



Steps to Discovery and Protection of Sensitive Data: Find IT. Search IT. Mask IT.

White Paper

Dataguise, Inc.

2201 Walnut Ave., Ste. 260

Fremont, CA 94538

(510) 824-1036 www.dataguise.com

Contents

- Introduction 2
- Need for improved data protection..... 2
- Challenges to implementing data protection 4
 - Knowing where data repositories reside..... 4
 - Knowing what sensitive data resides in repositories 5
 - Managing the risk of distributing application data 5
- Data masking for protecting non-production data..... 5
 - Overview of masking approaches 6
 - Benefits of applying masking for non-production data 7
- What to look for in a data masking solution..... 8
- The Dataguise solution for data protection..... 8
 - Find IT: DgDiscover 9
 - Search IT: DgDiscover.....10
 - Mask IT: DgMasker.....10
 - Benefits of the Dataguise solution for sensitive data protection11
- Conclusion12
- About Dataguise.....12

Introduction

Enterprise applications are the repositories for a wide variety of sensitive data. HR systems can contain information about employees and their dependents such as salaries, taxpayer IDs, names and addresses, and medical histories. Sales automation applications contain customer information such as credit card numbers, card expiration dates, addresses and telephone numbers. Supply chain applications contain proprietary information such as pricing and sales margins while financial applications contain financial performance data. The need to satisfy requirements for regulatory compliance, data theft prevention and sound corporate governance make it imperative that organizations implement the necessary controls to prevent exposure of this sensitive information.

Most enterprises take measures to protect the sensitive data in their production environments. These can include technical controls such as firewalls, two-factor authentication and column-level access controls, as well as policy controls such as user training, data handling procedures and regular audits. However, data from production applications is frequently copied to support application development, testing, QA and pilots. As a result, sensitive data managed in these applications can be inadvertently propagated throughout the extended enterprise and exposed to a wide variety of individuals who may not have a business need to access this information.

Dataguse addresses the problem of protecting sensitive data in enterprise environments. Dataguse's solutions help organizations identify repositories in their networks, search structured and unstructured repositories for potentially sensitive data, and automatically de-identify sensitive data with proven masking technologies. Dataguse's solutions are enterprise ready, deliver high performance, and integrate into secure software development lifecycles. They help organizations more effectively manage risks as well as address critical regulatory compliance requirements.

Need for improved data protection

The need for enterprises to take effective measures to protect their sensitive data has never been greater. With data thefts and breaches in the news, regulators and legislators are imposing higher standards for protecting sensitive personal and financial information, with higher fines and penalties for offenders. Three trends ensure that data privacy concerns will remain a top-of-mind issue for operational managers. These trends are: a heightened threat environment, regulatory compliance and privacy laws, and the increasing management complexity of the enterprise application environment.

There can be little doubt that the threat environment for information systems has changed significantly in the past few years. Enterprise applications have been connected to the internet for over a decade. However, tools and techniques for perpetrating data theft are now available to anyone in the world with a laptop and internet connection. The lucrative nature of data fraud has resulted in the emergence of loosely affiliated organized crime syndicates focused on stealing and trafficking in personal information. Today there are highly efficient marketplaces for buying and selling personal information. It is interesting to note that the black market price for a complete stolen identity has dropped from \$100-\$150 in 2005 to around \$14 to \$18 in 2009.ⁱ

This drop in price reflects the fact that there is a plentiful supply of stolen information now available on the market. Table 1 lists the top 10 reported data breaches since 1984.ⁱⁱ

Records Lost	Date Reported	Organizations
130,000,000	1/20/2009	Heartland Payment Systems
94,000,000	1/17/2007	TJX Companies, Inc.
90,000,000	1/6/1984	TRW, Sears Roebuck
76,000,000	5/10/2009	National Archives and Records Administration
40,000,000	6/19/2005	CardSystems, Visa, MasterCard, American Express
30,000,000	6/24/2004	America Online
26,500,000	5/22/2006	U.S. Department of Veterans Affairs
25,000,000	11/20/2007	HM Revenue and Customs, TNT
17,000,000	10/6/2008	T-Mobile, Deutsche Telekom
16,000,000	1/11/1986	Canada Revenue Agency

Table 1: The 10 largest reported breach incidents since 1984 have resulted in the loss of over half a billion records.

Increased publicity about financial fraud and data theft has fueled public concerns about corporate governance and how organizations handle private data. Regulators have responded with an ever expanding list of government and industry regulations. Legislation including HIPAA, GLBA, FISMA and HITECH mandate that organizations take steps to protect information systems and data. Industry requirements such as PCI (payment card industry) and NERC (energy industry) can carry their own fines and sanctions. Increasingly, state legislators are enacting privacy laws which can carry steep fines for non-compliance. All of these regulations have implications for how sensitive data is collected and managed in the enterprise.

Finally, the increasing complexity of the enterprise application environment makes protecting sensitive information more challenging. Engineering organizations are under pressures to streamline and compress their deployment cycles, driving them to run development and test activities in parallel and more frequently. Business units, pressured to “do more with less” are sending more of their development and testing activities to offshore or outsourced organizations. The result is less management visibility into where systems are deployed, who has access to

Steps to Discovery and Protection of Sensitive Data: Find IT. Search IT. Mask IT.

those systems, and what sensitive data might reside on them. Figure 1 shows an overview of a typical application development flow and the resulting mix of production and non-production systems.

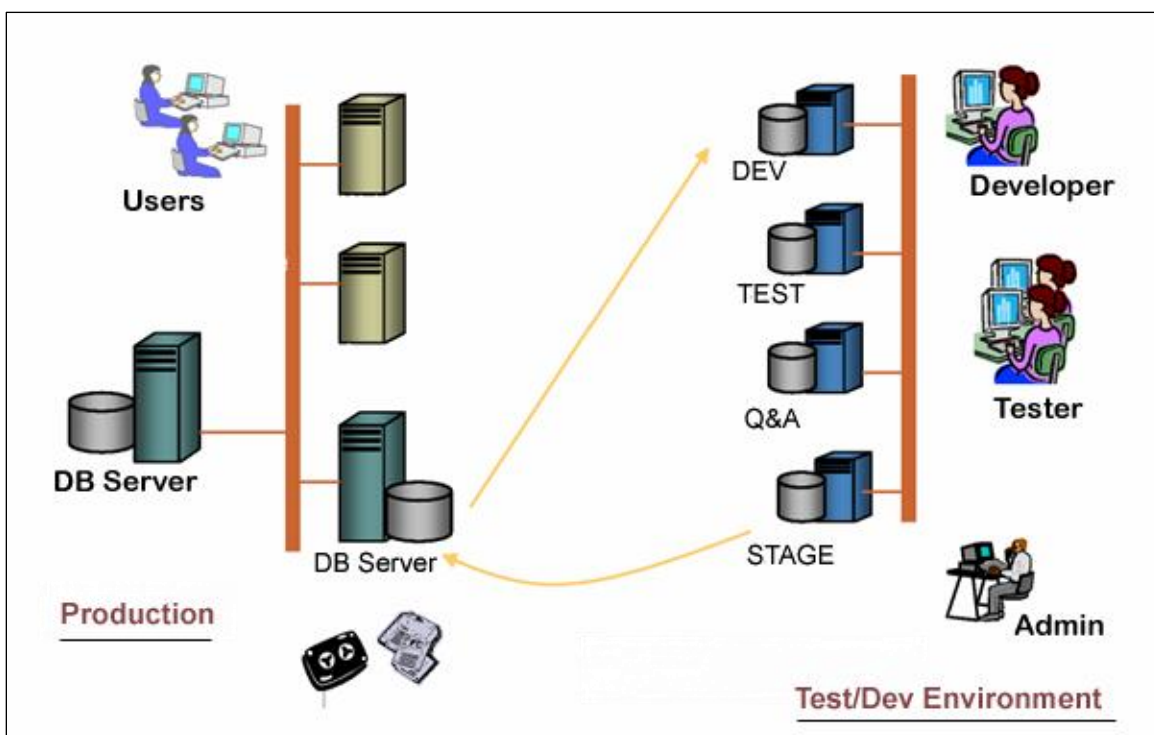


Figure 1: Enterprise application deployment involves the frequent cloning or replication of data to non-production, development and test environments.

Challenges to implementing data protection

The factors listed above are driving managers to reconsider their information security practices. From an execution perspective, there are three practical challenges managers and administrators face to implementing strategies for protecting the sensitive data under their responsibility:

- Knowing where data repositories reside in the enterprise
- Knowing what sensitive data resides in these repositories, and
- Managing the risk associated with distributing application data to support critical business processes

Knowing where data repositories reside

With today's rapid development and deployment practices data tends to proliferate. For example, parts or all of a production dataset may be cloned repeatedly to support non-production activities such as development, QA, and pre-deployment testing. Running many such activities in parallel results in multiple database instances deployed in the enterprise to support each

production application. Increased use of virtualization, especially for development and test environments, complicates the problem because additional application instances become very easy to deploy on demand. As a result, multiple new data repositories containing sensitive information can potentially emerge in the enterprise at any time.

Knowing what sensitive data resides in repositories

Nearly all production applications will contain sensitive information which must be protected from disclosure. Responsible organizations have taken measures to protect these applications and data as part of their normal information security practices. However, in the interest of supporting efficient and streamlined application development and deployment, sensitive information can end up replicated into non-production systems. Without detailed knowledge of the applications and business process flows which underlie these applications, it is very difficult to know where sensitive information resides. Data analysis and analyzing data flows can take months and a team of analysts to complete. At the end of the process, managers are often not sure that all the sensitive data has been identified.

Managing the risk of distributing application data

Application development, test, QA and training are critical enterprise functions and require real or realistic data in order to be effective. However, the users of these systems rarely have the legitimate business need to access all of the sensitive data that may be maintained in these production systems. Many times the access controls for these non-production environments will not be as robust as they are for the production environment. A case in point would be outsourced or offshore development, where management visibility into application and data controls can be minimal.

Encryption is an increasingly popular solution for protecting sensitive data in production environments. However, encryption is generally not an appropriate solution for protecting data in non-production environments. Encryption key management and distribution can be complex and labor intensive. Encryption, since it is a two-way function, can be broken either by technical or social means. Finally, encrypted data must be decrypted to be useful to applications. This means that sensitive information running in these environments has the potential to be exposed to a variety of internal and external users. The key lesson for managers and security analysts planning to rely solely on encryption to protect data in non-production systems is, "If you can see it, you can steal it".

Data masking for protecting non-production data

For making data available to applications and users in non-production environments, data masking has emerged as the preferred approach. Data masking works by performing a one-way transformation on the data, which hides or de-identifies sensitive data in the repository. This renders the data safe for use in non-production deployments. A diagram illustrating the concept of masking is in Figure 2.

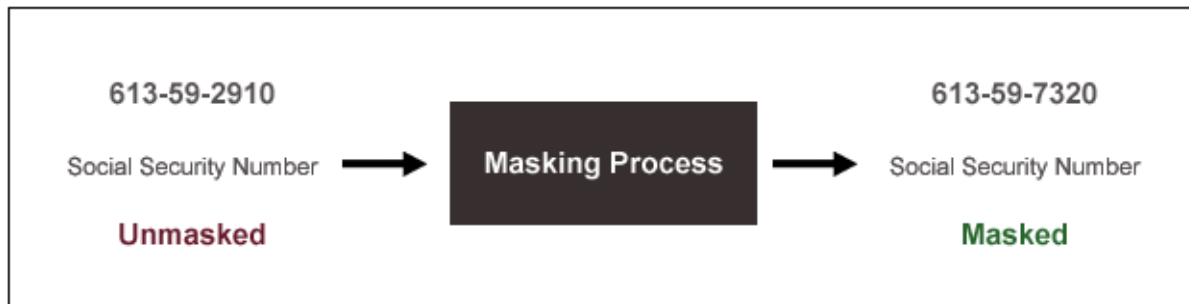


Figure 2: With data masking, sensitive information is selectively replaced with fictitious data.

Masking, or de-identification, applies one-way algorithms to the data so that sensitive data cannot be recovered. Masked data needs to be generated with proper business rules so as not to break applications. Also, care needs to be taken to preserve the relational integrity of data in structured sources. The best tools available for masking automatically determine structural relationships within the data and make transformations in such a way as to preserve the relational integrity of the source data.

Recently, the adoption of data masking technology has grown because of the need to protect private data in test environments, especially when supporting offshoring or outsourcing of application development. In addition, regulatory and legal requirements are demanding protection for private data regardless of where it is stored. The analyst firm Forrester estimates that 35% of enterprises will be implementing data masking by 2010, with financial services, healthcare, and government sectors leading the adoption.

Overview of masking approaches

There are several data masking approaches that can be applied for de-identifying data. These include:

Character masking: This technique replaces sensitive parts of a data field with a masking character such as a hash (#) or asterisk (*). A variant of this would be replacing sensitive parts of a data field with random numbers and/or characters, producing a character masked field capable of being processed by downstream applications.

Fictitious data: This technique replaces data with fictitious values, making the data look real when it is in fact bogus. For example, the customer name "John Barrow" could be substituted with the name "Jim Carlos." This technique can be transparent to application development or testing functions when it retains the business rules associated with the data.

Date aging: In this technique, a date field is either increased or decreased, based on policies defined for data masking. However, date ranges must be defined within acceptable boundaries so that the application is not impacted. An example of date aging would be moving the date of birth back by 2,000 days, which would change the date "12-Jan-1978" to "16-Mar-1972."

Numeric alternation: In this technique, a numeric value is increased or decreased based on a percentage. For example, a salary value could be increased by 10 per cent. This approach conceals the real value of the data, but if someone knows even one real value, they could decipher the entire pattern. While this is an easy technique to employ, it can also be easily decoded.

Shuffling data: In this approach, data is swapped between rows within a particular column, like shuffling a pack of cards, breaking the associations between each data record. An example might be moving an account number to a random row, so that the account number associated with a given customer's record is different from the original.

Table 2 illustrates how each of these algorithms works against original data.

Algorithm	Original data	Masked Data	Explanation
Character Masking	613-30-3291 (SSN)	613-30-#### (SSN)	Last four characters hashed out
Fictitious Data	John Barrow	Jim Arthur	Real name replaced with a random name
Date Aging	5/1/2006	3/1/2002	Date decreased by 4 years and 2 months
Numeric Alteration	10201	10401	Numeric increment by 200
Shuffling Data	Jack Mellon	Roger Smith	Name was shuffled

Table 2: Illustration of typical data masking algorithms.

Benefits of applying masking for non-production data

With masking, the need to control data in non-production environments is greatly reduced since sensitive information can be selectively scrubbed from the data permanently. Automated masking technology provides a number of benefits for supporting secure software development and deployment lifecycles. Masking tools make better use of administrator time by removing the need to write scripts and devise ad hoc procedures for implementing data masking algorithms. If data masking is too time consuming or difficult to perform, administrators may be tempted to skip it entirely thereby placing sensitive data at risk. Finally, without automated masking technology it is difficult to verify that ad hoc, DBA-written scripts are effectively masking all the sensitive data, and that these processes are being implemented consistently and in accordance with enterprise information security policies.

What to look for in a data masking solution

There are a number of important features which are essential for an enterprise-class data masking solution.

Finds data repositories in the network. In an enterprise, new database instances can be deployed at any time. This is especially true where virtualization technology is being used, which makes it possible to easily deploy new instances on demand. An effective sensitive data protection solution needs to have the capability to find new data repositories in the network.

Identifies sensitive data automatically. Sensitive information can reside anywhere in the enterprise, in a variety of structured and unstructured repositories. A sensitive data protection solution needs to have the ability to identify probable sensitive information based on factors such as character patterns, numerical sequences, data relationships and column labels.

Creates “real” data. Any datasets generated by a masking solution need to preserve application integrity. Data generated by the masking process must fit the appropriate business rules so that it won't break consuming applications running in development and test environments.

Preserves relational integrity. A data masking solution needs to be capable of automatically discerning the data relationships columns and tables. It then needs to perform the appropriate data transformations in such a way as to preserve the relational integrity of the data set.

Offers predefined rules for meeting compliance requirements. One of the design centers for any data masking solution is as a tool for addressing compliance requirements. A data masking solution should include predefined rules for identifying data related to specific compliance initiatives, specifying the appropriate masking strategies, and issuing management- and auditor-ready reports.

Supports the enterprise development lifecycle. Many organizations clone production data sources for nonproduction use on a weekly or even daily basis. A data masking solution needs to deliver high performance and support the various database management systems in use in the enterprise. It also must allow administrators to define masking policies and apply those masking policies routinely as new data sets are generated.

The Dataguise solution for data protection

The Dataguise solution for data protection is built around three critical data protection functionalities: Find IT, Search IT, and Mask IT. These are implemented with two tools, DgDiscover™ and DgMasker™. DgDiscover locates databases on the network, and uses advanced pattern matching capabilities to search structured and unstructured repositories for sensitive data. DgMasker applies proven data de-identification techniques to identify

Steps to Discovery and Protection of Sensitive Data: Find IT. Search IT. Mask IT.

relationships between the data and transform the data so that it is protected but still useable by applications.

Find IT: DgDiscover

The typical large enterprise will have thousands of development and test databases deployed in the environment. Virtualization technology now makes deploying application instances easier than ever, creating a challenge for managers and administrators who need to know where all of these various applications are located. DgDiscover's "Find IT" capabilities can identify Oracle, SQL Server, MySQL, DB2, and other database instances deployed in the network. This is a unique capability in the industry, and supports the secure development lifecycle by helping to ensure that new database instances deployed in the environment are identified in a timely fashion.

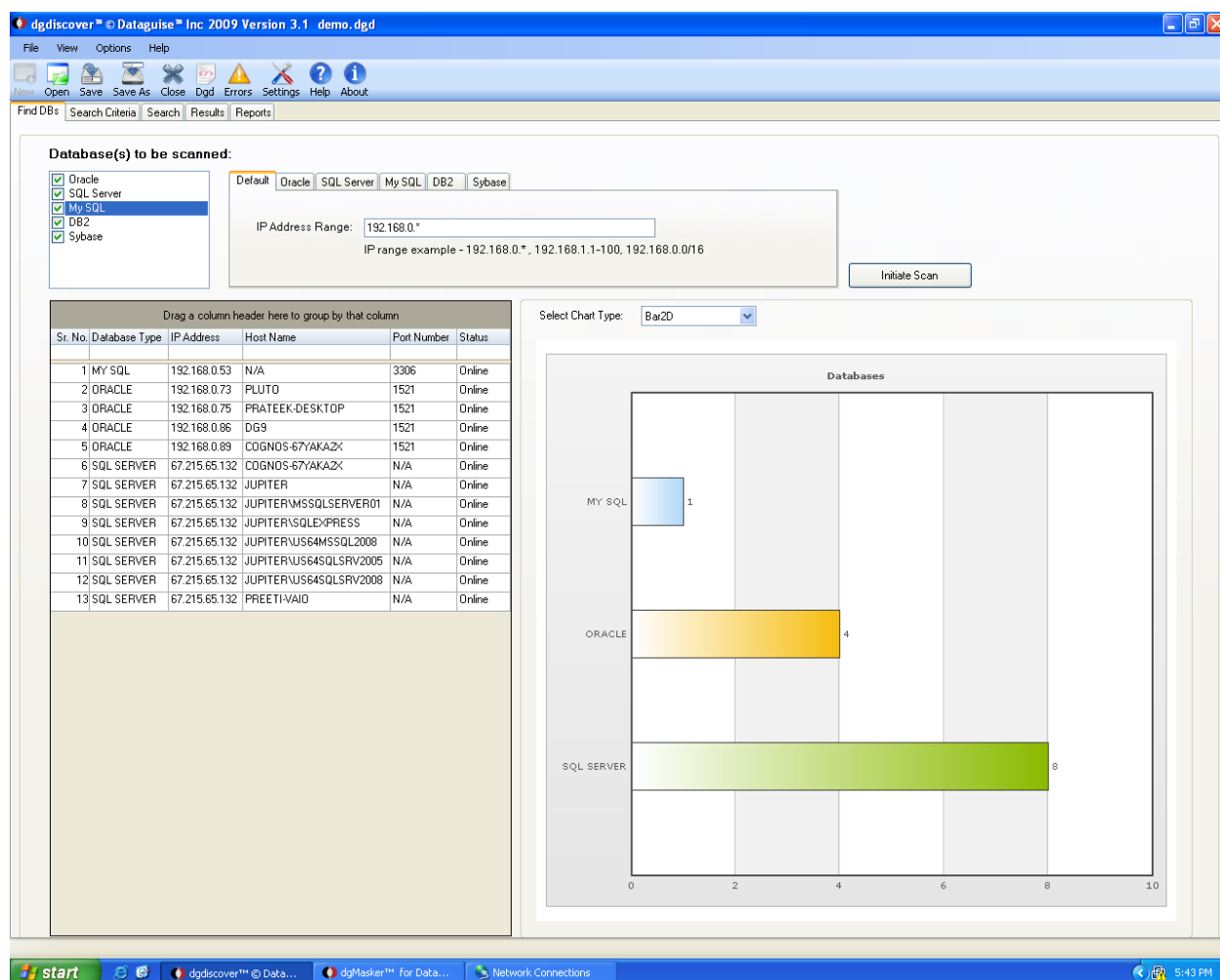


Figure 3: DgDiscover's "Find IT" functionality identifies new database instances deployed on the network.

Search IT: DgDiscover

DgDiscover's "Search IT" capability looks for sensitive and potentially sensitive data within structured and unstructured repositories. DgDiscover employs sophisticated pattern matching algorithms to automatically identify data such as credit card numbers, expiration dates, social security numbers, and phone numbers, and creates detailed reports showing where this information resides in database tables. In addition to databases, DgDiscover has the unique ability to apply the same rules to search unstructured sources such as text files, Word, Excel and PowerPoint documents, and other file formats. DgDiscover includes a library of predefined templates for conducting searches for data relevant to compliance initiatives such as PCI and HIPAA. It also provides a completely customizable search rule base which can be tailored to support specific use cases, and generates an easy-to-interpret graphical summary of search results.

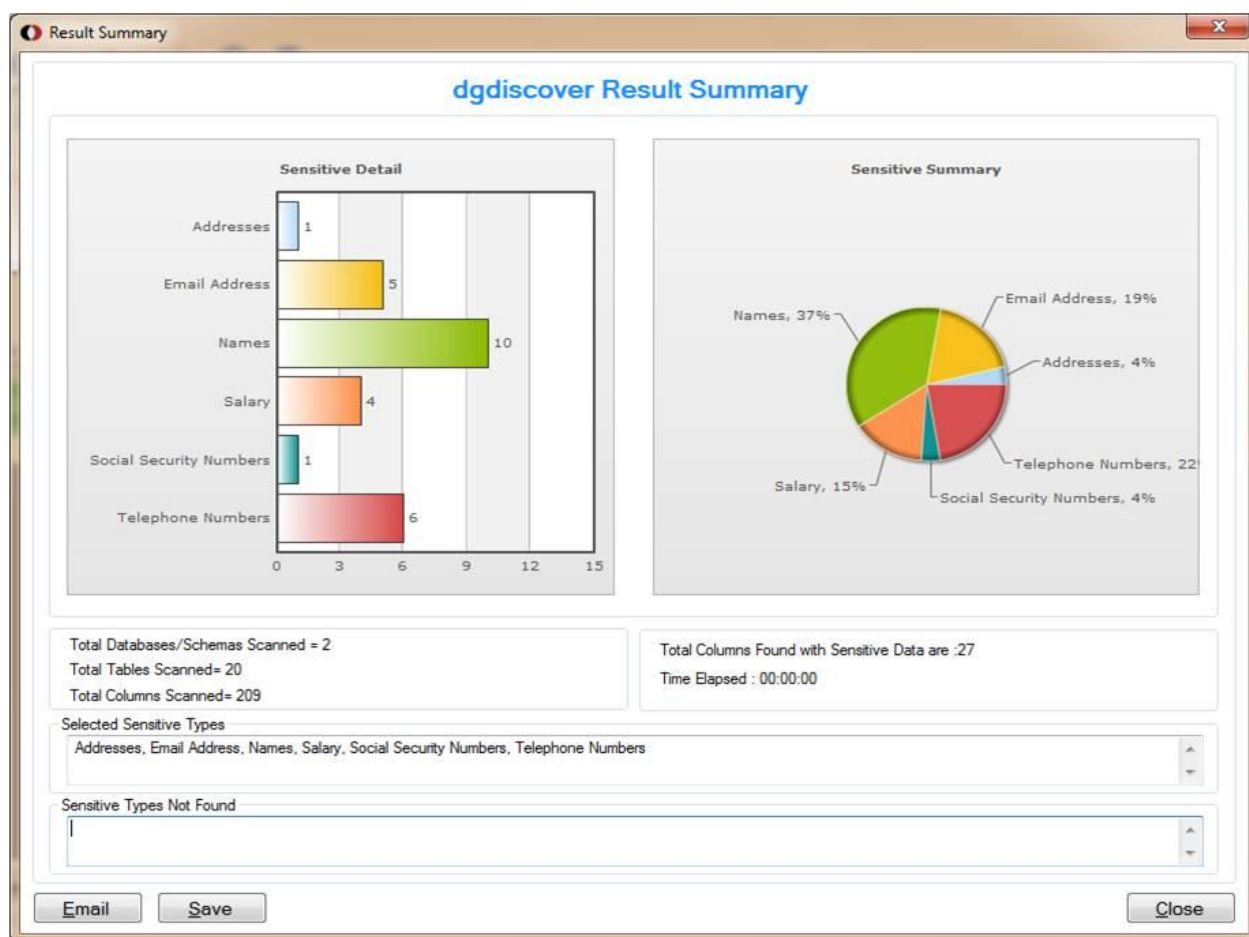


Figure 4: DgDiscover's "Search IT" functionality provides graphical reports of sensitive information.

Mask IT: DgMasker

DgMasker is an advanced data security solution which helps enterprises protect data and address compliance requirements by selectively masking, or de-identifying, sensitive data in

Steps to Discovery and Protection of Sensitive Data: Find IT. Search IT. Mask IT.

cloned databases. DgMasker automatically determines data relationships and ensures that generated masked data preserves data integrity and meets application business rules. Its unique, masking-in-place architecture gives high performance and can be run interactively at a client or in batch mode on a remote server to support automated software development processes.

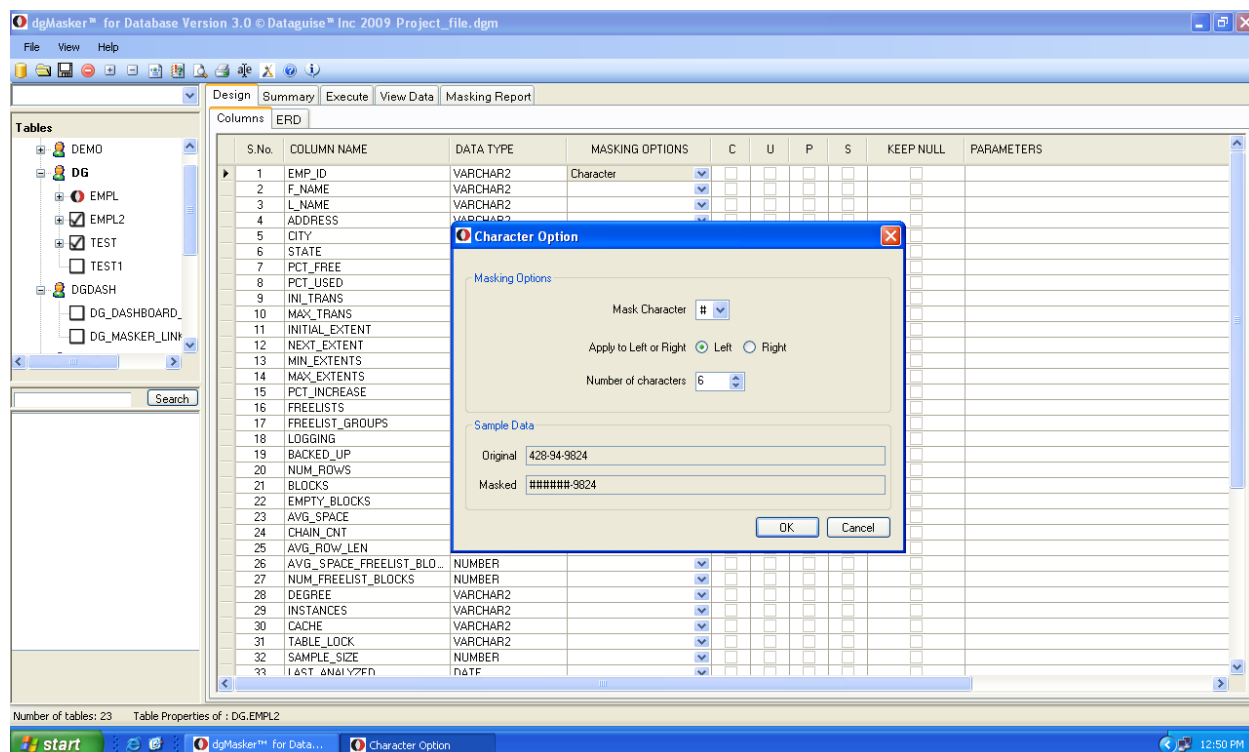


Figure 5: DgMasker's "Mask IT" functionality provides customized automatic data masking for sensitive information contained in databases.

Benefits of the Dataguisse solution for sensitive data protection

The Dataguisse solution delivers a number of unique benefits to enterprises seeking to protect sensitive data.

Unmatched time-to-value. The Dataguisse solution deploys quickly with no professional services required. Its easy-to-use, Windows-based user interface requires little training.

Actionable management reports and intelligence. DgDiscover produces graphical reports showing which repositories are deployed in the environment and where sensitive data reside. DgMasker's masking reports provide detailed, auditor-ready reports showing how data was protected in the environment.

High performance, enterprise-class architecture. DgDiscover supports and efficiently searches all of the on-line databases in the enterprise. DgMasker's unique masking-in-

Steps to Discovery and Protection of Sensitive Data: Find IT. Search IT. Mask IT.

place architecture delivers fast masking performance to support frequent cloning of large databases and fast deployment cycles through “Masking on Demand.™”

Complete solution for data protection. Only Dataguise delivers a complete solution for sensitive data protection based on its unique, “Find IT, Search IT, Mask IT” capabilities. Dataguise’s integrated tools help enterprises address some of their most pressing compliance and risk management concerns.

Conclusion

Protecting sensitive data in the enterprise has never been more important, or more challenging. An effective data protection strategy must start with finding all of the data repositories on the network and identifying all of the sensitive data contained in those repositories. For preventing exposure of sensitive data to non-production environments inside and outside the enterprise, data masking has emerged as the proven, best practice technique.

Dataguise provides a superior solution for performing enterprise-wide sensitive data identification and protection. Dataguise’s “Find IT, Search IT, Mask IT” capabilities find databases deployed on the network, search structured and unstructured repositories for sensitive data, and mask production data so it can be safely leveraged for development, test and analysis. This enables enterprises to better manage their risk and address their most pressing compliance requirements.

About Dataguise

Dataguise offers automated and advanced database security solutions to help ensure regulatory compliance and protect against data theft. Dataguise DgDiscover™ focuses on sensitive data discovery and classification across the enterprise and the company’s DgMasker provides secure masking of database content with unprecedented flexibility and functionality across heterogeneous environments. For more information, call 510-824-1036 or visit www.dataguise.com.

ⁱ Carey, Lisa. "The Black Market for Your Identity: What's Your Identity Worth". Associated Content. 10/7/2009 <http://www.associatedcontent.com/article/1726894/the_black_market_for_your_identity.html?cat=17>.

ⁱⁱ "datalosssdb". Open Security Foundation. 5/14/2010 <<http://datalosssdb.org/index/largest>>.