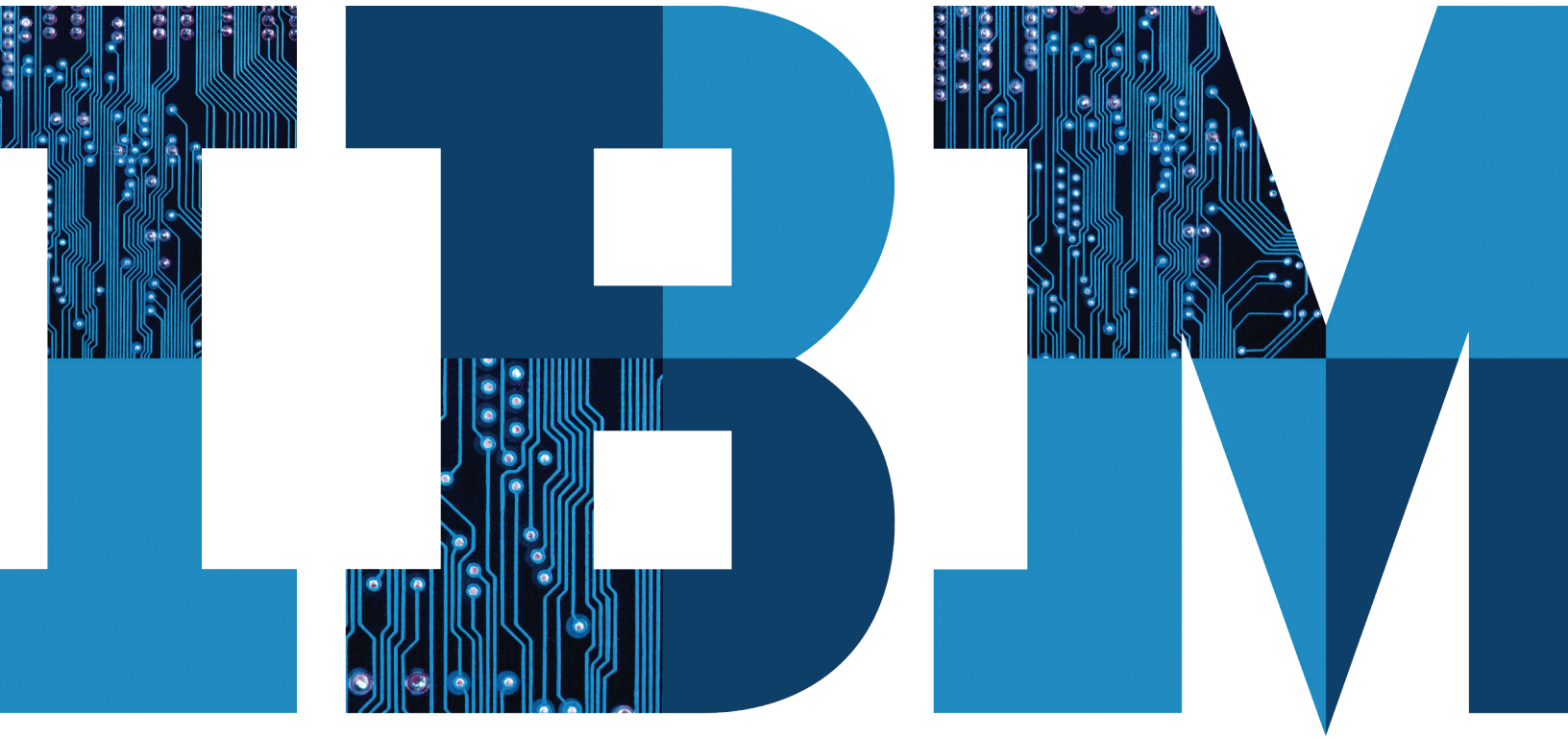
IBM Software
White Paper

July 2012

IBM InfoSphere Guardium Data Redaction: Reconciling openness with privacy

Document protection for regulatory compliance and risk reduction



IBM

Introduction: The market demand for redaction

Until recently, the need to delete sensitive information in documents was restricted to the national intelligence community. In business, government and nonprofits, the need for redaction was much rarer. But the need for redaction is growing quickly, as organizations face regulatory and business requirements to enforce document privacy—not just by controlling access to entire documents, but by selectively deleting private units of information. This means automated redaction solutions are rapidly becoming more important.

The conflict between openness and privacy

With the data explosion in IT today, storing more data is a challenge, but managing it is even more of a challenge. Proper data management ensures that the right people have access to information and that they are using it for legitimate purposes. This creates a conflict between ensuring that those entitled to see given types of information can access it easily, while simultaneously withholding access to information that they are not allowed to see (see Figure 1).

To take an example from the U.S. government: the Freedom of Information Act (FOIA) is intended to hold government organizations more accountable for their actions by making information about those actions available on demand. On the other hand, the same regulation requires that those ordering the documents must not see any sensitive personal or national security information.

Similarly, the Health Insurance Portability and Accountability Act (HIPAA) is designed to enhance sharing of documents between physicians, hospitals and insurers while preventing the unauthorized disclosure of individuals' personal healthcare information. For example, consulting physicians need access to individuals' electronic health records, but they do not need to see billing information that is unrelated to their job duties.

Regulation	FOIA	United States Federal Rules of Civil Procedure	HIPAA	PCI	Sarbanes-Oxley Act
Functional area	Government	Legal eDiscovery	Health	Bank cards	Corporate governance

Figure 1: Sample regulations by sector.

Both regulatory requirements and business pressures make redaction essential.

- Redaction can satisfy governmental regulations, including those in data privacy laws, without restricting the legitimate use of information—thus avoiding sanctions, penalties and costs associated with addressing compliance violations after the fact.
- Redaction can accelerate business interaction by sharing information with customers, partners and other third parties without exposing them to sensitive information that they should not see. Organizations often find strong business value in sharing information, but they must take care to limit exposure to the minimum needed, to avoid the embarrassment and competitive risk of leaks.

Traditional IT solutions are not granular enough

What can technology do to achieve this balance between openness and privacy? The first line of defense consists of familiar access controls for documents, often defined per role. Another part of the solution is data loss prevention (DLP) software, which can restrict the transmission (such as through email) of sensitive information from an authorized user to an unauthorized user. Encryption is also an important means of ensuring data confidentiality. But these approaches, with some exceptions, tend to be blunt instruments that too often restrict access more than is really necessary.

These forms of document security are essential; but for a more flexible, fine-grained approach, redaction is needed as well.

To do the job correctly, redaction software needs the following capabilities:

- The software must securely and completely delete all relevant data. Some ad hoc solutions layer a black rectangle over the data, leaving the private data underneath. At the trial of former Illinois Governor Rod Blagojevich in 2010, documents with information about U.S. President Barack Obama's relationship to the case were officially released with redactions—but the underlying text could be easily recovered by simply copying and pasting it.
- To comply with some regulations, redaction software must be able to retain the original pre-redaction version in a safe place; for other regulations, it must securely delete such versions.
- The tools should allow the labeling of redactions. Rather than simply masking text, some regulations require a meaningful label for the redacted text, such as the words "Social Security Number" or a regulatory section-number, for readability and justification of the redaction.

"Enterprises must now add to the basic characteristics of data protection—preservation, availability, responsiveness and confidentiality—the what and where of data."

-David Hill, Analyst, Mesabi Group

- To address the growing amount of electronically stored information, the solution must scale to large numbers of documents.
- The software should automatically identify suggested redactions, but also allow for manual review, so that a compliance officer can accept, reject or refine suggested redactions.
- The interface should allow secure online viewing, in addition to the creation of redacted documents. Viewing documents in the browser is more convenient and more secure than issuing a file that could more easily leak out.
- In some use cases, the web viewer must give users with proper permissions the ability to securely retrieve some types of redacted information as long as they specify a valid business reason and their access is logged. Without the flexibility of this feature, redaction policies must be either overcautious and redact information that the user may need, or too permissive, exposing information for the user's convenience but revealing more information than needed.
- The solution should log the information redacted, along with the documents, pages and text sections that were viewed, for future auditing.

IBM® InfoSphere® Guardium® Data Redaction addresses all these capabilities and more.

Privacy vs. security

Security and privacy are related, but they are distinct concepts.

Security is the infrastructure-level lockdown that prevents or grants access to data based on authorization. It is the realm of passwords and encryption. In contrast, privacy control validates that already-authenticated users have a legitimate business need to see specific information. These needs are usually specific to a job function and defined by regulatory or management policy.

There are many security solutions that prevent unauthorized user access. However, there are very few privacy solutions that protect sensitive data from improper use by employees and other authorized users who might pry into data that they have no legitimate business purpose to see.

Two recent cases illustrate this distinction: doctors and nurses at UCLA Medical Center were caught going through Britney Spears' medical records. And during the 2008 presidential campaign, U.S. State Department contractors viewed passport records of presidential candidates, including Barack Obama.

This was not hacking. These people had passwords; they needed access permissions as part of their day-to-day work. The problem was that they had no need-to-know. They accessed the records out of mere curiosity, not out of a legitimate functional need.

With a redaction-based web viewer, users see the documents redacted according to their roles: in a hospital, physicians and financial personnel will see different information; and in the military, combat officers will see different information from logistics specialists. These redactions can be made very conservative, redacting information if there is any doubt about whether it should be visible to users of a given role. Thus, where permitted, authorized users can state their business purposes and fill in some of the redacted information in their documents, knowing that all accesses are logged.

Automation is essential

As redaction takes on an important role across enterprise IT departments, manual redaction is insufficient, whether with a black marker or with ad hoc electronic solutions.

High volumes of electronic documents make manual redaction expensive and error-prone. People who have the regulatory knowledge to identify private information are too expensive for painstaking rote tasks, and even if they are assigned to such a task, they are only human; they are slow and make mistakes.

Even if the black marker method were feasible for isolated documents, managing the workflow involved makes it prohibitive for large document collections. The redaction process includes identifying the documents in the repository that need redaction; finding the sensitive information in each document; cross-referencing the semantic type of each unit of information to the role of the recipient and determining whether to redact it; creating the redacted copy; reviewing the redaction and then redacting it again if needed; and finally storing the redacted copy in a way that links to the original in the repository.

Automating the redaction process is essential for making this time-intensive activity more cost-effective, allowing organizations to better comply with regulations, preserve their competitive advantage, secure their intellectual property and safeguard their public reputation.

Free-text and forms

The IBM InfoSphere Guardium Data Redaction solution provides automated redaction that works in two ways, depending on whether the document is free-text or a structured form.

For free-text documents, the redaction engine automatically identifies and extracts relevant units of information (see Figure 2). Simply using text patterns is not enough—there is no formal pattern, for example, that captures personal names. Dictionaries are not sufficient either, since homonyms can disguise meanings (is "bush" a plant or a former U.S. president?). Instead, it is necessary to combine regular expressions and dictionaries with a syntactic analysis of the text surrounding the relevant information.

Structured forms, on the other hand, require a different technique, in which the known form layout is leveraged for accurate redaction. This allows even low-quality scans with

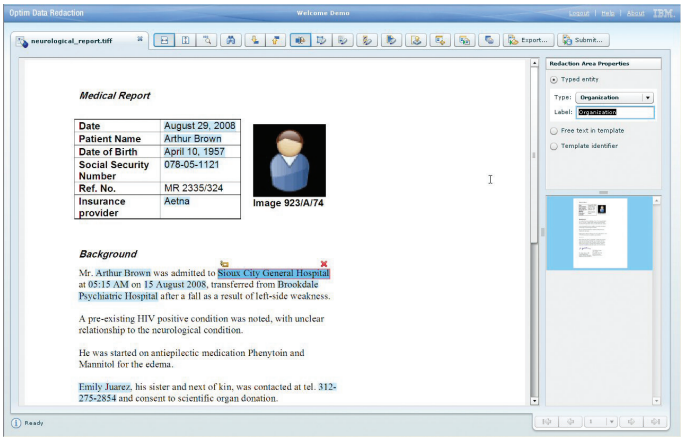


Figure 2: A free-text document in InfoSphere Guardium Data Redaction Manager; the text highlighted in blue is to be redacted.

handwritten text to be processed; if they are accidentally skewed or resized, they can be straightened and aligned with a template. To accomplish this, a reviewer begins by redacting a sample form (for example, a blank) and marking the sensitive fields to be redacted, together with elements that identify instances of the form, such as the form title or identification number. This creates a template for subsequent forms. The software redaction solution matches templates to forms, eliminating the costly presorting of different form types. Next, it applies a template to each form, precisely deleting the marked fields based on their position (see Figure 3).

Complete document format coverage

InfoSphere Guardium Data Redaction processes documents in many formats: PDF, TIFF, Microsoft Word, plain text, XML files and more.

Some input documents, such as Microsoft Word and many PDFs, carry text in them, but others like TIFF and some PDFs are pure images. For image files, the solution applies high-quality optical character recognition, and then processes the text. If there are any photographs or other graphics in a document, the solution preserves them as such.

Though most sensitive information arrives as text, images too can contain sensitive information. For example, an X-ray image may identify a patient's name, a portrait photograph may betray an identity or a satellite image may expose the location of a military unit. In the InfoSphere Guardium Data Redaction system, sensitive images can be located in a form using templates, or marked by a reviewer in the web-based Redaction Manager.

Each of the document formats—PDF, TIFF, Microsoft Word and plain text and XML files—can serve not only as input but also output, and the choice of output type is configurable. In some cases, regulations or business needs require the redacted document to be in "native" format, the original format of the

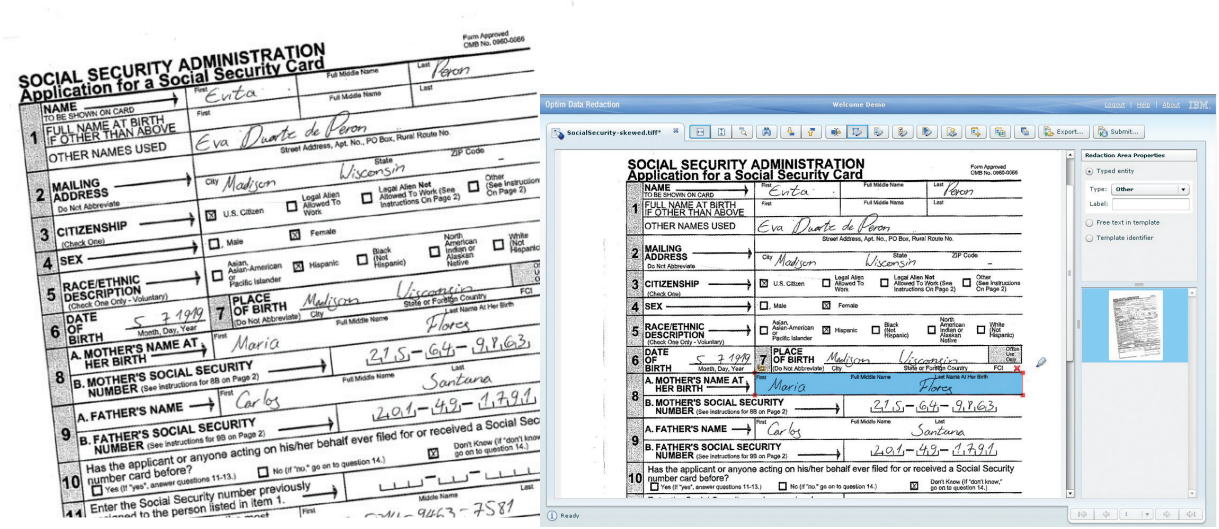


Figure 3: On the left is a skewed scan of a form; on the right is the automatic identification of the sensitive field in the form, as seen in the InfoSphere Guardium Data Redaction Manager.

input. In other cases, it is necessary to output all documents, regardless of the input, into standard graphical formats preserving the precise layout, such as TIFF or PDF. Alternatively, if further machine processing is needed, plain-text output can be specified for all input formats.

Finally, the InfoSphere Guardium Data Redaction system automatically removes the wide variety of hidden information that is often stored in PDFs and Microsoft Word documents, even without the user's knowledge. This includes hidden layers; comments and scripts; white text and tiny fonts; metadata such as the names of document editors and the creation date; and historical contents of a document preserved with editing features like Undo and Track Changes. All these are safely deleted as part of the redaction workflow.

Efficient workflow

With thousands of documents to redact, workflow management is essential. Simply printing out the pages and deleting the sensitive text won't do it—it would be impossible to keep track of the stacks of paper—and the same is true with ad hoc redaction of masses of electronic documents. The only solution is for redaction to fit into enterprise content management (ECM) processes. The InfoSphere Guardium Data Redaction system supports a variety of such workflows out of the box, and can read and write documents in ECM systems such as IBM FileNet® P8 and IBM Content Manager 8:

- **Batch redaction:** This workflow automatically redacts large numbers of documents in a repository. Depending on regulatory requirements, a reviewer can then examine from 0 to 100 percent of the redacted documents and approve, reject or refine the redaction as needed. In this way, the redaction solution combines the strengths of machine processing and human domain knowledge.
- **On-demand redaction:** A workflow used when individual documents must be processed as needed. For example, a business user may need to cleanse private information from a document before emailing it to a business partner. The sender can open the document in Redaction Manager, which instantly suggests text to be redacted. The sender can then refine this redaction before releasing the document.
- **Secure document viewing:** For this workflow, InfoSphere Guardium Data Redaction provides a document viewer. All documents, regardless of original format, are displayed to the user in a uniform way in the browser, with no need to download a document that could subsequently be leaked. Sensitive information in the document is securely deleted according to the recipient's job role. In accordance with regulations, this data is typically deleted in a way that does not allow it to be viewed; for some types of information, users may have permission to securely retrieve the redacted units after specifying their need to know (see Figure 4).

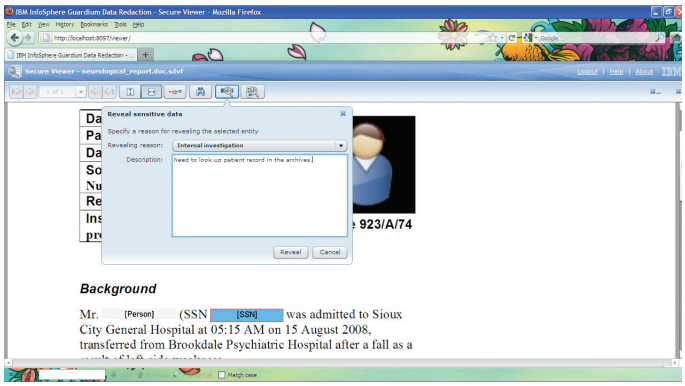


Figure 4: In this example, a user is providing a business justification to access redacted SSN information; the user's permission level determines whether or not the requested information will be revealed.

Policy-based redaction

To gain maximum business value in the redaction process while also minimizing deployment costs, the redaction solution must have the ability to quickly and easily implement the policies defined by regulatory frameworks, typically by cross-referencing the recipient's role against the type of information to be redacted.

The relevant roles are already defined in many enterprises. The redaction solution leverages these roles and links them to fine-grained permissions drawn from regulations, creating a privacy compliance system that directly meets requirements at minimum cost.

Thus, a physician might be allowed to see a patient's medical information, but not sensitive financial information, while the reverse is true for the hospital's billing clerk.

Likewise, in eDiscovery as part of legal cases, the United States Federal Rules of Civil Procedure specify that a litigant's attorney can see all client documents in full, including privileged information, while the opposing counsel can see the documents minus the attorney-client privileged information.

The InfoSphere Guardium Data Redaction architecture

InfoSphere Guardium Data Redaction is centered on a server that automatically identifies sensitive information and generates the redacted documents. The server controls the various workflows needed to manage the redaction process, typically accessing ECM systems with thousands or millions of documents. For maximum efficiency and better hardware utilization, the server can run multiple redaction sessions in parallel in multi-core CPUs or across multiple machines.

InfoSphere Guardium Data Redaction features a modular architecture, so the various components (such as repository connectors, information extractors, and authentication and policy libraries) can be plugged in as needed to support specific redaction requirements.

For programmatic access, the server exposes redaction services over SOAP or Java, or simply by placing files in a file system or ECM folder. This enables integration into existing enterprise redaction workflows. There are two graphical user interfaces: a web-based Redaction Manager that offers redaction review and refinement capabilities, and a secure document viewer that enables documents to be presented on the web without requiring software installation for the end user (see Figure 5).

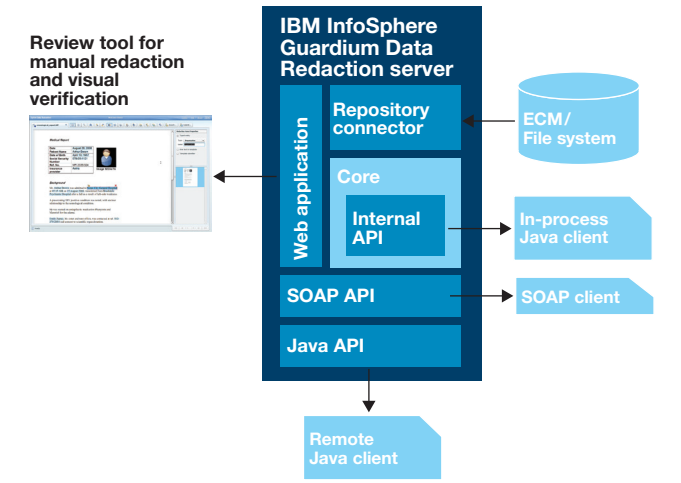


Figure 5: The InfoSphere Guardium Data Redaction solution architecture.

Case in point: Data redaction at a health insurer

A large private health insurer faced new regulations requiring it to share health records with its customers, healthcare providers and the National Health Service. However, the same laws required the insurer to carefully tailor the released personal health information according to the role of the document recipient. This insurer also faced financial laws that required it to delete credit card numbers stored in its archives.

The insurer needed to make documents accessible to independent insurance agents and other business partners. At the same time, preserving privacy was essential to maintaining the company's good reputation for respecting its customers' rights.

Previous privacy practices, such as manual redaction and document-level access controls, put the insurer at risk of regulatory violation. If the organization shared documents without checking them, it risked exposing customers' private information. This led to the opposite extreme: the insurer withheld documents from various recipients, even where this meant violating regulations or missing opportunities for business value.

As one part of the organization's requirements, the document management team needed to archive an extensive collection of insurance policy applications, most of which were low-quality scans of forms. These forms contained credit card information, and Payment Card Industry Data Security Standard (PCI-DSS) regulations required the insurer to keep credit card numbers out of its archives. Conversely, insurance regulations required these documents to be archived indefinitely—decades may pass before insurance claims are made. By batch-redacting the policy applications using the InfoSphere Guardium Data Redaction solution, the insurer was able to rapidly and efficiently control the private information in the archives. As new forms arrived, the same process was automatically applied before they were archived. The insurer is now able to smoothly share or archive documents while precisely withholding the information required by law.

InfoSphere Guardium Data Redaction: Share sensitive data securely

Regulation, best practices and data privacy laws are changing the rules for how organizations grant or deny access to information. Organizations must not only ensure the retention, availability and confidentiality of documents, but release or archive precisely the information allowed by regulations and business needs, taking into account intended readers of the documents.

IBM InfoSphere Guardium Data Redaction features

Enterprise integration

Out-of-the-box support for FileNet P8 and IBM Content Manager 8

Integration-ready capabilities for other document management products

Integration-ready for enterprise authentication and policy systems

Regulation-based policy model

Ease of use

Automatically identify sensitive information in documents, forms, images, text and more for review by security professionals or other stakeholders to set appropriate redaction or other privacy policies or for reporting

Richly functional, zero-install web interfaces

Workflow support; automated batch redaction with optional review; on-demand redaction; document review and forwarding

Secure role-based document viewer with optional flexible revealing by policy

Automated multi-format redaction

Free-text entity extraction with industry-leading library developed by IBM Research

Advanced form redaction, including low-quality, skewed or resized scans

Mixed free-text/form redaction

Support for multiple input and output document formats, including Microsoft Word, TIFF, plain text, XML and PNG

Graphical/textual redaction, simultaneously preserving layout and accurately analyzing the text

Support for English, German, French and Spanish textual entities

Complete removal of hidden data

Support for stamps such as Bates Number, Date, Document ID, Content Type, Repository Info, or other ways to uniquely identify a document

Support for watermarking to help with content identification and authentication as well as communication of ownership and copyrights

The InfoSphere Guardium Data Redaction solution was designed to do just this. It automatically identifies sensitive data in documents, and securely deletes sections of the document according to the regulatory or business policy while taking into account the semantics of the information and the role of the recipient.

InfoSphere Guardium Data Redaction supports many of today's document types, including scanned or originally electronic documents. It leverages unique entity extraction and optical character recognition techniques to identify sensitive data in documents, making the redaction process repeatable and reliable for organizations to manage, measure and trust. InfoSphere Guardium Data Redaction is part of the IBM Security framework for data and information, helping organizations meet the broader challenge of protecting sensitive data, no matter where it resides.

About IBM InfoSphere Guardium Solutions

Since data is a critical component of daily business operations, it is essential to ensure privacy and protect data no matter where it resides. Different types of information have different protection requirements; therefore, organizations must take a holistic approach to safeguarding information.

- **Understand where the data exists:** Organizations can't protect sensitive data unless they know where it resides and how it's related across the enterprise.

- **Safeguard sensitive data, both structured and unstructured:** Structured data contained in databases must be protected from unauthorized access. Unstructured data in documents and forms requires privacy policies to redact (remove) sensitive information while still allowing needed business data to be shared.
- **Protect non-production environments:** Data in nonproduction, development, training and quality assurance environments needs to be protected, yet still usable during the application development, testing and training processes.
- **Secure and continuously monitor access to the data:** Enterprise databases, data warehouses and file shares require real-time insight to ensure data access is protected and audited. Policy-based controls are required to rapidly detect unauthorized or suspicious activity and alert key personnel. In addition, databases and file shares need to be protectedagainst new threats or other malicious activity andcontinually monitored for weaknesses.
- **Demonstrate compliance to pass audits:** It's not enough to develop a holistic approach to data security and privacy. Organizations must also demonstrate and prove compliance to third party auditors.

IBM InfoSphere Guardium solutions for data security and compliance support this holistic approach, helping organizations protect against a complex threat landscape while remaining focused on their business goals.

About IBM InfoSphere

InfoSphere Guardium is a key part of the IBM InfoSphere portfolio. IBM InfoSphere software is an integrated platform for defining, integrating, protecting and managing trusted information across your systems. The InfoSphere platform provides all the foundational building blocks of trusted information, including data integration, data warehousing, master data management and information governance, all integrated around a core of shared metadata and models. The portfolio is modular, allowing you to start anywhere, and mix and match InfoSphere software building blocks with components from other vendors, or choose to deploy multiple building blocks together for increased acceleration and value. The InfoSphere platform provides an enterprise-class foundation for information-intensive projects, providing the performance, scalability, reliability and acceleration needed to simplify difficult challenges and deliver trusted information to your business faster.

For more information

To learn more about IBM InfoSphere, please contact your IBM sales representative or visit: **ibm.com**/software/data/infosphere

For more information about data privacy and IBM InfoSphere Guardium Data Redaction, please contact your IBM representative or visit: **ibm.com**/software/data/guardium/data-redaction/

IBM Corporation
Software Group
Route 100
Somers, NY 10589
U.S.A.

Produced in the United States of America
July 2012

IMW14307-USEN-03