacy attacks. Public cloud service providers and existing big data processing frameworks have no easy way of detecting or monitor such data leakage. Therefore, data auditing,[13,14] data protection,[15] and privacy preservation[16] have emerged as salient areas of inquiry for researchers from industry and academia.

Another potential future research opportunity is bringing together the inherent features of public clouds (scalability) and private clouds (security) to build a trustworthy big data processing platform.

A Trustworthy Hybrid Cloud for Big Data Processing

There has been a paradigm shift toward hosting big data applications in hybrid infrastructures consisting of private and public clouds. However, building trustworthy end-to-end big data processing platforms that exploit hybrid cloud infrastructures can be challenging for several reasons.

First, existing big data ingestion frameworks (such as Apache Kafka and Amazon Kinesis), data storage frameworks (such as MongoDB, BigTable, MySQL, and Cassandra), parallel and distributed programming frameworks (such as Apache Hadoop and Apache Storm), scalable data mining frameworks (such as Apache Mahout and GraphLab), and distributed file systems (such as the Hadoop Distributed File System and Google File System) might not guarantee trustworthy (secure and privacy-preserving) data processing. Most of these frameworks can't support encryption of big data without compromising their inherent scalability and performance.

In addition, traditional data and distributed system security and privacy-preserving techniques can't be automatically adapted to operate efficiently in a hybrid cloud infrastructure deployed with multiple big data processing frameworks to process big data with 5Vs characteristics. There are two core reasons for this. First, as noted in a previous column, most of the big data processing frameworks can only process data within a single private or public cloud datacenter.[5] Second, porting existing security and privacy-preserving techniques to multiple big data processing framework is a hard undertaking because they support diverse data programming abstractions (for example, MapReduce in Hadoop, continuous query operators in Storm, and transactional operators in MySQL and Cassandra) and perform computation on diverse dataflows (such as batch, streaming, and transactional)

Security and privacy controls in public cloud computing infrastructures support basic security feature such as public key infrastructure- (PKI)-based access control and authorization to VMs and binary storage resources. They might have limited capability to protect data and applications against security attacks such as denial-of-service (DoS) and Sybil attacks.

Data holders (such as healthcare providers) typically want to ensure that their data is protected from malicious insiders who might steal or exfiltrate the data for sale. Such data can then be used to derive direct clinical value or profit from possible insights.[7] Data movement and advanced encryption techniques (such as homomorphic encryption) can make it challenging to provide data holders full control of their data in hybrid clouds without affecting performance. This gets even more complicated when patients want full control of their own data.

Existing cryptographic schemes are unlikely to be suited to a hybrid cloud deployment because of computational efficiency limitations and other constraints. Attribute-based encryption (ABE), for example, is designed to provide the scalability and flexibility of real-time data sharing in computing environments, including the cloud.[17,18] However, in existing ABE schemes, user revocation remains challenging, particularly when there's a large number of users. In addition, existing ABE schemes require the cloud server to be fully trusted, and in the aftermath of Edward Snowden's revelations that the National Security Agency has been conducting wide-scale government surveillance, the requirement that all cloud servers in the deployment be trusted might be onerous. Therefore, it isn't surprising that cloud cryptography offering enhanced security without compromising usability and performance is an ongoing research topic.

In the event of a security breach, there might not be an easy way to conduct digital investigations, particularly across borders and between organizations, which would allow the victim to mitigate future risks and/or pursue the criminals through a criminal investigation or civil litigation. For example, would it even be possible to remotely collect evidence from a hybrid cloud in the event of a digital investigation or incident response[19]? In addition, as noted elsewhere, "it's currently unclear whether existing legislation, say in Australia, permits the use of such remote real-time evidence preservation and collection processes and tools to preserve evidential material stored or held overseas without a mutual assistance request."[20]

In the virtual laboratory approach, data is kept inside the private cloud.[21] The virtual laboratory hosts the data and supports a number of data processing algorithms. The output datasets are checked against all privacy rules before they're released. This approach isn't scalable because it's built for private cloud infrastructures. Furthermore, it doesn't support privacy-preserving computation over data from multiple, heterogeneous, and dynamic sources because the virtual laboratory is a trusted entity and resides within a defense perimeter.

> There's a need to balance the data's privacy and security against data sharing or performing scalable, efficient, near-real-time data analytics.

Therefore, there's a need to balance the data's privacy and security against data sharing or performing scalable, efficient, near-real-time data analytics. To this end, a data outsourcing approach has emerged.

The Data Outsourcing Approach and Encryption Techniques

Traditional access-control mechanisms have been successfully used for controlled data sharing in collaborative environments; however, the applicability of such mechanisms is limited in a hybrid cloud environment where some data can reside outside the defense perimeter (that is, organizational boundaries). An alternative is to use the PKI infrastructure supported by public clouds. In addition to its data security limitations, this approach isn't scalable. Recently, several encryption techniques have been developed to address existing security concerns.

Proxy reencryption (PRE) enables data encrypted using one user's public key to be transformed in such a way that it can be decrypted with another user's private key.[22] The basic idea is that two parties publish a proxy key that allows an untrusted intermediary to convert ciphertexts encrypted for the first party directly into ciphertexts that can be decrypted by the second.

Identity-based encryption (IBE) allows any pair of users to communicate securely and to verify each other's signatures without exchanging private or public keys, keeping key directories, or using the services of a third party.[23] This scheme is ideal for sharing information among closed groups of people (for example, within an organization). The idea is based on the public key cryptosystem, but the public keys are generated using attributes (such as company name or IP address) and individual users have corresponding private keys.

Attribute-based encryption (ABE) aims to overcome one of the limitations of earlier IBE schemes—that is, their use of string-based attributes.[24] ABE is one of two applications of Fuzzy IBE, introduced by Amit Sahai and Brent Waters, which allows attributes to take value from a domain other than strings (the other application is IBE that uses biometric identities).[24] In an ABE system, a user's keys and ciphertexts are labelled with sets of descriptive attributes, and a particular key can decrypt a particular ciphertext only if there's a match between the attributes of the ciphertext and the user's key. Sahai and Waters' cryptosystem allows for decryption when a ciphertext and a private key share at least k attributes. Although this primitive was shown to be useful for error-tolerant encryption with biometrics, the lack of expressibility limits its applicability to larger systems.

Existing solutions based on ABE and PRE introduce a heavy computation overhead on the data owner so don't scale well when fine-grained data access control is desired. To address this problem, a combination of ABE and PRE schemes have been proposed in the cloud security and cryptography literature to exploit the benefits of both schemes.

Moreover, existing data sharing techniques do not support the data analytics. A different branch of research has recently emerged in which the computation can be performed on encrypted data in the cloud. Craig Gentry introduced the first fully homomorphic encryption scheme in 2009.[25] This was a revolutionary cryptographic achievement, but the scheme was far too inefficient for any practical use, especially because of its computational complexity (running time). Since 2009, several works have improved upon Gentry's technique, leading to significant reductions in running time. Although many researchers have improved the processing time, homomorphic encryption has other limitations. For instance, it requires that all recipients have access to the same key to encrypt the inputs and decrypt the results, which might be difficult to arrange if they belong to different organizations. This also doesn't support computation over data from multiple sources. Furthermore, current fully homomorphic encryption solutions are limited to a small number of operations or their performance isn't suitable for real-time and complicated analysis. In addition to numerical operations, all data mining operations must be performed over encrypted data. An encrypted data versioning system is also needed. These challenges offer great opportunities for future research.

> Data sharing approaches should be combined with data analytics approaches to support end-to-end trustworthy data sharing.

Data sharing approaches should therefore be combined with data analytics approaches to support end-to-end trustworthy data sharing and processing platforms in public clouds. This question leads to further research on secure multiparty computation. MPC takes private input data from multiple parties and carries out a joint computation on them while ensuring that the input data remains private to their owners during the computation process.

The focus so far has been on data privacy in a private or virtual cloud. Some applications require a hybrid cloud approach, in which privacy-sensitive data is kept in the private cloud and de-identified data is kept in the public cloud. This approach works well in the health domain, where the de-identified data can be shared with collaborators and processed in the collaborators/public cloud environment. However, segregating private and public data, moving public data, and integrating results after processing are some of the challenging issues requiring further research.

To ensure the privacy of personally identifiable information (PII) and other sensitive healthcare data in a (hybrid) cloud environment (despite the varying legal requirements in different jurisdictions), it's necessary to ensure the security of the underlying cloud architecture or ecosystem—for example, through the use of cryptography and privacy-enhancing or preserving technologies. Therefore, we need efficient cloud cryptography as well as privacy-enhancing or preserving systems that can be deployed in healthcare settings. We must also ensure that the underlying cloud architecture or ecosystem is designed to facilitate the identification, preservation, and collection of evidential data in the investigation of a data breach incident.

Developing techniques and APIs that can guarantee data security and privacy and computation across a hybrid cloud ecosystem consisting of multiple private and public cloud datacenters remains an open and difficult research problem. Future efforts also need to focus on designing and developing computationally efficient privacy-preserving techniques that seamlessly scale across multiple big data processing frameworks by exploiting the elasticity of hybrid (multiple private and public) cloud infrastructures while adapting to uncertain data volume, data velocity, and data variety. This could be achieved by exploiting the inherent software-level configuration of big data processing frameworks for scaling existing security and privacy-preserving techniques.

In summary, efforts need to focus on the development of security and privacy techniques that can deal with changing volume, velocity, and variety of heterogeneous dataflow (batch, streaming, transactional); be ported to diverse big data programming frameworks (Apache Hadoop, Apache Storm, Apache Hive); deal with variable computational complexity due to heterogeneous VM, storage, and network configurations across multiple clouds; and be seamlessly implemented in multicloud orchestration APIs such as jclouds.

Developing techniques and APIs that can guarantee data security and privacy and computation across a hybrid cloud ecosystem.

References

1. R. Pepper and J. Garrity, "The Internet of Everything: How the Network Unleashes the Benefits of Big Data," *Global Information Technology Report 2014*, Cisco Systems, 2014; http://blogs.cisco.com/wp-content/uploads/GITR-2014-Cisco-Chapter.pdf.
2. "Bringing Big Data to the Enterprise," IBM, www-01.ibm.com/software/in/data/bigdata.
3. J. Gantz et al., "The Digital Universe in 2020: Big Data, Bigger Digital Shadows, and Biggest Growth in the Far East," IDC iView, 2012; www.emc.com/leadership/digital-universe/2012iview/index.htm.
4. R. Ranjan, "Streaming Big Data Processing in Datacenter Clouds," *IEEE Cloud Computing*, vol. 1, no. 1, 2014, pp. 78–83.
5. L. Wang and R. Ranjan, "Processing Distributed Internet of Things Data in Clouds," *IEEE Cloud Computing*, vol. 2, no. 1, 2015, pp. 76–80.
6. V. Shaffer, *Agenda Overview for Healthcare*, Gartner report G00270705, 2015; www.gartner.com/doc/2995217/agenda-overview-healthcare-.
7. O. Badawi et al., "Making Big Data Useful for Health Care: A Summary of the Inaugural MIT Critical Data Conference," *JMIR Medical Informatics*, vol. 2, no. 2, 2014, e(22); doi:10.2196/medinform.3447.
8. D. Borghino, "New Medical Imaging Technique Delivers Streaming Video at Molecular Level," *Gizmag*, 7 Dec. 2010; www.gizmag.com/medical-imaging-tracking-molecules-live-tissue-video-rate/17202.
9. *Medical Imaging in the Cloud*, AT&T tech. report, 2012; www.corp.att.com/healthcare/docs/medical_imaging_cloud.pdf.
10. J. Sun and C. Reddy, "Big Data Analytics for Healthcare," *Proc. 19th ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining*, 2013;

doi:10.1145/2487575.2506178; http://dmkd.cs.wayne.edu/TUTORIAL/Healthcare.

11. F. Dalyrimple, "Postgres Meets HIPAA in the Cloud," blog, 31 Mar. 2015; www.enterprisedb.com/postgres-plus-edb-blog/fred-dalrymple/postgres-meets-hipaa-cloud.
12. Nat'l Inst. of Standards and Technology, *DRAFT NIST Big Data Interoperability Framework: Volume 1, Definitions*, NIST, 2015; http://bigdatawg.nist.gov/_uploadfiles/BD_Vol1-Definitions_V1Draft_Pre-release.pdf.
13. C. Liu et al., "MuR-DPA: Top-Down Levelled Multi-replica Merkle Hash Tree Based Secure Public Auditing for Dynamic Big Data Storage on Cloud," *IEEE Trans. Computers*, preprint, doi: 10.1109/TC.2014.2375190.
14. C. Liu et al., "Authorized Public Auditing of Dynamic Big Data Storage on Cloud with Efficient Verifiable Fine-Grained Updates," *IEEE Trans. Parallel and Distributed Systems*, preprint, doi: 10.1109/TPDS.2013.191.
15. J. Yao et al., "TrustStore: Making Amazon S3 Trustworthy with Services Composition," *Proc. 10th IEEE/ACM Int'l Conf. Cluster, Cloud and Grid Computing* (CCGRID 10), 2010, pp. 600–605.
16. X. Zhang et al., "A Privacy Leakage Upper Bound Constraint-Based Approach for Cost-Effective Privacy Preserving Intermediate Data Sets in Clouds," *IEEE Trans. Parallel and Distributed Systems*, vol. 24, no. 6, 2013, pp. 1192–1202.
17. V. Goyal et al., "Attribute-Based Encryption for Fine-Grained Access Control of Encrypted Data," *Proc. 13th ACM Conf. Computer and Comm. Security* (CCS 06), 2006, pp. 89–98.
18. J. Bethencourt, A. Sahai, and B. Waters, "Ciphertext-Policy Attribute-Based Encryption," *Proc. IEEE Symp. Security and Privacy* (SP 07), 2007, pp. 321–344.
19. D. Quick, B. Martini, and K.-K.R Choo, *Cloud Storage Forensics*, Syngress/Elsevier, 2014.
20. B. Martini and K.-K.R Choo, "Cloud Forensic Technical Challenges and Solutions: A Snapshot," *IEEE Cloud Computing*, vol. 1, no. 4, 2014, pp. 20–25.
21. C.M. O'Keefe et al., "Protecting Confidentiality in Statistical Analysis Outputs from a Virtual Data Centre," *Proc. Joint UNECE/Eurostat Work Session on Statistical Data Confidentiality*, 2013; http://www.unece.org/stats/documents/2013.10.confidentiality.html.
22. M. Blaze, G. Bleumer, and M. Strauss, "Divertible Protocols and Atomic Proxy Cryptography," *Advances in Cryptology—EUROCRYPT 98*, LNCS 1403, Springer, 1998, pp. 127–144.
23. R.L. Rivest, A. Shamir, and L. Adlemann "A Method for Obtaining Digital Signatures and Public-Key Cryptosystems," *Comm. ACM*, vol. 21, no. 2, 1978, pp. 120–126.
24. A. Sahai and B. Waters, "Fuzzy Identity-Based Encryption," *Proc. 24th Ann. Int'l Conf. Theory and Applications of Cryptographic Techniques* (EUROCRYPT 05), 2005, pp. 457–473.
25. C. Gentry, *A Fully Homomorphic Encryption Scheme*, PhD Thesis, Stanford Univ., 2009; https://crypto.stanford.edu/craig/craig-thesis.pdf.

SURYA NEPAL *is a principal research scientist in the Digital Productivity Flagship at the Commonwealth Scientific and Industrial Research Organization (CSIRO), Australia. His research interests include cloud computing, big data, and cybersecurity. Nepal has a PhD in computer science from Royal Melbourne Institute of Technology, Australia. Contact him at surya.nepal@csiro.au.*

RAJIV RANJAN *is in the Digital Productivity Flagship at the Commonwealth Scientific and Industrial Research Organization (CSIRO), Australia, where he's also a senior research scientist, Julius Fellow, and project leader. At CSIRO, he leads research projects related to cloud computing, content delivery networks, and big data analytics for Internet of Things (IoT) and multimedia applications. Ranjan has a PhD in computer science and software engineering from the University of Melbourne. He has published more than 120 scientific papers. Contact him at rajiv.ranjan@csiro.au or http://rajivranjan.net.*

KIM-KWANG RAYMOND CHOO *is a senior lecturer in the School of Information Technology and Mathematical Science at the University of South Australia. His research interests include cyber and information security and digital forensics. Choo has a PhD in information security from Queensland University of Technology, Australia. Contact him at raymond.choo@fulbrightmail.org or https://sites.google.com/site/raymondchooau.*