MarketUpdate

Test Data Management 2024

Introduction

Testing, and therefore test data, is an essential part of effective and compliant software development. But test data needs to be anonymised, to avoid exposing sensitive data; representative of your production data, so that your testing is meaningful; and easily accessible for your testing teams, to prevent bottlenecks in the testing process. Accomplishing all of this is the domain of Test Data Management (TDM).

There are a handful of naïve approaches to test data, such as leveraging whole, raw copies of your production databases, but these are generally a bad idea. There are several reasons for this, the greatest being scale: most production databases contain far more data than is practical to expediently, and repeatedly, distribute and test. Accordingly, the TDM space is concerned with more efficient alternatives, primarily data subsetting, synthetic data generation, and database virtualisation. These are often used alongside data masking and sensitive data discovery in order to provide the aforementioned anonymisation. These methods do not necessarily stand alone, and there are good arguments for having access to more than one, as they each tend to excel in different use cases. We describe them in more detail in the following section, then follow up with a discussion of the space as a whole and the vendors that operate within it.

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 the most mature method available for test data management.

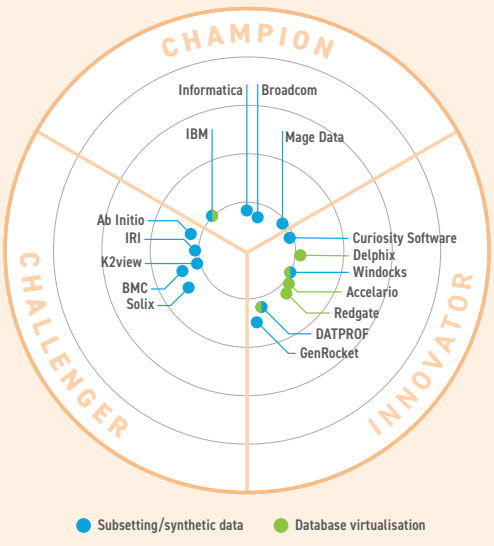
That said, it does pose some challenges, most notably in how you take your subset: taking a random sample will rarely result in a useful test data set. Instead, you will want your subset to be representative of your data as a whole, to ensure that all important scenarios are tested. This means that it will need to contain all conceptually meaningful test data points and combinations that are present in your production 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 have been solved by most 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 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 (and often containerised) copies of your

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.



databases. These virtual copies reference a master dataset, or are a delta store, or have some other means of being lightweight and easy to move around. This makes distribution much faster and easier, to the point that it can give each of your testers a personalised test data set to play with, and makes representation a nonissue.

Database virtualisation is somewhat less mature than synthetic data generation and (especially) data subsetting. 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. There are also potential scalability and cost issues, particularly when operating in the cloud. That said, the vendors that offer database virtualisation are largely aware of these issues, and – for the most part – are either working to address them or have already done so.

Data discovery and masking

Although neither data discovery nor data masking are TDM 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 falls foul of many, if not all, compliance mandates.

Therefore, unless you intend to leverage synthetic data exclusively, you will want to use 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 data masking) are sometimes used as well, but usually for ancillary purposes.

Data subsetting and database virtualisation vendors often offer discovery and masking functionality as part of their solutions, in order to allow them to function without relying on third-party products. At the same time, although discovery and (especially) masking are fairly mature capabilities in and of themselves, they are often not the primary focus within the TDM 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 helping you to directly leverage your production data for testing, it allows you to create your own "synthetic" test data in an automated fashion, often – but not always – based on your production data. Conceptually, synthetic data is data that "looks real, but isn't".

This has several advantages, including complete control of your test data set (for example, if you want to create a scenario that hasn't come up in production); better support for greenfield environments, where production data isn't present in a significant quantity (or at all); and the total absence of sensitive data, removing any need for discovery and masking as well as any possibility of deidentification.

Its most notable difficulties are representation and onboarding. What's more, these issues are linked. The more sophisticated synthetic data solutions analyse your production data, detect the trends and patterns within it (often using AI – see the next section), and create synthetic data that contains those same patterns ("maintaining statistical integrity"). 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.

These are not necessarily problems if synthetic data is present in a secondary capacity, which is common in solutions that primarily leverage subsetting or virtualisation. 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 full synthetic data solution if it could not reasonably be used standalone.

Market trends

TDM is, as ever, a changing space. Moreover, it is a space in two halves. On the one hand, TDM technology is increasingly popular and widespread among enterprises. In many cases, it is even seen as a requirement. A large part of this can be attributed to a greater recognition of the need for regulatory compliance, which has spurred on the desire for both test data management in general (especially synthetic data) and data masking in specific. The latter is increasingly seen as important for data security (in data breach prevention, for example) when deployed across the enterprise, and although data masking and TDM are separate spaces, the substantial overlap between them has created a knock-on effect for TDM. In some cases, TDM tools can also contribute to other data tasks, such as more general data provenance, versioning, provisioning, and so on, because the techniques used for curating test data are not always that different from the ones needed for curating production data.

On the other hand, many enterprises are, to quote one vendor we spoke to, "still in the stone age" when it comes to TDM. Production data is still widely used for testing (by as many as 60% of companies, according to one survey), in spite of both compliance demands to the contrary and the sheer quantity of testing tools (for TDM and more general test automation) available. We have to wonder what, if anything, it will take to convince these enterprises to see the value in test data, and we suspect that most vendors in the space have – with good reason – effectively given up on them.

The biggest new technology in TDM is the same as in most every other data space right now: generative AI. This is already in use by several vendors, and is proving to have some substantial applications in TDM, and especially synthetic data generation (more on this below). There is also an increased capacity for and desire to centralise test data (at least from a user perspective) and make it more easily accessible, usually via some sort of self-service data access.

At the same time, from an implementation perspective testing has long been moving away from the centralised

"centre of excellence" model to a more distributed architecture. This trend has continued, and hybrid solutions are common. In addition, it is increasingly normal for enterprises to integrate their TDM processes into a common workflow alongside various *Ops processes (DevOps, DataOps, MLOps, and so on). The complexity this integration adds to test data pipelines, as well as the increased complexity of test pipelines in general, has made the ability to systematically operate on your test data after its creation – but, crucially, before it enters the wider data environment – more necessary as of late. It is also worth noting that the emphasis on compliance that is now usually present in TDM initiatives often exists in tension with the desire to automate test data processes and make them as easy as possible to engage with. It is entirely possible that in your efforts to ensure compliance, you will inadvertently make your system a bear to work with. This is obviously undesirable and should be avoided or ameliorated as much as possible.

As for the three principle TDM techniques (data subsetting, synthetic data generation, and database virtualisation) synthetic data continues to rise in popularity, in part due to the ever-growing concerns around compliance, while the buzz around database virtualisation appears to have somewhat tapered off. There are a number of reasons why this could be the case – compliance concerns around distributing even masked copies of entire production databases, "hidden" additional costs when deploying database virtualisation to the cloud, and Delphix inadvertently convincing the market that database virtualisation must be very expensive, to name a few – but regardless, it is clear that much of the market has demurred from the technique. Indeed, this also applies more generally, with a number of vendors losing interest in other subsections of the TDM space or even the space itself (often by implication only, of course – watch out for products that are poorly supported or that are simply treading water).

This seems particularly poor timing given the recent excitement around AI (and specifically generative AI and LLMs (Large Language Models)) that has swept up every data space, most certainly including TDM. In fact, we know at least one TDM vendor that considers AI to be the key issue for the modern enterprise, and frankly, we would be hard-pressed to disagree. Within testing, we have seen AI used as an accelerator and copilot, interfacing with an LLM to help you build and provision your test assets, and indeed your test data, more quickly and more effectively. Several vendors use AI to generate synthetic data sets that mimic the attributes of production data sets, for example. Conversely, we have also seen vendors feed LLMs their test data (and other test assets) in order to provide them with additional context and rigor. In short, it is clear there is a lot of potential for leveraging AI in the TDM space. Indeed, there is still a lot of potential for TDM as a whole.

Vendors

TDM vendors need to possess at least one of data subsetting, synthetic data generation, and database virtualisation as a basic requirement for being a part of the space. In the former and latter cases, static data masking is also essential, as is the ability to discover sensitive data to mask in the first place. These capabilities should be offered at a competent standard: masking needs to retain referential integrity, subsets and synthetic data sets need to be representative of your production data, and sensitive data discovery needs to offer at least a few reasonably sophisticated discovery methods (column name matching alone is not sufficient, for instance). This is just the bare minimum, and more depth of functionality in these areas, functionality in more than one area, and/or additional, adjacent capabilities (such as dynamic data masking) are definite plusses.

In fact, many vendors differentiate themselves by offering significantly greater-than-average depth of capability in one or more TDM or TDM-related areas, often achieved by capitalising on new technologies such as AI. This usually applies to synthetic data generation, database virtualisation, and/or data discovery – it is less common with relatively mature technologies like data subsetting and (to a lesser extent) data masking. A second (often complementary) strategy employed by a similarly large number of vendors is to position their products as highly integrable and automated TDM platforms. Indeed, this is sufficiently ubiquitous that it has almost ceased to be a meaningful differentiator. On the other hand, there are still vendors (albeit vanishingly few of them) that maintain a deliberately narrow focus, limiting their scope while enabling them to develop technological advantages that may put them ahead of wider-reaching competitors within their chosen niches. Another approach is to offer one or more of improved automation, ease of use or performance, or reduced costs, compared to the other vendors in the market. This is easier said than done (performance advantages especially are difficult to prove), although ease of use and reduced costs (via a lower price point) tend to be something that younger solutions do well at, for obvious reasons.

From a capability perspective, practically every vendor in the space now offers data subsetting, synthetic data generation, sensitive data discovery, and static data masking. Some also offer database virtualisation, and those that don't tend to integrate with those that do. In this Market Update, we have primarily looked at the TDM offerings from Broadcom, Curiosity Software, DATPROF, IRI, K2view, Mage Data (née MENTIS), Redgate Software, and Windocks. Of these, DATPROF, Redgate and Windocks provide database virtualisation natively, while Curiosity Software and IRI offer it via official partnership (and integration) with Windocks. Various other products can similarly integrate with Windocks and/or other virtualisation solutions. Broadcom also technically offers a database virtualisation solution, but it is almost entirely deemphasised. Moreover, almost all of these vendors offer what they describe as a complete solution for test data management. Database virtualisation aside, this is by and large an accurate assessment: TDM vendors are now distinguished not by what they can do, but by the efficacy with which they do it. When it comes to generative AI, currently the most prominent trend in data, the vendors differ somewhat in their approach: some have dived straight into the deep end, while others are still testing the waters. Both attitudes have their merits, but we would hazard a guess that a middle ground will prove most fruitful in the medium-long term.

In terms of market movement and major rebranding efforts, MENTIS is now Mage Data (as you will have already realised from the above paragraph). K2view has risen to some prominence in the space, emerging out of the ETL market, and has been included in this report for the first time. DATPROF and Redgate have launched new products, DATPROF Virtualize and Redgate Test Data Manager, respectively. Broadcom has sold BlazeMeter to Perforce, both because it was competing with Broadcom Test Data Manager and because it purportedly didn't mesh well the company's enterprise-level customer base. Perforce has also recently completed its acquisition of Delphix, almost inarguably the most entrenched database virtualisation solution. This goes some way to explaining why Delphix was unwilling to talk to us during this report's research phase (hence why it is not prominently featured). More interestingly, Perforce appears to be putting together the pieces for a tidy TDM solution – one wonders if it will feature in the next version of this Market Update. In less rosy news, several large vendors seem to be stepping away from TDM. We have heard that IBM and Informatica are ceasing their support for dynamic data masking, which is not particularly relevant to TDM in and of itself, but they both also demurred from engaging in the research process for this report, either by declining the opportunity or by simply being uncommunicative. The same is true, but more so, for Solix, which outright stated it does not consider TDM a high-priority area (and hence has not updated its TDM solution since 2021). That said, if these vendors are distancing themselves from the space, we are confident that it will mainly serve to make room for new growth.

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 capabilities and lineage. That said, as the space has continued to grow and its demands have crystalised – especially around regulatory compliance – there has been a certain degree of homogenisation. In previous versions of this report, we described the way in which each vendor tended to push just one of the methods we've described as the solution for TDM. These days, it is much more common for vendors to offer all of them and, notably, position them on roughly equal footing, with the main reason to use one over the other its applicability to a particular set of use cases. There are exceptions, and many of the old biases still remain – albeit to a lesser degree – but overall, the space is much less partisan than it once was.

In short, the TDM space has matured, both in terms of the capabilities provided by its vendors and its reception in the wider market. Even if many (perhaps even most) companies still refuse to invest in a TDM solution, enterprise-level organisations are increasingly and acutely aware of the benefits such a solution can provide.