BLOOR RESEARCH
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

Sensitive Data Discovery in IRI Voracity

The company

IRI is a privately-owned independent software vendor founded in 1978. Its offices are in Florida and it relies on a partner network of resellers for international coverage in 40 locations throughout the world.

The company's first product, CoSort, was – and is – a high-performance data transformation utility that has since been spun off into various data management offerings that collectively cover data integration, quality, migration, masking, and synthetic data generation, among other things. Moreover, the CoSort engine remains at the heart of IRI's structured data processing. Notably, it powers Voracity, a broad data management platform designed to accelerate and consolidate common work in data discovery, integration, migration, governance, and analytics.

What is it?

IRI Voracity is a data management platform that offers its core capabilities through two product suites: IRI Data Manager Suite, and IRI Data Protector Suite. In particular, the latter provides a selection of data masking products (namely IRI FieldShield, CellShield EE, and DarkShield, plus a services option that leverages them called DMaaS) which come equipped with significant data discovery capabilities. These tools serve a variety of purposes, not the least of which is to find and protect your sensitive data.

Each product in the suite is designed for different use cases. FieldShield, for example, works primarily with structured databases and flat files, although it additionally supports Excel sheets, COBOL files, and a few other file types. It is most often used for test data management in relational environments, but also provides re-identification risk scoring (for compliance purposes) and powers Ripcurrent, a Voracity component for enabling CDC (Change Data Capture) on your relational data, allowing you to incrementally mask and refresh it. CellShield works exclusively with and within Excel spreadsheets, while DarkShield – the newest of these offerings – can work with structured, semi-structured and unstructured data sources, on-premises or in the cloud.

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.

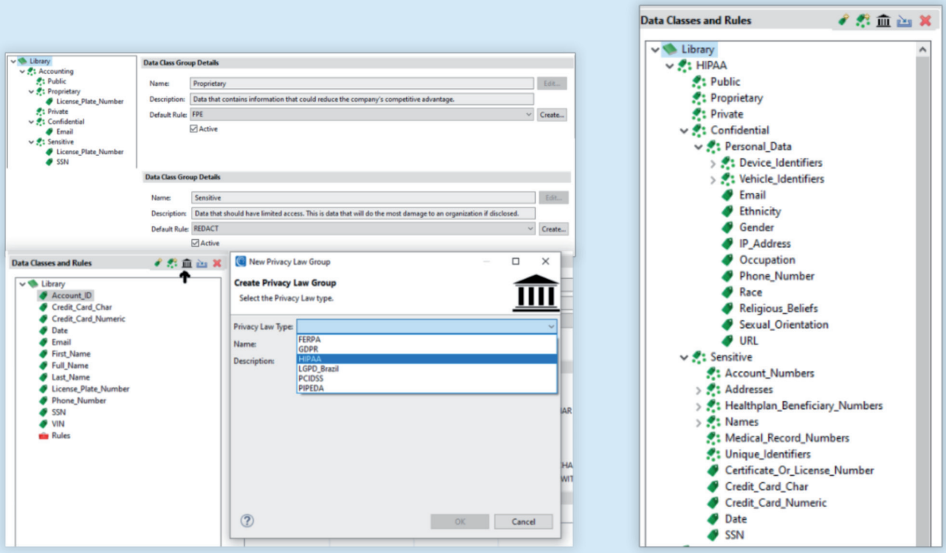

Figure 1 *– Data classification in IRI Workbench*

The latter is typically the product of choice for achieving more general, enterprise-wide data compliance, or for integrating discovery and masking into your existing applications.

The Voracity platform, including the above products, can be accessed through either IRI Workbench, a largely wizard-driven (and relatively friendly) Eclipse interface backed by graphical modelling, or via APIs. These APIs allow you to fold sensitive data discovery and masking into your application, testing, and other data pipelines. FieldShield and DarkShield jobs can also be executed from the command line. Licensing is flexible, with options available for Voracity as a whole as well as individual products and APIs. IRI also partners (and integrates) with a number of other vendors, adding additional capabilities to its offering.

> *"Our experience with millions of unstructured files confirms the need to identify and mitigate the data privacy risks within them."*
> GDPR Tech

What does it do?

The Data Protector Suite products provide various data discovery and profiling capabilities. In general, they enable you to classify your data against a centralised library of data classes shared between all of the Shield products, which can in turn be married to masking or other anonymisation rules (such as redaction or encryption) when they correspond to sensitive data (see ***Figure 1***).

These rules are acted on at execution time, ensuring that the associated sensitive data is protected. Each data class can be discovered against using data and/or location matchers, which are used to find matching data in your system by examining the contents and structure of your data assets respectively. This means that IRI can automate the process of both finding and anonymising your sensitive data, by discovering it using one or more matchers, associating it with the appropriate data class, and anonymising it at execution time.

There are also considerations for performance that have been built in. For instance, tables that have already been scanned will be skipped during repeated discovery phases, and you can choose to exclude specific tables or data classes from the process entirely. Multi-threading and load-balancing are also supported, for vertical and horizontal job scaling as the data volume expands.

An impressive range of discovery methods can be used as part of these capabilities, including lookup value or pattern matching, NER (Named Entity Recognition), column name matching, fuzzy or exact dictionary matching, path searching, cell matching, font matching, character recognition, and coordinate matching (the latter two mostly for images). NER in particular uses semi-supervised machine learning – based on your choice of Apache OpenNLP, PyTorch, or TensorFlow – to enable more sophisticated and effective language analysis of highly unstructured data. This can be necessary for more complex discovery use cases: for example, examining person data from a hospital and distinguishing between staff and patients.

Any number of these methods can be used in concert with each other to improve the accuracy of your results and reduce false positives, although note that some of them are only applicable to certain kinds of data and/or are only applicable in DarkShield. There is also a configurable matching threshold for discovery, allowing you specify how sure you want to be before settling on a result. Moreover, IRI is committed to developing new matching methods. Most recently, it implemented AI-powered database classification and signature detection in collaboration with DeepLobe, an AI platform that makes extensive use of computer vision.

In terms of the data classes themselves, a default set – including commonly-used but generic classes like

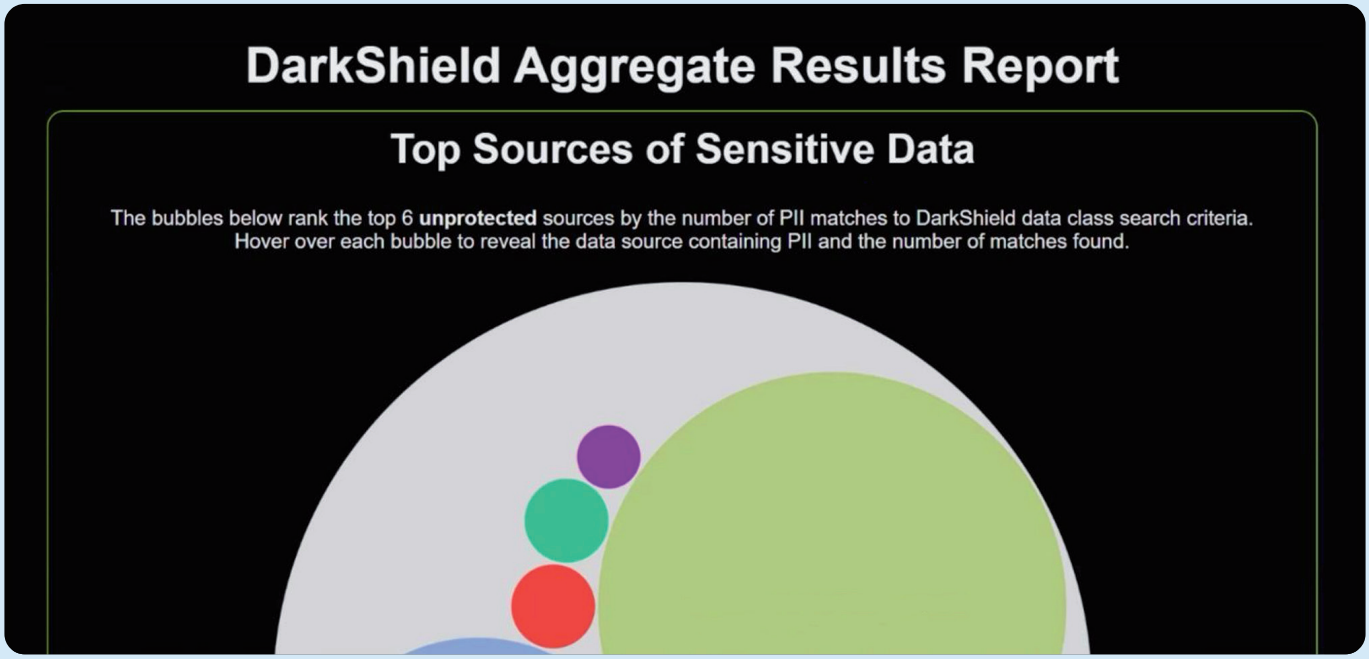

Figure 2 *– IRI DarkShield discovery report*

name, email, credit card number, and so on – are provided automatically. In addition, the product can automatically generate data classes applicable to a number of regulations, such as GDPR and HIPAA.

Data classes can be grouped together, and data class groups can be categorised further by assigning them specific sensitivity levels and/or associating them with one or more of the aforementioned regulations. Data class masking rules can be given a priority level, which determines which functions takes precedence if a given data asset seems to fit into multiple categories at once.

> *"Fortunately, the search methods and masking functions in IRI DarkShield specifically and Voracity generally help us get control of these hidden risks."*
> GDPR Tech

Both FieldShield and DarkShield offer the additional benefit of a robust set of data discovery reports, as well as audit trails for your data masking jobs. For instance, the DarkShield discovery dashboard automatically aggregates and organises classification information into a customisable HTML report (see ***Figure 2***). This report is highly visual, prominently displaying various bubble and pie charts that highlight your top sources of sensitive data, your most common data classifications, and so on. You can also export search results to third-party products for display, analysis, or other actions.

Why should you care?

IRI Voracity uses a robust data classification infrastructure that manages sensitive data definitions, and assigns discovery and masking methods to them, centrally. It offers a healthy range of discovery methods ranging from the simple to the sophisticated, and its applicability to highly unstructured data such as image files is particularly notable.

Moreover, Voracity is billed as a total data management platform, and to that end it offers a wealth of additional capabilities – data integration, migration, quality, wrangling, reporting, and so on – that will frequently tie into, and either augment or be augmented by, data discovery (and, to a lesser extent, masking) in one way or another. These capabilities 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.

These advantages carry over to data discovery and data masking, at least if you plan to leverage these technologies through Workbench. That said, even if you don't, you can simply benefit from the flexibility, integration and automation offered by an API-driven approach instead. By way of example, data discovery through the DarkShield API can be coupled with test data generation using IRI RowGen to replace values in images and documents with synthetic (but realistic) data and fonts, providing more safety for applications and processes that handle those sorts of files.

IRI takes a moderate but reasonable approach to generative AI and LLMs. While it does not offer any particular capabilities for handling data flowing into or out of an LLM, it argues that it does not need to, as it is as capable of discovering and protecting sensitive data in that context just as in any other. Indeed, Voracity jobs can be integrated into AI-driven workflows via either command line or API calls, and can thereby be used to discover and mask sensitive data before it is fed into various AI processes. This includes, but is not limited to, LLMs and chatbots. At the same time, the company is certainly willing to take advantage of AI, shown in its support for various types of AI-driven recognition models, which cover sensitive relational data, named entities, signatures, and soon – we are told – handwriting and multinational addresses.

The bottom line

IRI justifiably positions Voracity as a total data management platform. As a solution for data masking and data discovery, either for sensitive data or not, it is both highly competent and rather flexible in how you can interact with it. Its approach to AI, particularly for sensitive data discovery in disparate source contexts, is both practical and expanding. In short, whether you want a discovery solution that comes integrated into a larger platform, or one that works as a standalone solution, IRI Voracity is likely to satisfy.