Test Data Management

MarketUpdate

Introduction

In our 2017 Market Update on Test Data Management (to which this report is an update), we described five basic ways of provisioning test data: taking a copy or snapshot of a production database, provisioning it manually or via a spreadsheet, generating a subset of a production database (or databases), deriving a virtual copy of your production database (again, or databases), or creating synthetic data. We also commented that copying an entire database, or provisioning test data manually, did not fall under the jurisdiction of test data management, since there is effectively no management involved, and hence enunciated three test data management methodologies: data subsetting, data virtualisation, and synthetic data generation.

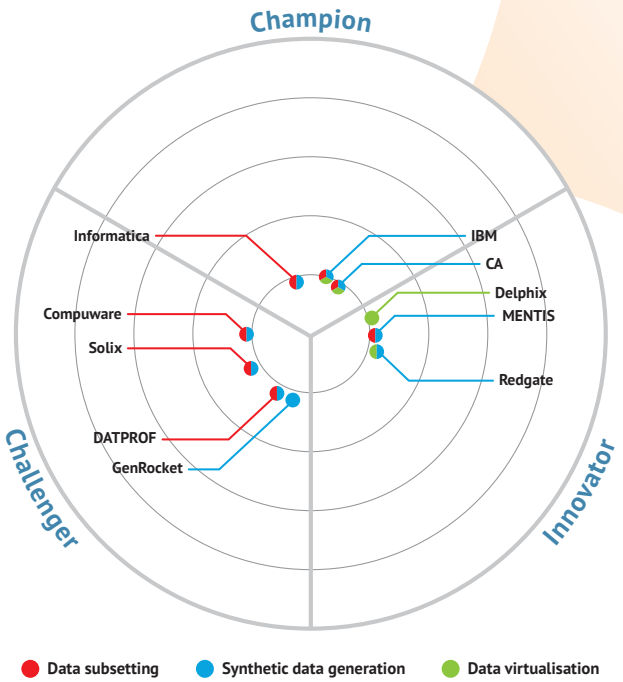
This is all still essentially true. However, that does not mean there have been no changes in the space. On the contrary, although the fundamental concepts have remained the same, the three test data management methods described above have seen significant shifts, either driven from within the space or from without. In this paper, we will seek to elucidate these shifts, as well as the trends within and overall state of the test data management space.

Market Basics

As discussed above, we recognise three distinct methods of test data management: data subsetting, data virtualisation, and synthetic data generation. Data subsetting consists of taking a subset from one or more production databases, usually of a much smaller size than the database(s) as a whole. This small size is a significant advantage, since it makes both test data distribution and testing much faster than a complete database clone. There are some challenges with this approach. For example, you must have a way of ensuring that your subset is representative of your entire dataset, and it must be referentially intact. However, data subsetting is a very mature subspace, and these problems have been solved by any solution worth talking about. It is by far the most stable element of test data management, and relatively little has changed since our last report.

Data virtualisation has a similar motivation to data subsetting, at its core: take large production databases and make them easy and efficient to distribute and test with. However, where data subsetting does this by reducing the amount of data being bandied around, data virtualisation does it by allowing you to create virtual copies of your databases. These virtual copies are (initially, at least) just referencing a master dataset, and are therefore incredibly lightweight and easy to move around. This makes it much easier to distribute your test data. This approach also offers a number of other advantages, not the least of which is that, with data virtualisation, you never have to worry whether your test dataset is representative, because it consists of your entire dataset.

Figure 1: The highest scoring companies are nearest the centre. The analyst then defines a benchmark score for a domain leading company from their overall ratings and all those above that are in the champions segment. Those that remain are placed in the Innovator or Challenger segments, depending on their innovation score. The exact position in each segment is calculated based on their combined innovation and overall score. It is important to note that colour coded products have been scored relative to other products with the same colour coding.



Key: products are colour coded to ensure that readers do not compare apples with pears. Note that all vendors in this Bullseye are consider best-of-breed.

It's also worth noting that data virtualisation as a test data management tool is significantly less mature than either data subsetting or synthetic data generation. In fact, in our last report, there was only one company in the space leveraging this technology. Although this is no longer true, it remains the rarest capability of the three methods described.

Finally, synthetic data generation breaks with data subsetting and data virtualisation by opting to disregard your production data for use as test data. Instead, it allows you to create your own *'synthetic'* test data in an automated fashion. The general idea is that this test data will look real – and will be in some sense representative of your production data – while, at the same time, being entirely fake. The biggest concern in regards to synthetic data is how to achieve this, and to make sure your test data covers a range of relevant test cases. A second concern is how to do this without making the process so laborious that it loses any benefit over manual creation of test data. Synthetic data is the area which has perhaps shown the most growth over the last couple of years. While at the time of our last report only a handful of vendors offered synthetic data as a capability, now almost every vendor we describe in this report self-reports as providing synthetic data in one form or another (although the actual capabilities vary). We discuss this further in the *'Market Trends'* section below.

It is also important to note the relevance of sensitive data to test data management. In the cases of data subsetting and data virtualisation, you will be distributing and exposing your production data to your testers. This cannot happen (for various reasons, primarily relating to data security and regulatory compliance) if any of that data is sensitive, which will be the case the lion's share of the time. This is, in fact, one of the chief advantages of synthetic data: since synthetic data is fake, none of it can be sensitive. In every other case, however, you will need to **a)** find and **b)** obscure any personal or otherwise sensitive information within your test data before supplying it to your testers. This is usually achieved via data discovery and static data masking capabilities, respectively. Consequently, discovery and masking functionality is offered by practically every vendor that is offering a subsetting or virtualisation solution. However, the efficacy of these capabilities can serve as a significant differentiator, particularly for data subsetting products. Dynamic data masking and/or encryption may also be offered as ancillary capabilities.

Market Trends

Interest in synthetic data has been high since our last report. This can most easily be seen by looking at the synthetic data capabilities that were offered in 2017, and that are on offer now. In 2017, only a small handful of vendors even offered synthetic data. Now, the opposite is true: all but one of the products covered in this paper provide some form of synthetic data generation. However, it is worth noting that several vendors recommend its use in a strictly ancillary capability, or as a last resort when other methods cannot be used (such as when production data is unavailable). Suffice it to say that these vendors are not giving synthetic data generation top billing. Consequently, we expect the market to continue along these lines, with a handful of vendors providing mature synthetic data generation to the organisations that want to leverage it as a primary solution, while the rest offer it as a purely complementary capability.

Subsetting capabilities, on other hand, are by and large capable, mature, and to a very significant degree, homogenous. We see little or no innovation in this area in terms of feature set, and we do not expect this to change. Masking (but pointedly not discovery) appears to be heading in the same direction, although it is not quite there yet. Consequently, we no longer see the specifics of data subsetting as a meaningful differentiator.

Sensitive data is of extreme importance right now, thanks to GDPR as well as upcoming regulations. This has produced a great deal of interest in the data security space, and this has had a knock-on effect on test data management and particularly data masking. As we suggested in our 2017 paper, data masking is increasingly seen as a data security capability, and will be sold as such. In fact, we would expect to find it increasingly common for test data management solutions to be sold as extensions of offerings in data security and related spaces (such as GDPR compliance). This is very much in the favour of those companies that provide both data security and test data management capabilities, for obvious reasons. We also believe that this increased concern over sensitive data has driven (and will continue to drive) interest in synthetic data. Afterall, one of the chief advantages of synthetic data is that it is fundamentally not real, and therefore cannot be sensitive.

Within the space, we also see an increased emphasis on test data provisioning, as opposed to merely test data management. This is likely due to continued interest in DevOps practices, as well

as Agile and continuous testing. The idea is to provide not only a way to create test data, but a method of distributing it effectively and efficiently to your testers, often by means of self-service. The advantage here is a significant improvement to the tester experience and to testing efficiency, thus (one hopes) preventing test data as a whole from becoming a bottleneck to your continuous testing, test automation, or DevOps pipelines. It's also worth noting that test data provisioning particularly benefits from a data virtualisation capability. This may be why so many more vendors are offering data virtualisation compared to a few years ago.

Vendors

In terms of the vendors themselves, there have been some notable developments since our last report. Most prominently, the number of products offering data virtualisation has increased dramatically. At the time of our previous report, Delphix was the only game in town when it came to data virtualisation. It has now been joined by three more vendors: Redgate, CA and IBM. The former has provided database cloning technology for SQLServer for some years, and has now combined its capabilities with offerings from Net2000 (whom it has recently acquired) to provide a data virtualisation solution for test data management. IBM and CA, on the other hand, have released their own data virtualisation solutions as part of their overall platforms (in IBM's case, in partnership with Actifio). Delphix, Redgate, CA and IBM all provide data masking in addition to virtualisation, and CA and IBM also provide data subsetting (among other things).

In our previous report we highlighted those few vendors that, at the time, offered synthetic data capabilities, and contrasted them against vendors which did not. As of this report, almost every vendor we have covered offers synthetic data generation. Of course, that does not mean every product provides equivalent capabilities (far from it) but the distinction is now one of degree rather than one of kind. Of note, GenRocket is the only product we have examined that focuses entirely on synthetic data generation, choosing to omit subsetting capabilities entirely. Delphix takes the opposite approach, being the only vendor in the report not to offer any synthetic data capabilities whatsoever. Every other vendor covered offers synthetic data along with either data virtualisation (Redgate), subsetting (CA, Compuware, DATPROF, Informatica, MENTIS, and Solix) or both (IBM). It is also worth bearing in mind that a number of these vendors – Compuware, DATPROF and Redgate – only recommend leveraging synthetic data generation as a secondary capability, either in support of data masking, or when production data is unavailable.

In terms of market development, apart from the aforementioned acquisition of Net2000 by Redgate, the most major event within the space has been the arrival of GDPR, as discussed in the previous section. In addition to that, SQLServer has recently added a native static data masking capability. This may prove to be a competitor to dedicated data masking products, but it is unlikely to affect much change for test data management: data masking alone is insufficient for its purposes, while at the same time ubiquitous within the space.

It's also worth discussing (briefly) the vendors we have not included in this report. Tricentis and Oracle provide test data management offerings, however they are both severely restricted in terms of the environments they support: Tricentis is only available as part of an entire test automation framework, while Oracle requires the use of an Oracle database. Although Redgate shares somewhat similar restrictions, relying on the use of SQLServer, it also provides data virtualisation, a notable capability within the space which justifies its inclusion. In addition, Actifio provides a test data management solution, however the company has resisted our attempts at contact. In any case, a significant part of its offering is resold by IBM, which we have covered.

Scoring

To score the various vendors/products discussed in this report we have used the following metrics:

- **Data masking –** the breadth and depth of data masking functionality provided by the product. We are primarily concerned with static data masking, but dynamic data masking may also contribute, particularly if the two can be combined in some way.
- **Data virtualisation –** the efficacy and range of features offered by the product's data virtualisation capability.
- **Ease of use –** how easy the product is to use. This mostly concerns the user interface, but can also cover the degree of automation offered and ease of deployment.
- **Sensitive data discovery –** the strength of the product's ability to discover sensitive data within your system, in support of data masking. This is usually achieved via data profiling and/or data classification.
- **Synthetic data generation –** the ability to generate synthetic or fabricated test data, particularly the ability to do so automatically, at scale, and/or in such a way that your synthetic data is representative of your production data.
- **Test data provisioning –** how easily and effectively the product can distribute test data. Products which have data virtualisation capabilities will inevitably score well here, but lacking data virtualisation does not preclude a product from having effective test data provisioning. Self-service, for example, goes a long way. Similarly, there is some overlap with ease of use, but the emphasis here is placed on the functionality rather than the interface.

Note that for the data masking, data virtualisation, sensitive data discovery and synthetic data generation categories, products only receive a score if they have the corresponding capability. Note also that data subsetting is conspicuously absent from our metrics: this is intentional. As we mention above, we view subsetting as a mature capability that no longer functions as a differentiator between products that offer it. Hence, we distinguish products with subsetting capabilities on the Bullseye diagram but do not use it as a scoring category.

We recognise that some aspects of these requirements will be more important for some users than others. So, while all of the scores for individual products are included in the detailed descriptions that follow later, the tables below represent the comparative scoring for each of the areas set out above. Note that each score is out of 5 but, unlike Amazon or Trip Advisor, it is impossible to score 5 on any topic. A score of 5 would represent a *"perfect"* product at this time. As we do not believe in perfection no product can be awarded a maximum score. In addition, when multiple products are awarded the same score in a given category, those products are displayed in alphabetical order.

The scores below are solely related to the products under evaluation. However, the positioning on the Bullseye diagram, as well as the *"mutable"* diagrams accompanying each vendor evaluation, also encompasses company issues such as support, geographic presence, stability and so on; as well as factors like innovation and the ability to support moves towards a data-driven enterprise.

Scores for test data management solutions

VENDOR	DATA MASKING
Mentis	4.5
CA	4.0
Compuware	4.0
IBM	4.0
Informatica	4.0
Redgate	4.0
Solix	4.0
Delphix	3.5
DatProf	3.5

VENDOR	DATA VIRTUALISATION
Delphix	4.5
CA	4.0
IBM	4.0
Redgate	4.0

VENDOR	EASE OF USE
DatProf	4.5
CA	4.0
Delphix	4.0
Compuware	3.5
Genrocket	3.5
Informatica	3.5
Mentis	3.5
Redgate	3.5
Solix	3.5
IBM	3.5

VENDOR	SENSITIVE DATA DISCOVERY
Mentis	4.5
Informatica	4.5
CA	4.0
Compuware	4.0
Delphix	4.0
IBM	4.0
Solix	4.0
DatProf	3.5
Redgate	3.5

VENDOR	SYNTHETIC DATA GENERATION
CA	4.5
IBM	4.5
Genrocket	4.0
Informatica	4.0
Solix	3.5
Mentis	3.5
Compuware	3.0
DatProf	3.0
Redgate	3.0

VENDOR	TEST DATA PROVISIONING
Delphix	4.5
Redgate	4.5
CA	4.5
DatProf	4.0
IBM	4.0
Informatica	4.0
Genrocket	3.5
Compuware	3.0
Mentis	3.0
Solix	3.0

Conclusion

As with the test data management space itself, most of the solutions described in this report are mature. All of them are competent, and each has their area (or areas) of expertise. Which solution is best suited to your organisation will depend on which of those areas you care to take advantage of, as well as which of the three major test data management methodologies you intend to use.

Bloor

Bloor Research International Ltd
20–22 Wenlock Road
LONDON N1 7GU
United Kingdom

Tel: **+44 (0)20 7043 9750**
Web: **www.Bloor.eu**
email: **info@Bloor.eu**

About the authors

DANIEL HOWARD
Senior Researcher

Daniel started in the IT industry relatively recently, in only 2014. Following the completion of his Masters in Mathematics at the University of Bath, he started working as a developer and tester at IPL (now part of Civica Group). His work there included all manner of software and web development and testing, usually in an Agile environment and usually to a high standard, including a stint working at an 'innovation lab' at Nationwide.

In the summer of 2016, Daniel's father, Philip Howard, approached him with a piece of work that he thought would be enriched by the development and testing experience that Daniel could bring to the table. Shortly afterward, Daniel left IPL to work for Bloor Research as a researcher and the rest (so far, at least) is history.

Daniel primarily (although by no means exclusively) works alongside his father, providing technical expertise, insight and the 'on-the-ground' perspective of a (former) developer, in the form of both verbal explanation and written articles. His area of research is principally DevOps, where his previous experience can be put to the most use, but he is increasingly branching into related areas.

Outside of work, Daniel enjoys latin and ballroom dancing, skiing, cooking and playing the guitar.

Bloor overview

Technology is enabling rapid business evolution. The opportunities are immense but if you do not adapt then you will not survive. So in the age of Mutable business Evolution is Essential to your success.

We'll show you the future and help you deliver it.

Bloor brings fresh technological thinking to help you navigate complex business situations, converting challenges into new opportunities for real growth, profitability and impact.

We provide actionable strategic insight through our innovative independent technology research, advisory and consulting services. We assist companies throughout their transformation journeys to stay relevant, bringing fresh thinking to complex business situations and turning challenges into new opportunities for real growth and profitability.

For over 25 years, Bloor has assisted companies to intelligently evolve: by embracing technology to adjust their strategies and achieve the best possible outcomes. At Bloor, we will help you challenge assumptions to consistently improve and succeed.

Copyright and disclaimer

Bloor Research International Ltd
20–22 Wenlock Road
LONDON N1 7GU
United Kingdom

Tel: **+44 (0)20 7043 9750**
Web: **www.Bloor.eu**
Email: **info@Bloor.eu**