Featured Articles

Privacy-preserving Analysis Technique for Secure, Cloud-based Big Data Analytics

Ken Naganuma
Masayuki Yoshino, Ph.D.
Hisayoshi Sato, Ph.D.
Yoshinori Sato

OVERVIEW: Big data analytics, the process of collecting and analyzing large amounts of data to obtain new knowledge, is being applied in a wide range of fields, including the analysis of purchase histories, medical data, and sensor data. While this has been accompanied by a growth in services that offer to perform these analyses in the cloud, it has also been recognized that analyzing data on a third-party cloud server runs the risk of information leaks due to unauthorized access or criminal activity within the service provider. To overcome this problem, Hitachi has proposed a privacy-preserving analysis technique that uses searchable encryption, which can perform text matching of encrypted text, to perform tasks such as statistical analysis and analysis of correlation rules without decrypting the data. This technique reduces the risk of information leaks because it allows data analysis to be outsourced without divulging the content of the data to the service provider conducting the analysis.

INTRODUCTION

BIG data analytics, the process of collecting and analyzing large amounts of data to obtain new knowledge, is being used to analyze data such as purchase histories, medical data, and sensor data. One common technique for analyzing purchase histories is the analysis of correlation rules (also known as association rule learning). This involves identifying correlations between a particular product and other products that can be utilized in marketing, such as noting that customers who purchase diapers also often purchase beer. Big data analytics identifies knowledge like this that is hidden in large quantities of data. As a current trend in information technology (IT), it is being used in market analysis and a wide variety of other fields.

Complementing this, it is anticipated that software-as-a-service (SaaS) services that analyze customers' data on cloud servers will become widely used. A recognized problem when having analysis performed on a third-party cloud server, however, is the risk of information leaks due to unauthorized data access or criminal activity within the service provider, and therefore the challenge is to develop secure ways of performing this data analysis. Accordingly, Hitachi is working on the research and development of a privacy-preserving analysis technique that can analyze encrypted data without having to decrypt it.

This article describes a technique that can perform two of the most basic forms of analysis, namely statistical analysis and the analysis of correlation rules, without decrypting the data being analyzed. The technique reduces the risk of information leaks by allowing cloud users to outsource analyses such as these by supplying the third-party cloud service provider with data in encrypted form to avoid divulging its content. Along with ensuring the security of the encrypted data, Hitachi also focused on processing efficiency when developing this technique to ensure that it would be capable of analyzing large quantities of data. Experimental testing confirmed that the privacy-preserving analysis technique is practical for use on medium-sized quantities of data, with correlation rule analysis of 100,000 records of encrypted data completing in approximately 600 seconds (10 minutes).

PRIVACY-PRESERVING ANALYSIS TECHNIQUE

Fig. 1 shows the system configuration for privacy-preserving analysis provided by a third-party cloud service. The privacy-preserving analysis system

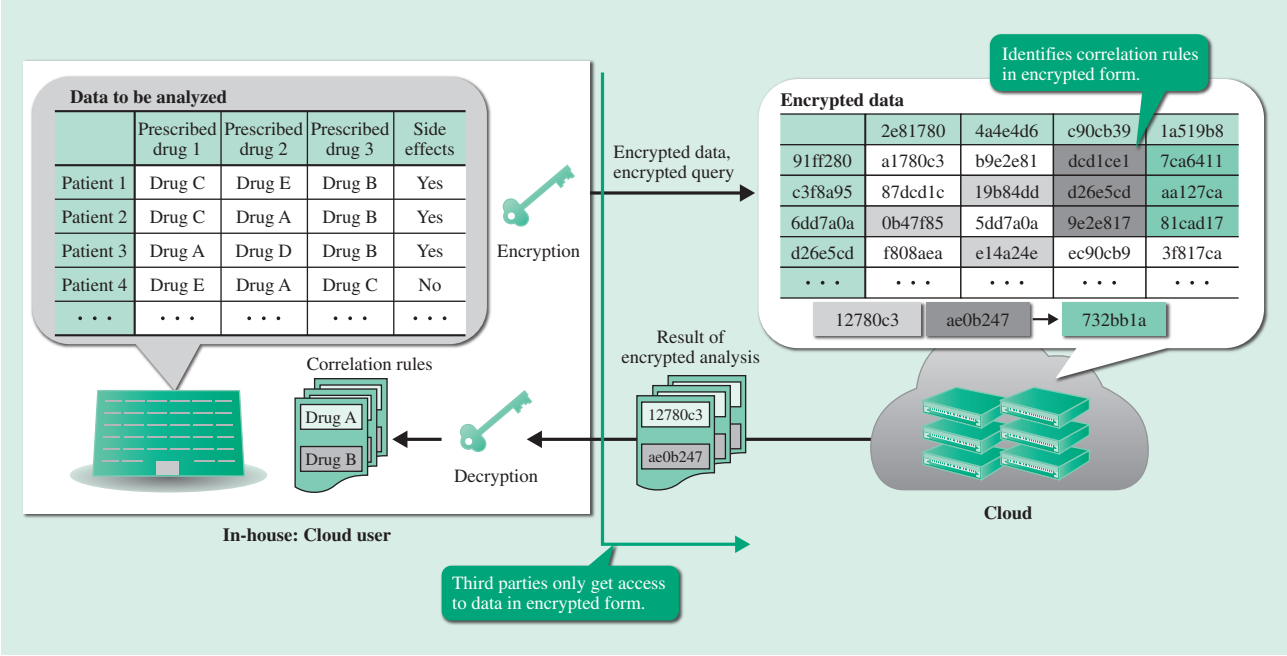

Fig. 1—System Configuration for Privacy-preserving Analysis.
Data is stored in the cloud in encrypted form. To perform an analysis, an encrypted analysis query is passed to the cloud. Storing the cloud data in encrypted form reduces the risk in the event of an information leak.

works by the cloud user with the data to be analyzed (indicated by "in-house" in the figure) encrypting the data using their own key and then supplying it to the third-party cloud service. The user then issues an encrypted analysis query (instruction) to the cloud service. The service uses the encrypted query to perform an analysis on the encrypted data, obtaining a result, also in encrypted form, that is returned to the user. The entire analysis is conducted in encrypted form, with neither the data nor the query being decrypted by the cloud service at any point. Finally, the cloud user decrypts the result to obtain the desired information.

The requirement in the past for data to be decrypted for processing when performing an analysis raised the risk of information leaks. With the privacy-preserving analysis technique, on the other hand, because the data in the cloud remains encrypted, the risk when an information leak occurs is reduced.

The following sections describe techniques for performing statistical analysis and correlation rule analysis of encrypted data using searchable encryption, which is able to perform text matching on encrypted text (testing for an exact match between the plain text versions of the original text and query text). Statistical analysis and correlation rule analysis have been chosen as example analysis applications because they are the most basic forms of data mining.

Searchable Encryption

"Searchable encryption" is a generic term for encryption techniques that allow not only conventional encryption and decryption, but that can also perform text matching using an encrypted query on encrypted text. While encryption and decryption keys respectively are required for encryption and decryption, text matching does not require any special information and therefore can be performed by a cloud service that does not have the keys. However, some techniques also have a separate private key for text matching so that it can only be performed by authorized users.

A number of searchable encryption techniques exist[(1), (2), (3), (4), (5)], and can broadly be divided into those that use a common key system and those that use a public key system. Common key systems use the same key for both encryption and decryption. They are best suited to large quantities of data because they tend to execute more efficiently than public key systems. While public key systems have separate encryption and decryption keys, meaning the encryption key can be made publically available without compromising security, they require greater computing resources for encryption and decryption because they tend to require more complex processing than common key systems.

For reasons of efficiency, Hitachi chose to develop the privacy-preserving analysis technique based on its common-key searchable encryption technique[(5)] so

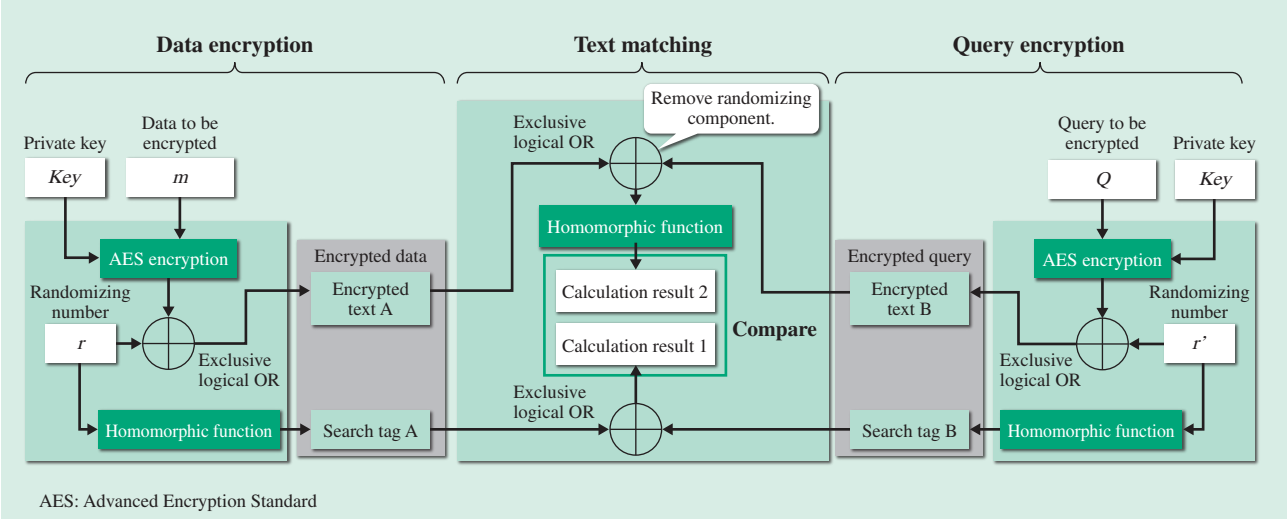

Fig. 2—Overview of Data Encryption, Query Encryption, and Text Matching Using Hitachi's Proposed Searchable Encryption Technique.
The security of the encryption algorithm is enhanced by using random numbers to randomize both data and query encryption. Also, with each step being designed using high-speed encryption primitives, the algorithm can execute searches about 1,000 times faster than searchable encryption algorithms based on public key encryption.

that it would be capable of analyzing large quantities of data. The searchable encryption algorithm has a high level of security, using random numbers to randomize both data and query encryption to encrypt the same plain text (or plain text query) differently each time (see Fig. 2). Since each step of the algorithm is designed using high-speed encryption primitives, such as Advanced Encryption Standard (AES), the most standardized form of common key encryption, it can execute searches about 1,000 times faster than searchable encryption algorithms based on public key encryption[5].

Privacy-preserving Analysis Using Searchable Encryption

By using the above text matching function for searchable encryption, a cloud service can determine the frequency of occurrence of an encrypted query in a database without decrypting the encrypted data. By submitting the appropriate encrypted query to the cloud service, the user can run data mining algorithms on encrypted text that are able to work using only the frequency of occurrence of items. Examples of such data mining algorithms include simple statistical analysis and correlation rule analysis (described below). In terms of the assumptions underlying the use of these searchable encryption techniques, it is important to note that, although the plain text is not divulged to the cloud service, it is possible to obtain frequency of occurrence information from the encrypted data. On the other hand, both the plain text and frequency of occurrence information for the encrypted data are kept confidential from third parties who do not have the encrypted query.

Correlation Rule Analysis

Correlation rule analysis is a data analysis technique for identifying relationships between phenomena from tabular transaction data such as purchase histories[6]. That is, it identifies cases in which if one particular phenomenon occurs (the antecedent), another particular phenomenon has a high probability of occurring also (the consequent). The following describes how correlation rule analysis is performed for a table of transaction data listing the drugs prescribed to a number of patients (see Table 1).

Table 1. Transaction Data
Each line indicates the drugs prescribed to a patient and whether or not they suffered side effects.

	Prescribed drug 1	Prescribed drug 2	Prescribed drug 3	Side effects?
Patient 1	Drug A	Drug B	Drug C	Yes
Patient 2	Drug B	Drug A	Drug F	Yes
Patient 3	Drug B	Drug D	Drug E	No
Patient 4	Drug C	Drug E	Drug F	No
Patient 5	Drug E	Drug A	Drug B	Yes
Patient 6	Drug A	Drug D	Drug E	No
Patient 7	Drug C	Drug B	Drug A	Yes
Patient 8	Drug C	Drug E	Drug F	Yes

Each line (transaction) in the table lists the drugs prescribed to a particular patient. For example, patient 1 was prescribed drugs A, B, and C. In this case, the aim is to look for a rule that states that, if a patient is prescribed drug A (the antecedent), then they are likely also to be prescribed drug B (the consequent). This is represented below by the notation, "correlation rule A ⇒ B." Each correlation rule is evaluated in terms of three indicators: the "support," "confidence," and "lift." These are defined as follows.

The support for correlation rule A ⇒ B is:

$$\text{Supp}\ (A \Rightarrow B) = \frac{\text{Total number of transactions containing A and B}}{\text{Total number of transactions}}$$

The confidence for correlation rule A ⇒ B is:

$$\text{Conf}\ (A \Rightarrow B) = \frac{\text{Total number of transactions containing A and B}}{\text{Total number of transactions containing A}}$$

The lift for correlation rule A ⇒ B is:

$$\text{Lift}\ (A \Rightarrow B) = \frac{\text{conf}\ (A \Rightarrow B)}{\text{supp}\ (B)}$$

For example, the support, confidence, and lift for the correlation rule "Drug A ⇒ Has side effects" in Table 1 are:

Supp (Drug A ⇒ Has side effects) = $\frac{4}{8} = 0.5$

Conf (Drug A ⇒ Has side effects) = $\frac{4}{5} = 0.8$

Lift (Drug A ⇒ Has side effects) = $\frac{0.8}{0.625} = 1.28$

The following section describes the meaning of these indicators.

The aim of correlation rule analysis is to identify the relationships between phenomena that occur frequently in a transaction table, and the support indicates the probability of occurrence in the table. Typically an analysis will focus on those phenomena with high support values. Confidence indicates the probability that the consequent will occur given that the antecedent has occurred. That is, it can be interpreted as the conditional probability. The lift is this conditional probability divided by the probability of occurrence of the consequent. If the lift is significantly greater than one, it indicates a strong correlation between the antecedent and consequent. In the above example, the lift for "Drug A ⇒ Has side effects" is 1.28. This means that a patient who has been prescribed drug A is 1.28 times more likely to have side effects than a patient selected at random. Actual analysis involves identifying correlation rules with high values for support, confidence, and lift, so that further analysis can be performed to determine why there should be a correlation between these phenomena, and the information is provided as feedback for marketing or other activities.

An important point to note with correlation rule analysis is that the definitions of support, confidence, and lift mean that the analysis can be performed using frequency of occurrence information (number of transactions that contain a specific item). It is this fact that makes it possible to use searchable encryption to perform correlation rule analysis on encrypted data.

Privacy-preserving Statistical Analysis and Correlation Rule Analysis

This section describes how searchable encryption is used to perform statistical analysis and correlation rule analysis on encrypted data.

The table at the top of Fig. 3 lists transaction data from Table 1 that has been encrypted using Hitachi's proposed searchable encryption technique. For example, "drug A," the prescribed drug 1 for patient 1

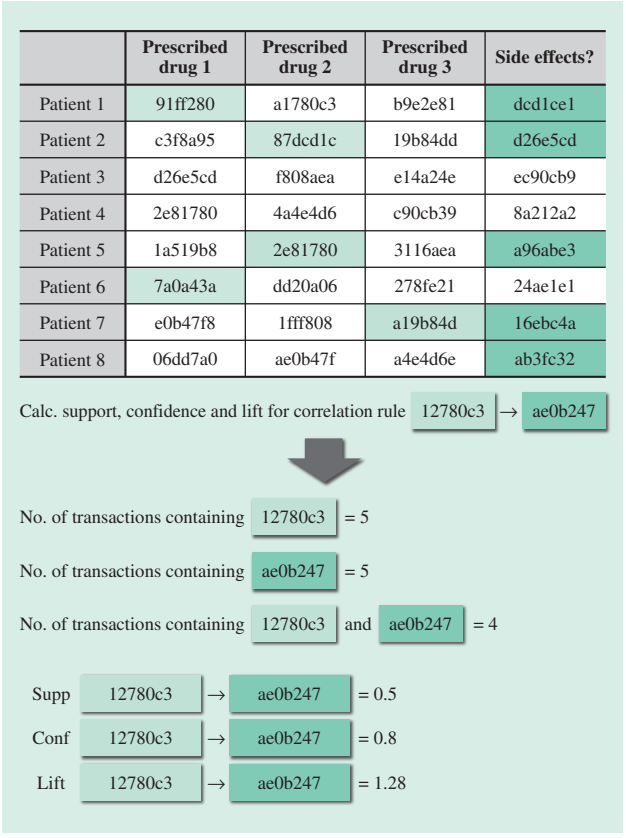

Fig. 3—Correlation Rule Analysis of Encrypted Transaction Data.
The diagram shows how the support, confidence and lift are calculated for the encrypted queries (12780c3 and ae0b247).

in Table 1, is encrypted as "91ff280." Note also that the prescribed drug 2 for patient 2, also "drug A," is encrypted differently, as "87dcd1c." That is, Hitachi's proposed searchable encryption technique encrypts different instances of the same plain text differently, making the encrypted text on its own difficult to distinguish from a random string. Accordingly, even if the encrypted transaction data were leaked to a third party, they would not be able to use it to reconstruct the plain text version.

Now, consider the case of a cloud user who wants to use the cloud to execute the calculations needed to determine the support, confidence, and lift values for the correlation rule A ⇒ B for phenomena A and B. As explained above, the point to note about this calculation is that it only requires four values from the table: the "number of transactions containing A," "number of transactions containing B," "number of transactions containing both A and B," and "total number of transactions." The total number of transactions is known from the table size.

The cloud user first uses searchable encryption to encrypt A and B and generate the encrypted queries, Query (A) and Query (B). Next, these encrypted queries are submitted to the cloud service where the searchable encryption text matching function is used on the encrypted transaction data and queries to calculate the number of transactions containing A, containing B, containing both A and B, and the total number of transactions. These are then used to obtain the support, confidence, and lift values. That is, the support, confidence, and lift values for the encrypted correlation rule Query (A) ⇒ Query (B) are obtained by using searchable encryption text matching in place of conventional text matching. Statistical analysis can be performed using this text matching function in the same way.

Fig. 3 shows how searchable encryption text matching is used to determine the number of transactions that contain the respective encrypted queries, Query (drug A) (12780c3) and Query (has side effects) (ae0b247), and to calculate the support, confidence, and lift values for the correlation rule Query (drug A) ⇒ Query (has side effects). The point to note is that this process uses only encrypted data, with no need to pass plain text information to the cloud. Furthermore, rather than calculating the support, confidence, and lift values for a specific correlation rule A ⇒ B, conventional correlation rule analysis identifies all correlation rules for which the support, confidence, and lift are above specified thresholds. To achieve this, the cloud user needs to create encrypted queries for all phenomena [Query (A), Query (B), and Query (C), for phenomena A, B, and C, and so on] and request the cloud service to perform all the associated analyses.

Sequence of Steps for Privacy-preserving Analysis

Fig. 4 shows a flow chart for the privacy-preserving correlation rule analysis described above when it is used by a cloud user and cloud service. The point to note is that only encrypted data is provided to the cloud.

(1) The cloud user uses common key searchable encryption to encrypt the transaction data (T) and submits the resulting encrypted transaction data [E(T)] to the cloud database.

(2) The cloud user uses searchable encryption to generate encrypted queries {Q(A), Q(B), Q(C), …} for the set of items {A, B, C, …} for which to conduct the correlation rule analysis (or statistical analysis) of the transaction data, and then forwards these to the cloud service in the form of an analysis request. The

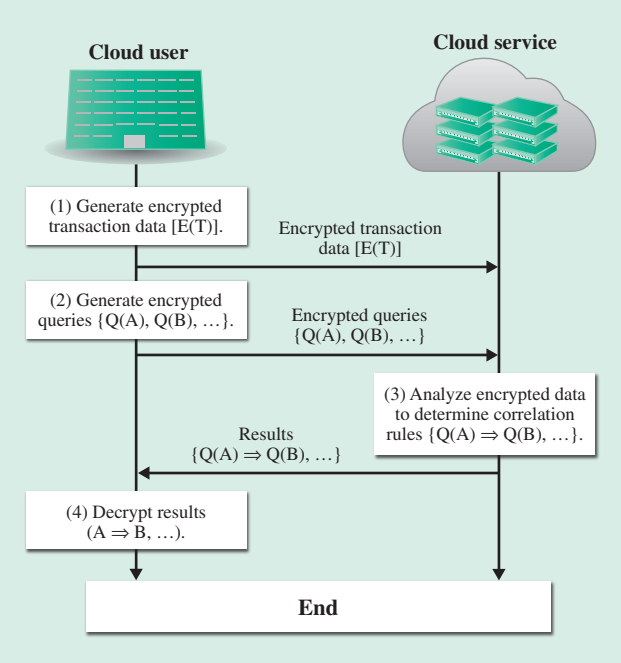

Fig. 4—Flowchart for Correlation Rule Analysis of Encrypted Transaction Data.
(1) Cloud user encrypts transaction data and stores in cloud. (2) Cloud user encrypts queries, which represent the set of items for which to perform correlation rule analysis, and submits these to the cloud. (3) Cloud service determines correlation rules without decrypting and returns the results to the cloud user. (4) Cloud user decrypts the results to obtain the correlation rules.

cloud user also specifies a set of thresholds for the support, confidence, and lift (a, b, c).
(3) The cloud service applies the correlation rule analysis technique for encrypted text (which uses searchable encryption text matching) to determine which of the correlation rules represented by the received queries {Q(A), Q(B), Q(C), …} have support, confidence, and lift greater than or equal to the respective thresholds (a, b, c). The set of those rules that satisfy this criterion {Q(*) ⇒ Q(*), …} is then returned to the cloud user. Statistical analysis works in the same way.
(4) Because the cloud user already knows the correspondence between the set of items {A, B, C, …} and set of associated queries {Q(A), Q(B), Q(C), …}, they can decode the received set of rules {Q(*) ⇒ Q(*), …} and obtain the desired analysis result. That is, the correlation rules for which the support, confidence, and lift exceed the specified thresholds (a, b, c).

Through this procedure, the cloud user can perform a correlation rule or statistical analysis on the cloud without divulging plain text data.

Prototype Performance Evaluation Trial

A trial analysis for determining correlation rules from encrypted test data was conducted using the procedure described above. As in the reference quoted below[6], the trial used 100,000 transactions of test data, each containing ten items of data on average, and with a total of 1,000 different item values.

Hitachi's proposed technique was used for searchable encryption. This searchable encryption technique can perform text matching about 1,000 times faster than searchable encryption based on public key systems. Since determining the total number of transactions containing each item takes up the majority of the execution time during correlation rule analysis, the analysis execution time is approximately one-thousandth of that when public key searchable encryption is used. In a trial run on a conventional personal computer (PC), determining the correlation rules in 100,000 encrypted transactions completed in approximately 600 seconds (10 minutes).

This result demonstrates that the privacy-preserving analysis technique is practical for use on medium-sized quantities of data containing several tens of thousands of records. Meanwhile, although not part of the trials reported here, because the execution time can be expected to scale with the quantity of data, further enhancements such as speed improvement or parallelization will be needed if analyses are to be performed on large data sets containing between several million and several hundred million records. Hitachi hopes to make dealing with large quantities of data the subject of future research.

CONCLUSIONS

This article has described a privacy-preserving analysis technique that can be used to analyze data in encrypted form to provide data security when performing big data analysis on third-party cloud servers.

The core of the proposed method is a searchable encryption technique that permits searching of data in encrypted form and can be used for statistical or correlation rule analysis of encrypted data. Because this privacy-preserving analysis technique only requires encrypted data and encrypted queries, it reduces the risk in the event of unauthorized access or a data leak.

In the future, Hitachi intends to continue with the research and development of advanced security techniques that support both the robust protection and utilization of big data, and to supply highly secure solutions.

REFERENCES

(1) D. Boneh et al., "Public Key Encryption with Keyword Search," EUROCRYPT 2004, pp. 506–522 (2004).
(2) D. Boneh et al., "Conjunctive, Subset, and Range Queries on Encrypted Data," TCC 2007, pp. 535–554 (2007).
(3) R. Curtmola et al., "Searchable Symmetric Encryption: Improved Definitions and Efficient Constructions," CCS 2006, pp. 79–88 (2006).
(4) D. X. Song et al., "Practical Techniques for Searches on Encrypted Data." Security and Privacy, 2000, S&P 2000, Proceedings, 2000 IEEE Symposium on, pp. 44–55 (2000).
(5) M. Yoshino et al., "Study of Searchable Encryption Technique for DBs (2)," The 2011 Symposium on Cryptography and Information Security (2011) in Japanese.
(6) R. Agrawal et al., "Mining Association Rules between Sets of Items in Large Databases," Proceedings of the ACM SIGMOD International Conference on Management of Data, Washington, D.C. (May 1993).

ABOUT THE AUTHORS

Ken Naganuma

Service Innovation Research Department, Yokohama Research Laboratory, Hitachi, Ltd. He is currently engaged in research and development of security technologies for cloud computing and big data analytics.

Masayuki Yoshino, Ph.D.

Service Innovation Research Department, Yokohama Research Laboratory, Hitachi, Ltd. He is currently engaged in research and development of security technologies for cloud computing and big data analytics. Dr. Yoshino is a member of the Information Processing Society of Japan (IPSJ) and The Institute of Electronics, Information and Communication Engineers (IEICE).

Hisayoshi Sato, Ph.D.

Service Innovation Research Department, Yokohama Research Laboratory, Hitachi, Ltd. He is currently engaged in research and development of security technologies for cloud computing and big data analytics. Dr. Sato is a member of the IEICE.

Yoshinori Sato

Service Innovation Research Department, Yokohama Research Laboratory, Hitachi, Ltd. He is currently engaged in research and development of security technologies for cloud computing and big data analytics. Mr. Sato is a member of the IPSJ.