FORTRA™

WHITE PAPER (DIGITAL GUARDIAN)

Why Database Record Matching™ is the Most Effective Method of Identifying Regulated PHI

A Technical Overview for Healthcare IT Executives

Protecting PHI

This paper is addressed to Healthcare Industry executives responsible for maintaining compliance with Personal Health Information (PHI) regulations. In particular, the important role of Database Record Matching (DBRM), as the most effective and accurate method of identifying regulated PHI in a Data Loss Prevention (DLP) solution is described.

Every organization handling Personal Health Information (PHI) today is being held to increasingly higher standards to avoid breaches and leaks that may expose this data. Loss of control of this information can lead to regulatory financial repercussions as well as significant damage to the organization.

Compliance in meeting these standards, begins with the ability to accurately identify any records pinpointing a specific person and their private medical data.

Many products currently exist in the market that can help scan for PHI in outbound email and web transactions, on internal file stores, and exiting to external devices. But these solutions are often incomplete and may be unsatisfactory in their ability to accurately detect true PHI wherever it may be.

Traditional methods used to find regulated or other senstive information are prone to false positives and false negatives.

Tranditional Methods of Detecting Regulated Data

Traditional methods used to find regulated or other sensitive information are prone to false positives and false negatives, have high staffing costs, and often result in drawn out implementations that don't lead to accurate actionable events. Three methods are most commonly employed:

1. Pattern or Regular Expression matching — The most popular methods of finding sensitive data in use today inspect the format of the data in question. Examples include US Social Security Numbers, which often follow a 9 digit "###-##-####" format. In the case of a medical record or patient identification, it will depend on the method of assignment employed by the particular organization.

The main problem with the pattern matching method is the high rate of false positives and false negatives that are generated. If the matching is tuned to allow for a liberal formatting of a medical record number, for example, such as identifying any sequence of the right number of digits as a potential MRN, then the typical result will be an extremely high numbers of false positives to contend with. If tuned for stricter formatting, such as requiring dashes in appropriate places, then the risk increases of missing a MRN not presented, for some reason, in that form.

2. Dictionaries — This detection approach uses a collection, or "dictionary", of words surrounding the potential sensitive information to try to determine context. For example, if a lot of medical terms are used in a transmission, or if terms like "diagnosis code" or "insurance number" are used nearby, the suspected data may be flagged as potentially sensitive and flagged for further inspection.

This method also suffers from false positive and false negative issues. For example, flagging suspected protected information based on general and highly used terms would result in many false positives in instances where those terms are being used conversationally or without necessarily being indicative of personal medical data. A variation of this method is to require extreme specificity, such as looking for the term "patient number" or "MRN" preceding any sequence of the appropriate number of digits. If this is done the false positive rate will be very low, but legitimate numbers can slip by undetected (false negatives) if they aren't expressly labeled.

The real goal in a PHI inspection engine should be to accurately detect the actual pieces of information that are being sought.

3. Combinations — Variants that combine forms of pattern matching and the use of dictionaries have been developed over the years spent trying to build improved detection engines. Unfortunately, these combinations will necessarily present a trade off to the organization deploying such a system. Users are forced to choose between: a) wading through a sea of false positives, or, b) allowing more false negatives in order to avoid getting overwhelmed with manual inspections. History shows that organizations will commonly opt to not spend an inordinate amount of time manually weeding out false positives and, instead, simply accept failing to identify significant instances of regulated information leaving the organization.

4. The Problem with Compromises — For these reasons, many compliance solution vendors have simply chosen to err on the side of accepting false negatives in order to reduce the burden on the user to have to check false positives. Their logic being that many organizations handling PHI will not be aware of, won't see, and, therefore won't complain, about, false negatives – until such time as a loss of data becomes public.

Thus, end results obtained from the use of these traditional solutions may be quite unsatisfactory:

- Too many events that are false positives and require manual inspection
- Too much time spent tuning the system
- Expensive ongoing professional services requirements
- Irritated end users
- Acceptance of gaps in risk

Despite these problems, each of these methods may be applicable in limited use cases. However, a significantly stronger method is available to more accurately detect particular types of sensitive data.

Database Record Matching (DBRM)

Purpose

The real goal in a PHI inspection engine should be to accurately detect the actual pieces of information that are being sought. While matching to certain formats or noting relevant surrounding text as described above may provide an alert to a higher possibility of finding sensitive data, a more accurate method is to look for matches to the actual data itself. And, that is what Database Record Matching does.

Database Record Matching (DBRM) is a method of creating mathematical hashes of the true sensitive data, and using those hashes to look for that exactly identical data by ashing the target data when inspecting other sources such as an email, a file share, the cloud, a web posting; anywhere that true data would be problematic if found there.

Creating Fingerprints

The DBRM process begins with querying an internal database table known to contain complete and accurate records identifying personal information. This is usually a string of characters, such as SSN, MRN, Policy ID, Account number, Member number, etc.

This is typically a simple query performed against a data warehouse or reporting database. It is only important

that data known to be accurate (true) is obtained. Once established, this process is automated to requery the database on a daily or other appropriate regular basis so that current values will always be incorporated. In practice this is typically setup in less than an hour with someone normally responsible for report generation or business intelligence.

At this point the DBRM engine creates one way hashes, often called "fingerprints", of each individual field of protected data, and stores these fingerprints within the engine. For security, the procedure does not keep the original (readable) data, only the hashes are used. These fingerprints will then be used to find instances of the exact same data if it exists in an inspected target file.

Inspecting Data

At this point, the DBRM engine is ready to find sensitive data elements inside operational data. The inspected content might be an email, a web posting, in the cloud, a file on a network device, a file being copied to a USB drive, or anything else being inspected by the overall DLP solution.

Database Record Matching™ (DBRM), exclusive to Digital Guardian for Compliance, is an extremely accurate method to detect an actual policy ID in all inspected text.

The content to be inspected is serially run through the same DBRM hashing algorithm that was used to create the fingerprints of the actual data. When fingerprints (hashes) match, then that exact sensitive data element has been accurately identified.

DBRM is thus able determine which elements in the inspected record matched the actual sensitive data. In addition, multiple elements from the same actual records can be used for further confidence. This could include, for example, requiring that the corresponding patient last name is seen somewhere nearby a potential sensitive MRN discovered in the target data.

An Example with Social Security Numbers

An extreme example using SSNs may illustrate how DBRM differs from simple use of patterns, in reducing the potential for false positives:

- There are 1,000,000,000 possible 9-digit numbers.
- Of these, about 900,000,000 are in valid issuable ranges (as of changes made in 2011).
- If there are only 10,000 actual patient records in a hospital's database, then the odds that merely finding a sequence of 9 digits actually identified a hospital patient would be unacceptably low. In other words, a high number of false positives will be generated by this method.
- With DBRM, on the other hand, the source data for creating the hashes will be the hospital's actual 10,000 patient SSNs. Hence the record being inspected is either related to that patient or that record only coincidentally happens to contain that 9 digit sequence in some other context.

The DBRM system may eventually come upon 9 digit sequences that are only coincidentally equal to, but not actually representing a patient's SSN within the record. To improve upon this the DBRM system can be tuned to reduce the generation of such false positives by requiring another true database element to be present. For instance, by requiring the corresponding person's actual last name (hashed) to be somewhere nearby the SSN, such false positives will be virtually eliminated.

An Example with Medical Records

Medical Record or Medical Insurance policy "numbers" don't follow the same format across institutions, leading to an array of differing possible formats. Some of these are numeric. Some are alphanumeric. And, they can differ in length and other formatting aspects. Thus, pre-built pattern matching solutions can require significant tailoring to produce reliable results with such data. In contrast, DBRM, by its nature, easily provides an extremely accurate means to detect an actual ID in all inspected alphanumeric forms.

DBRM gains more reliable matching for account numbers with fewer digits, while Pattern matching suffers even worse false positive problems than with longer values. With DBRM, detecting account numbers as low as 6 digits (and even sometimes 5 digits on lower volume inspection streams) may be reliably performed.

DBRM is appropriate for any unique database elements, regardless of format. This makes it an ideal method for identifying data protected by regulatory compliance requirements.

Pattern matching will always sacrifice real discovery of sensitive data while producing more false negatives.

Resilience to Variations in Format

The inherent flexibility of DBRM allows the implementation of many improvements worth mentioning here. For example, an MRN such as "257121234" may need to match to "257-12-1234" in the inspected text. Similarly, any identifiers may optionally be presented with any alternate formatting required.

For international use, DBRM is able to handle registration and inspection of non-ASCII data, including double byte language characters. Further, non-ASCII data in the inspected text does not interfere with inspection of ASCII or non-ASCII elements.

The End Result

In customer implementations of DBRM, Code Green Networks has observed organizations experiencing low implementation time requirements, low ongoing time requirements, more actionable results, and lower risk of loss of sensitive data.

A PHI Example

DBRM was recently employed at a medical facility, monitoring transmissions for the approximately 3000 network users. The installation ran a parallel evaluation of a competitive product that employed various pattern matching and dictionary methods. The competing product had been tuned by the vendor for approximately two weeks to achieve the best results it was capable of attaining.

Using patterns and dictionaries and two weeks of tuning, the competitive method produced approximately 1800 events per day, of which approximately 100 were correct matches. In contrast, using DBRM with only 2 hours of setup generated approximately 800 events per day, of which approximately 750 were correct matches.

In this side by side comparison, the traditional methods without DBRM missed at least 87% (650/750) of real PHI, while still producing 94% (1700/1800) false positives.

This facility was able to get to a level of accuracy that was production quality within a day of setup, without requiring a long tuning time, and without wasting employee time wading through false positives. Further, the real risk was reduced in that they were able to find a sizable percentage of the actual PHI leaving the organization, where competing methods failed to catch even the majority of those events.

In a side by side comparison, the traditional methods without DBRM missed at least 87% (650/750) of real PHI, while still producing 94%(1700/1800) false positives.

The New Standard in PHI Protection

Having full time employees chase down false positives and continually tune compliance inspection solutions is now unacceptable. What once passed for "acceptable" may often now require employee resources that are not available, and, most importantly does not deliver meaningful reductions to risk!

DBRM as part of an Enterprise DLP solution fundamentally leads to actionable detection of PHI. Employee time spent will be spent actually correcting behavior and investigating sensitive data leaks, rather than on the technical tedium of trying to make a product work. This allows compliance and risk management owners to participate in event management, rather than assigning information research and remediation tasks to people not involved in the risk management process.

With the application of appropriate technology it is possible to substantially reduce the risk of PHI data loss. Several products are currently available on the market that can help identify PHI in important but limited applications such as outbound email and web transactions. However, the many successful Healthcare Industry deployments of DLP using Database Record Matching have demonstrated that this method can be employed to most accurately identify regulated data throughout the enterprise network and including the cloud with minimal ongoing time commitments and with measurable success.

About Digital Guardian

Fortra™'s Digital Guardian® is the only data aware security platform designed to stop data theft. The Digital Guardian platform performs across traditional endpoints, mobile devices and cloud applications to make it easier to see and stop all threats to sensitive data. For more than 10 years we've enabled data-rich organizations to protect their most valuable assets with an on premise deployment or an outsourced managed security program (MSP). Our unique data awareness and transformative endpoint visibility, combined with behavioral threat detection and response, let you protect data without slowing the pace of your business.

FORTRA™

Fortra.com

About Fortra

Fortra is a cybersecurity company like no other. We're creating a simpler, stronger future for our customers. Our trusted experts and portfolio of integrated, scalable solutions bring balance and control to organizations around the world. We're the positive changemakers and your relentless ally to provide peace of mind through every step of your cybersecurity journey. Learn more at fortra.com.