A Guided Approach to Data Masking

Abstract

Data Masking has assumed a lot more significance in the recent past than before. While Compliance needs has made the process of Data Obfuscation a necessity, the realization on the Implementation challenges for an Enterprise seems to be growing steady. In a relatively immature solution arena, there are a variety of players trying to establish themselves based on their respective areas of strength. This paper is a discussion on the variety of problems in identifying an enterprise wide solution and emphasizing on why point solutions suffer. Specific pointers are discussed to help choose the right set of solutions based on organization specific requirements. To conclude, we present an in house framework, focusing on keeping it simple and extensible, as a robust option for common considerations in this space.

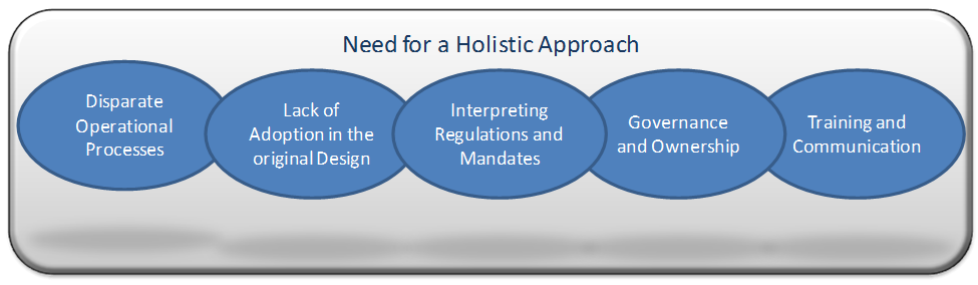
Disclaimer

The contents of this Document are not intended to constitute professional advice of any description. The information presented is for informational purposes only.

The Problem Space

The width in the Problem Space is often overlooked in an effort to implementing a pure technology oriented and point solution. An analysis of projects that have failed to take off even after enormous effort and cost presents us with an important set of parameters that make this discussion. Challenges have been broadly classified under Technology and Process.

Process Challenges



a. Need for a Holistic Approach

Data Masking, as an atomic problem might appear small and is often initiated as a Requirement from a specific line of Business or for a specific set of Applications due to sudden awareness on Compliance Mandates and Risk postures. Implementations become specific to a particular system or a set of systems and expose the need for a Holistic Approach given the commonality in the Problem space. The challenge lies in the diversity of an organization, lack of synchronized communication and single points of ownership for this common problem.

b. Interpreting the Regulations and Mandates

To best define a boundary and scope for any project, it is important to adopt a 'top down' approach in clearly articulating the Regulations and Mandates that apply to a particular application/ business context. The challenge with respect to Interpretation lies once again in the Diversity of the organization (when distributed across geographies present a variety of Regulations), the application of the Regulations specific to a business process or context (for example, in an outsourced scenario, how can the vendor demonstrate adherence to adopting the same laws that apply to the parent organization?) and the myriad of Regulations themselves.

c. Governance and Ownership

As emphasized in the first point in this section, Organizations have to treat Data Security as a common problem across Business Units and Systems. Most organizations do not have a clear Organization hierarchy to assume ownership for Data Obfuscation and track this through a single Dashboard.

d. Different needs for different processes

Regulatory mandates and Contractual obligations (with customers) tend to distinguish operations across Production and Non Production environments. For example, specific contracts in Germany mandate that Test data cannot be created through masking of Production data for any of the Non Production environments.

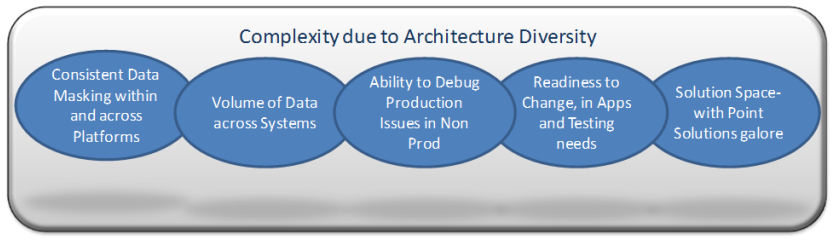
e. Adopting Data Security in Application Design

Data Masking is always a retro fitted solution. Applications, Data Models and Data stores have evolved with time with no intrinsic capability to address Data Masking across environments. Any attempt to adding a Data Masking capability within a given system is always a custom engineered plug in that does not guarantee robustness or present itself as a complete solution. It is a huge challenge to even think that systems built from here on have to take up Data Masking as one of the key use cases and design for the same as part of the Process.

f. Training and Communication

A combination of many other challenges leads to insufficient Training and Communication. Some of them include- absence of a Governing body for interpreting and sharing the implications of Regulations and Mandates specific to a given context, Data classification itself (once again in context of the operations performed by a given team) and the absence of a Data Security checkpoint in a Development lifecycle.

Technology Challenges



a. **Architecture complexity.** Myriad of Legacy, Mid range and distributed systems has resulted in an assortment of databases and applications under the same problem space. Given that systems design was never originally built with an intent of 'hiding' sensitive data from specific environments, the analysis of the dependencies on each of the sensitive elements takes us through application requirements, development and testing needs, target user community (that has access to this data), geography in context and compliance mandates to name a few. Schema complexities and license mandates that prevent data manipulation in the data layer present additional

Enterprise Risk & Security Solutions
An EBA Initiative

Cognizant

challenges as well.

b. **Consistent Data Masking within and across Systems-** the most common use case seen across a distributed environment is the need for preserving a masking logic across systems for consistency. Given the variety in the technology, it becomes a challenge to adopt a single solution that can provide for a consistent output across the systems.

c. **Volume of data involved-** it is mind boggling at times to even consider adopting an inter-mediate process called data masking within an existing data refresh cycle. While organizations have to start rethinking on whether they always need a complete refresh for all their development/ testing needs, it remains a fact that most processes today handle very large volumes and require a cost and cpu efficient solution to be plugged in as a data masking procedure in between.

d. **Debugging Production issues in Non Production environments-** Performance tests in Non Production environments require data volumes in closest resemblance to the Production environments. Besides, the ability to reproduce specific scenarios for debugging in Non Production environments will mean that the data distribution, integrity and uniqueness has to be preserved to the best extent possible with sanitized data.

e. **Preparation for Changes to the Applications and Testing needs-** Repeatability is not just an atomic concern. The ability to embrace Change, in adopting newer applications, testing needs, B2B integrations etc. is a constant challenge that has not had the prominence that it should.

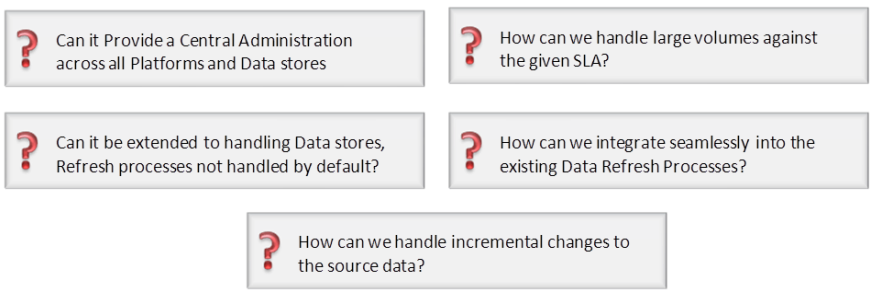
f. **Point Solutions galore-** Breaking down the masking problem to specific systems and elements actually does make it sound very simple. Development and Testing teams are often prompted to adopt point solutions within small boundaries that are eventually rendered ineffective due to the natural flow of data across platforms.

The Solution Arena

Majority of the Solutions in the market address the Data masking problem either at the Data Layer or by presenting themselves as an intermediate Proxy. Using this as a parameter for classifying these solutions, the below section has a few additional points on how these solutions fit into the scheme of things today.

Data Obfuscation Solutions

Products have often evolved based on one central system as their core focus area. It is tough to find a single Product vendor with a comprehensive Enterprise Implementation case study. Products remain point solutions at best, attending either to specific data stores (such as Oracle, SQL server or VSAM etc.) or specific application architectures (such as the Oracle HRMS or Peoplesoft). Common challenges with these solutions include



I. Inability to act as a central and consistent data obfuscation solution across all types of data stores. While some of the products do promise connectors and capabilities across data stores, they do not automatically translate to assurances with respect to algorithm consistencies across these platforms.

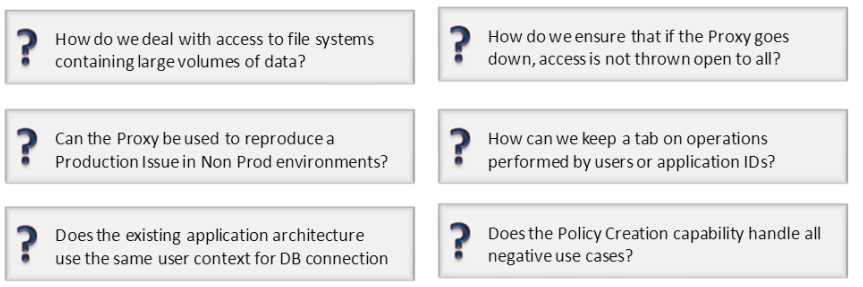
II. Extensibility- to the architecture and the algorithms. The absence of connectors and algorithms for specific requirements is definitely not uncommon. However, the ability of the solution to allow for easy extensions to newer data stores or algorithms that can be plugged into remains a concern.

III. Ability to handle large volumes- The Challenge comes from the Design itself. For example, some Products tend to operate on data being extracted to their own schema structures, which might not always be tuned to handle the largest of data volumes from stores such as a Warehouse. They might limit the need for a specific staging environment by acting as one, but their ability to address all kinds of data stores and volumes remains a concern. Solutions that do not perform their execution on native systems present two key challenges, namely - need for maintaining an additional infrastructure and the constant dependency on a vendor should a process fail.

iv. Integration with an existing Refresh Process- Products act as silos and most times remain the single point of execution for the masking process. However, the existing Data Refresh processes across organizations could be complex and inter dependent with other systems, demanding a repeatable and 'callable' solution for Masking.

v. Support for Incremental Masking- Daily and incremental changes to data require 'nimble' policy making capabilities that do not seem to be a common feature across most solutions today.

Proxy based Solutions



There are innovative solutions in the market that make use of the underlying protocol to establish them as a Proxy to all data access in the Database. Their greatest advantage comes from not having to deal with sanitizing large volumes of data or tampering with current refresh processes etc. Just as with Data Obfuscation solutions, they too cannot be considered as the final solution to all kinds of requirements. Some of the common challenges with such solutions are discussed below.

a. Proxy solutions are only for Databases. There are a host of file systems in Non production environments that require data level obfuscation solutions.

b. They are Runtime solutions- consequently they require constant monitoring and attention.

c. Their ability to help debug a production scenario in Non Production environment needs to be evaluated. For example, in a Data obfuscation scenario, the name "John" could have been consistently de-identified to "Alan" and to reproduce a Production situation, it might've been sufficient to troubleshoot with a pseudo ID in "Alan". However, with a Proxy solution in place, the ability of the solution to take in a pseudo value from the Application, and replace it with the real value is something that needs to be explored.

d. Monitoring and Audit capabilities- there is a constant threat to the underlying data since it remains in its original form. The need for auditing access to the lowest level of detail (including context of user, nature of operations, time of access etc.) assumes significance here. Ability to raise alarms in exceptional situations also becomes mandatory.

e. Changes to Application Architectures might be required- For example, in situations where Applications use a single user Id to authenticate and operate with a database (say using a connection pool), the Proxy solution wouldn't have any ways of distinguishing between users. It becomes necessary to change Application architectures to distinguish user context based on various parameters such as source of access, user designation, etc. Sometimes it might also be required to create new instances or environments based on the nature of operations.

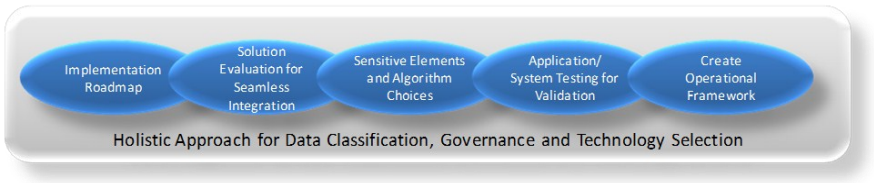
f. Ability to handle large volumes of requests needs to be evaluated.

g. Granular Policy Creation capabilities- With varied access and masking (at runtime) requirements across applications, operations, databases and end users, comprehensive policies have to be first established as a pre requisite to this Implementation. The ability of the solutions to address this variety needs to be explored.

To summarize, it would be fair to state that the variety in technology landscape and Operational requirements necessitate a high degree of flexibility and customization to any given solution in the marketplace today.

Approach to Implementing a Data Masking Solution

The above sections highlight the need for identifying the right set of solutions based on many different factors specific to a given organization. To approach the Solution space, a broad set of guidelines are discussed below.



1. Plan and Adopt a Holistic Approach

Most organizations have large enough Business Units that operate independent of each other, when it comes to Technology Evaluations and Vendor selections. However Data Masking remains a common denominator across Platforms, necessitating a consistent Approach and high degree of reuse for better Optimization and cost savings. After all, the sensitive elements within an organization are the same and the variance in the choice of Algorithms cannot be infinite. A holistic Approach will have to plan for-

a. Sensitive Data Classification, based on Policies, Guidelines and Mandates as applicable to a given Application/ Business.

b. Establishing a Data Security Org chart- with clear roles and responsibilities for identified Owners across Applications, Business units, Information Security teams, Testing teams, QA managers etc.

c. Delineate Production and Non Production problems. Production environments require access to real data at runtime which is significantly different from the Data Masking needs of a Non Production environment. This paper is a discussion on Data Masking purely for Test Data which are copies of sensitive data from Production environments.

2. Prioritize and Sequence the Implementation Roadmap

An inventory of Impacted Processes, Data Stores, Applications and the data flow will help drive an Implementation Roadmap. Focus should remain on identifying source systems and prioritizing Implementation based on optimal categorization, either across Applications or Data Stores. For example, if an Organization has most systems dependent on data from a Peoplesoft system, it makes for a suitable case to sanitize data for all of Peoplesoft Applications first before handling downstream systems as required. Alternatively, if a group of custom Applications operate on large volumes of Mainframe or Oracle data sets, it might be beneficial to categorize them based on user stores and prioritize an Implementation roadmap based on Applications that create data, use data and share data across Platforms.

3. Solution Evaluation with focus on Seamless integration

As discussed above, there are a variety of solutions in the market today. It is important to recognize the fact that a single solution might not be an option for all kinds of Problems. It is important that the chosen solution can best reuse People, Processes and Technology for easy integration. Simple criteria that can be used to establish a close knit mix of Solution options include

a. Evaluating the percentage of Requirements that can be handled 'out of the box' by a solution against the percentage that requires customization.

b. Evaluating Solution extensibility- for operating on proprietary data sets, choice of algorithms, integration requirements (for refresh processes) etc.

c. Evaluating Effort and Cost for Implementation- While cost is a clear differentiator, the effort required for Implementing a solution needs to be in sync with Organization priorities, Test cycles, Deadlines and Maintenance overhead.

d. Skills required for execution and maintenance- Ability to reuse existing Organization skills across Data Stores and File systems is critical to seamless integration, maintenance and support of a data masking system. Proprietary and Non transparent tools will also have to adopt changes and customizations which will become a management overhead from a cost and support point of view.

4. Optimize choice of Sensitive elements

Compliance Mandates and Laws of the land clearly identify elements that can be considered sensitive. However, most Implementations fail to recognize the fact that some of the elements in isolation are not really sensitive, but present a privacy issue only when observed in conjunction with other elements. This is a context specific issue. For example, first names, by themselves appear harmless, until they are viewed along with other business specific identifiers such as credit card numbers or account numbers. If the application's intent is not to operate on first names along with other unique identifiers together, it might be optimal to discard sanitizing names in this context. The lesser the number of policies, the easier it is to maintain. Caveat- Beware of data distributions and statistical analysis that can divulge sensitive data. For example, if a salary field is considered non sensitive in isolation (as it has only a bunch of numbers), someone could be prompted to identify the max (salary) and conclude the identity of a senior member in the Organization.

5. Remain conservative with Algorithm choices

This guideline is also a recommendation on adopting a Holistic Approach even for Algorithms. The choice of Masking algorithm should be first based on the Risk associated with the Information exposure and then take into consideration the Application and System dependencies. In some cases, it could be prudent to mandate specific algorithm choices even if the Test cycles or Applications themselves have to go through a minor change.

6. Automate Data Validation test procedures

The Implementation cycle can be made more efficient only by introducing automation for Data Validation processes. There will be a high degree of reuse in creating simple data validation scripts based on the rules and conditions that went into the Masking process.

7. Identify Impacted Use cases for Application and System Testing

Based on the choice of elements and the Algorithms implemented, a parallel effort to track the Impact on the specific Application use cases that need to be validated during System testing needs to be in place. The need for an Operational framework (with all these processes defined) is brought out in next guideline.

8. Creating an Operational framework

The existing process of creating data in Non Production environments has gone through a change due to the intervention of a Data Masking routine. Lessons learnt across modular Implementations should make way for an Operational framework that describes

a. The process of handling a new Application or a Data store

b. The process of identifying an Algorithm for a sensitive element, either based on the available suite of options or guidelines for creating one

c. Identifying Data Validation related Requirements

d. Evaluating Performance Needs, based on the current stats as observed

e. Capturing Integration touch points, for executing Data Masking as batch jobs or offline processes, along with an existing Refresh process

Summary

Data Masking demands the need for a consistent and holistic approach across Organizations today. The existing Design, Development and Testing processes have to go through a revision to embrace Data Masking. Process changes including the need for a Data Security Organization and the Interpretation of Data Security regulations in the context of a given System will go a long way in optimizing the choice of the right set of Solutions. With such a vast diversity amongst systems today, it becomes imperative that the chosen solution makes the best out of existing investments, promotes reuse and extends itself to specific situations in a seamless manner.

For more on how Cognizant solved similar problems across Enterprise wide Implementations, please refer to other whitepapers in this section or reach out to DataObscure@cognizant.com

Enterprise Risk & Security Solutions
An EBA Initiative

Cognizant

About Cognizant

Cognizant (NASDAQ: CTSH) is a leading provider of information technology, consulting, and business process outsourcing services, dedicated to helping the world's leading companies build stronger businesses. Headquartered in Teaneck, New Jersey (U.S.), Cognizant combines a passion for client satisfaction, technology innovation, deep industry and business process expertise, and a global, collaborative workforce that embodies the future of work. With over 50 delivery centers worldwide and approximately 150,400 employees as of September 30, 2012, Cognizant is a member of the NASDAQ-100, the S&P 500, the Forbes Global 2000, and the Fortune 500 and is ranked among the top performing and fastest growing companies in the world. Visit us online at www.cognizant.com or follow us on Twitter: Cognizant.

World Headquarters

500 Frank W. Burr Blvd.
Teaneck, NJ 07666 USA
Phone: +1 201 801 0233
Fax: +1 201 801 0243
Toll Free: +1 888 937 3277
Email: inquiry@cognizant.com

European Headquarters

1 Kingdom Street
Paddington Central
London W2 6BD
Phone: +44 207 297 7600
Fax: +44 207 121 0102
Email: infouk@cognizant.com

India Operations Headquarters

#5/535, Old Mahabalipuram Road
Okkiyam Pettai, Thoraipakkam
Chennai, 600 096 India
Phone: +91 (0) 44 4209 6000
Fax: +91 (0) 44 4209 6060
Email: inquiryindia@cognizant.com

Enterprise Risk & Security Solutions
An EBA Initiative