# Test Data Management



## Introduction

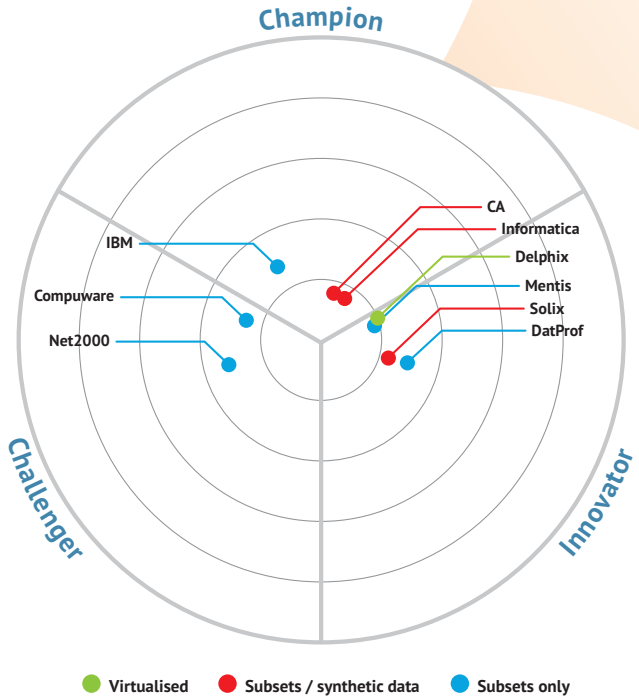
There are five ways to provision test data. You can copy or take a snapshot of your production database or databases. You can provision data manually or via a spreadsheet. You can derive virtual copies of your production database(s). You can generate subsets of your production database(s). And you can generate synthetic data that is representative of your production data but is not actually real. Of course, the first four examples assume that the data you need for testing purposes is available to you from your production databases. If this is not the case, then only manual or synthetic data provision is a viable option.

In practice, neither copying production data nor manually provisioning data can be called "test data management". There is no management involved. We cannot imagine recommending the use of either approach for anything but the very smallest projects: in the case of taking copies, because of the costs involved; and as for manual methods, it is too onerous and error-prone.

The other three approaches can all genuinely be described as valid test data management (TDM) methods. It is worth detailing some major features that we expect from a TDM product. Firstly, it should be able to deal with sensitive data. That is, either the product needs to include data masking capabilities or it needs to be able to generate synthetic data, or both. Secondly, the solution should implement the principles of DevOps and support agile environments. This means putting as little strain as possible on database administrators and, in practice, this either means generating virtual copies of your source database(s) or running off a test data warehouse. In either case, the idea is to allow the re-provisioning of (amended) test data on demand, without having to go back to the production data or the administrators thereof. In effect, providing a self-service test data environment for developers and testers. Thirdly, you would like your TDM solution to integrate with other tools within the testing environment so that, for example, test data can be automatically provisioned to relevant test cases. Lastly, you would like your TDM solution to be not just representative of your real data but also to include outliers, boundary conditions and erroneous data that might not – almost certainly will not – appear in your production data.

Finally, it is worth briefly commenting on the pluses and minuses of the different approaches to TDM. Taking virtual copies of your production data means that you do not have to ensure that your data is representative of your production data, because it is the production data (even if virtualised). Conversely, you have to profile your production data sources if you want to subset your data or generate synthetic data, in order to ensure that the results are representative of the data as a whole. While tools are typically provided to do this, some manual effort will be involved. Secondly, test data warehouses can be expensive in storage terms compared to taking virtual copies.

**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 segment if their innovation rating is over 2.5 and Challenger if it is less than 2.5. 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.

On the other hand, virtual copies are just that – copies – and production data does not typically include the outliers and boundary conditions that are most likely to break your code.

In so far as comparing synthetic data generation with other methods is concerned, the disadvantage is that even more in-depth profiling of production sources is required but, conversely, you don't have to do any data masking. And, as we have mentioned, synthetic data generation will likely be required (perhaps in conjunction with service virtualisation software) if data is not available. This can occur for a variety of reasons: security protocols prevent access, this is a new development and data is not available, the data at run-time will be provisioned by third parties and it is not accessible for testing purposes, and so on.

## Market trends

Perhaps the most significant trend in the market is towards greater interest in synthetic data generation. This is recognised by almost all vendors, even those that do not currently offer such capability. This is driven primarily by security and compliance requirements, especially with new laws – such as the EU's GDPR (general data protection regulation) – being put in place. Almost all the vendors not currently offering synthetic data generation plan to do so in the future and we expect several relevant announcements during the course of 2017. Currently, only CA, Informatica and Solix (plus Tricentis and GenRocket – see below) provide synthetic data generation.

On a related topic, it is worth commenting that while data masking and test data management have traditionally been seen as going hand in hand – and they still do, if synthetic data generation is not available – they are increasingly perceived as divergent. That is, TDM is about productivity and efficiency in testing and development (usually, though TDM can also be used to provision, for example, demonstration data), while data masking is more about security and has a role that goes beyond test data management. For this reason, Bloor Research will shortly be publishing a companion report to this one, specifically on (static) data masking as a stand-alone function. In this respect, it is worth noting that a few companies are advocating the use of format preserving encryption (FPE) as an alternative to data masking. Mostly this is not relevant to test data management. However, HPE (soon to be Micro Focus), has suggested to us that FPE is suitable for use in conjunction with test data although the company does not offer an explicit TDM product (though sub-setting is available within its Structured Data Manager product, which is aimed primarily at archival). It is worth adding that other vendors in the data masking space, who do not offer TDM, have quoted test data as a common use case for their technology.

A further trend is towards greater adoption of DevOps principles. This is apparent not only in the increased deployment of test data warehouses (and test data marts) and other techniques for refreshing test data that do not require the intervention of operations personnel, but also in increasing integration with wider testing environments, either directly or through open APIs. While neither of these characteristics is universal there has been growth in both areas and we expect to see more of the same in the months to come.

## The vendors

From a vendor standpoint, there is Delphix and there is everybody else. Delphix provisions virtual copies of your production data (masked if required) to developers and testers. Everybody else either offers subsetting or synthetic data generation or both. We have rehearsed the chief arguments for and against such approaches. One major distinguishing feature between products is the range of platforms they support. While this doesn't impact on the quality of the respective products it does have implications as to when you might want to use them, so we have separated platform support out as a separate consideration. Note that all of the products support file systems, but this is not true of databases. In particular, there is a significant trend towards support for unstructured data environments, both for data masking and test data management. Specifically:

- **CA** offers support for a wide range of databases, including NoSQL environments. There is also significant support for unstructured data. Mainframe support is provided.
- **Compuware** focuses on mainframe environments and databases though distributed databases such as SQL Server, Oracle, DB2 and Sybase are also supported.
- **DatProf** hosts its metadata in an Oracle, SQL Server or DB2 database and while it has been used to support test data from other environments, these are not a focus.
- **Delphix** supports the leading relational database products as well as Amazon RedShift and MariaDB.
- **IBM** provides support for both mainframe and distributed environments. The company aims to be generic rather than IBM-specific when it comes to supporting a range of databases and it supports unstructured as well as structured environments.

- **Informatica**, as befits a data integration specialist, supports a very wide range of databases, in both mainframe and distributed environments. The company is rapidly increasing its support for unstructured data including support for databases such as MongoDB and Cassandra.
- **Mentis** supports leading relational products in both mainframe and distributed environments. Plans are in place to extend support to IMS, Teradata and Hadoop/Hive. These are already supported by the company's data masking products.
- **Net2000** is available for use with Oracle and Microsoft SQL Server.
- **Solix** supports traditional relational products but not NoSQL environments as yet. It has especially strong (native PL/SQL) support for Oracle databases.

From a market development perspective, not a great deal has happened: Grid-Tools was acquired by CA in 2015 and HPE (though it is no longer a TDM vendor: what used to be its TDM product has been dropped) is to become part of Micro Focus later this year. However, the TDM market is not extensive enough to warrant consolidation and we do not expect that to happen.

There are three products we are aware of that we have not included in this Market Update. These are from Tricentis, GenRocket and Oracle respectively. In the case of the former, this is the only product we know of that only generates synthetic data and does not support subsetting. The reason we have not included it here is that (in practical terms) it is not available as a stand-alone product but only in conjunction with Tricentis Tosca TestSuite. Similarly, Oracle Data Masking and Subsetting (which describes the product exactly) requires the use of an Oracle database, because you have to stage third party data into that database: from a generic standpoint, we regard this as a significant minus.

Finally, we have excluded GenRocket because we see this as a test data generation tool rather than a test data management product. The company offers synthetic data generation (and can work from production data) as a service but it lacks the broader capabilities we would expect of a true management product. It will be useful for small projects and where the data elements involved are limited.

## Conclusion

Research suggests that while the use of TDM products is increasing, a majority of companies do not currently use formal TDM methodologies or tools. In practice, we believe that there are two direct reasons and three indirect ones, to suppose that TDM products will capture a greater and greater share of the market. The direct reasons are that TDM is more cost-effective and productive than simply copying production data. The first two indirect reasons are that TDM is an intrinsic part of DevOps, and that it facilitates agile development approaches through support for automated testing. Finally, the third indirect driver is that we expect privacy requirements (such as the EU's GDPR) to encourage the increased use of products that provide synthetic data generation.

From a product perspective, there are only three vendors – CA, Informatica and Solix – that support synthetic data generation as well as subsetting. Solix has only recently introduced support for synthetic data and the product does not yet have the maturity of either CA or Informatica, so we would class these two as clear market leaders. IBM has strong subsetting capabilities but is not easy to use and we rank Mentis as the highest rated product of those that only support subsetting. Delphix, of course, is in a class of its own. Compuware Topaz Workbench, through which users can initiate File-AID along with Compuware's other solutions and capabilities, is available at no additional charge to maintenance paying customers. Users will no doubt find this attractive. In our view, DatProf wins the prize for the easiest to use offering though both it and Net2000 lack some of the bells and whistles provided by at least some of the other vendors. On the other hand, these may well represent more cost effective solutions.

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**

# CA Technologies

CA Test Data Manager (TDM) is a test data management tool that provides a standardised set of data that covers all possible tests, remains up-to-date, can be provisioned on-demand, and contains no sensitive data. CA TDM can profile your production data, creating a view of what exists and where, while exposing and visualising any relationships that exist within your data. This aids in understanding your data and will inform your use of the product.

CA TDM supports both subsetting and the generation of synthetic data. For the former, the functions provided ensure that all relevant attributes are covered. Where the data is sensitive, CA provides data masking so that subsetted data can be masked, and the product includes over eighty distinct masking options. It automatically discovers data that looks like it might be sensitive, keeps the data referentially intact, is fully auditable, can mask millions of rows in a matter of minutes either in place or in flight, and it's demonstrably compliant with EU GDPR, GLBA, HIPAA, PCI DSS, PIPEDA and other regulations.

Notwithstanding that, CA TDM's standout feature is its ability to generate synthetic data. The process relies on a user-created model of valid test data attributes (leveraging built-in profiling capabilities), and once this model is built, CA TDM can automatically generate synthetic data that achieves one hundred percent coverage – according to your test model – while still being representative of your production data. This includes covering outliers, unexpected results, boundary conditions and negative paths.

In addition, CA TDM also makes test data easily accessible. Using CA's Test Data On Demand portal you can set up self-service and automated delivery of test data to testing teams. Any data that gets generated can also be stored as a reusable asset in a test data warehouse (or mart). This warehouse then provides a central library of test data that can be reused on demand. In addition, there is a 'test matching' feature that allows decoupling of test data from test cases so that testers can dynamically find and reserve test data as required. TDM integrates with CA Agile Requirements Designer, CA's test case modelling tool, and supports both mainframe and distributed environments.

## Strengths

- CA TDM's support for synthetic data is market leading. Relatively few test data management tools support synthetic data generation.
- CA TDM emphasises the efficiency of generating and obtaining test data. The test data warehouse, test matching, and the Test Data On-Demand feature are all valuable capabilities that facilitate reuse and make it quicker and easier to set up automated tests.

**CA Technologies**
Islandia, USA

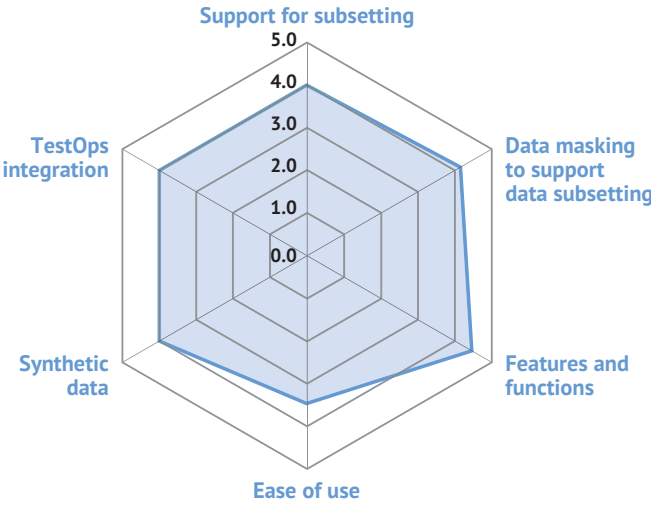
**www.ca.com**

## Threats

- To start generating synthetic data, CA TDM has a start-up cost in the form of creating the test data attribute model. Conversely, there is no need to mask and subset data from production.
- While CA is rightly focused on testers and teams, some other suppliers approach test data management more from a data management perspective. These competitors will likely be marketing to different people within user organisations, so there is a marketing challenge for CA in ensuring that it gets its message across to all interested parties.

## Summary

CA TDM is a test data management tool with an emphasis on efficiency, compliance, security and data quality issues, particularly using synthetic data. This was the first product in the market to support synthetic data generation and the product remains one of the clear market leaders.



# Compuware

Compuware's test data management leverages its file and data management solution File-AID, which is readily accessible from within the Topaz Workbench, an Eclipse-based IDE that provides a common framework and integrated user interface from which to initiate Compuware's array of mainframe development, testing and maintenance tools. The Topaz Workbench is available at no additional charge to maintenance paying Compuware customers.

From the Topaz Workbench, it is easy to browse and edit data in a variety of preset or custom layouts (including raw data, for those who need it). Additional functionality is available within Topaz Workbench such as comparing and subsetting data, visualizing data relationships, visualizing extract executions, and copying data from one mainframe LPAR to another.

Compuware's Data Privacy solution provides the ability to identify and mask sensitive data. The Data Privacy interface within Topaz Workbench is used to create and manage privacy definitions. This tool allows you to abstract each type of data in your database (for instance, 'name' data) into a data element that can then have masking rules applied to it. Fields are matched to the data elements (by matching the field name) and the disguise rules defined for the data element are dynamically built and applied at execution time. A coverage view is available that will display which fields are mapped to which data elements, allowing a user to adjust them if appropriate. Relational integrity is always maintained and an audit log is generated upon execution. Compuware's patented Composite processing finds data within a larger field and ensures that differently formatted values will be properly masked. For instance, it will recognize that "Mary Jane Smith" and "Smith, Mary Jane" are the same name and mask them appropriately (for instance, to "London XXXX Klein" and "Klein, London XXXX", respectively). Compuware's Eclipse-based interface disguises data where it resides, either on the mainframe or in the distributed files or databases in which it exists.

## Strengths

- Topaz Workbench is very easy to use, especially bearing in mind that it manages mainframe data. Compuware's capabilities for test data management and data privacy work across distributed environments as well, so that masking can be deployed in a consistent fashion across data sources.
- The abstract nature of data elements in the Data Privacy definition makes it easy to mask several different but conceptually similar fields using the same set of rules. More to the point, it minimises maintenance issues if those rules are ever changed.

**Compuware**
1 Campus Martius, Detroit, Michigan 48226
**www.compuware.com**

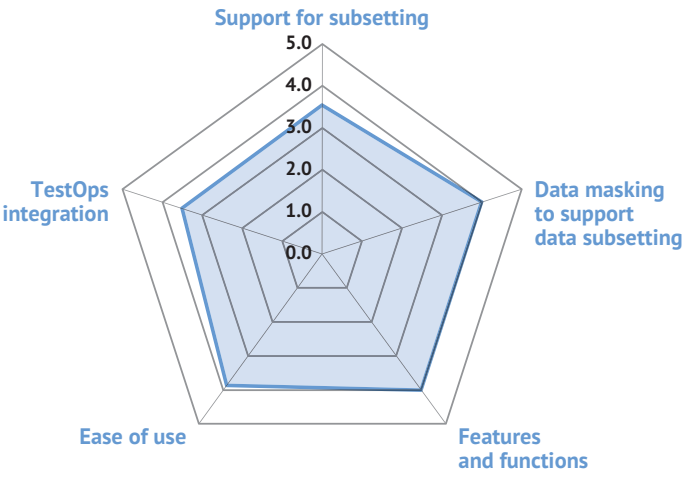
- The fact that maintenance paying Compuware customers receive the Topaz Workbench at no additional charge is a major plus point.

## Threats

- Matching fields to data elements based on their names is not always ideal. In practice, fields tend to be named arbitrarily and not necessarily with forethought. This is mitigated somewhat by the ease of adding fields to the data element and use of the coverage view to preview data identification and rule assignment.
- Compuware provides test data generation capabilities but these are based on selection criteria rather than database profiling, so it will be up to the user to ensure that test data is representative.

## Summary

The File-AID functionality used within the Topaz Workbench makes it relatively simple to apply test data management and test data masking within mainframe and distributed environments. Mainframe and distributed data residing in files, relational databases or IMS will all be disguised consistently using Compuware's Data Privacy solution.



# DataBee

Net2000's DataBee is a fast and effective data subsetting and test data management product that supports SQL Server and Oracle. You begin the subsetting process by fetching data and loading in tables – that may be excluded on a case-by-case basis if necessary – from your source database. Then you must specify a driver table that will form the core of your subset and extraction driver rules that determine how this driver table is sampled. When the extraction is run, DataBee will build and extract a referentially intact subset of your database based on the rows sampled from your driver table. It is then a simple task for DataBee to load this extract into your target database. The most difficult part of this process is creating your extraction rules, and accordingly DataBee provides some guidance to help you do this. First and foremost, it allows you to make a plan for your subset, allowing you to decide whether, ideally, the contents of each table in your database should be included, subsetted, or left out entirely. You can then compare your extract to your plan and see how well it measures up. Further, DataBee provides the Chain Finder and the Who-Loads-What tools that allow you to view the dependencies in your data, respectively taking a narrow and a broad view, making it much easier to gauge the impact of each existing extraction rule as well as any new rules.

Data subsetting is only one half of test data management, the other being data masking. Consequently, DataBee is supported by its sister software Data Masker, a static data masking product also offered by Net2000. Like DataBee, Data Masker is fast (processes are run in parallel) and can mask millions of rows an hour. It has a dual focus on compliance (for instance, with HIPAA) and maintaining the credibility of masked data. The latter is done by ensuring that correlated values (for instance, age and date of birth) remain consistent after masking, and in addition this can be done between databases or even database instances. Notably, masking in Data Masker always retains relational integrity and includes the capability to mask primary or foreign keys without a join operation. Data Masker also provides a column finder that allows you to search your database based on column name.

Net2000

**Net2000 Ltd**
Llangunllo, Knighton, Powys
United Kingdom LD7 1SP

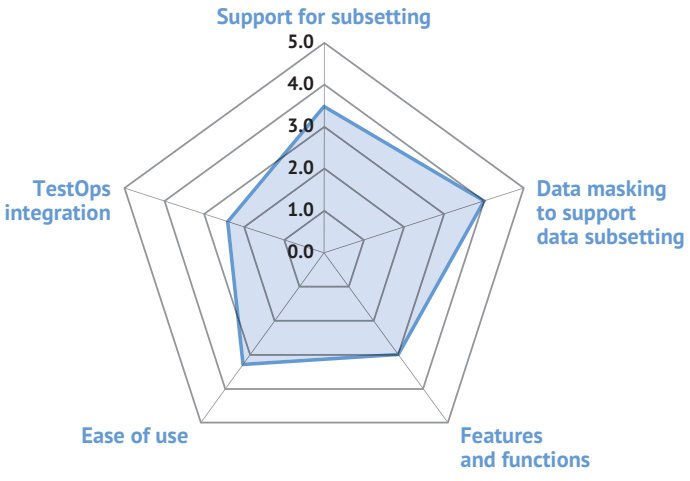
**www.net2000ltd.com**

## Strengths

- DataBee and Data Masker are fast, able to process millions of rows an hour. Data Masker's support for parallel processing is particularly impressive.
- Both tools are lightweight and have an intuitive interface. In particular, neither tool has extraneous functionality that makes them difficult or overcomplicated to work with.
- The Who-Loads-What and Chain Finder tools provide useful insight into your data and the dependencies therein.

## Threats

- Building your subset in DataBee can be difficult due to the requirement for a referentially intact subset. However, this is mitigated somewhat by the guidance provided by the Chain Finder and Who-Loads-What tools.
- The column finder is a useful feature, but only searching by column name is limited. We would like to see something more sophisticated.



# DATPROF

DATPROF offers three products relevant to test data management: DATPROF Subset is used to create subsets of production data for testing purposes. DATPROF Analyze (which is currently in beta) is used to ensure that subsets are representative of the production data as a whole, and is also used to discover sensitive data within those subsets. Finally, DATPROF Privacy is used to mask sensitive data. The DATPROF tools can work with (in theory) any source database but require an Oracle, SQL Server or IBM DB2 database to process test data, although only metadata is ever actually extracted from your system. The area where DATPROF most excels is ease of use: the various products, although technically separate tools, feel like a single product that is exceptionally easy and intuitive to work with, and doesn't require any significant training. This means that you can discover, subset or mask your data easily and – more importantly – quickly.

Using Subset, you can visualise your database as either (or both) a data and process model and this is extremely helpful for understanding the structure of your database (and therefore how to conduct the sub-setting process). From a functional point of view, we would say that DATPROF's capabilities are good without being exceptional. As we have stated, ease of use is the primary selling point, and in this respect the products are exceptional. We particularly like the option that Subset provides that will append newly generated test data onto already populated tables in your test database. Additionally, options for recreating or refilling existing tables are available. We also like the custom masking rules, constraints on existing rules and rule dependencies offered by Privacy. Privacy also maintains its own audit log. Both tools complete their deployment operations, either sub-setting or masking, by generating and running a SQL script. This ensures that they remain performant.

## Strengths

- Ease of use and user experience are exceptional, and a cut above competing products in the space.
- An important feature to emphasise is that DATPROF never physically extracts data, only metadata.

**Datprof**
Friesestraatweg 211, 9743 AD Groningen
The Netherlands

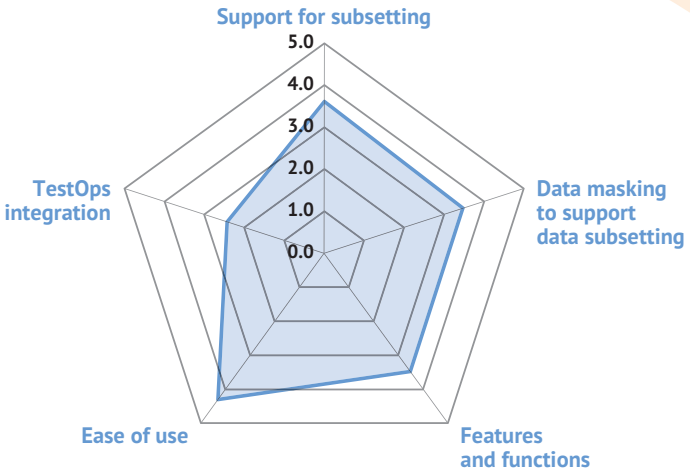
**www.datprof.com**

## Threats

- DATPROF does not have the bells and whistles that some other vendors offer. For example, it doesn't support 80 different masking methods. On the other hand, do you really need such complexity?
- DATPROF is a relatively small company and not well-known outside of its native environment (the Netherlands)

## Summary

DATPROF lack some of the more sophisticated capabilities that are offered by more well-known competitors. On the other hand, its products are exemplars of user experience, ease of use and intuitive design: they are the easiest to use products we have seen.



# Delphix

Historically, there have been four ways of provisioning data for test and development purposes. The most obvious is simply to take (multiple) physical copies or snapshots of production databases. Secondly, for small projects you could define test data manually or in a spreadsheet. More productive and efficient methods involve generating subsets of your database or generating synthetic data. In both of these last two cases, care has to be taken to ensure that the data is representative of your data as a whole.

Delphix offers a fifth way: data virtualization. Delphix holds a single, continuously updated, copy of your production databases (including all binaries, configuration files and so forth). It can then provision complete virtual copies of this data, as required. The automation and self-service that is provided fits well as part of agile and DevOps initiatives, with features that include the ability to refresh, reset and rewind data, as well as the ability to bookmark, branch and share data. Different databases can even be provisioned in a synchronised fashion. Delphix software can be hosted on-premises or in the cloud (AWS or, shortly, Microsoft Azure) and leverages your existing storage. Supported data sources include Oracle, MS SQL, Sybase, PostgreSQL, MySQL, DB2 as well as file system data, and the company is adding support for other Amazon supported databases such as MariaDB. Support for other data sources is under consideration.

If your test data is potentially sensitive (discovered through using the Delphix Data Profiler) then it will need to be masked, and in 2015 Delphix acquired one of its partners: Axis Technologies, which specialised in data masking. How this works is that you take your copy of the production database and then you create a masked virtualised version of that and then provision virtual copies from that source. If you already have your own data masking solution then you can use that instead. Another option is to use internally written scripts for masking, though this is not an approach that we would recommend. As far as features are concerned, whilst Axis was targeted at the mid-market, Delphix has now successfully deployed their masking in their enterprise customer base. Stand-alone (without data virtualisation) masking support is also available for mainframe (iSeries, VSAM and z/OS) systems. Delphix also supports a tokenisation capability. Audit and compliance reporting is provided out of the box.

**Delphix**
1400A Seaport Blvd, Suite 200,
Redwood City, CA 94063

**www.delphix.com**

## Strengths

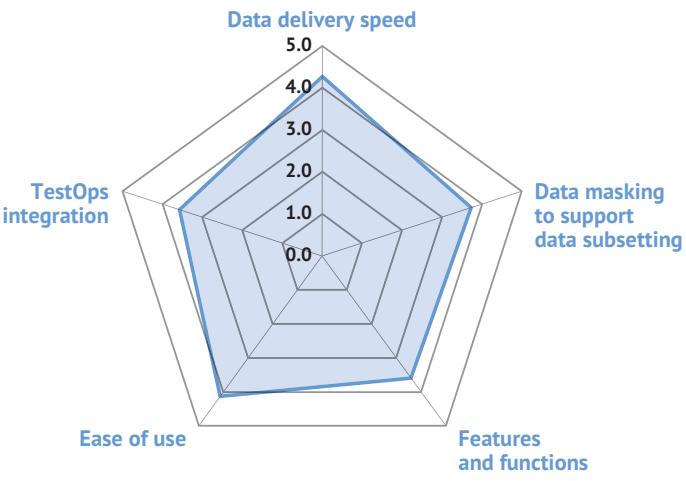
- Delphix offers a genuine alternative to other approaches to test data management. While there are other products that do bits of the same things that Delphix does, it has no real competitors.
- The approach offered by Delphix is rapid and automated. Unlike other approaches to TDM no effort (manual or otherwise) is required to ensure that test data is representative of your data as a whole.

## Threats

- The data masking provided is good without being outstanding, though it is improving. But you can always use a third-party data masking product if you want to.
- The range of databases supported by Delphix is relatively small compared to some of its competitors.

## Summary

Delphix is unique. From a test data management perspective, it's not so much about the product it's more about the approach. If you like the approach you'll love the product.



# IBM

IBM offers Test Data Management (TDM) and Data Masking as separate products within its Optim suite of tools. Historically, this was not the case and the products were bundled together. This makes sense since TDM is essentially about improving efficiency and productivity, while data masking is about reducing risk and ensuring compliance. Nevertheless, they will often be used in conjunction.

TDM is based on sub-setting and the product has advanced features in this respect, with Information Analyzer being used to discover the data model in use and you can extract representative subsets of the data. These can be right-sized by test type. For example, you might want a different sized subset for, say, unit testing as opposed to integration testing. Facilities are provided to refresh test data when required and also to compare data subsets. The product also works alongside IBM's Greenhat product for service virtualisation, there is a plug-in for UrbanCode and integration with some of the Rational tools. The generation of synthetic data (which IBM refers to as test data fabrication) is on the company's roadmap. However, it has been there for the last two years, so we are not holding our breath.

From a masking point of view, Information Analyzer can this time be used to determine what data needs to be masked, while masking itself is comprehensive. Notable features include affinity masking (for example, maintaining case), consistent masking across multiple platforms, and semantic masking, though the latter is not easy to use and is not enabled for test data management. The company has also implemented in-database masking in a number of its environments (for instance, Netezza) and is actively extending this to others (such as DataWorks). This is supported via user defined functions. IBM is also actively integrating masking into other environments (for example, there is a Data Masking Stage in DataStage). You can use Optim, which supports both mainframe and distributed environments, for dynamic data masking but this is more usually the domain of IBM Guardium.

## Strengths

- The subsetting capability offered by IBM is very strong and we especially like the support for right-sizing.
- Data masking capabilities are extensive. It is interesting to note that IBM is seeing more customers interested in masking unstructured as well as structured data, and it is pleasing that the company is developing capabilities to support these environments.

**IBM**
Armonk, USA
**www.ibm.com**

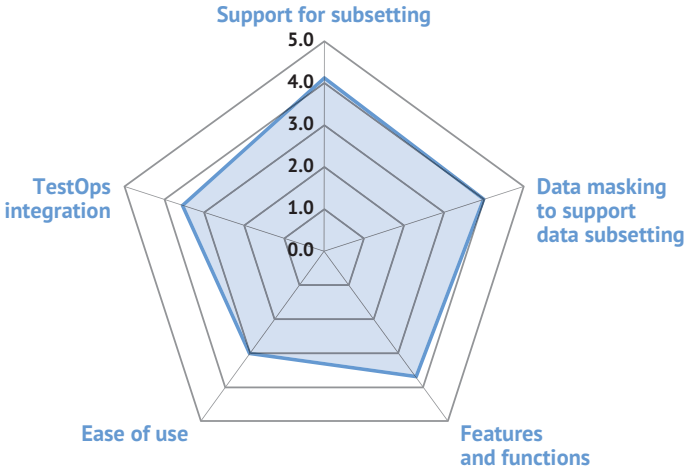
- IBM is implementing the ability to call masking functions from Java. This will be released during the course of 2017 and will support the Hadoop "menagerie".

## Threats

- IBM's offering is not as easy to use as we would like.
- We are disappointed that test data fabrication has not yet been released. We hope that it will be made available during the course of 2017.
- We get the impression that the Rational/Optim interface is not as close as it might be, and that the company is losing opportunities to leverage synergistic capabilities.

## Summary

IBM has a strong offering in this space. However, it is hampered, in our opinion, by the fact that development and testing is in one division (Rational) while Optim is in another. We also get the sense that, within its own division it does not get the attention it deserves, because it is not a "Watson" product.



# Informatica

For test data management (TDM) Informatica offers support for both sub-setting and synthetic data generation. In the case of the former the company offers a variety of methods for subsetting, while there is increasing demand for synthetic data generation, for a variety of use cases. Historically, synthetic data generation has been to support environments where there is no access to production data, or where that is not suitable or for new functions. Compliance is also an important driver and Informatica ships with out-of-the-box policies to support PCI, PII and PHI compliance. GDPR, which is expected to be a major driver for synthetic data generation, is planned.

A major capability, not shared by many of its competitors, is the provision of a test data warehouse, that can be used to provision updated test data on demand and without troubling the production environment. This links into DevOps environments and there is integration with HPE ALM. Using its command line interface, provisioning of test data can be integrated with DevOps tools such as Jenkins. There is also a joint development that integrates with Cognizant's ADPART. In addition, there is support for flat files, as well as databases. There is a graphical test data coverage capability that allows you to see whether you have sufficient data to provide the level of coverage you require.

As far as data masking is concerned, there are extensive options, including support for masking unstructured data and federated masking to ensure consistency across multiple datasets. Dynamic masking – usually used with production data – is available but requires a separate license. A notable capability is provided by the company's Secure@Source product, which provides lineage against masked data and allows you to see where data has been masked as it flows through the enterprise. Full integration with Secure@Source is planned for later during 2017.

## Strengths

- Relatively few vendors currently offer both sub-setting and synthetic data generation.
- We especially like the graphical capabilities provided for test data coverage and, indeed, that the product actually provides test data coverage in the first place.
- The test data warehouse capabilities are extensive and a significant differentiator.

**Informatica**
2100 Seaport Blvd, Redwood City
California, USA, 94063

**www.informatica.com**

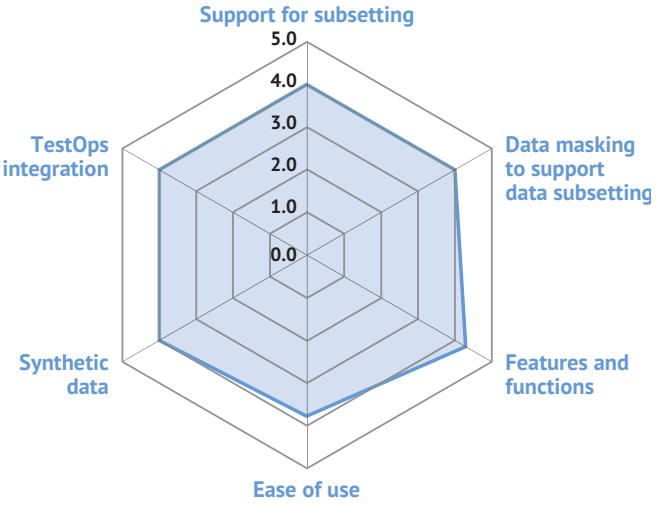
- We have criticised Informatica in the past for taking a stance to TDM that was too data-centric so we are pleased that the company is taking a more tester-centric approach. It would probably be a good idea to broaden the scope of its DevOps integrations. In particular, HPE ALM is losing its previously dominant position and integration with other tools would be an advantage.

## Threats

We would like to see the lineage available in Secure@Source directly available within the data masking product and without requiring a license for the former. Apart from this minor quibble we are impressed with Informatica's offering and, while it has the usual competition, there is no obvious weakness in its products.

## Summary

Informatica is – justifiably – a leading player in this market and we do not expect that position to change.



# Mentis

Mentis focuses on "sensitive data lifecycle management" and it offers a number of relevant and complementary products for this purpose. These include iDiscover, which is used to discover sensitive data that needs to be masked, using either iScramble for static data masking or iMask for dynamic data masking. Complementary products include iSubset for test data management (TDM) and iRetire, which can be used to support the "right to be forgotten". In so far as TDM is concerned the main products that will be used are iDiscover and iScramble, in addition to iSubset.

As its name implies, iSubset provides a subset-based approach to generating test data. The company does not offer synthetic data generation. When creating subsets of your data to support testing, Mentis offers a number of options: specifically, you can create a subset across all applications within your database or you can create a subset that is relevant to a single application only. Further, these subsets can be generated either on a percentage basis or by time slice (for example, the last 100 days).

Of course, not all data is sensitive, so sometimes you would use iSubset on its own. More commonly, it will be used in conjunction with iDiscover and iScramble. With respect to the former, Mentis goes further, in our opinion, than any other supplier in its facilities for discovering sensitive data. In particular, in addition to pattern recognition and similar profiling capabilities the software has the ability to introspect business rules written in SQL (for example, PL/SQL) that may identify sensitive data. As far as iScramble is concerned Mentis has a number of neat features. For example, it was the first company, as far as we know, to introduce conditional masking based on location, whereby you can apply different masking or access rules, according to the location of the user.

## Strengths

- The discovery capabilities offered by Mentis are excellent, and market leading.
- The company offers a number of unique or near unique features such as location-based, conditional masking and time slice based subsetting. Its iRetire product will be especially beneficial in regulated environments such as the EU's GDPR (general data protection regulation).
- The company's offering (which also includes iMonitor and iProtect) is much broader in scope than most of its competitors. Only IBM could offer something comparable but that would be in diverse products that are not well integrated.

**Mentis**
3 Columbus Circle, 15th Floor
New York, NY 10019

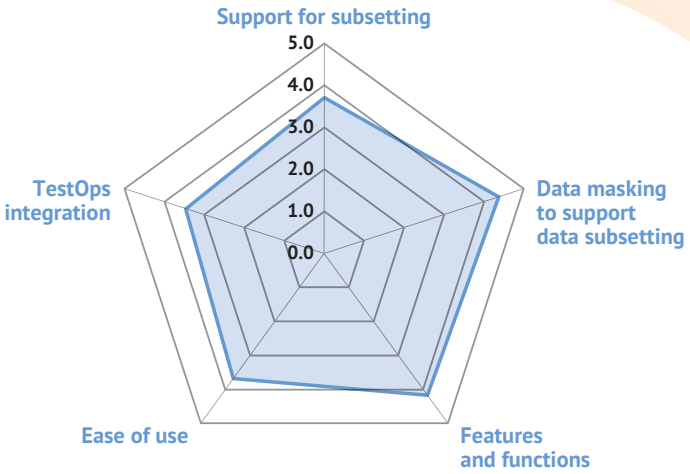
**www.mentisoftware.com**

## Threats

- While Mentis has good support for structured database and other products (in both mainframe and distributed environments) it is less strong when it comes to unstructured data, compared to some other vendors.
- There is no support for synthetic data generation, which we expect to become increasingly popular.

## Summary

We are particularly impressed with the Discovery capabilities offered by Mentis. Combined with robust masking capabilities and relative ease of use, and the add-on capabilities of iRetire, this makes Mentis the leading technical provider of solutions that purely offer subsetting.



# Solix

Solix has two solution suites: the Big Data Suite and the Enterprise Data Management Suite (EDMS). Within the latter are the company's offerings for both test data management (TDM) and data masking.

Solix's test data management offering is one of only three that we know of, that offers synthetic test data generation as well as subsetting capability. In so far as synthetic data is concerned there is tooling provided to ensure that test data is representative of your real data and there is an embedded rules engine which defines how data should be generated. For subsetting, two approaches are supported. The first of these is conventional – you literally extract a subset against defined parameters – but in the second you take a back-up of your production database and then create a golden copy from this that has transactional data stripped out of it: you then populate this with production data (again, based on relevant parameters) as required. The advantage of this approach is that performance can be significantly improved. The parameters on which subsets can be derived can be either vertical or horizontal across the database so that, for example, you can subset by time or operating unit as opposed to subsetting by table.

As far as data masking is concerned (which may be required to test subsets), Solix offers the normal sorts of masking algorithms you would expect. These are complemented by a discovery tool that both looks at metadata (column names) and actual data (looking for patterns) to discover sensitive data. To reduce false positives it uses sampling to help users decide what is and is not sensitive. A notable differentiator is that Solix offers pre-packaged masking capabilities for Oracle and PeopleSoft application environments. While the Solix products run generically across most leading relational databases the company has implemented its masking algorithms natively in Oracle PL/SQL.

## Strengths

- Solix is one of only three vendors and the only pure-play provider to offer synthetic test data generation as well as subsetting.
- The pre-packaged capabilities and specific support for Oracle environments will be beneficial when it is Oracle-based data that needs to be generated or subsetted.

## Threats

- Solix does not currently support test data generation for NoSQL environments. However, the fact that the company's Big Data Suite includes capabilities for data lake management mean that this is likely to appear sooner rather than later.

SOLIX® Empowering Data Management

**Solix**
4701 Patrick Henry Dr., Bldg 20
Santa Clara, CA 95054

**www.solix.com/**

- The synthetic data generation offered by Solix is a relatively recent addition to the company's product line. As such, it is not as richly developed as some more mature competitive offerings. No doubt Solix will add the bells and whistles of its competitors in the fullness of time.

## Summary

Solix is ahead of many of its competitors in offering synthetic data generation. On the other hand, it is behind a number of its rivals in terms of the database environments it supports. The company clearly has the expertise needed to support NoSQL and other non-relational environments for test data: it is a question of when that will become reality.