# Test Data Management

MarketUpdate

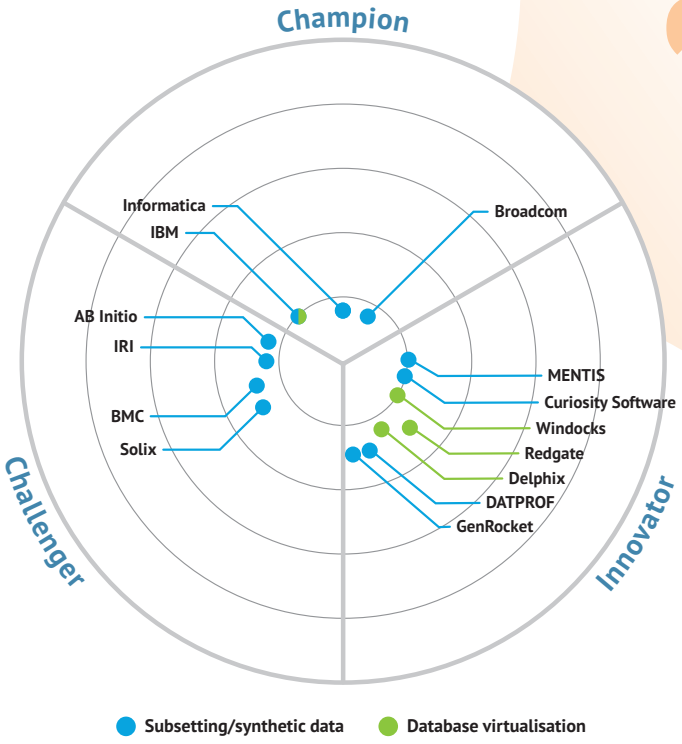
## Introduction

Testing is an essential part of development, and to test effectively, you will need test data. What's more, you will (among other things) want that test data to be anonymised, to avoid unnecessarily exposing sensitive data; for it to be representative of your production data, so that your testing is meaningful; and for it to be easily accessible by (or even deliverable to) your testing teams, in order to prevent bottlenecks in the testing process. Accomplishing all of this is the domain of test data management.

While there a handful of naïve approaches you could take to test data, such as leveraging whole, raw copies of your production databases, these are generally a bad idea. There are several potential reasons for this, the greatest of which is simply scale: most production databases contain far more data than is practical to expediently, and repeatedly, distribute and test. Accordingly, the test data management space is concerned with more efficient alternatives, namely data subsetting, synthetic data generation, and database virtualisation, often alongside data masking to provide the aforementioned anonymisation. These methods do not necessarily stand alone – there is certainly an argument to be made for using more than one – and we explore each of them in more detail in the following section.

The vendors within the space are a similarly eclectic bunch. They run the gamut from small, focused pure-plays to expansive, seemingly all-encompassing platforms, and as you might expect take quite radically different approaches to the space. Some provide two or even all three of the above methodologies without preference, while others are staunch proponents of a particular one. Some approach the space from the standpoint of development and DevOps, while others are more focused around data privacy, security, and/or governance. It is, to put it shortly, a large and varied space, of equal parts bold innovation and time-tested maturity. We, and this report, are here to guide you through it.

**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.




**Note:**
We have colour-coded vendors according to their capabilities in order to ensure we compare apples with apples, and we highlight database virtualisation in particular because, with the sole exception of IBM, it partitions the space into two discrete subsets. By contrast, it is normal for vendors that don't provide database virtualisation to offer both subsetting and synthetic data capabilities, albeit to differing degrees. This is not in any way a value judgement.

# Methods for managing test data



## Data subsetting

Data subsetting consists of taking a subset from one or more of your production databases, usually of a much smaller size than the database(s) as a whole. This small size enables much more efficient distribution and testing than a complete database clone, and has been the standard tool for managing test data for much of the space's history. Accordingly, it is by far the most mature method available for test data management.

That said, it does pose some challenges, most notably in how you take your subset: taking a random sample will rarely result in a useful test data set. For example, you will want your subset to be representative of your data as a whole, to ensure that all important scenarios are tested. This means that it will need to contain all conceptually meaningful test data points and combinations that are present in your data. You will therefore need a way to analyse your data, determine what these points and combinations are, and extract data that includes them. You will also want your subset to carry forward any relationships present (between tables, for instance) and hence be referentially intact. That said, as you would expect from such a mature sub-area, these problems – particularly the latter – have been solved by any solution worth talking about.

## Database virtualisation

Database virtualisation (sometimes referred to as simply data virtualisation; we prefer the former, due to the fact that the latter, as a term, has become severely overloaded) has a similar motivation to data subsetting: take large production databases and make them easy and efficient to distribute and test with. However, where data subsetting does this by simply reducing the amount of data being bandied around, database virtualisation takes the original data and virtualises it, creating fully-fledged virtual copies of your databases. These virtual copies reference a master dataset, or are a delta store, or have some other means of being incredibly lightweight and easy to move around. This makes distribution much faster and easier, and you never need to worry about representation.

It's worth noting that database virtualisation is significantly less mature than either data subsetting or synthetic data generation. It can be difficult to implement, it sometimes struggles with limited compatibility, and it has an inherent inability to mix real and virtualised data. In addition, using entire production data sets in your tests can be unwieldy, and potentially result in overtesting, even if database virtualisation makes them much easier to provision. That said, at least some of the vendors that offer it are aware of these issues, and are either working to address them or have already done so. In the latter case, they are certainly ahead of their competition. In short, it is clear there is significant, and growing, interest in database virtualisation, and that it is currently the most volatile – and innovative – part of the space.

## Data discovery and masking

Although neither data discovery nor data masking are test data management methods in and of themselves, they are still vitally important to the space. Without them, both data subsetting and database virtualisation leave your sensitive data unprotected and exposed during the testing process. This is dangerous, unnecessary, and almost certainly noncompliant.

Therefore, unless you intend to leverage synthetic data (see below) exclusively, you will want to use these (sensitive) data discovery and (static) data masking to a) find and b) anonymise any personal or otherwise sensitive information within your test data before supplying it to your testers. Other techniques (such as obfuscation, encryption and dynamic masking) are sometimes available, although usually for ancillary purposes.

Data subsetting and database virtualisation vendors usually offer discovery and masking functionality as well, in order to allow their solutions to function without relying on third-party products. At the same time, although masking is a fairly mature capability in and of itself, it is often not the primary focus with the test data management space. This means that the efficacy of discovery and masking can vary substantially from vendor to vendor. A particularly robust masking (or discovery) solution can therefore serve as quite the differentiator.

## Synthetic data generation

Synthetic data generation breaks with data subsetting and database virtualisation, in that instead of allowing you to leverage your production data for testing, it allows you to create your own *'synthetic'* test data in an automated fashion, ideally based on your production data. The idea is that synthetic data looks real – but isn't.

Synthetic data has several advantages, notably including complete control of your test data set (for example, if you want to add in a particularly specific use case that has yet to come up in production), better support for greenfield environments where production data isn't present in a significant quantity, and perhaps most of all, removing the need for masking alongside any possibility of deidentification.

Its most notable difficulties are representation and onboarding. These issues are linked. To wit, the more sophisticated synthetic data solutions will analyse your production data, bring out the trends and patterns therein, and thence create synthetic data that contains those same patterns (that *"maintains statistical integrity"*, as one vendor put it to us). On the other hand, vendors that lack this capability will usually leave it to you to specify the particulars of how to generate your synthetic data set. This can be a laborious process – hence the difficulty of onboarding – and -leaves representation entirely up to the user.

Neither of these are necessarily problems if synthetic data is present in a secondary capacity. This is common within subsetting, masking, and virtualisation solutions, where the idea is usually that you can use test data generation to fill in gaps in your production data, or to generate convincing replacement data as part of masking. This is a useful capability, but for our money it is not a synthetic data solution if it could not reasonably be used standalone.

# Market Trends

Data subsetting, data masking, and synthetic data generation are (and remain) widely supported within the space, although the quality and extensiveness of the latter is far more variable than either of the former. Database virtualisation, on the other hand, appears to be taking off. Although it has been present in the space for years, it was previously confined to a small handful of vendors, usually with quite specialised demographics. Despite the fact that only one new database virtualisation vendor has entered the space, and the space as a whole does not (yet) seem keen on building database virtualisation into their solutions directly, several new partnerships have sprung up between vendors that offer database virtualisation and vendors that don't. This has largely been spurred on by Windocks, the aforementioned newcomer, which is actively – and fervently – seeking out new partners. They are not the only ones, however, and Delphix have been busy on the same front (although to a much lesser extent). Regardless, this means that there are far more options to choose from if you want to leverage database virtualisation alongside more traditional test data management methods. Moreover, we are frequently seeing smaller vendors, usually with point solutions, partnering together to jointly offer a more complete solution. Although the level of integration provided can differ, in the most robust cases these combined solutions may even be able to rival the big boys in the space.

Data discovery and masking are also particularly important at the moment, due to continued interested in regulatory compliance, most notably in terms of GDPR but also more recent, and assuredly forthcoming, legislation around the world. California, New Zealand, and Brazil, for instance, have all released their own data privacy acts in the wake left by GDPR. Within test data management itself, little has changed on this front – to no-one's surprise, you still don't want to test with sensitive data – but it has lent discovery and masking a significantly greater applicability. It is now not uncommon to leverage what might once have been considered test data management technologies to protect your production data. Data masking in particular is increasingly seen as a data privacy first and foremost. This certainly benefits vendors that approach test data management from this angle. The recent popularity of the cloud, and corresponding demand for cloud migrations, has also played into this. Moving from a trusted to an untrusted environment (which is to say, from your in-house server to the cloud) demands a certain level of data security, and data masking is regularly used for this. This growth in cloud migrations, and in digital transformations more generally, can at least partially be attributed to necessity following COVID-19 (although frankly, the less said about that, the better).

Following compliance, we also see automation and DevOps as significant drivers for the space. Test data vendors are keen to automate the creation and distribution of test data sets to the point that they can be ready wherever and whenever they're needed within your development pipeline, making test data bottlenecks a thing of the past and maybe even allowing your testing to keep pace you're your development. Some solutions do this better than others, but the results are, generally, positive. A few vendors have even taken to referring to what they do as test data automation, rather than just test data management. Semantics, of course, but not necessarily unwarranted. In terms of what gets automated, the creation of data sets is not so much the issue as is the provisioning of those sets: automating the former is de rigueur; the latter much less so. For test data provisioning, then, self-service and collaboration are standard. No-one seems to be relying on the old request/receive model, and for good reason. More advanced solutions might bake test data into your other processes (usually CI/CD) or even your test scripts themselves, automatically deliver up-to-date test data to your testing teams, integrate test data management processes and utilities directly into your development pipelines, or generate and allocate test data entirely on the fly. Needless to say, expediating these processes can only be a good thing.

# Vendors

There are a number of different ways to look at the vendor makeup of the test data management space. Perhaps the most obvious is in terms of the methods mentioned above, which vendors support each, and to what extent. This last point muddies the waters significantly: it is not unusual for vendors to support one or two methods extensively, then the remaining one or two at a much more basic level of functionality. Therefore, rather than simply running down a list of nominal capabilities offered by each vendor, we prefer to highlight the methods each vendor emphasises. For instance, for database virtualisation it is relatively clear cut: Delphix, Redgate, Windocks, and IBM all make a point of offering database virtualisation as a primary method for test data management (although in IBM's case it is a white-labelled solution developed by Actifio); the other vendors in this report do not. For synthetic data, on the other hand, the line between primary and secondary use is much fuzzier. That said, we think it is fair to say that it is a particular focus for Curiosity Software, GenRocket, and Broadcom. Other vendors tend to focus on data subsetting – offering synthetic data as an ancillary capability – or split the difference and offer the two without emphasis. BMC, DATPROF, Informatica, IRI, Solix, Ab Initio and MENTIS all fall into this camp. Note that the broader vendor offerings, in particular, tend to blur the lines between these categories. For example, Broadcom and Informatica both offer highly capable data subsetting and synthetic data generation, while IBM provides subsetting, virtualisation, and synthetic data without particularly emphasising any of them. Note also that some form of data masking is more or less a constant within the space. This makes sense, given that data subsetting and database virtualisation require it, and synthetic data generation can be closely connected to it.

Size and scope of offering can also be a useful means of distinguishing between products. Solutions from the larger vendors in the space – Informatica, Broadcom, and IBM in particular – are inevitably broad, expensive, and just one part of a suite of data offerings. On the other hand, companies like Curiosity Software, Windocks, GenRocket, and DATPROF are much closer to pure-play test data vendors, and accordingly offer relatively narrow but highly fit-for-purpose products. The remaining vendors are somewhere in the middle. Moreover, where a company plays outside of test data management can significantly influence their appeal within the space. For example, if you are approaching test data management as an outcropping of data security, vendors that operate in data security (MENTIS, say) will likely offer an appropriately holistic solution. The same is true for data privacy, data governance, DevOps and so on.

There has been significant market movement over the last few years: a number of new names have appeared, both due to acquisition and the arrival of fresh faces. CA has been acquired by Broadcom, Compuware by BMC, and Actifio (although we do not include them formally in this report) by Google. The latter in particular may spell trouble for IBM, who as we have noted, resell Actifio's solution for database virtualisation. Given that Google is presently leveraging Actifio as part of GCP for that very purpose, and IBM's forthcoming plans to offer its solution as part of its Cloud Pak framework, we have to wonder if there isn't a conflict of interest at play. On the face of it, IBM is blasé; we remain concerned. We have also included IRI and Ab Initio for the first time, both old hands but with interesting solutions, as well as Windocks, an up-and-coming database virtualisation vendor that looks like it might be poised to shake up the space. It has already made waves by partnering with IRI and Curiosity, and moreover, by attempting to partner with almost everyone. It is easy to see its appeal: for the first time within the space, it offers database virtualisation that is both widely compatible and economically priced. By all accounts, it is also highly integrable. Accordingly, we expect several more official partnerships to be forthcoming. That said, Windocks is not the only vendor that has been busy networking: GenRocket and Delphix have also announced their partnership. Between these two examples, it is already becoming apparent that many vendors in the space are keen to leverage database virtualisation, but not to develop it themselves (perhaps speaking to the difficulty of doing so). The upshot is that by taking a pair of these solutions – Delphix and GenRocket, Windocks and Curioisty, Windocks and IRI, or Windocks and whoever else they end up partnered with – you can, at least in principle, build an integrated solution that offers comparable breadth of functionality to the space's big boys, thereby creating a competitive alternative. This is not something to be dismissed lightly, although whether it pans out remains to be seen.

# Metrics

We have identified eight capabilities that we have used to evaluate the products included in this report alongside more generic concerns such as geographic presence, stability, support, and innovation. Conceptually, they are split into two groups. The first consists of the following:

- Data subsetting
- Synthetic data generation
- Database virtualisation

These are, as you may have already noted, the three methods for managing your test data that we have described in this report. As such, they form the three primary use cases for licensing a test data management product. We have evaluated each product in the report based on their overall applicability to each of them, noting that many of the additional capabilities we describe below can play a part as well.

In several cases – particularly in regard to database virtualisation – support for one or more of these is essentially nil, as some products simply do not support them. We have not held this against them in terms of their placement on the Bullseye diagram, because we acknowledge that lack of capability in one area is not a factor if you do not intend to use it (or if you intend to leverage a second product – perhaps a partner – for that purpose). Instead, we have colour-coded the Bullseye according to which use cases each vendor supports.

The second group is slightly more ethereal. It is as follows:

- Data discovery
- Data masking
- Automation
- Ease of use
- Integration

These capabilities dig into the finer points of each product, although as mentioned they will also feed into the above. We have already talked at some length about data discovery, data masking, and the three use cases in our first grouping of capabilities. As such, we do not belabour their descriptions here, except to say that we have considered vendors based on the both the breadth of options they make available, the efficacy of those options, and any relevant additional features.

Automation is also fairly self-explanatory: how much automation can a given product enable? More prosaically, this is about how much more automated and expedient your test data processes can become once a given product has been integrated into your system. This means that automated test data provisioning, in addition to automated test data creation, is taken into account. On the other hand, self-service and onboarding are not: by our metric, they fall under ease of use instead.

Speaking of, ease of use is as it sounds. It encompasses the user interface, collaboration, self-service, and so on as you would expect, as well as – perhaps more importantly – the difficulty of getting a product up and running. If a product is notably fast to deploy and boasts a low time to value, it will likely score well here. Note that it is entirely plausible for a product to be high on automation but low on ease of use. This might mean, for example, that the product produces extremely well-oiled, automated channels for producing and delivering test data, but only after a lengthy, laborious, and manual setup process. The reverse, of course, is also true.

Finally, integration is a combination of the range and degree of connectivity a product offers, the breadth of additional test data capabilities it can provide via partnerships, and any closely related functionality the vendor themselves are able to offer owing to the solution's place as part of a suite. This means that robustly partnering with other vendors in the space to shore up your own capabilities is a recognised good, as is offering a test data management solution as part of a more general, integrated data management or security platform.

# Conclusion

Test data management is a broad space. As both a space in its own right and as the meeting point between data privacy and test automation, it contains a substantial number of vendors, many of whom approach the space from dramatically different angles depending on their own lineage. Accordingly, the way each vendor tackles the space can vary significantly, even beyond which of the three distinct methodologies for test data management they support.

What's more, the space is equal parts mature and innovative. Data subsetting, masking, and increasingly synthetic data, are more or less universally offered – though not always to the same extent – but at the same time, database virtualisation looks like it might only just be taking off. It is certainly an exciting time to exist in its periphery.

In short, it can be a difficult space to talk about. Regardless, we feel confident that whatever your use case for test data management is, at least one of the products we've included in this report will be able to address it.

## About the authors

**DANIEL HOWARD**
**Research Director, Information Management**

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)1494 291 992**
web: **www.Bloorresearch.com**
email: **info@Bloorresearch.com**