Bloor

InBrief

Daniel Howard – Senior Analyst

IRI
Total Data Management

www.iri.com

2194 Highway A1A, 3rd Floor, Melbourne, FL 32937-4932, USA
Tel: +1 321 777 8889 Email: info@iri.com

Test Data Management in IRI Voracity

The company

IRI is a privately-owned independent software vendor. It was founded in 1978, and has international coverage. Its first product, CoSort, is a high-performance data transformation utility that remains at the heart of the company's offerings today, including IRI Voracity, a "total data management" platform that spans data discovery, masking, integration, migration, governance, and analytics.

What is it?

IRI Voracity contains two product suites that are relevant to test data management (TDM): IRI Data Manager Suite and IRI Data Protector Suite. The latter provides a selection of masking products (IRI FieldShield, CellShield EE, and DarkShield) suitable for various use cases, including TDM, that also come equipped with significant data discovery and classification capabilities. It also offers data classification, discovery, and masking as a professional service, aptly named Data Masking as a Service, or DMaaS. The former, on the other hand, contains IRI RowGen, which can be used to generate synthetic test data. In principle it also provides data subsetting, but in practice this is more typically delivered as part of the platform's broader data integration capabilities.

The Voracity platform, including the above products, is accessed through either IRI Workbench, a largely wizard-driven Eclipse interface backed by graphical modelling (displayed in ***Figure 1***), or via APIs. Licensing is flexible, with options available for Voracity as a whole as well as individual products and APIs. Database virtualisation is not offered directly, but is provided through integration with partner vendors Windocks and Actifio. Other partnerships support integration with provisioning and CI/CD pipelines – among other things – and recent collaborative efforts with Cigniti and ValueLabs are resulting in those company's more workflow-oriented front-ends being applied to the core Voracity engine, resulting in a smoother experience when they are deployed together (at least for organisations that require extensive approval processes as part of their data access).

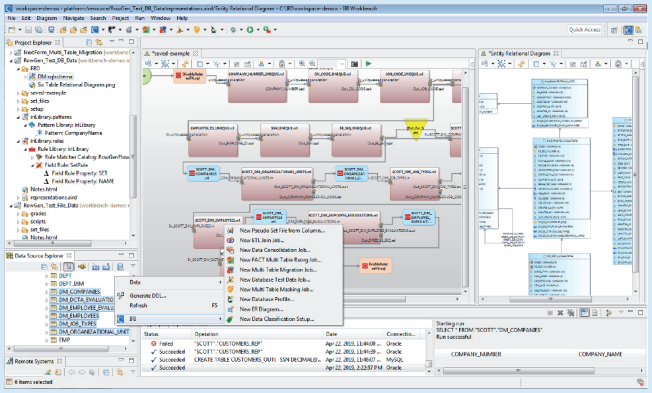
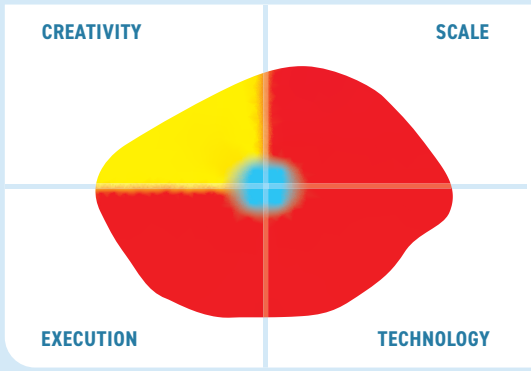


Figure 1 *– Test data synthesis of a relational schema using RowGen in IRI Workbench*



This **Mutable Quadrant** is derived from 13 high level metrics, the more the image covers a section the better. **Execution** metrics relate to the company, **Technology** to the product, **Creativity** to both technical and business innovation and **Scale** covers the potential business and market impact.

What does it do?

The Data Protector Suite provides (sensitive) data discovery and classification facilities in support of data profiling and masking operations (and thence TDM). It categorises your data against an extensible library of pre-configured or bespoke data classes, which can be tagged with varying levels of sensitivity and then married to appropriate masking or test data generation rules that are acted on at execution time. In this way, you can use Voracity to find and protect your sensitive data, allowing it to be used for testing.

Various discovery methods are available, including pattern matching, named entity recognition (which in turn leverages semi-supervised machine learning), column name matching, fuzzy and exact dictionary matching, path searching, font matching, character recognition, and coordinate matching. Any number of these methods can be used together for additional accuracy, and validation scripts can be employed to reduce false positives. Discovery results can be rendered as graphical reports; an example of this is shown in ***Figure 2***.

Masking is powered by the CoSort engine. FieldShield masks relational databases and flat files, CellShield masks Microsoft Excel data, and DarkShield masks structured, semi-structured and unstructured data (including images and documents) simultaneously and consistently. Static and dynamic masking are available, as is support for a variety of data sources. Various masking functions are provided out of the box, and you can build your own functions externally and integrate them via an API. You can also combine multiple discovery methods and/or masking functions together and apply them simultaneously. Masked data is consistent across all sources, while referential integrity is always maintained.

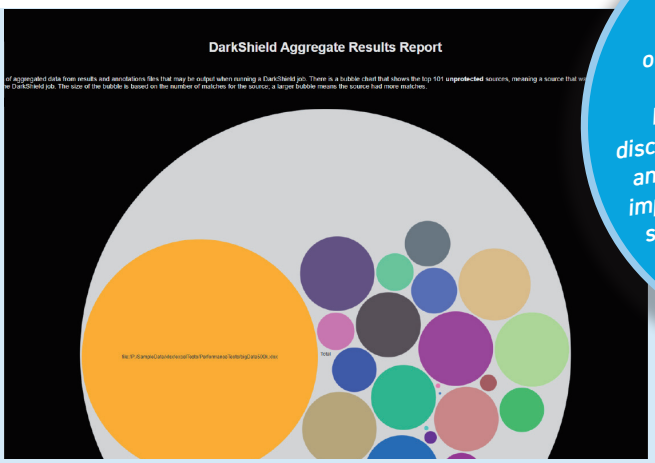


***Figure 2** – Aggregate results report for sensitive data discovery in IRI DarkShield*

> "Test data management (TDM) is a critical part of our agile SDLC, and is subject to data privacy regulations. Integrated data classification, discovery, anonymization, subsetting, and synthesis functions in Voracity improve our time-to-market delivery strategy, and help us comply with GDPR and similar laws."
>
> Capgemini Technology Services

RowGen provides synthetic data generation. It emphasises the customisation of test data, giving you fine-grained control over what, how and where your data is generated. For instance, it can generate test data using parameters you provide to it (including which class of data it should belong to) or select data randomly from one or more "set files" that have been prepared ahead of time, creating a holistic data profile for a person or other entity that does not exist but that has realistic attributes drawn directly from your data. Moreover, this extends past just what data you are generating and also encompasses how and where you are generating it (which means that you could, say, generate test data within a CI/CD pipeline).

Various generation functions are available for creating test data sets, including both the specific – such as national ID number generation – and the generic – generating data according to a predefined, weighted statistical distribution. There are multiple ways to customise the end results of these functions: test data can be generated in such a way that each value is unique, each value in a set file can be mandated to be used exactly once, and so on. You can even define your own compound data formats. Regardless, its production characteristics – including original data formats and sizes, value ranges, key relationships, and frequency distributions – are preserved. You can also generate test data in a variety of unstructured formats, including images and PDFs, based on predefined templates.

Subsetting is delivered via either RowGen or with Voracity's data integration capabilities. In either case, you can specify a driver table and trace its foreign key relationships to create a self-contained subset. Voracity gives you the option to follow these relationships "downhill" – only moving from parent to child – or to move through them in either direction. The former is faster, but the latter is more comprehensive. In addition to quantitative subsetting based purely on volume, you can also employ more qualitative methods that apply conditions to the initial data set in order to create a coherent subset (which will, again, be self-contained).All of this functionality can be executed as individual scripts or batch jobs, which can be created using various wizards, form editors, and mapping diagrams. They can then be executed from within IRI Workbench, the command line, or a partnered database virtualisation environment such as Windocks. In the latter case in particular, Voracity and Windocks can be used to create sanitised clones of your production data in on-demand, self-service, containerised, and virtualized repositories.

APIs are provided, meaning that Voracity TDM functions are also operable as part of an external pipeline, and can be invoked directly from within your CI/CD platform processes, either on-premises or in the cloud. The test data created by Voracity's processes can be exported to many databases and file formats, including spreadsheets, PDFs and images.

Finally, IRI Ripcurrent, a real-time database event processing module, was recently added to Voracity. Ripcurrent offers incremental data replication by detecting and acting on changes to relational database tables in real-time. This works by monitoring log events for inserts, updates, deletes and schema structural changes, then mapping the data on-the-fly and/or issuing alerts. Applied to TDM, it can be used to refresh your test data environment by carrying out both data replication and masking processes automatically as soon as a corresponding production environment is changed.

Why should you care?

IRI's subsetting, masking and synthetic data generation capabilities are all highly competent. The ability to create representative synthetic data sets via analysis is particularly notable and useful, as is Ripcurrent's automatic and real-time refresh of your test data. That said, in this paper we have only been able to scratch the surface of the product's capabilities. There is a significant depth of functionality here: IRI has been organically growing its technology for over 40 years, and it shows. If you would like to learn more, we refer you to our recent series of articles on IRI and Voracity, which explores several of the topics touched on in this report in greater detail. We are also told that IRI is working on implementing generative AI as part of its sensitive data discovery and synthetic data generation capabilities, although the details of this have yet to be announced.

What is more, TDM is only one aspect of Voracity. It is billed as a total data management platform, and to that end it offers a wealth of other capabilities – data integration, governance, quality, and so on – that stretch beyond just TDM. Moreover, these capabilities (including TDM) are offered through a unified and user-friendly interface, complete with wizards, visual programming and so on. This makes it easy to use each individual product and to shift your attention from one product to another. Integration with CI/CD pipelines is also a useful feature, enabling Voracity to automate both the production and consumption of test data.

The company's partnership with containerised database virtualisation vendor Windocks is particularly notable, and its other relevant partners, including Actifio, CommVault, Cigniti, and ValueLabs, should be considered as well.

The bottom line

IRI Voracity is both a data management platform and TDM solution, with many elements of the former being highly applicable to the latter. Ripcurrent is a particularly compelling example of this kind of applicability. The end result is an effective and versatile solution for TDM.