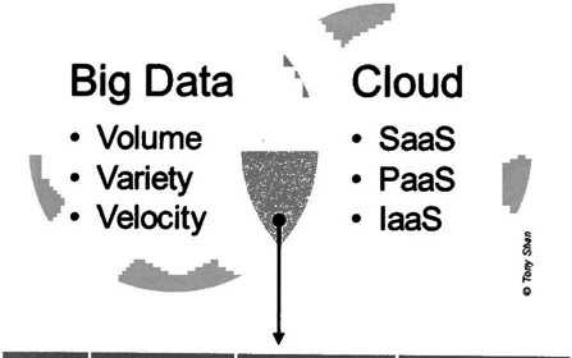
Bảo mật cho dữ liệu lớn trên điện toán đám mây

DƯƠNG THỊ THANH TÚ,
ĐỖ MINH HIỆP,
ĐỖ THỊ THU THỦY

1. Giới thiệu chung

Việc số hóa thông tin ngày này qua ngày khác trong điện thoại thông minh, máy tính, đồng hồ kết nối Internet, mạng xã hội, các đối tượng kết nối và công nghệ trực tuyến... tạo ra một khối lượng lớn nguồn thông tin kỹ thuật số, tăng lên theo cấp số nhân mỗi ngày. Nguồn thông tin này là gọi là dữ liệu lớn (Big Data), được nhiều tổ chức sử dụng trong các lĩnh vực khác nhau để tự động trích xuất thông tin trong thời gian thích hợp [1].

Dữ liệu lớn đang trở thành một phần không thể thiếu trong các công ty hàng đầu để đạt được mục tiêu của mình, bằng cách thích ứng với các danh mục đầu tư sao cho phù hợp với nhu cầu của khách hàng. Tuy vậy, việc xử lý và phân tích số lượng lớn và không đồng nhất dữ liệu như vậy, không thể thực hiện được bằng cách sử dụng cơ sở dữ liệu có cấu trúc và phương pháp thông thường.



	Volume	Variety	Velocity
SaaS	Semantics	Visualization	Real-time
PaaS	Distributed processing	Schemaless	Integration on the fly
IaaS	Scalable storage	Federated store	On-demand resources

Hình 1: Mối quan hệ giữa Big data và điện toán đám mây

Điện toán đám mây là một giải pháp toàn diện cung cấp công nghệ thông tin như một dịch vụ; một giải pháp điện toán dựa trên Internet trong đó các máy tính trong đám mây được cấu hình để làm việc cùng nhau. Điện toán đám mây cho phép người dùng lưu trữ và phân tích dữ liệu của họ bằng cách sử dụng tài nguyên máy tính được chia sẻ, đồng thời dễ dàng xử lý sự biến đổi lượng và tốc độ của dữ liệu.

Chính vì thế, điện toán đám mây nhanh chóng trở thành một công cụ cho việc xử lý và phân tích dữ liệu lớn với ưu điểm giảm giá thành, dễ dàng mở rộng việc kết nối cho hệ thống, xác định dịch vụ,... [2]. Tuy nhiên, điện toán đám mây cũng tạo thêm nhiều rủi ro bởi cơ sơ hạ tầng máy tính được chia sẻ - điều chưa từng tồn tại trong kiến trúc tính toán truyền thống. Thêm vào đó, những nhà cung cấp và người sử dụng đám mây có thể là thực thể không đáng tin cậy – những người cố tình làm xáo trộn việc lưu trữ hay tính toán dữ liệu. Vì vậy, bảo mật cho dữ liệu lớn trong môi trường điện toán đám mây gần đây đã thu hút được rất nhiều sự quan tâm nghiên cứu.

2. Những vấn đề bảo mật cho dữ liệu lớn sử dụng điện toán đám mây

2.1. Thách thức an ninh và đảm bảo tính riêng tư cho dữ liệu trong dữ liệu lớn

Dữ liệu lớn là một cơ hội to lớn cho nhiều các ngành công nghiệp và các nhà sản xuất, nhưng kèm theo đó là thách thức trong việc đảm bảo sự riêng tư và các vấn đề an ninh. Thách thức này phát sinh từ thực tế, việc sử dụng các công cụ phân tích bao gồm lưu trữ, quản lý và phân tích hiệu quả dữ liệu đa dạng được tập hợp từ tất cả các nguồn có thể hoặc có sẵn. Hậu quả là dữ liệu người sử dụng trở thành các mục tiêu dễ bị ngắm tới bởi vì tính kết hợp và khai thác dữ liệu hành vi cụ thể. Nghĩa là kẻ tấn công có thể thu thập nhiều dữ liệu hơn so với quyền hạn của mình dẫn đến vi phạm một loạt vấn đề an ninh và riêng tư.

Có một tập hợp các mối quan tâm riêng tư và bảo mật cần phải xem xét trước khi xây dựng một môi trường dữ liệu lớn. Dưới đây là một số thách thức quan trọng nhất nên được xem xét cẩn thận khi xử lý dữ liệu lớn:

Phân bố ngẫu nhiên: Khái niệm về Big Data Analytics chủ yếu dựa trên các phương pháp song song, điều này khiến các dữ liệu lớn phải được lưu trữ và xử lý tại cụm khác nhau, đó là một tập hợp các máy chủ phân phối vòng quanh thế giới và hoạt động như một trạm. Vấn đề chính với cấu trúc này là rất khó để biết chính xác vị trí lưu trữ và xử lý dẫn đến khó có thể đảm bảo an ninh trước các hành vi vi phạm quy định.

Tính riêng tư: Thách thức chính với Big Data Analytics là khó có thể phân phối lưu trữ và xử lý theo quy định đối với các dữ liệu nhạy cảm. Các công nghệ phân tích dữ liệu lớn hiện hành đối xử với tất cả các dữ liệu với cùng độ ưu tiên giống như mã hóa hoặc xử lý giống nhau với tất cả các loại dữ

liệu [3]. Như vậy, một hacker hay một nút độc hại có thể cưỡng chế truy cập vào các cụm để dễ dàng ăn cắp, khai thác trái phép hoặc thay đổi các nguồn thông tin.

Xử lý: Ý tưởng chính đằng sau dữ liệu lớn là để trích xuất những thông tin hữu ích cho các hoạt động xử lý cụ thể. Tuy nhiên, điều quan trọng là đảm bảo an toàn và bảo vệ những xử lý để tránh bất kỳ rủi ro hay các hành vi cố gắng thay đổi hoặc do thám các kết quả trích xuất.

Tính toàn vẹn: Trong một bối cảnh mở dữ liệu lớn, một khối lượng lớn nội dung không phải luôn đưa ra một chỉ số tốt cho chất lượng kết quả trích xuất. Do đó, trước khi tìm kiếm thông tin và ra quyết định dựa trên dữ liệu lớn, điều quan trọng là phải đảm bảo tính hợp lệ và mức độ tin cậy của dữ liệu, để tránh tin tưởng nhầm vào một bản ghi dữ liệu đáng nghi hoặc đã bị cưỡng chiếm.

Truyền thông: Dữ liệu lớn được lưu trữ trong một số các nút thuộc nhiều cụm được phân phối trên toàn thế giới. Tất cả các thông tin liên lạc giữa các cụm và các nút được đảm bảo thông qua các mạng công cộng và các mạng riêng. Tuy nhiên, nếu ai đó có thể thay đổi truyền thông liên nút sẽ dễ dàng trích xuất thông tin có giá trị. Vì vậy, một thách thức khác cho các công cụ dữ liệu lớn thông qua các giao thức mạng là đảm bảo an toàn để bảo vệ tương tác giữa các bên khác nhau.

Quản lý truy cập: Trong một bối cảnh dữ liệu lớn, truy cập vào các dữ liệu cần được quản lý bởi một hệ thống kiểm soát truy cập mạnh mẽ để bất kỳ các thành phần không được phép truy cập đến các máy chủ lưu trữ sẽ bị từ chối. Từ đó chỉ các nút với quyền quản trị đầy đủ mới có khả năng quản lý và xử lý bất kỳ thông tin nào. Hơn nữa, bất kỳ thay đổi trong trạng thái cụm như bổ sung hoặc xóa các nút phải được giám sát bởi một cơ chế xác thực để bảo vệ hệ thống khỏi các nút độc hại.

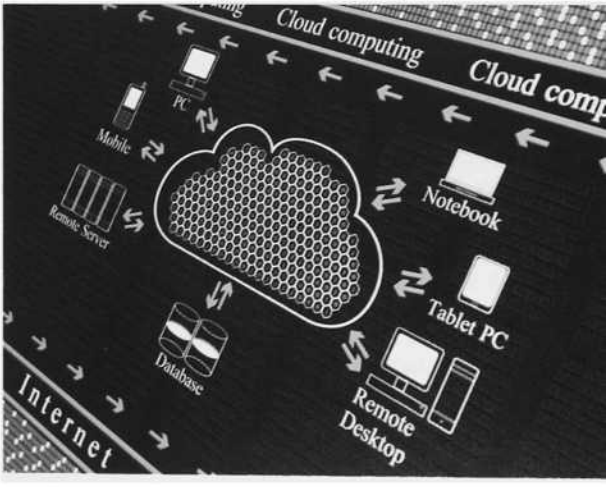
2.2. Bảo mật dữ liệu trong điện toán đám mây

Điện toán đám mây có thể tạo ra rủi ro với các thông tin nhạy cảm, những rủi ro này xuất phát từ nhu cầu giao phó việc bảo vệ dữ liệu cho nhà cung cấp điện toán đám mây. Điểm khác biệt lớn so với các môi trường truyền thống là trong môi trường điện toán đám mây, việc xử lý dữ liệu có thể bị điều khiển hoặc quản lý bởi các bên không tin cậy khác nhau và có thể bị tổn thương do sự tấn công từ người thuê điện toán đám mây khác. Hay nói cách khác, đối tượng độc hại có thể xuất hiện từ bên trong lẫn bên ngoài.

Khi người sở hữu dữ liệu buông lỏng sự kiểm soát dữ liệu của mình trong môi trường đám mây, họ yêu cầu rằng dữ liệu của họ vẫn được bảo vệ tốt. Ngày nay, sự đảm bảo này là những lời hứa pháp lý mà những nhà cung cấp điện toán đám mây đưa ra cho người sử dụng – thỏa thuận cấp độ dịch vụ (Service Level Agreement - SLA). Mật mã hóa là một giải pháp cho phép người chủ dữ liệu bảo vệ tài nguyên của họ một cách chủ động thay vì phản hồi đơn độc trên thỏa thuận pháp lý. Ba mục đích bảo mật truyền thống dưới đây luôn được xem xét với mỗi đề xuất bảo mật sử dụng mã hóa trong điện toán đám mây [4]:

- ***Độ tin tưởng:*** Tất cả các dữ liệu nhạy cảm (đầu vào máy tính, đầu ra hay trạng thái trung gian) được giữ bí mật khỏi bất kì mối nguy hại hay đối tượng không đáng tin cậy.

Tính toàn vẹn: Có khả năng phát hiện bất kì sự sửa đổi trái phép hoặc sửa đổi dữ liệu nhạy cảm nào. ⊕



Giá trị: Người sở hữu dữ liệu (hay người tiếp nhận kết quả) được đảm bảo tiếp cận với dữ liệu của họ và tài nguyên tính toán.

3. Kỹ thuật mã hóa dữ liệu đảm bảo an toàn cho dữ liệu lớn trong điện toán đám mây

Mã hóa luôn là một kỹ thuật tốt để bảo vệ dữ liệu nhạy cảm. Trong trường hợp dữ liệu lớn, mã hóa có thể được sử dụng trong tất cả các khâu: lưu trữ, tính toán và truyền thông.

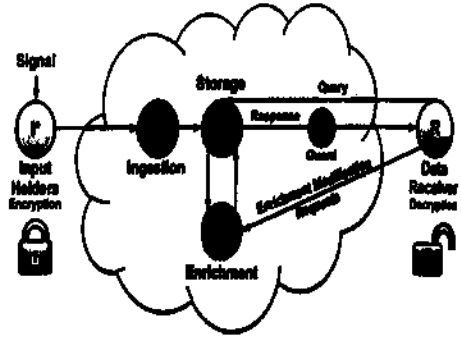
Đối với môi trường điện toán đám mây, có nhiều kĩ thuật mã hóa khác đã từng được sử dụng trong bảo mật như mã hóa dựa trên nhận dạng và mã hóa dựa trên thuộc tính. Tuy nhiên trong nội dung giới hạn của bài báo chỉ giới thiệu 3 kỹ thuật mã hóa có khả năng ứng dụng để đạt được bảo mật dữ liệu lớn trong điện toán đám mây. Đó là: mã hóa đồng nhất (homomorphic encryption - HE), tính toán xác thực (Verifiable computation - VC) và tính toán bảo mật đa chiều (secure multi-party computation - MPC).

3.1 Mã hóa Homomorphic

Mã hóa Homomorphic là một kiểu mã hóa cho phép chức năng xử lý dữ liệu được mã hóa mà không cần giải mã nó. Một cách chính thức, Ek(m) là một thông điệp mã hóa m với chìa khóa là k. Một chương trình mã hóa gọi là homomorphic đối với một hàm f nếu tồn tại một hàm f' tương ứng sao cho $D_k(f'(E_k(m)) = f(m)$, trong đó D_k là thuật toán giải mã dưới khóa k.

Hình 2 minh họa phương pháp mã hóa này trong đó I để cập tới nút đầu vào, C để cập đến nút tính toán, S kí hiệu cho nút lưu trữ, R kí hiệu cho nút kết quả, và X+ là 1 hay nhiều nút giao tiếp cho phép $X \in \{I,C,S,R\}$. Những điểm nút trên điện toán đám mây không được tin cậy để bảo vệ tính bí mật, người gửi dữ liệu sẽ mã hóa dữ liệu trước khi đưa vào điện toán đám mây, người nhận dữ liệu sẽ thực hiện giải mã sau khi dữ liệu rời khỏi đám mây. Những chiếc khóa kí hiệu cho mã hóa và giải mã.

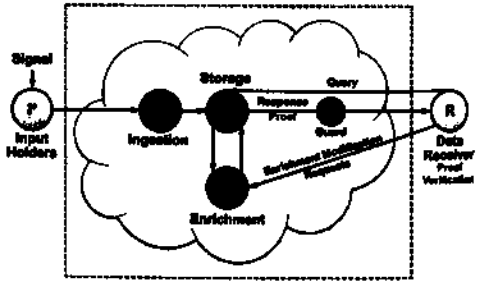
Chú ý rằng mã hóa Homomorphic chỉ đảm bảo độ tin tưởng của dữ liệu mà không đảm bảo tính toàn vẹn của dữ liệu. Tuy nhiên, nó có thể được kết hợp với các tính toán xác thực (được giới thiệu trong phần 3.2). Sự kết hợp mã hóa và tính toán xác thực này cho phép tính toán an toàn, thậm chí trên điện toán đám mây hoàn toàn không được tin tưởng.



Hình 2: Sơ đồ cấu trúc mã hóa Homomorphic

3.2 Xác thực

Tính toán xác thực (VC) cho phép người sở hữu dữ liệu kiểm tra tính toàn vẹn của việc tính toán. Trong một chương trình tính toán xác thực, VC cho phép chủ sở hữu gửi dữ liệu của mình cùng với một bản liệt kê những kỹ thuật tính toán mong muốn mà chúng ta gọi là Prover. Các Prover đảm bảo kết quả đầu ra là kết quả của việc tính toán quy định, cùng với một số "lập luận thuyết phục" hoặc "minh chứng" rằng dữ liệu này là chính xác với yêu cầu đã đưa ra trước đó.

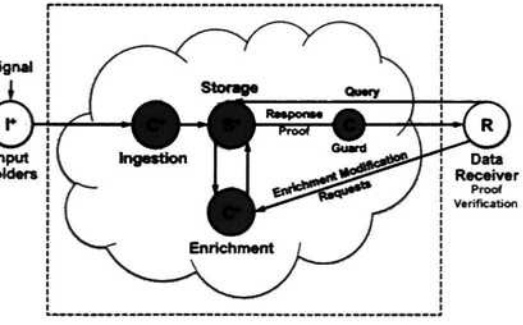


Hình 3: Sơ đồ tính toán xác thực

Hình 3 minh họa phương pháp tính toán xác thực. Trong đó, các nút trên điện toán đám mây không được tin cậy để bảo vệ tính toàn vẹn của dữ liệu. Nút tính toán cung cấp các bằng chứng về tính chính xác và người tiếp nhận dữ liệu sẽ kiểm chứng các bằng chứng ấy. Các nét đứt biểu thị cách ly vật lý với hệ thống xung quanh.

3.3 Bảo mật tính toán đa chiều (MPC)

Việc bảo mật cho tính toán đa chiều (nhiều bên) thích hợp để tận dụng việc thiết lập các đám mây bán tin tưởng. MPC thúc đẩy sự hiện diện của các bên một cách trung thực mà không nhất thiết phải biết bên nào trung thực, để đạt được độ tin tưởng và tính toàn vẹn của dữ liệu và việc tính toán. Trong MPC, không tồn tại duy nhất một đối tượng nào học được bất cứ điều gì về dữ liệu. Tuy nhiên, nếu nhiều bên cùng bị tấn công bởi 1 đối thủ và góp chung thông tin, chúng có thể phá vỡ độ tin tưởng của hệ thống MPC. Hình 4 minh họa hoạt động của MPC trên đám mây. Trong đó, điện toán đám mây là bán tin cậy. Người giữ đầu vào chia sẻ bí mật dữ liệu giữa các nút tính toán, nơi thực hiện tính toán nhiều bên. Người nhận kết quả tái cấu trúc dữ liệu đầu ra.



Hình 4: Bảo mật việc tính toán đa chiều

4. Kết luận

Sự phát triển mạnh mẽ về nhu cầu của các tổ chức, cá nhân trong thời đại công nghệ thông tin đã tạo ra một khái niệm mới, đó là dữ liệu lớn. Dữ liệu lớn đem lại sức mạnh về lưu trữ, sự tối ưu và sự cải thiện trong thời đại công nghệ thông tin. Tuy nhiên với trách nhiệm đảm đương khối dữ liệu khổng lồ ấy thì dữ liệu lớn cũng tự biến mình trở thành mục tiêu tấn công nhằm đánh cắp, thay đổi thông tin của một hệ thống hay tổ chức cá nhân, đặc biệt khi triển khai trên mô hình điện toán đám mây nhằm mang lại sự tiện lợi cũng như giảm chi phí đầu tư. Để giải quyết vấn đề này, sự kết hợp các kỹ thuật mã hóa bao gồm mã hóa đồng nhất (homomorphic encryption - HE), tính toán xác thực (verifiable computation -VC) và tính toán bảo mật đa chiều (secure multi-party computation - MPC) được biết đến như là những công cụ tốt nhất để bảo vệ an toàn thông tin trong kỉ nguyên của dữ liệu lớn với môi trường điện toán đám mây.

Tài liệu tham khảo:

1. YOUSSEF GAHI, "Big Data Analytics: Security and Privacy Challenges ", IEEE Symposium on Computers and Communication (ISCC), 2016.

2. "Big Data in the Cloud: Converging Technologies" , Intel IT Center, April 2015.

3. NATALIA MILOSLAVSKAYA, "Survey of Big Data Information Security ", 4th International Conference on Future Internet of Things and Cloud Workshop, 2016.

4. SOPHIA YAKOUBOV, "A Survey of Cryptographic Approaches to Securing Big-Data Analytics in the Cloud", IEEE 2014.