

InDetail



Grid-Tools Test Data Management

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

As far as we know, Grid-Tools is the only specialist vendor in this space. In our view, Datamaker is the most extensive and most complete test data management product that is available on the market today.

Philip Howard

Executive introduction: major issues in test data management

Test Data Management (TDM) is about the provisioning of data for non-production environments, especially for test purposes but also for development, training, quality assurance, the creation of demonstration or other dummy data, and so on. There are essentially three ways to derive such data: you can take a copy of your production database and, where appropriate, mask any sensitive data; or you can subset your data and mask it; or you can generate synthetic data, based on an understanding of the underlying data model, which means that no masking is required. Each of these approaches to TDM has both advantages and disadvantages.

Copying your database has the advantage of being relatively simple. However, it is expensive in terms of hardware, license and support costs to have multiple copies of the same database. It is not unknown for companies with large development shops to have upwards of twenty different copies of the same database for development and testing purposes.

Sub-setting your database is less expensive. However, it suffers from the same problems as all sampling processes in that you can miss outliers. This is particularly important in development environments because outliers may cause the system to break whereas normal results do not, so it is important that outliers are properly tested. Therefore, if you are using sub-setting then you need to ensure that outliers are captured and represented within the process of creating your subset. However, this pre-supposes that those outliers are present in the production database at the point at which the data is sub-setted. Since it is unlikely that all possible outliers are present at any one time this means that a copied or sub-setted database can never be fully representative of what you need to test.

A further issue that occurs with both full and partial database copies is that any sensitive data may need to be masked. Of course, there are many applications that address data that is not sensitive or subject to data privacy laws. On the other hand, there are also large numbers of applications where it is necessary to de-identify data to meet compliance requirements or to protect intellectual property. Where that is the case then data will need to be masked. This is not as simple as it may appear at first sight. In the first instance, you need to identify data that needs to be hidden: typically, this is done by looking for relevant patterns of

information such as credit card numbers ending xxxx. This can be done manually but it is an onerous process that is better automated.

Related to this point is the question of how the masking is to 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-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, simple shuffling techniques (for example, replacing zip code 12345 with 54321) will not work if your application requires a valid zip code. For test data management you will need to mask in such a way that the data remains valid.

Further, it may not be simply a question of identifying what data needs to be masked and then hiding it. This is because you need to ensure that data relationships remain intact during the masking process, otherwise application testing may break down. This will, of course, be dependent on the application but in complex environments it can be critical. 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 may be essential.

It should also be borne in mind that masking is never perfect. In healthcare environments, to continue the preceding example, 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.

The third alternative is to generate synthetic data. From an a priori perspective this is preferable to using either of the other two approaches 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

Executive introduction: major issues in test data management

data, which means no impact on operational performance. However, in order to generate 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. In other words you need some sort of discovery process; but then you need a comparable capability to do a good job of masking, for the reasons just discussed. You will also want to be able to include errors within your synthetic data generation, 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 using either copy or subset-based approach but which should be possible with synthetic data creation.

We should also note that some vendors claim to be able to generate synthetic data based on subsetting or copying the data and then repeatedly masking it. While this can be used to support load testing that is probably the limit of its value and we would not describe this as synthetic data generation in any real sense of that term. Otherwise, claims for synthetic data generation may be based on nothing more than having seed tables. If you are interested in synthetic data generation you will therefore need to be wary of different vendors calling different things by the same name.

It is also worth noting that it is a common misconception that tools such as HP's QuickTest Professional generates data: they do not.

Finally, another major issue in test data management and, indeed, testing in general, is that of coverage. What you would really like to achieve is testing of every possible code path with every possible combination of data with a minimum of tests. Unfortunately, that is very far from the experience of most testers. Taking a database copy, for example, often supports no more than 30% coverage and often much less. There are mechanisms (which we will discuss later) available to improve this percentage and reduce duplicated testing but this cannot be eliminated because of the very

nature of the production data, not least because production data is not representative of all possible data, as previously discussed. The same problem also applies to sub-setted data.

Conversely, the aim of synthetic data is to provide a truly representative dataset but without duplication. When combined with appropriate mechanisms it is possible to get as much as 100% functional coverage and 90% code coverage with an absolute minimum of tests. This is much more thorough than typical development environments (where 50% coverage is nearer the norm) and should result in the production of better code in less time and at less cost, because of the reduced number of tests that need to be run. Anecdotal evidence suggests that the use of synthetic data can reduce testing cycles by as much as one third.

Fast facts

Grid-Tools provides a complete suite of Test Data Management tools, which we will describe in detail in this report. These can be licensed en masse as the single product Datamaker or on a modular basis. As far as we know the company is the only vendor to support all of the methods just discussed for managing test data. Grid-Tools works with data in both flat files and databases and it can also be used to support SOA and user interface test environments.

Key findings

In the opinion of Bloor Research the following represent the key facts of which prospective users should be aware with respect to Grid-Tools Test Data Management:

- Grid-Tools supports database copying with masking, sub-setting combined with masking and the generation of synthetic data with full support for referential integrity.
- Test data, however derived, is stored in a test data warehouse. This allows you to manipulate, filter and, in the case of synthetic data, re-generate the data without having to access the production database. Re-generation, in particular, supports an agile approach to development and testing.

Executive introduction: major issues in test data management

- To support both masking and synthetic data creation, as well as archival, Grid-Tools provides data profiling capabilities in order to understand data relationships that exist within the data, regardless of whether these are explicit or implicit. The product can also link to third party tools that can expose relevant data models.
- The product has advanced coverage support.
- Synthetic data may include a defined percentage of errors for testing purposes.
- Two different data masking products are provided. One uses native database connectors for popular databases and generic connectors otherwise. As a result the former is very fast at masking but the latter not so. The former also has extended functionality. We expect (and hope) to see more data sources moving into the native category as time progresses.
- There is a module designed specifically to support SOA testing.
- A workflow engine is built into the product.
- There is a module available to identify trends within your data (how its profile is changing) that can be used in the generation of data to cater for future requirements.

The bottom line

Most vendors offering test data management either do so as an appendage to software suites that are primarily focused on application testing, or as an extension to archival and information lifecycle management products. As far as we know, Grid-Tools is the only specialist vendor in this space. It is not surprising, therefore, that it has capabilities that go beyond any of its competitors. Moreover, that is something of an under-statement. It has capabilities that go a long way beyond those of most of its competitors. Indeed, one very well-known company that you might think of as a major competitor to Grid-Tools (which it is in some instances) is also a partner, prepared to bring Grid-Tools into its customer base when it cannot meet the needs of those customers. That probably says as much as you need to know about Grid-Tools: if even one of its arch-rivals is prepared to concede that its offering is more extensive and complete than theirs then it truly deserves the tag of market leading.

Product availability



Figure 1: The range of capabilities provided by Datamaker

Datamaker is the company's flagship product and it includes a variety of different modules that cover the various technical areas illustrated in Figure 1. Note that although the diagram shows data quality as an aspect of Datamaker the product does not have data cleansing capabilities, though it does include data profiling and when you generate data you can deliberately include errors and duplicates for testing purposes. We will discuss these features in due course.

The individual modules are available separately if required. What will not be obvious from this diagram is that there is a module specifically for SOA environments that will simulate relevant services along with data creation. Also noteworthy is the fact that there are 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, Oracle, SQL Server and Teradata) while the latter is generic. On this topic, Data Archive supports DB2, MySQL, Oracle, SQL Server and Sybase while Datamaker as a whole also supports DB2 UDB, DB2 400, Ingres, PostgreSQL, Sybase ASE (IQ is planned), Informix and InterSystems Caché plus various flat file formats including Excel, VSAM/ISAM, CSV, TXT, SQL and fixed definition files. Synthetic data can be generated for HTML files but these cannot be sub-setted or masked. XML files may be created and masked but not sub-setted and this is also true for HIPAA (40-10, 50-10 and X12), EDI and SWIFT files.

The products run on Windows, Linux, UNIX, IBM i-Series and IBM System z. The current version number of the products is 2.6.

Product details

As noted previously, there are basically three approaches to test data: copy your database, subset it or generate synthetic data. Both of the first two will often require masking but in order to understand what needs masking you will need to profile your data, and you will similarly need to profile your data in order to support the creation of synthetic data, though for somewhat different reasons: in order to understand the data model underlying your environment. So, we will start with a discussion of Datamaker's profiling capabilities, followed by a consideration of its masking and sub-setting capabilities. If you are using a subset-based approach then you will normally subset first and then mask or do them both concurrently. However, if you simply want to work from a copy of your production database then you will only require masking. For this reason we discuss masking before sub-setting even though that may appear illogical. Finally, we will go on to discuss the various requirements of synthetic data generation as well as other aspects of Grid-Tools' offering.

Data profiling

Stand-alone data profiling tools are widely used for two purposes: to discover the quality of data and where there are errors, as a precursor to data cleansing; and to discover relationships that exist within a dataset regardless of whether these relationships are implicit or explicit. As may be imagined, some of these capabilities are specific to data quality environments while some, including all the relationship discovery features, are also relevant to test data management. For obvious reasons, Grid-Tools focuses on the latter and is not concerned with the former.

In practice, the data profiling features provided by Grid-Tools look and feel very much like those of data quality vendors. You can, for example, discover the 'shape' of the data. That is, identify maximum and minimum values, calculate standard deviations, and so on. While we will discuss other uses for data shapes later in this paper for the time being it is important to note that this is where data profiling tools usually stop. Conversely, with Datamaker you can compare columns in a database to see if they have the same or very similar shapes, which might suggest a relationship. Data quality tools are typically limited to looking for things like primary/foreign key relationships and, sometimes, the ability to identify potentially redundant columns (all of which Datamaker

can do also). Datamaker also includes filters that allow you to automatically recognise column types such as dates, credit card numbers, social security numbers, emails, addresses, columns that contain spaces, columns that contain special characters and so on.

Where synthetic data is going to be generated, Datamaker's profiling capabilities are designed to supplement any information about the structure of your data that you may have. This may depend on what facilities you already have in place. For example, if you use a data modeling tool such as ERWin, ER/Studio or PowerDesigner, then you can import relevant entity-relationship diagrams from those environments. Otherwise you will need to reverse engineer your existing database schema using the facilities provided by Datamaker. This explicit information can then be supplemented with any implicit details discovered during the profiling phase. In some instances, for example in SAP or Oracle (including Siebel, JD Edwards and PeopleSoft) environments, it may make sense to use specialist third-party tools that can generate relevant data models and then import these into the Datamaker environment.

Data masking

We have already noted that there are a number of masking techniques that are not suitable for use in a test data environment. Datamaker supports these for those who simply have a data privacy concern and want to simply license a data masking product and nothing else. However, that is not the focus of this report and we will not be discussing these techniques, which include blanking, encryption and simple shuffling, amongst others.

As we have noted there are two data masking products: Fast Data Masking and Simple Data Masking. In addition to the distinction between native and generic data drivers (which means that fast Data Masking is around 80 times faster, typically masking over 100 million rows per hour) there are also some functional differences. This also applies when masking flat files as opposed to database tables.

Both data masking products come with multiple seed tables, with internationalised versions of these tables where appropriate (such as names) and 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

Product details

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 data 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, which is a technique that is also known as dynamic data masking, 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 profiling (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 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 some 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 any 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.

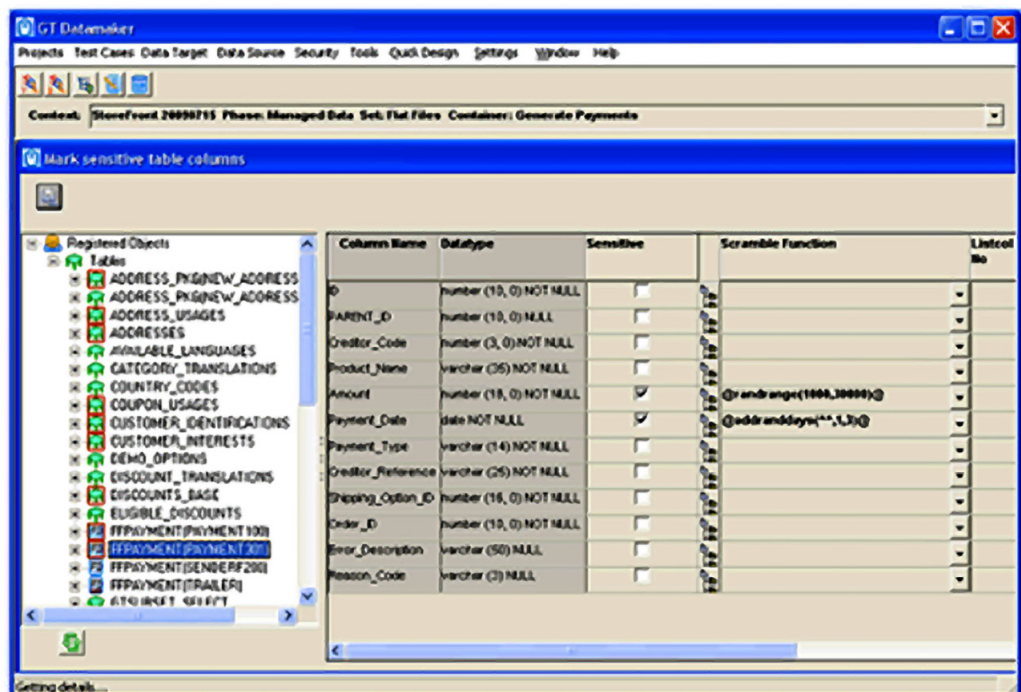


Figure 2: The Datamaker interface: marking sensitive table columns

Product details

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 mentioned, Grid-Tools' products can be deployed in a dynamic as well as a static fashion.

Sub-setting

There are several approaches provided for sub-setting but basically you define a driving table, a view or a temporary table (typically a subset of a driving table). Assuming a driving table, this will be extended along with any related tables defined by a primary-foreign key relationship (though these can also be selectively removed) and you can then add any others that might be linked in application terms, by defining a relevant join condition. You can view the data as you design your subsets by means of a graphical user interface.

Features include the ability to shrink a database in situ so that you can reorganise your tables into different tablespaces and compress your data; the ability to validate data quality during testing; to compare data before and after a transaction, regardless of whether

changes were expected or not; and date shift capabilities that will keep time-sensitive data up-to-date.

The actual process of sub-setting uses the native capabilities of the relevant data source such as Oracle Data Pump, DB2 import/export and so on.

Coverage

Once you have sub-setted and masked your data, thought should be given to coverage. The same will also apply if generating completely synthetic data.

Coverage in Datamaker is provided through the Data Design module. This is, in fact, BenderRTM (formerly CaliberRTM), which has been OEM'd by Grid-Tools, integrated into its environment, and extended, most notably by the use of Spotfire (TIBCO) technology for visualising coverage maps.

Data Design actually offers two test design engines: Quick Design, which is based on orthogonal or optimised pairs; and Cause-Effect Graphing, which is the more advanced of the two engines. Figure 3 shows the graphical design interface that you use in conjunction with the latter engine. As can be seen, you use

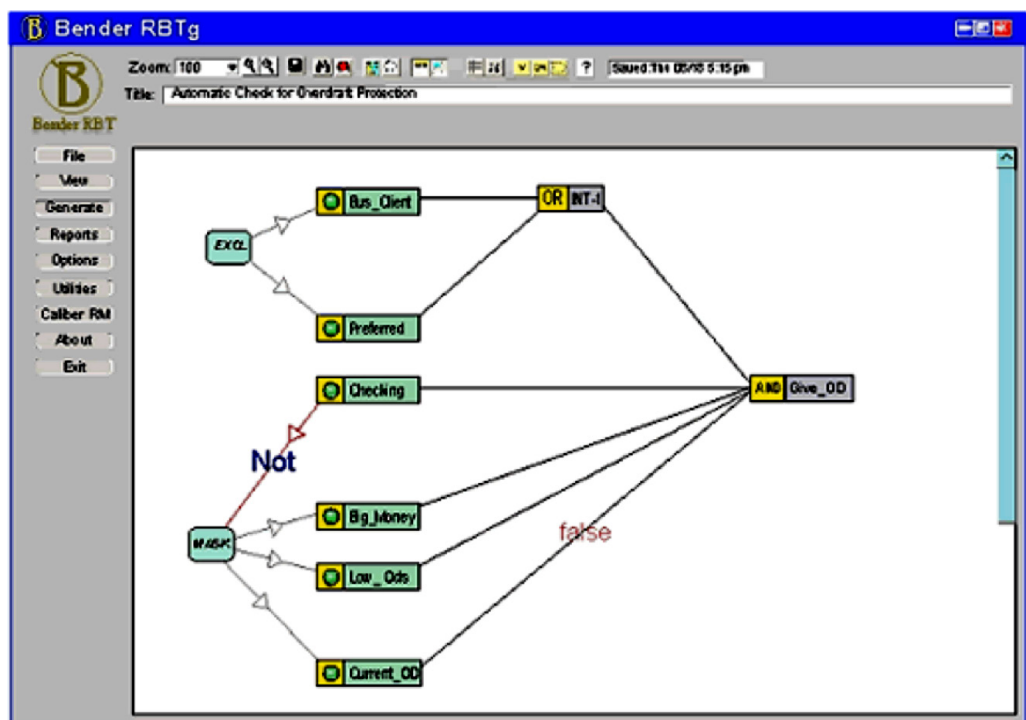


Figure 3: The Datamaker interface for Cause-Effect Graphing

Product details

this to define relationships, constraints and attributes and then the software will generate relevant test cases based on the design you have created.

Options include the ability to re-run old tests, to generate new test cases from the same diagram, and there are also facilities to link in any existing test cases that may have been created outside of the environment. Reports generated include a Functional Variation report (which will highlight any logical inconsistencies) and a Script Test Definition report as well as various matrices and the aforementioned coverage map.

Synthetic data

The basic principle behind the generation of synthetic data is that you discover your data structures, define the characteristics of the data you want to generate, generate the synthetic data and then store the results in what Grid-Tools calls a Test Data Warehouse, which is shown in Figure 4.

Once you have a data model that the generated data needs to match, as discussed previously, you now need to define the characteristics of the synthetic data. At a mundane level you can select the assignment of attributes: for example, what percentage of female first names versus males, which can then be generated from seed tables. Then you can define what

percentage of errors you want in the data and what sorts of errors. You can also, using the product's Data Shaper module, identify trends and patterns in the existing data that you may want to include, or expand upon, in the synthetic data. For instance, using Data Shaper you might have identified that the shape of your data is gradually changing (perhaps because your online business is growing rapidly and typical order sizes and values are different from conventional retail sales) and you might want to project that, going forward, for testing purposes. Data Shaper has a large number of parametric tests it can use for identifying such trends, including Weibull, Pareto, Chi-Squared, Gaussian Bell-curve and other tests. There is also support for time series analysis though, unlike the other tests, this is not currently automated (the company is working on this).

Once those definitions are complete you can assemble these objects (an object consisting of data together with relevant transformations and selection criteria) into 'scenarios', where a scenario might represent customers and their interactions, or products and their relationships, or other types of business entity. Scenarios may be arranged hierarchically so that one scenario can inherit characteristics from another. You manipulate scenarios via a graphical interface (portal). Also within this portal you can perform coverage analysis, as discussed previously.

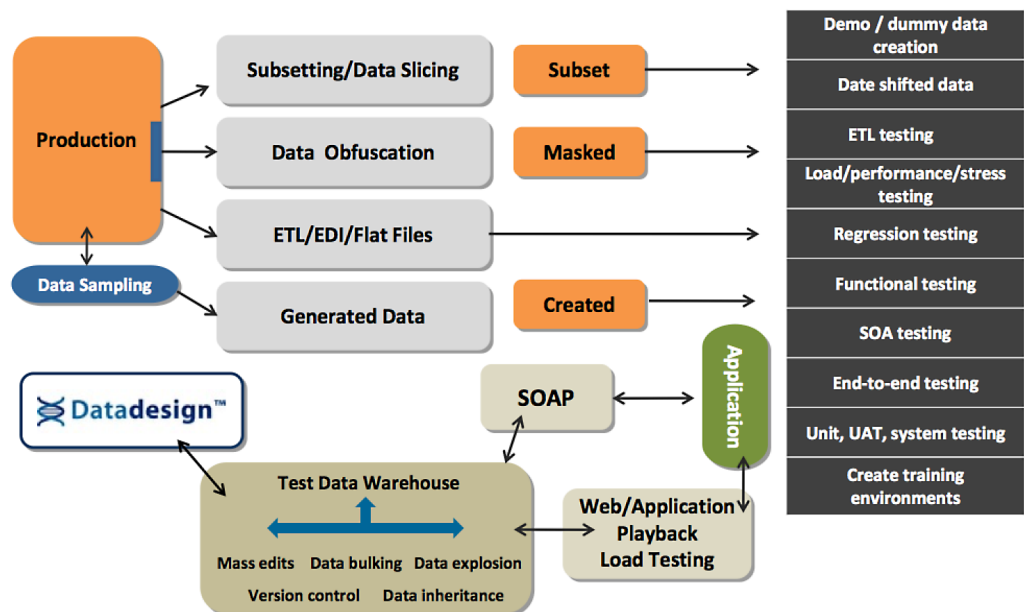


Figure 4: Datamaker architecture

Product details

Once you are happy with the preceding processes you select the relevant scenario and generate the data for it. Note that Datamaker has transformation capabilities (similar to those within an ETL tool). The results are stored in the Test Data Warehouse and may be re-generated on demand. This is a really important feature. During development, specifications change. This often means that you need to change the dataset that you are testing against, which can be an onerous process in conventional environments. With Datamaker you simply change the relevant attributes within your scenario and re-generate the data. In agile development environments, in particular, this means that the test data can be as agile as the development and application testing, which is not usually the case. Full version control is maintained by the Test Data Warehouse, so that you can go back to previous versions if necessary.

One final point that is worthy of mention is that Datamaker is not limited to generating database-specific data. It can also be used, for example, to populate the spreadsheets that are typically used with tools such as HP's QuickTest Professional and LoadRunner products.

SOA Data Pro

SOA Data Pro is used to provide the back-end data needed to service SOA requests so that testing becomes more automated and less manual. It includes a repository that allows you to store and freeze generated test data, in the former case so that this can be referenced by downstream requests and in the latter case to support the testing of multiple SOA requests. You can also store spreadsheets within the repository, thereby providing a central point of control. SOA Data Pro integrates with products like soapUI Pro from Eviware, SOA Test from ParaSoft, HP's Service Test and iTKO's LISA.

Data Archive

Data Archive is a stand-alone product that is not part of Datamaker but we include it here for the sake of completeness. For archival purposes you need to understand your data in much the same way as you need to for generating synthetic data and Data Archive reuses the capabilities already described in Datamaker for that purpose. Similarly, the product reuses Datamaker's sub-setting capabilities for extracting the data to be archived, along with join views that span the archived and live data. The software will generate delete scripts to support purging and, of course, there are masking options. The product integrates with various third party archival vendors including Bridgehead, Mobius (ASG) and SAND Technology.

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

In our view, Datamaker is the most extensive and most complete test data management product that is available on the market today. In particular, it can do things that other players in the market cannot. On the downside, Grid-Tools is a relatively small organisation in a market that is dominated by major vendors. In that sort of scenario the only way to stay above water is to be technically more advanced as well as generally being smarter and more agile. That is exactly what Grid-Tools is doing and there is every indication that it will continue to do so.

Further Information

Further information about this subject is available from
<http://www.BloorResearch.com/update/2080>

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