Bloor

InBrief

Daniel Howard – Senior Analyst, Information Management and DevOps

Sensitive Data Discovery and Masking in IRI Voracity

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

The company

IRI is a privately-owned ISV. 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.

> "The multiple source- and silo-compatible search methods available in IRI Voracity platform "shield" software are proving to be effective mechanisms for finding and reporting on the locations of sensitive data enterprise-wide. The integrated classification and de-identification of discovered data also supports key regulatory compliance, breach mitigation, and DevOps initiatives."
>
> SecureITLab

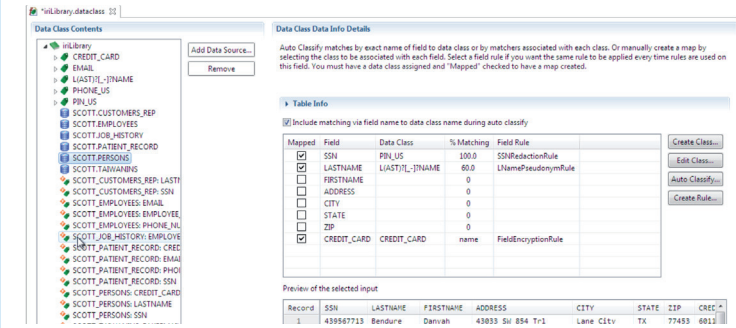

Figure 1 – Data classification in IRI Workbench

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) that also come equipped with significant data discovery capabilities. This functionality can be used for a variety of purposes, not the least of which is to find and protect your sensitive data.

The Voracity platform, including the above products, can be accessed through either IRI Workbench, a largely wizard-driven Eclipse interface backed by graphical modelling, or via APIs. 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. These can variously add additional capabilities to the IRI offering as well as provide enhanced support for provisioning and CI/CD pipelines.

What does it do?

Masking in Voracity is rule-based and powered by the CoSort engine. FieldShield masks structured databases and flat files, CellShield masks Excel sheets, and DarkShield can search and mask structured, semi-structured and unstructured data sources simultaneously. Several dozen static masking functions are available for FieldShield and DarkShield, and about half of those are available in CellShield as well. In static operations, masked data is kept consistent across multiple data sources so that referential integrity is always maintained. Dynamic data masking is also available.

In addition to data masking, the various Data Protector Suite products provide data discovery and profiling capabilities. This enables you to classify your data against a centralised library of either pre-configured or bespoke data classes shared between all of the shield products, which can in turn be married to masking rules 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 also be equipped with a search methodology that is used

> "Our experience with millions of unstructured files confirms the need to identify and mitigate the data privacy risks within them. Standalone and embedded spreadsheets, Word and PDF documents, image files in multiple formats, as well as logs and emails, are strewn with PII unknown to our customers. These needles in historical or operational customer haystacks need to be found and blunted. Fortunately, the search methods and masking functions in IRI DarkShield specifically and Voracity generally help us get control of these hidden risks."
>
> GDPR Tech

to locate matching data in your system. This means that when set up correctly IRI can effectively automate the process of finding and anonymising your sensitive data: it will discover your sensitive data using the aforementioned search, associate it with the appropriate data class, and mask 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.

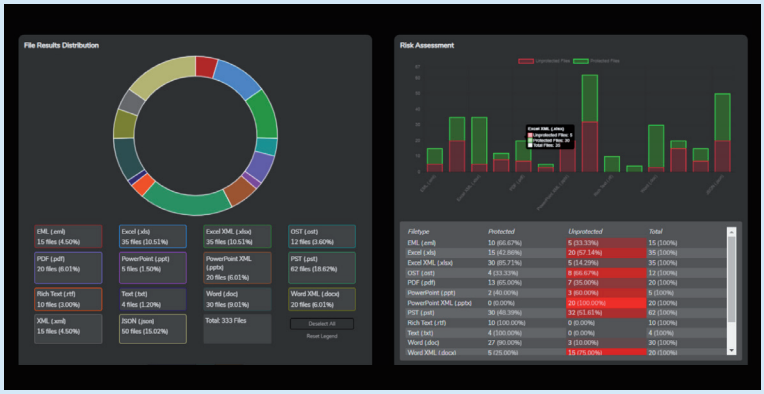

Figure 2 – IRI Voracity dashboards

An impressive range of discovery methods can be used as part of this these capabilities, including lookup value or pattern matching, NER (Named Entity Recognition), column name matching, fuzzy or exact dictionary matching, path searching, facial recognition matching, font matching, character recognition, and coordinate matching (the latter two mostly for images). NER in particular uses semi-supervised machine learning to enable more sophisticated and effective language analysis of highly unstructured data. In addition, any number of these methods can be used in concert with each other to improve the accuracy of your results. There is also a configurable matching threshold for discovery, allowing you specify how sure you want to be before settling on a result.

Moreover, there are two ways to consume Voracity's discovery and masking capabilities. You can go through Workbench – which has the advantage of a relatively friendly, wizard-driven user interface coupled with visualised reporting, as shown in ***Figure 2*** – or you can leverage them directly through an API. In the latter case, this essentially allows you to use Voracity as a discovery and masking engine that underpins your other data pipelines. This has obvious (and positive) implications for integration and automation.

Why should you care?

IRI Voracity uses a robust architecture for managing your data classes that both manages data class definitions, and assigns discovery and masking methods to them, centrally. It offers a healthy range of discovery methods running 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, governance, quality, 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 will 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 these sorts of files.

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. In short, whether you want a solution that comes integrated into a larger platform, or one that works as a standalone engine, IRI Voracity should satisfy.

FOR FURTHER INFORMATION AND RESEARCH CLICK HERE