

InDetail



Grid-Tools data masking

An InDetail Paper by Bloor Research
Author : Philip Howard
Publish date : July 2011

Where Grid-Tools has an edge over pure data masking solutions is in its discovery and profiling capabilities

[Philip Howard](#)

Executive summary

All over the world regulators require organisations to comply with data protection and privacy laws. The terms of the relevant legislation may vary but the intent is the same: to protect information about individuals from those unauthorised to see that information. However, it is not just the fear of fines for non-compliance that is an issue for organisations but also the reputational damage and loss of customer trust that can occur when data is not adequately protected and when knowledge of that fact comes into the public domain (which it increasingly does). In addition, organisations may want to protect sensitive information for intellectual property reasons that go beyond the scope of relevant legislation.

In some situations encrypting the data can provide the necessary protection, but in operational, testing and development environments this is not a practical solution. Authorised users need to be able to see the data and the applications that process that data are not designed to run against encrypted data. It is therefore necessary to be able to hide what is known as technically as 'personally identifiable information' (PII) in some way, whilst maintaining the referential integrity of the original environment. There are various ways in which this can be done and the general term for this process is data masking, which is otherwise known as de-identification or obfuscation.

Tools for data masking are widely available. However, this does not mean that they are widely deployed. In a recent (2011) survey conducted by Bloor Research into data migration, one of the questions we asked was with respect to how users were coping with privacy issues that arose during data migration processes. Over 70% of respondents used manual methods for masking and just under 10% ignored the problem completely. That is, they broke the law. Less than 20% used a tool for data masking.

This paper argues that using manual methods for data masking will typically be inadequate, is time consuming and costly. While we will discuss data masking specifically with respect to the offering made by Grid-Tools, the arguments in favour of tools for this purpose go beyond that of any particular product, though our discussion of what is available from Grid-Tools will describe the sort of facilities that should be looked for in a data masking product.

Fast facts

Grid-Tools is a specialist provider of test data management software of which data masking forms a part. It has two different data masking products: Fast Data Masking and Simple Data Masking. The former is faster, provides extended capabilities and is available for the most popular data sources. The company is happy in principle to develop Fast capabilities for other sources, upon request.

Key findings

In the opinion of Bloor Research, the following represent the key facts of which prospective users should be aware:

- Grid-Tools offers a complete range of static masking options.
- In addition to masking, and as an alternative to it, Grid-Tools also offers synthetic test data creation as a part of its Test Data Management Suite. This is not discussed in detail in this paper, but details can be found in a previous paper on test data management published by Bloor Research.
- The ability to extend from masking to a full Test Data Management solution is a major advantage for Grid-Tools when compared to suppliers that only offer the former and not the latter.
- The company offers discovery and profiling capabilities that go beyond those of conventional data profiling tools with respect to data masking, though they may be more limited in support of data quality processes.
- Grid-Tools does not offer dynamic data masking as the term is usually understood. That is, a technique by which SQL is intercepted and data dynamically masked. In our view, this approach raises security issues that have not yet been addressed by data masking vendors. Grid-Tools does, however, offer dynamic capabilities through the use of masked views and shadow tables within its Fast Data Masking product.

Executive summary

- Unlike some other vendors (and analysts) Grid-Tools believes that performance is an important consideration for data masking. In particular, this solution supports agile development and extends agility to the data itself within a test data management environment.

The bottom line

There are lots of data masking vendors and support for different approaches to data masking is unlikely to prove a differentiator. Where Grid-Tools has an edge over pure data masking solutions is in its discovery and profiling capabilities that can significantly reduce the amount of time that needs to be spent on this part of the exercise. The company's emphasis on performance is also important.

Conversely, there are only a handful of companies that offer full test data management capabilities. If you are absolutely sure that you will never ever want to go beyond pure data masking then by all means ignore the fact that Grid-Tools' data masking offerings are part of a broader suite of products. However, if you cannot give that guarantee then this fact represents a significant advantage for Grid-Tools.

The product

Grid-Tools offers two data masking products: Fast Data Masking and Simple Data Masking where the main difference (there are others, which we will discuss) is that the former includes native drivers (for DB2 z/OS, IMS, VSAM/ISAM, Oracle, SQL Server and Teradata) while the latter is generic. Also, Fast Data Masking makes use of native database utilities where possible. As a result the former is around 80 times faster, typically masking over 100 million rows per hour. The Simple Data Masking solution applies not only to databases other than the ones listed but also to Excel, XML, CSV, TXT, EDI, SWIFT and fixed definition files as well as HIPAA 40-10, 50-10 and X12 formats. Simple Data Masking is also available for all formats that are supported by Fast Data Masking, except IMS.

Data Masking is actually part of Grid-Tools' broader Datamaker Test Data Management suite of products, as illustrated in Figure 1. The products run on Windows, Linux, UNIX, IBM i-Series and IBM System z. The current version number of the products is 2.6.

Pre-masking

Before you mask data you need to know or discover the sensitive data you need to mask. In theory you can do this manually. However, this is not only time consuming and expensive in terms of manpower but also very tedious. The end result is highly error-prone and liable to lead to fields that should be protected being missed. An additional consideration is that during the masking of the data it is necessary to retain the integrity of the data because otherwise participating applications will not run. For example, if the same data appears in multiple tables within a database then it should be masked in the same way in each case. This is especially true where primary/foreign keys need to be masked and, in general, relationships that exist across data elements may need to be retained through the masking process and you need to discover what these relationships are. For example, a patient has a disease, which has a treatment, which has a consulting physician who practices in a particular hospital and uses a designated operating theatre. If you scramble the data so that a patient with flu ends up having open heart surgery then your software may break down simply because your masking routines have not ensured that important relationships remain intact. So, discovery of these relationships



Figure 1: The Datamaker suite

may be essential. Further, some applications may access multiple data sources and the data needs to be masked consistently across those sources. All of this means that manual discovery of what needs to be masked is virtually impossible. Of course, there will always be some manual effort involved but you would like to automate as much of the process as possible.

An alternative approach is to profile the data using a stand-alone data profiling tool. Using such an approach you can automate the discovery of fields with a particular format such as xxxx-xxxx-xxxx-xxxx for credit card numbers and you can also discover relationships that exist between different data elements. However, such tools have primarily been designed to support data quality initiatives and therefore have a lot of extraneous capabilities that are not required for data masking purposes.

Grid-Tools provides its own discovery (profiling) capabilities. It ships with some 20 pre-built algorithms to discover social security numbers, mixed cases fields, data of birth fields, credit card numbers and so on, as well as allowing you to define your own such rules. Further, the product provides what might be termed "discovery by example". Here you can put in an example of actual sensitive data and then have the software check for other such examples. You can also re-check as development (if this is the scenario) progresses. A notable feature of Grid-Tools' discovery capabilities is the ability to pre-audit what is sensitive, using parameters, filters and the like so that you can narrow down the fields that you need to look at in order to discover what is and is not sensitive.

The product

Masking

There are multiple ways in which masking can be achieved. This will depend, at least in part, on why you are doing the masking. For example, you could simply hide a credit card number by replacing each digit with an x (xxxx-xxxx-xxxx), which will be fine if you are only concerned with data protection. However, if you want to test a payment application then you will need to work with real (pseudo-) numbers in order to test your applications. Similarly, you can redact data, which is where the data is completely obscured, a typical example being where data is censored. This technique applies usually to text and it is not suitable for testing purposes since the data cannot be read. Randomisation is another technique, whereby you simply replace a postal code (say) with random letters and numbers (as appropriate). However, this technique will not work if your application requires a valid zip code.

While Grid-Tools supports the preceding methods, for test data management you will need to mask in such a way that the data remains valid. Perhaps the most common technique used is shuffling. As an example of how this works suppose that you have Adam Jones, Richard Hills and John Smith in your database then shuffling might result in masked records for Adam Hills, Richard Smith and John Jones: in other words you shuffle all the first names and all the last names as two separate processes. Its advantages are that you ensure realistic data and the chances of reconstitution are very small provided the initial dataset is large enough (say, 5,000+ records). For smaller datasets you will need to use some other form of masking.

It should be borne in mind that masking is never perfect. In healthcare environments, to continue the example quoted in the previous section, a determined hacker may still be able to identify individuals, precisely because of the need to retain relationships. In addition, and as another example, your largest customer will still be your largest customer even if he, she or it is not immediately identifiable by name.

An alternative to masking that is offered by Grid-Tools (and, as far as we know, by nobody else) is to generate synthetic data. From an a priori perspective this is preferable to using masking because the dataset can be relatively small, assuming that it is representative, thereby keeping costs down and because there is no requirement for masking. Moreover,

there is no requirement to access production data, which means no impact on operational performance. However, in order to create representative synthetic data you do need to have a good understanding of the data relationships that are not only embedded with the database schema (or file system) but also those relationships that are implicit within the data but which are not formally detailed within the schema. You can use Grid-Tools' discovery capabilities for this purpose. You can also include errors within your synthetic data creation as you will wish to test the software in this respect. A further point is that the world does not stand still: trading patterns change over time and you may want to discover such trends that already exist within your data and project those forward to test against patterns of data that may be applicable in two or three years' time. This is clearly something that you cannot do by using only existing data but Grid-Tools supports this capability along with synthetic data creation in its Datamaker product, which is a superset of its data masking products. This is discussed in the Bloor Research paper on Grid-Tools Test Data Management.

Both Grid-Tools' data masking products come with multiple seed tables, with internationalised versions of these tables where appropriate (such as names). In both cases you can also add your own seed tables and there is multi-column capability so that, for example, state and zip code will match. Both products also include cross-reference management so that you can, say, retain the same transformations across runs or databases. Similarly, both products provide auditing, allow you to define your own functions and support flat file masking, though in the case of Simple Data Masking this only supports delimited files whereas Fast Data Masking supports all file types. Again, both products can update tables directly within the database but Fast Data Masking also offers the option of extracting data into a staging area, passing it through a masked view and then building shadow tables. Finally, both products support updating of primary keys, though in the case of Simple Data Masking these will need to be disabled whereas they can be rebuilt automatically using Fast Data Masking. Other features built into Fast Data Masking that are not in the Simple product include data discovery (as discussed previously), version control and difference management, common column discovery to ensure that the same mask is applied to matched columns, and the ability to incorporate sub-setting

The product

within the masking process. Note that the use of Simple Data Masking does not preclude any of these functions, simply that you would have to do them manually or through the use of an additional (possibly third party) tool.

It should be clear from this that Fast Data Masking is significantly richer than Simple Data Masking. One would therefore wish that the former was available for use with a much broader range of sources than is currently the case. Grid-Tools is very pragmatic about this: if there is appropriate demand for Fast Data Masking for a particular data source then they will build the relevant connector.

File-based masking is often required in conjunction with database masking. For example, you might have scrambled the social security numbers in the target database. However, your input file could now contain non-matching social security numbers and the load will fail. Using the cross reference table, or the hash routines employed by Grid-Tools as a part of the masking process, you can ensure that this mismatch doesn't happen.

In practice, in terms of the actual steps used in file-based masking you register the file definitions, identify (profile) the internal structure of the file and its relationships, import a sample file to make sure that the file definition and sample file match, define any sensitivity and data manipulation functions, and then run the scramble utility. As an example of the interface used for this process, Figure 2 illustrates how this works along with the general look and feel of the Datamaker product.

Finally, it is worth noting that there is an emerging distinction between static data masking, which is used in test and development environments and which we have largely been describing, and dynamic data masking. The latter is used in conjunction with operational data in real-time. As understood by most of the market this involves dynamically intercepting SQL requests to the database and masking

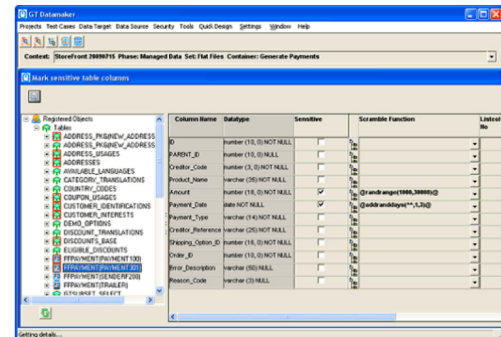


Figure 2: Marking sensitive columns

any sensitive data. This is complementary to, or a replacement of, the sort of role-based access controls that are in use in many environments to prevent, for example, HR employees seeing the salary levels of executives. In principle this approach seems like a good idea. However, it has potential security implications if you allow SQL to be intercepted in this way and you would need to have additional controls around the process such as might be provided by database activity monitoring. Unfortunately, at least at the time of writing, such capabilities are not available from any vendor. Having said all of this, the view-based capabilities and shadow tables provided by Grid-Tools in its Fast product can be deployed in a dynamic fashion.

Audit

Finally, it is important for data governance and compliance reasons that you can prove that you have appropriate processes in place to ensure the integrity of personally identifiable information. Grid-Tools provides workflow capabilities that allow you to define relevant procedures with, for example, checked, validated and approved stages to the identification of which data is to be masked. This lets relevant people look at the data profiling information and confirm that the data has correctly been identified as PII data or not as PII data. This provides a very rigorous audit trail of the due diligence taken by you in order to identify which data needs to be masked.

The vendor

Grid-Tools was founded in 2004 though, in a sense, its foundations go further back than that, since its founders had previously built up and then sold BitByBit to OuterBay (since acquired by HP), so the company has a depth of experience in this area. This also explains why the company is privately owned and self-financing.

The company's headquarters are in the UK and it also has offices in the United States and India. It has an extensive partner programme with trained and certified staff covering Australia, Austria, Belgium, Canada, France, Germany, Ireland, Israel, Italy, Latin America, New Zealand, Portugal, Scandinavia, Singapore, South Africa, South Korea, Spain, Switzerland and the Netherlands as well as the countries in which it maintains offices (where it also has partners). Partners tend to be consulting houses and systems integrators, both local and international. The latter category includes Cap Gemini, Cognizant, EDS, Infosys, Birlasoft, SQS and CSC, amongst others.

On the technical side Grid-Tools has partnerships with MySQL (Oracle), Oracle (which is also a customer), Compuware, Bridgehead Software, InterSystems, Silwood Technology, SAND Technology and various specialist testing vendors. The company also OEMs technology from Spotfire (TIBCO), Bender and Pervasive.

Web site: www.Grid-Tools.com

Summary

Grid-Tools is in an interesting position. It is the only pure play test data management vendor in the market and it is the only supplier in that space to offer synthetic data creation. This fact alone, in our opinion, means that it is a leader in the market for test data management. However, data masking is a subset of this space and there are a plethora of companies addressing this requirement without having adequate facilities for test data management more generally. No doubt many of the solutions will be relatively inexpensive, not to say cheap; Grid-Tools will be best positioned with organisations that require data masking now but can see the potential to expand into a full test data management solution in the future.

Further Information

Further information about this subject is available from <http://www.BloorResearch.com/update/2095>

Bloor Research overview

Bloor Research is one of Europe's leading IT research, analysis and consultancy organisations. We explain how to bring greater Agility to corporate IT systems through the effective governance, management and leverage of Information. We have built a reputation for 'telling the right story' with independent, intelligent, well-articulated communications content and publications on all aspects of the ICT industry. We believe the objective of telling the right story is to:

- Describe the technology in context to its business value and the other systems and processes it interacts with.
- Understand how new and innovative technologies fit in with existing ICT investments.
- Look at the whole market and explain all the solutions available and how they can be more effectively evaluated.
- Filter "noise" and make it easier to find the additional information or news that supports both investment and implementation.
- Ensure all our content is available through the most appropriate channel.

Founded in 1989, we have spent over two decades distributing research and analysis to IT user and vendor organisations throughout the world via online subscriptions, tailored research services, events and consultancy projects. We are committed to turning our knowledge into business value for you.

About the author

Philip Howard

Research Director - Data

Philip started in the computer industry way back in 1973 and has variously worked as a systems analyst, programmer and salesperson, as well as in marketing and product management, for a variety of companies including GEC Marconi, GPT, Philips Data Systems, Raytheon and NCR.



After a quarter of a century of not being his own boss Philip set up what is now P3ST (Wordsmiths) Ltd in 1992 and his first client was Bloor Research (then ButlerBloor), with Philip working for the company as an associate analyst. His relationship with Bloor Research has continued since that time and he is now Research Director. His practice area encompasses anything to do with data and content and he has five further analysts working with him in this area. While maintaining an overview of the whole space Philip himself specialises in databases, data management, data integration, data quality, data federation, master data management, data governance and data warehousing. He also has an interest in event stream/complex event processing.

In addition to the numerous reports Philip has written on behalf of Bloor Research, Philip also contributes regularly to www.IT-Director.com and www.IT-Analysis.com and was previously the editor of both "Application Development News" and "Operating System News" on behalf of Cambridge Market Intelligence (CMI). He has also contributed to various magazines and published a number of reports published by companies such as CMI and The Financial Times.

Away from work, Philip's primary leisure activities are canal boats, skiing, playing Bridge (at which he is a Life Master) and walking the dog.

Copyright & disclaimer

This document is copyright © 2011 Bloor Research. No part of this publication may be reproduced by any method whatsoever without the prior consent of Bloor Research.

Due to the nature of this material, numerous hardware and software products have been mentioned by name. In the majority, if not all, of the cases, these product names are claimed as trademarks by the companies that manufacture the products. It is not Bloor Research's intent to claim these names or trademarks as our own. Likewise, company logos, graphics or screen shots have been reproduced with the consent of the owner and are subject to that owner's copyright.

Whilst every care has been taken in the preparation of this document to ensure that the information is correct, the publishers cannot accept responsibility for any errors or omissions.



2nd Floor,
145-157 St John Street
LONDON,
EC1V 4PY, United Kingdom

Tel: +44 (0)207 043 9750
Fax: +44 (0)207 043 9748
Web: www.BloorResearch.com
email: info@BloorResearch.com