IRI FieldShield
PII / PHI Classification & Masking

Getting Started with FieldShield

IRI
Total Data Management

What is FieldShield?

IRI FieldShield® is Gartner-advised software for the discovery, classification, deterministic masking, re-ID risk scoring, and auditing of Personally Identifiable Information (PII) and other sensitive data in 1NF databases (DBs) and structured files, at rest or in motion. FieldShield is used for a variety of objectives, including: data protection and breach nullification, data privacy law compliance, test data provisioning, and data governance. More information on the case for data masking is provided in this white paper.

FieldShield technology was first developed in 2007 when IRI added field-level data masking functions to IRI CoSort data transformation jobs. The CoSort utility has been in worldwide use since 1978 to perform and speed large legacy sort, database, DW ETL, and BI operations.

IRI spun off FieldShield from CoSort in 2011, where today it is one of three organically grown static data masking products in the IRI Data Protector suite along with IRI CellShield EE for Excel sheets and IRI DarkShield for semi-structured and unstructured data sources. Also available from IRI is "Data Masking as a Service" where these technologies are leveraged by experienced HQ security engineers who remotely access client data under NDA to perform the same functions at standard rates.

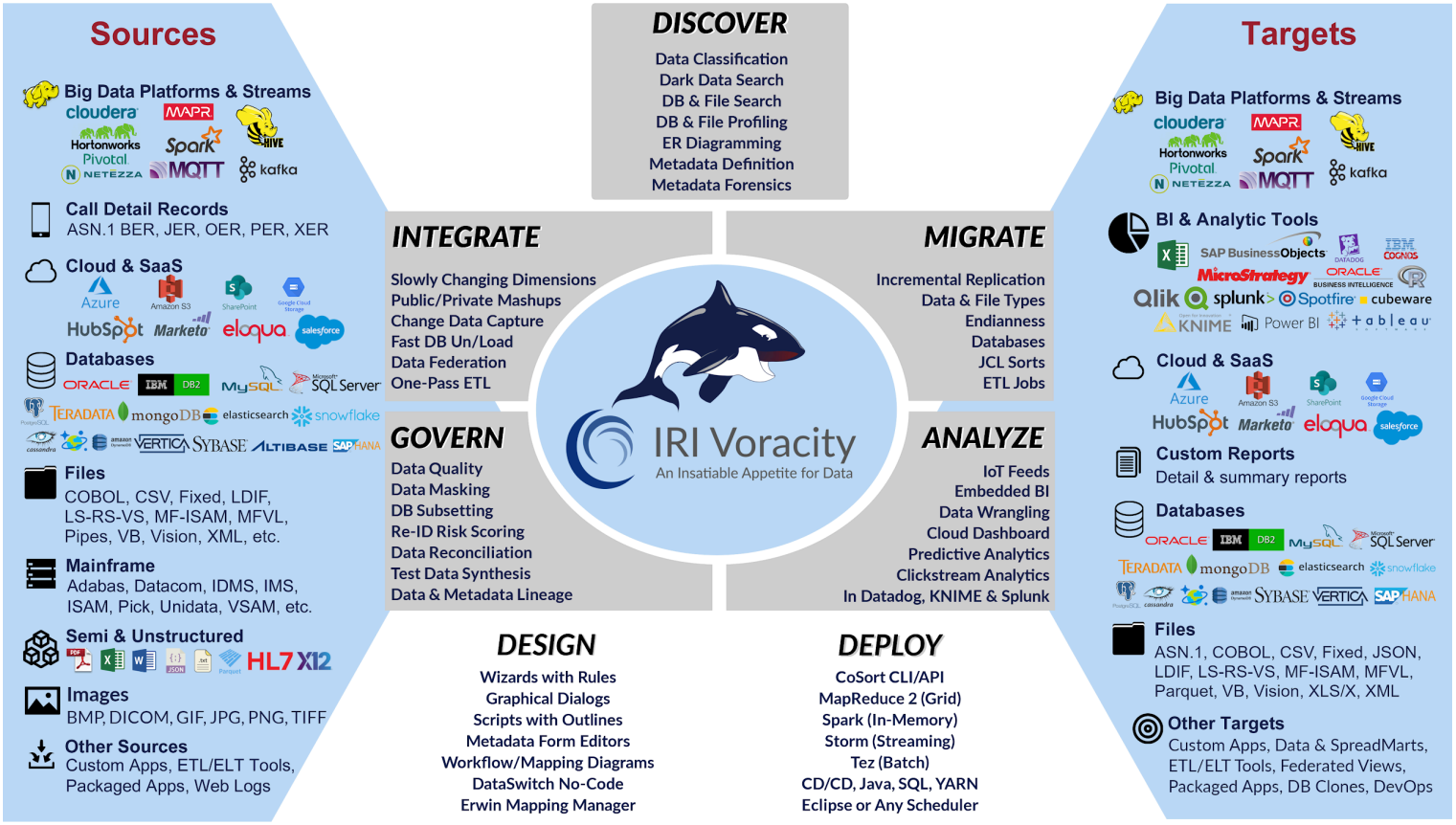
Data Masking in Broader Contexts

All three 'shield' products are also included in the IRI Voracity® total data management platform supporting the discovery, integration, migration, governance, and analytics of data in multiple sources. These products share with Voracity the same Eclipse™ IDE, IRI Workbench, for PII/data discovery and job design, and common un/masking functions for multi-source use.

In addition to being a unified job design, deployment, and metadata management environment for multiple data-stakeholders, Workbench is also a convenient place for cross-database profiling and SQL administration, and to coordinate activities with other Eclipse plugins like Git.

More importantly, the same scripts with FieldShield data layouts and masking specifications can also be used in -- and run in the same job and I/O pass with -- any type of Voracity data manipulation and movement operation. This means that you can add data masking to data transformation and reporting jobs, or mask data: in DW ETL and BI data wrangling operations; during data cleansing and validation; or, as you migrate and replicate data, subset databases or synthesize test data, pivot, display deltas, etc.

The Voracity platform diagram below shows the inclusion of FieldShield-related functionality under DISCOVER and GOVERN (data masking and re-ID risk scoring), as well as DarkShield and CellShield, per the many data sources and targets supported in its larger environment:

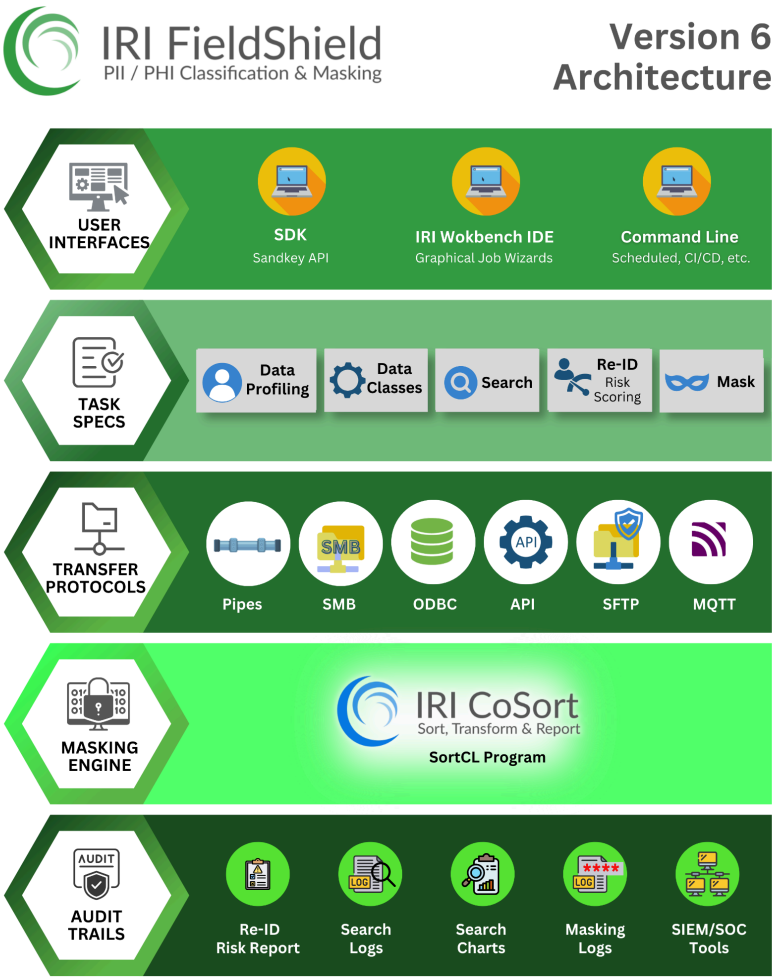


For the purposes of this solution summary however, only FieldShield features addressing common structured and static data masking requirements are covered. Contact IRI for more information about the use of FieldShield in dynamic data masking (DDM) scenarios.

Information on broader FieldShield-compatible data management operations is provided in this overview booklet on the IRI Voracity platform.

FieldShield Architecture

FieldShield has three main components:



1. IRI Workbench, a free graphical user interface (GUI) and Integrated Development Environment (IDE) built on Eclipse, used to: profile and administer DBs via JDBC; discover and classify PII in multiple data sources; auto-create, modify, manage, and optionally run or schedule 4GL data masking job (task or batch scripts); score the risk of re-identification after masking; store and use data classes and masking rules; manage metadata assets; and, configure all other tasks supported in IRI Voracity;

2. The CoSort SortCL command-line data transformation program written in C for parsing and executing the FieldShield 4GL data masking scripts. These scripts also define the layout of each structured source/target and many other data manipulation functions like ETL, cleansing and reporting you can uniquely execute *as* you mask the data.

3. An optional Software Development Kit (SDK) with libraries and documentation for optional dynamic masking calls in C/C++, Java or .NET programs, using the same encryption, hashing, and redaction functions in the main package.

Recommended Hardware

IRI Workbench runs on Windows, x86 Linux, and MacOS, while the FieldShield masking engine runs on all Windows, Linux and Unix systems, on-premise or in cloud systems like Amazon EC2, OCI, or Microsoft Azure. A minimum of 4GB of RAM is recommended on both client and server machines, per this case. Users with thousands of tables (e.g., with PeopleSoft) should upgrade server RAM to 64GB.

Compatible Technologies

FieldShield metadata is generated by erwin Mapping Manager, DataSwitch, and the Meta Integration Model Bridge (MIMB), and is also exposed for application use through a Java API ("IRI Gulfstream"). FieldShield job scripts can also mask data during Actifio, Commvault, and Windocks DB cloning ops. FieldShield encryption keys can be managed in Azure Key Vault and Townsend Alliance Key Manager. Contact IRI if you require key management integration with another cloud key vault or HSM technology.

FieldShield Workflow

Following are standard activity steps, a summary description, and a linked schematic diagram:

Download & Installation

1. Once you have installed IRI Workbench, and licensed the FieldShield masking executable, make both JDBC and ODBC connections to each database to be masked. For files, identify local or remote folder locations.

Data Profiling

2. Use wizards in the data discovery menu to profile your data sources, build E-R diagrams to model RDB schema, perform referential integrity checks, and search for specific items matching a pattern. You can find much more however using data classification (see below).

Data Classification

3. Define data classes (e.g., passport #, name, phone), or class groups (e.g., citizen PII) as needed, and assign to them search methods and masking functions. Run search wizards to find PII across file folders or DB schema. The search jobs produce log reports, dashboards, and a Data Class Map you can review before the bulk data masking wizards use it to consistently apply your rules to your data classes (which ensures referential integrity).

Data Masking

4. Run the *New Data Class Map DB Masking Job ...* or the *New Data Class File Map Masking Job ...* wizard to apply your chosen data masking functions to each data class depending on its need for reversibility, realism, uniqueness, consistency, and security. You can also configure FieldShield data masking jobs in mapping diagrams, 4GL scripts, or an IRI API.

Review Masking Results

5. Jobs once created are easily run and modified for iterative testing, where output can first be virtualized to the console, file or sample tables for review, modification, sharing, and re-use. SQL query and update logic can be inside scripts or workflows for real time or incremental masking operations.

Scoring Re-ID Risk

6. An included risk-scoring wizard statistically measures the likelihood that a masked data set (table) can still be used to identify an individual based on quasi-identifying information in the row.

Remediation

7. Further masking jobs can address the risks exposed in the scoring report by blurring or bucketing quasi-identifying values like birthdate and age. These techniques render the data anonymous, but still fit for purpose.

Manage Job Artifacts

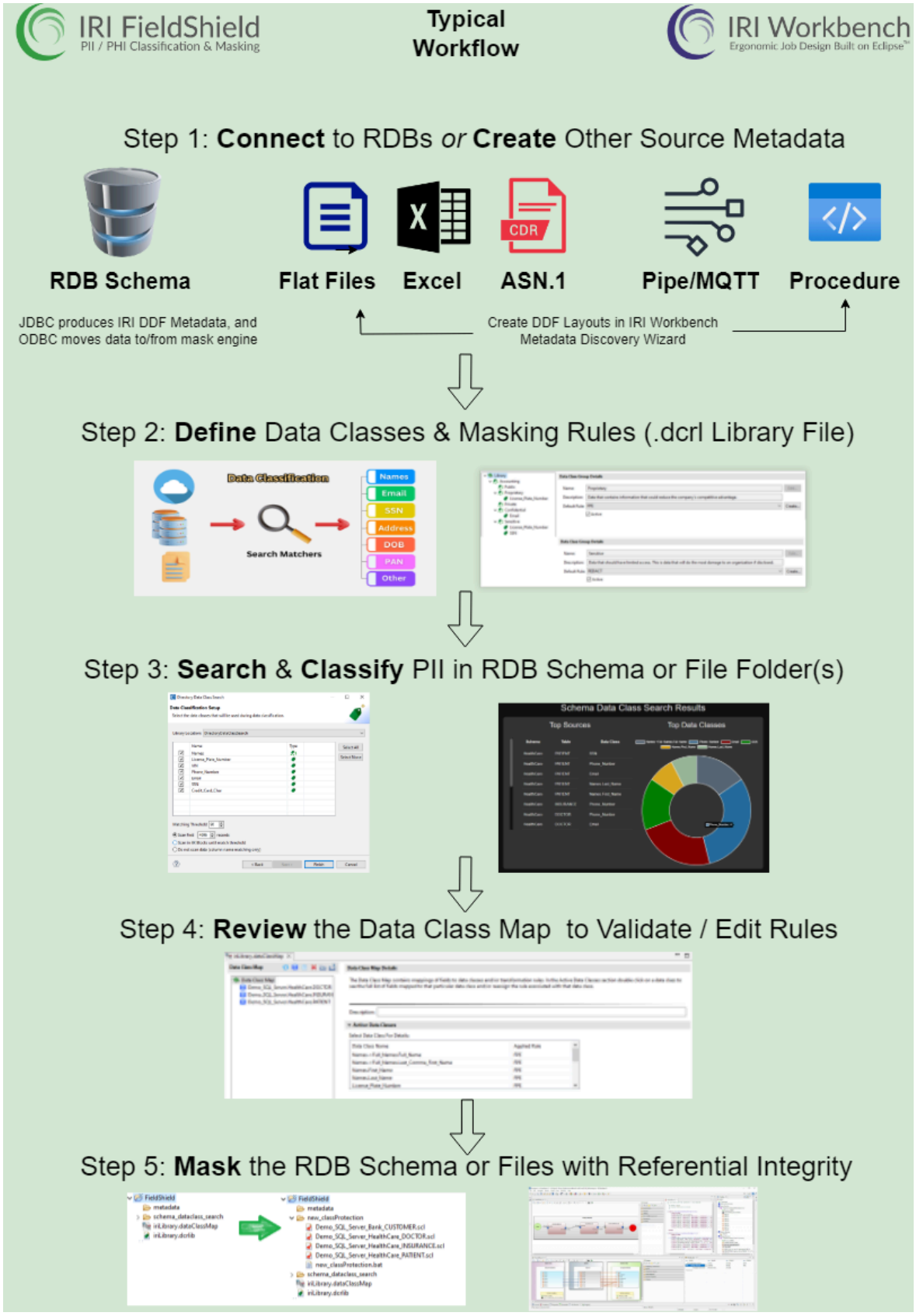
8. As you gain success and confidence in the above processes, and begin to satisfy production data masking needs, you should also take stock of your job performance and metadata assets (source layouts, task and batch scripts, encryption keys, etc.). Use an IRI Workbench plugin like Git to secure, share, version control, and track changes to your metadata.

Deployment, Monitoring & Maintenance

9. During and after the go-live dates, follow IRI recommendations to monitor performance, manage updates to data, conditions and software and address new use cases or audit requirements.

FieldShield Operational Schematic

The diagram below represents the first half of, and most commonly performed, steps described above.



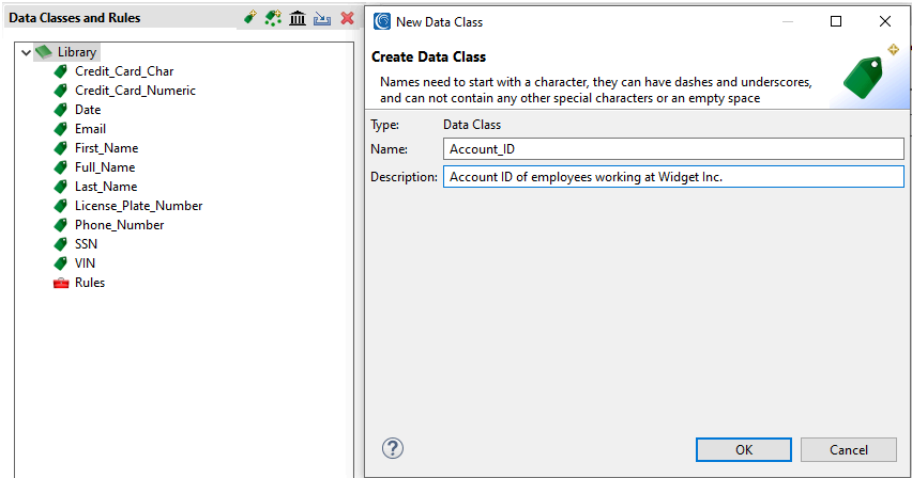
The sections that follow describe these capabilities in broad terms. They also provide links to more details and how-to articles which are indexed at https://www.iri.com/services/training/courseware. The most relevant sections for FieldShield users are Data Discovery & Classification, and Data Masking.

A separate .pdf FieldShield User Manual documents data masking script syntax in full detail.

FieldShield evaluators and licensees also have access to online help in IRI Workbench which includes Welcome section content, in-context dialog help, getting-started cheat sheets, user guides and troubleshooting. An offline .pdf manual on masking script syntax is also provided.

Various data classification and masking examples are also on YouTube. Guided vendor support is also provided through email and regular interactive online sessions arranged here.

PII Discovery and Classification



Data Classification usually will take place before a FieldShield Job is created through either a *Schema Data Class Search* **or** *Directory Data Class Search* wizard.

These wizards use data classes extracted from your Data Class & Rule Library to perform data classification.

The results of data classification is a file called a Data Class Map (.dataClassMap). The Data Class Map file stores mappings of fields from structured sources (table or file) to masking rules. These mappings of masking rules to fields are defined as a result of the data classification process where Data Classes match on PII and allocate the masking rule associated with the Data Class in question.

After a Data Class Map is produced, the *New Data Class Map DB Masking Job* wizard or *New Data Class Map File Masking Job* wizard will extract the mappings to produce FieldShield job scripts.

Note that a Data Class Map is *not* used in FieldShield jobs scripts built in the first, *New Masking Job …* wizard. That wizard is more appropriate if you only have one or two file or table sources, because:

1) the metadata for each source must first be defined, which gets more time-consuming process as the number of sources increases;
2) you want to assign masking rules manually in the Target Field Layout Editor dialog; or,
3) only a limited form of data classification (matching location matchers to column names) can be used as an alternative, or in addition, to manual rule assignment in the Target Field Layout Editor.

Data Masking Jobs

FieldShield static data masking jobs are usually designed and managed in -- and can run from – IRI Workbench. This is also where you can manage your database administration, profiling, modeling, classification, and searching work, plus other tasks you can combine with FieldShield (like ETL).

Whether you have pre-defined your data classes and masking rules beforehand or not, you can build FieldShield jobs either automatically (via wizard) or manually (via scripts, API, or Workbench palette). The masking jobs serialize into self-documenting task scripts (or a batch of several), written in the data definition and manipulation 4GL common to all IRI structured data processing jobs, called SortCL.

Each job can specify the layout (and masking) of one or more data sources and targets (e.g., RDB tables) at a time. Masking rules can be defined ad hoc in the wizards or scripts and saved. Rules mapped to fields from preceding data classification steps, or applied in-wizard to like-named columns, are translated into FieldShield Job scripts (.fcl) with transformation methods applied at the field level.

Masking Job Prerequisites: Data Classification

1. Each FieldShield project (folder) project must contain a Data Class & Rule Library.
2. Use default pregenerated Data Classes and Rules, or provide custom entries into the library.
3. Depending on the FieldShield wizard a Data Class Map may be required. To create a Data Class Map, perform either a Schema Data Class Search or a Directory Data Class Search to generate a .dataClassMap file applying rules to fields based on your data classifications.

Masking Job Creation Methods

1. There are three FieldShield Wizards in IRI Workbench that generate job scripts and a batch file. The first wizard, New Masking Job wizard, requires a Data Class & Rule Library with Data Classes and Masking Rules present to perform either manual Rule assignment to fields or a simplified form of Data Classification (Location Matchers only). From the assignments in the wizard, a FieldShield Job script is created.
2. The New Data Class Map DB Job and New Data Class Map File Job wizards are best for multi-source maskings and require a Data Class Map to produce the FieldShield Job scripts.
3. in the ETL-style workflow and mapping diagram design palette;
4. in the color-coded, syntax-aware editing window or any other text editor; or,
5. an API for driving parameters automatically into Workbench scripts from Java programs.

Masking Job Modification Options

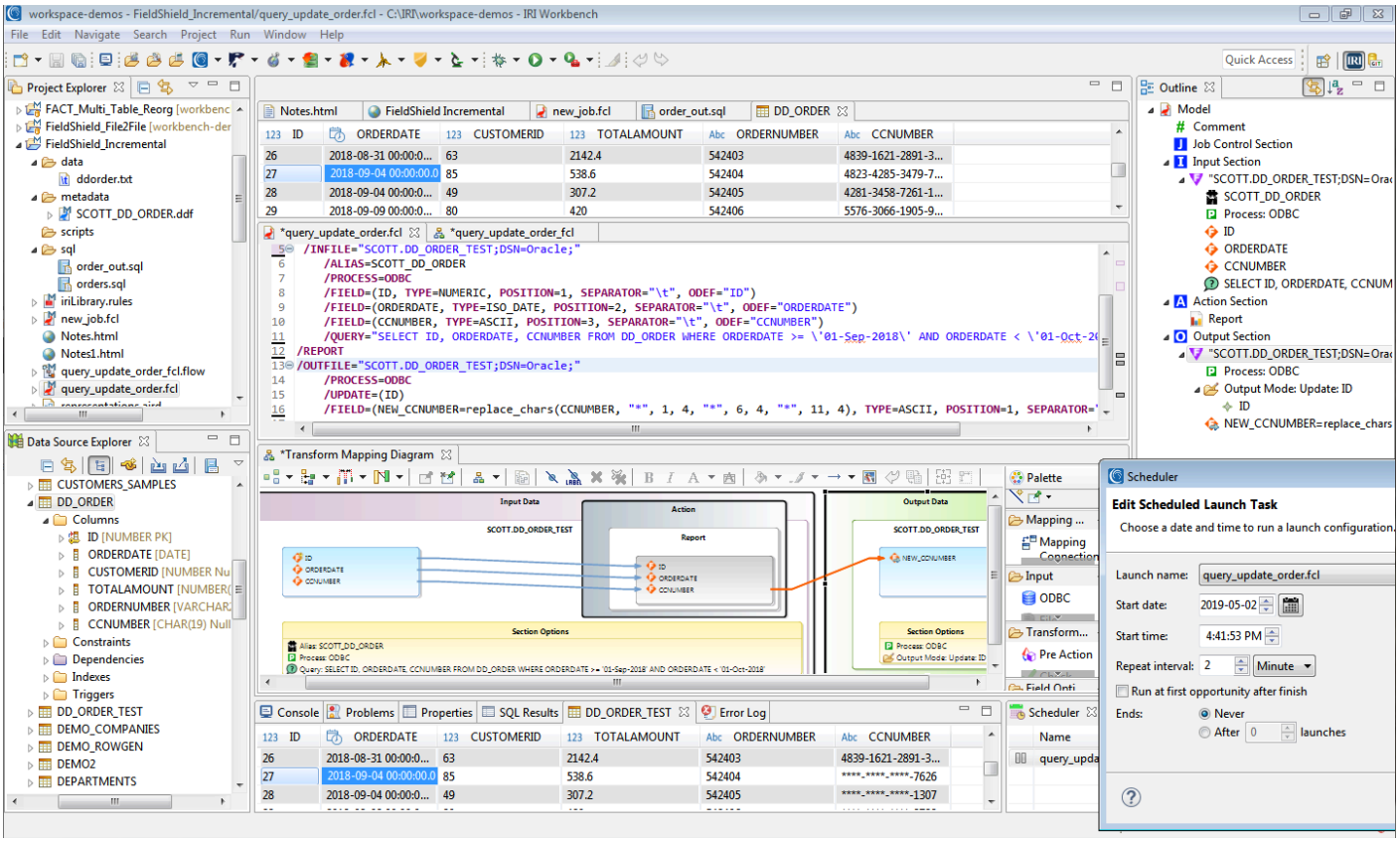
1. Manually, using the Workbench script and DDF editor or any plain text editor;
2. From the Workbench job outline views that interact with scripts or DDFs (or a DDF form editor);
3. Fom context-specific graphical dialogs (with help) launched via right-click in the script view; and,
4. In the Workbench transform mapping diagram and its properties view.

Masking Job Execution Options

Again, masking scripts are parsed by a local or remote FieldShield (SortCL) executable licensed on any physical or virtual machine running Windows, Linux or Unix, on-premise or in the cloud. As demonstrated in articles here, FieldShield task or batch scripts can *run* from:

1. the command line
2. any program that can make a system call (e.g., a batch program, ETL tool, or CI/CD pipeline)
3. IRI Workbench, ad hoc or scheduled
4. any other job scheduler (e.g., Stonebranch UAC, Autosys, cron)

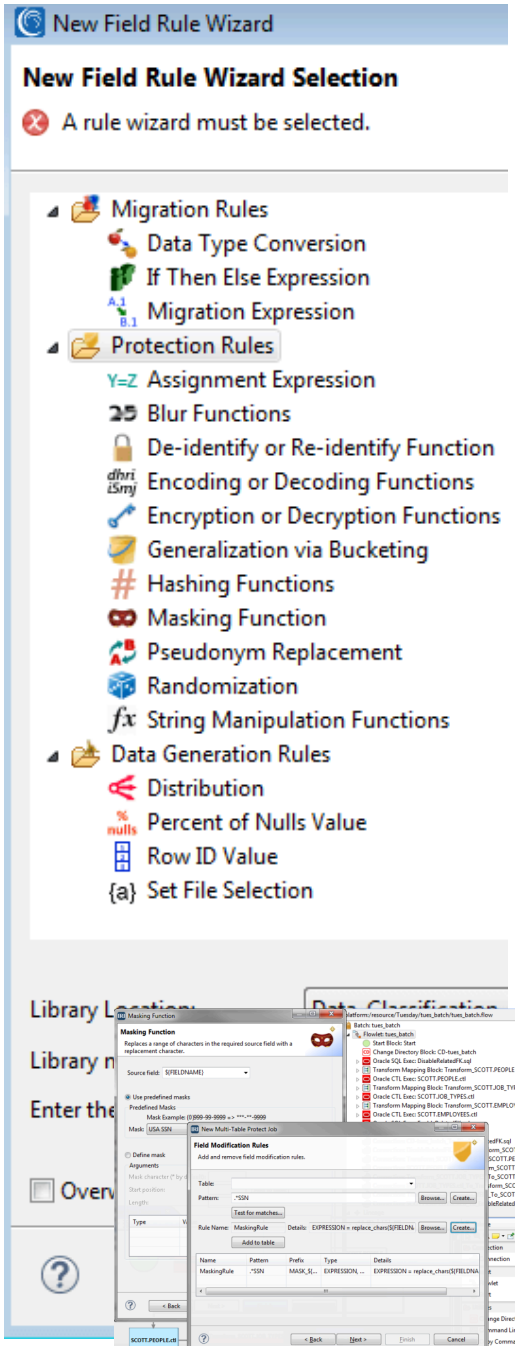
All masking jobs use artifacts like job scripts, data definition files, data classes, and masking rules, and can generate an XML audit log, application statistics, and both error and performance logs. Such assets are managed easily in IRI Workbench and plugins like EGit.



Data Masking Functions

The protection method you choose for each field depends on its need for security, reversibility, appearance, and speed. Available functions are within multiple categories:

1. multiple, NSA Suite B and FIPS-compliant encryption (and decryption) algorithms, including *format-preserving* encryption
2. MD5, SHA-1, SHA-2 hashing
3. ASCII de-ID (bit scrambling)
4. binary encoding and decoding (less secure)
5. blurring and bucketing (anonymization for HIPAA)
6. random value generation or selection
7. redaction (full or partial string masking)
8. reversible and non-reversible pseudonymization
9. filtering or omission (erasure for GDPR)
10. Default literal or lookup value replacement
11. byte shifting and (sub)string functions
12. tokenization (for PCI DSS compliance)
13. custom (user written and linked) routines

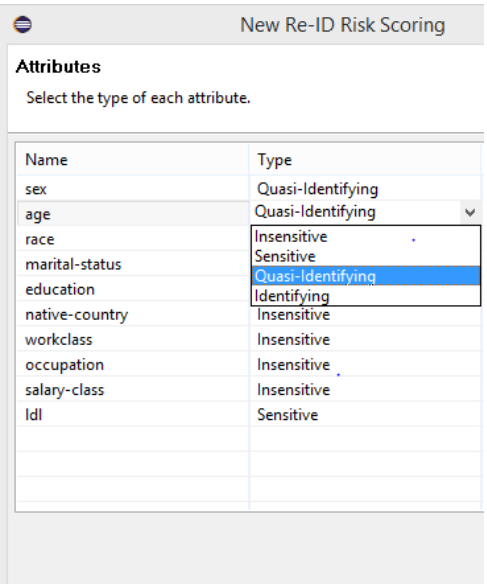


Both reversible and non-reversible techniques are available. Non-reversible methods like data redaction or filtering removes any computational basis for deriving the original value.

The rules in the selection wizard shown to the right can be applied in data classification and new masking job wizards, as well as ad hoc to fields treated individually in transform mapping diagrams or the Target Field Layout Editor. The latter is invoked from the job script editing window or dynamically linked script outline view.

To summarize, masking functions can be used ad hoc or saved as rules and applied to fields, defined and searched data classes, or across pattern-matching column names.

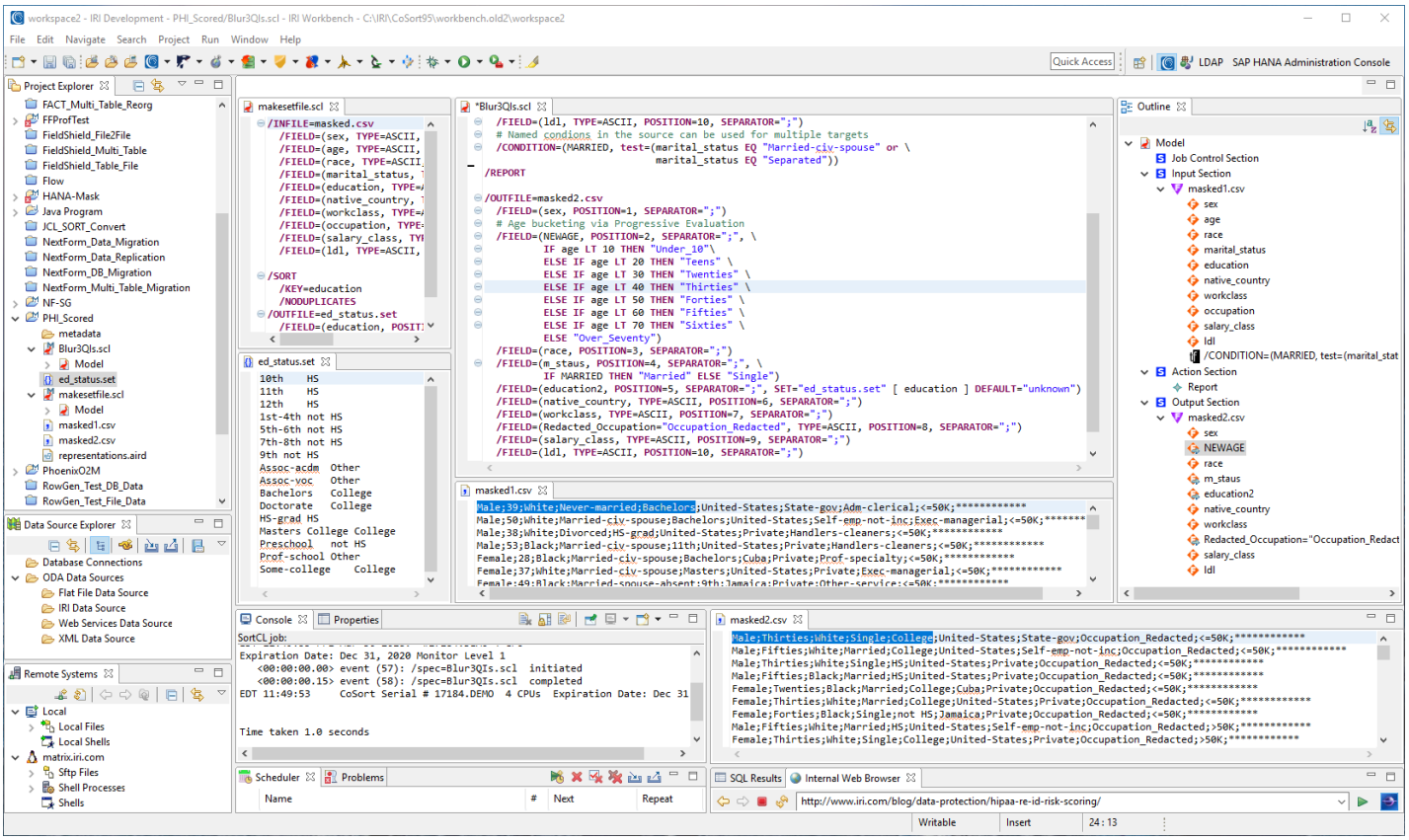
Risk Scoring & Anonymization



Despite masking direct identifiers (PII), you may decide to leave certain columns unmasked, including indirect or quasi-identifiers (demographic) columns like nationality, age, salary or gender. For this reason, FieldShield also leverages peer-reviewed statistical re-ID assessment functions in a standalone risk determination wizard that scores the likelihood of one's identity being revealed through a combination of quasi-identifiers.

The wizard produces detailed statistical measures and reports on the risk of re-identification in three modes of attack, plus the number of rows in each "equivalence class." It also provides an interactive look at different combinations of quasi-identifiers and their separation and distinction values to assess their capacity to re-identify a record.

For the 'riskier' quasi-identifying columns, additional FieldShield static data masking jobs can be run with anonymization functions applied to improve the risk score while retaining the utility of the data. For example, you could apply data blurring to add "random noise" values to ages and birth dates, or bucket specific values like Freshman into a category like Undergrad.



The new result set can now be re-run through the risk scoring wizard to produce another determination of re-identification risk based on now less distinct quasi-identifying attributes.

Post-Evaluation Licensing

IRI FieldShield can be licensed standalone for perpetual use CapEx), or on a subscription basis (OpEx) within the larger IRI Voracity data management platform. Both FieldShield and Voracity licenses account only for the number of hostnames where the masking engine is installed.

Most users license a FieldShield executable for use on at least one Workbench machine for development and convenience, but later deploy additional licenses on remote (database) servers to minimize LAN delays between the source / target tables and the masking engine. Please see this FAQ to help determine how many hostname licenses you may need.

Multi-node license fees are quoted by, and can be negotiated in confidence with, IRI or its authorized commercial and support representatives around the world.

Please contact fieldshield@iri.com for any technical or commercial questions arising from this document or your evaluation of the tool set.