IRI, The CoSort Company

Vendor Background

- Specializing in fast data management and data-centric security
- Privately owned and profitable since 1978
- Sales and support in more than 40 cities worldwide
- 8 of 9 products share 1 metadata and Eclipse IDE
- Featured in: CIO Review (top GRC and Compliance vendors); DBTA; Gartner Market Guide to Data Masking Tools; and in the QY, Markets & Markets, and Research & Markets forecast reports on Data Masking, DB Security, Data Classification, Data Governance
Selected IRI Data Masking Customers & Partners

Most IRI data masking customers profile and protect PII in databases, flat files and Excel sheets on premise, or in the cloud. Recent engagements also involve NoSQL DBs, documents, images, and faces. Streaming and Hadoop sources are also supported. Sites doing IRI mask/test work include:
GOVERN
Data Quality
Data Masking
Re-ID Risk Scoring
Data Reconciliation
Test Data Generation
Data & Metadata Lineage

DESIGN
Wizards with Rules
Graphical Dialogs
Scripts with Outlines
Workflow Palettes
Mapping Diagrams
Metadata Form Editors
Erwin Mapping Manager

DISCOVER
Data Classification
Dark Data Search
DB & File Search
DB & File Profiling
ER Diagramming
Metadata Definition
Metadata Forensics

INTEGRATE
Slowly Changing Dimensions
Public/Private Mashups
Change Data Capture
Fast DB Un/Load
Data Federation
One-Pass ETL

MIGRATE
Incremental Replication
Data & File Types
Endianness
Databases
JCL Sorts
ETL Jobs

ANALYZE
Embedded BI
Data Wrangling
Cloud Dashboard
Predictive Analytics
Clickstream Analytics
BIRT, KNIME & Splunk Feeds

DEPLOY
CoSort CLI/API
MapReduce 2 (Grd)
Spark (In-Memory)
Storm (Streaming)
Tez (Batch)
Java, SQL, YARN
Eclipse & More Job Launchers

TARGETS

SOURCES

Files
COBOL, CSV, Fixed, JSON, LDIF, LS-RS-VS, MF-ISAM, MFVL, Pipes, VB, Vision, XML, etc.

Mainframe
Adabas, Datacom, IDMS, IMS, ISAM, Pick, Unidata, VSAM, etc.

Semi & Unstructured
BMP, GIF, JPG, PNG, TIFF

Other Sources
Custom Apps, ETL/ELT Tools, Packaged Apps, Web Logs

Cloud & SaaS
AWS, Microsoft Azure, Salesforce, HubSpot, Marketo, Eloqua

Databases
Oracle, MySQL, SQL Server, MongoDB, Snowflake, Teradata, Sybase, Stata, SAS, etc.

Call Detail Records
ASIN, TAP3

Big Data Platforms & Streams
Cloudera, MapR, Hortonworks, Pivotal, Netezza, Spark, Kafka

Cloud & SaaS
Amazon, Microsoft Azure, Salesforce, HubSpot, Marketo, Eloqua

Files
COBOL, CSV, Fixed, JSON, LDIF, LS-RS-VS, MF-ISAM, MFVL, Pipes, VB, Vision, XML, etc.

Mainframe
Adabas, Datacom, IDMS, IMS, ISAM, Pick, Unidata, VSAM, etc.

Semi & Unstructured
BMP, GIF, JPG, PNG, TIFF

Other Sources
Custom Apps, ETL/ELT Tools, Packaged Apps, Web Logs

Cloud & SaaS
AWS, Microsoft Azure, Salesforce, HubSpot, Marketo, Eloqua

Databases
Oracle, MySQL, SQL Server, MongoDB, Snowflake, Teradata, Sybase, Stata, SAS, etc.
IRI Data Masking Tool Architectures

Both structured and unstructured data discovery functions -- including classification, search, and metadata creation -- are performed in IRI Workbench data discovery wizards.

Static data masking (SDM) jobs are usually built in IRI Workbench, while user-specific dynamic data masking (DDM) is demonstrated today in API calls and in a new governance mode (see Slide 33 for all DDM options).

Voracity data manipulation and masking jobs use the IRI CoSort (SortCL) engine on commodity LUN hardware, on premise or in the cloud. No database or cloud API is needed. This reduces runtime overhead, administrative complexity, and risk. The executable is also metadata-compatible with, and masks within data, integration, cleansing, and reporting jobs, too.

Finally, no server framework is required to orchestrate jobs. This fosters more granular allocation and tuning of resources, and the sharing of metadata artifacts.
Multiple Masking Job Design Options

Voracity data masking, cleansing, transformation, migration, reporting, and wrangling jobs can be created and run inside or outside of IRI Workbench.

Job design methods supported inside:

1) Job creation wizards
2) Color-coded syntax-aware job script editor with outline
3) Form Editors
4) Graphical parameters Dialogs
5) Mapping Diagrams

Job design methods supported outside:

6) Erwin Mapping Manager
7) Any external text editor
8) 3GL app (system or API calls)
1) 4GL scripts on command line or in batch.
2) From 3rd party automation tools like Stonebranch UAC, cron, etc.
3) Directly in BIRT or KNIME in Eclipse, or a Splunk add-on app, as you report or index.
4) Some jobs run without code changes in Hadoop via MR2, Spark, Spark Stream, Storm or Tez.
5) Use graphical run configuration dialogs or the built-in task scheduler to launch local, remote, or HDFS jobs from IRI Workbench
6) System or API calls from 3GL programs
IAM/RBAC Now & Later

Today, you can assign permissions to data (file) sources, IRI masking programs (sortcl.exe), and the scripts they run (spec.fcl) in LUW file systems using central LDAP/AD settings. You can optionally control them via Apache Directory Studio in IRI Workbench:

Soon, the IRI client/server governance system illustrated to the right will assign and enforce RBACs to the same elements and to more granular elements like field names (mapped from data classes), functions, and even specific data values or ranges of values.
Cloud Data & Systems Support

**FieldShield** can read/mask/write data in cloud databases like Snowflake, MS SQL in Azure, AWS Redshift, et al, via J/ODBC, and data streaming through URLs. **DarkShield** will support cloud sources like Amazon, Google, FB, et al soon (but right now it’s local/SMB only). Both operate on local, remote, or cloud systems running Windows, Linux or Unix.
**Metadata Integrations**

1. Voracity tooling *consumes* metadata from any structured source for data classification, profiling, search, de-ID, ETL etc.
2. FieldShield & RowGen job scripts also *produce* metadata for several **DB load utilities** in multi-DB masking & test data jobs.
3. Their data definition file metadata can also be *exported* (e.g. target field layouts) in CSV for catalog tools like **Collibra**.
4. DarkShield reads **attribute** metadata about source files, and produces artifactual metadata from its search and mask ops and it can auto-forward or populate Splunk ES with that information for analysis, dashboarding, or adaptive responses.
5. **MIMB** and **Erwin** mapping manager can hub and feed FieldShield DDF and .FCL specs based on third-party metadata:

### Bridge Mapping

<table>
<thead>
<tr>
<th>Meta Integration Repository (MIR) Metamodel (based on the OMG CWM standard)</th>
<th>&quot;IRI CoSORT SortCL Data Definition File&quot; Metamodel</th>
<th>Mapping Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>Attribute</td>
<td>/FIELD</td>
<td># Comments in the DDF file</td>
</tr>
<tr>
<td>Comment</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Description</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Name</td>
<td>field name (if physical name is not specified)</td>
<td></td>
</tr>
<tr>
<td>PhysicalName</td>
<td>field name</td>
<td></td>
</tr>
<tr>
<td>Position</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Data Type</td>
<td>Data Type name</td>
<td></td>
</tr>
<tr>
<td>Length</td>
<td>Field length</td>
<td></td>
</tr>
<tr>
<td>Scale</td>
<td>Field scale</td>
<td></td>
</tr>
</tbody>
</table>

6. All IRI metadata -- including data source/target layouts, job/task speci and batch files, worfklows and metamodels, discovery configurations, search matchers and masking rules -- can also be team shared, secured and version controlled in **Git** et al
IRI Data Protector Suite

IRI FieldShield
PII / PHI Classification & Masking

IRI DarkShield
Unstructured Data Search & Security

IRI CellShield
PII / PHI Search & Mask in Excel

IRI DMaaS
Data Masking as a Service
What FieldShield Does

- Connects and interacts with multiple sources and targets, on-premise or cloud
- Discovers and classifies sensitive data in DB, flat-file, and dark-data sources
- Protects fields with PII, PHI, etc. via 14 built-in masking function categories
- Addresses multiple protections and recipients in one job script, one I/O
- Applies masking rules across tables to preserve referential integrity
- Secures data conditionally, i.e. based on patterns, values, or ranges
- Delivers data for portability and redact it for GDPR compliance
- Masks DB data dynamically using C/C++, Java, or .NET SDK functions
- Retains data realism (e.g. FPE and pseudonyms) for testing and outsourcing
- Masks inside big data BI/analytics, ETL, migration, sub-setting, and test data jobs
- Determines statistical likelihood of re-identification (risk scoring) for HIPAA compliance
- Logs job and system runtime details to an XML audit file to verify compliance
- Supports streaming input and Hadoop execution paradigms within Voracity
# FieldShield Data Sources (Standard)

<table>
<thead>
<tr>
<th>Source Type</th>
<th>Delimited</th>
<th>Line Sequential</th>
<th>SQL Server</th>
</tr>
</thead>
<tbody>
<tr>
<td>Acucobol Vision</td>
<td>Derby (WB)</td>
<td>MaxDB</td>
<td>SQL Server</td>
</tr>
<tr>
<td>Altibase (FACT)</td>
<td>Derby (WB)</td>
<td>MaxDB</td>
<td>SQLite</td>
</tr>
<tr>
<td>ASN.1 TAP3</td>
<td>ESDS</td>
<td>MF-ISAM</td>
<td>Sybase ASA/E &amp; IQ</td>
</tr>
<tr>
<td>BIRT DB (WB)</td>
<td>Excel (WB)</td>
<td>WF Var. Length</td>
<td>Tibero (WB)</td>
</tr>
<tr>
<td>BIRT Hive (WB)</td>
<td>ELF web logs</td>
<td>MySQL</td>
<td>Teradata (WB)</td>
</tr>
<tr>
<td>BIRT JDBC (WB)</td>
<td>Fixed</td>
<td>Oracle</td>
<td>Text</td>
</tr>
<tr>
<td>BIRT POJO (WB)</td>
<td>Heap / print</td>
<td>Outlook (WB/DS)</td>
<td>UTF-8 &amp; 16</td>
</tr>
<tr>
<td>C-ISAM</td>
<td>HSQLDB (WB)</td>
<td>PDF (WB/DS)</td>
<td>Variable Block</td>
</tr>
<tr>
<td>CLF web logs</td>
<td>IDX 3, 4 &amp; 8</td>
<td>PostgreSQL/EDB</td>
<td>Variable Sequential</td>
</tr>
<tr>
<td>CSV</td>
<td>Informix</td>
<td>Powerpoint (WB)</td>
<td>VSAM MVS (UniKix)</td>
</tr>
<tr>
<td>DB2 (UDB)</td>
<td>Ingres</td>
<td>Record Sequential</td>
<td>Web Services (WB)</td>
</tr>
<tr>
<td>DB2 for i5/OS (WB)</td>
<td>JSON</td>
<td>RTF (WB)</td>
<td>Word (WB)</td>
</tr>
<tr>
<td>DB2 for z/OS (WB)</td>
<td>LDIF</td>
<td>SQL Anywhere</td>
<td>XML</td>
</tr>
<tr>
<td>DB2 for z/OS (WB)</td>
<td>LDIF</td>
<td>SQL Anywhere</td>
<td>XML</td>
</tr>
</tbody>
</table>

**FACT:** requires IRI Fast Extract (FACT)  
**DS:** requires IRI DarkShield  
**WB:** requires IRI Workbench, the free Eclipse GUI for FieldShield, etc.
# FieldShield Data Sources (Legacy)

<table>
<thead>
<tr>
<th>Access</th>
<th>D3</th>
<th>GA-Power 95, R91</th>
<th>K-ISAM</th>
<th>Pathway</th>
<th>RMS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adabas</td>
<td>Datacom</td>
<td>Gemstone</td>
<td>Knowledgeman</td>
<td>PDS</td>
<td>Reality/X</td>
</tr>
<tr>
<td>Advanced Pick</td>
<td>Dataflex</td>
<td>GENESIS</td>
<td>KSDS</td>
<td>PervasiveSQL</td>
<td>RRDS</td>
</tr>
<tr>
<td>ALLBASE</td>
<td>Db4o</td>
<td>Gigabase</td>
<td>Lotus</td>
<td>Pick/Pick64+</td>
<td>SAP HANA</td>
</tr>
<tr>
<td>Alpha5</td>
<td>dBase</td>
<td>H2</td>
<td>Manman</td>
<td>PI-Open</td>
<td>Sequoia</td>
</tr>
<tr>
<td>Amazon RDS</td>
<td>Desktop Adapter</td>
<td>IDMS</td>
<td>Mentor / pro</td>
<td>Powerflex</td>
<td>Sharebase</td>
</tr>
<tr>
<td>Azure</td>
<td>DL/1</td>
<td>IDS</td>
<td>MO</td>
<td>Powerhouse</td>
<td>Supra</td>
</tr>
<tr>
<td>BizTalk</td>
<td>DSM</td>
<td>Image</td>
<td>Model 204</td>
<td>Progress</td>
<td>Terracotta</td>
</tr>
<tr>
<td>Cache</td>
<td>Enscribe</td>
<td>IMS</td>
<td>Mumps</td>
<td>QueryObject</td>
<td>Total</td>
</tr>
<tr>
<td>Clipper</td>
<td>Enterprise Adapter</td>
<td>Interbase</td>
<td>MyBase</td>
<td>rBase</td>
<td>Ultimate</td>
</tr>
<tr>
<td>Codasyl</td>
<td>FileMaker</td>
<td>Intersystems</td>
<td>Netezza</td>
<td>R83</td>
<td>UltPlus</td>
</tr>
<tr>
<td>CorVision</td>
<td>Firebird</td>
<td>ISM</td>
<td>NonStop SQL</td>
<td>Rdb</td>
<td>Unidata</td>
</tr>
<tr>
<td>ConceptBase</td>
<td>Focus</td>
<td>Jasmine</td>
<td>ObjectStore</td>
<td>REALITY</td>
<td>Universe</td>
</tr>
<tr>
<td>D-ISAM</td>
<td>FoxPro</td>
<td>JBase</td>
<td>Paradox</td>
<td>Red Brick</td>
<td>VSAM VSE</td>
</tr>
</tbody>
</table>

Accessible via IRI partner (CONNX) J/ODBC drivers
## FieldShield Data Sources (Modern, in Voracity)

<table>
<thead>
<tr>
<th>Data Source</th>
<th>FinancialForce</th>
<th>Marketo</th>
<th>Pivotal Greenplum</th>
</tr>
</thead>
<tbody>
<tr>
<td>Amazon EMR Hive</td>
<td>Force.com apps</td>
<td>MongoDB</td>
<td>Pivotal HD Hive</td>
</tr>
<tr>
<td>Apache Cassandra</td>
<td>Hortonworks Hive</td>
<td>MS Dynamics CRM</td>
<td>Salesforce.com</td>
</tr>
<tr>
<td>Apache Hadoop Hive</td>
<td>Hubspot</td>
<td>MS SQL Azure</td>
<td>ServiceMAX</td>
</tr>
<tr>
<td>Cloudera CDH Hive</td>
<td>Lightning Connect</td>
<td>Oracle Eloqua</td>
<td>Spark SQL</td>
</tr>
<tr>
<td>Cloudera Impala</td>
<td>MapR Hive</td>
<td>Oracle Service Cloud</td>
<td>Veeva CRM</td>
</tr>
<tr>
<td>Database.com</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Sensitive Data Discovery - Multiple Wizards

To facilitate data masking, IRI FieldShield includes: PII definition (cataloging through data classes); discovery through string (literal or dictionary), pattern, and fuzzy-logic searches; statistical reporting; and, automatic metadata creation.

Fit-for-purpose GUI wizards deliver:

- DB and file data classification, with search and masking rule matchers
- DB profiling, ERDs, and table searches
- Flat-file profiling and value searches
- Data class searches through schema and directories for bulk discovery
- Metadata discovery and definition
- Dark data search and structuring, with metadata reporting (see DarkShield)
Static Data Masking Functions (1-3 of 15)

Character Scrambling

- For ASCII data
- Less secure
- Reversible

Encoding / Decoding

- Converts binary to ASCII
- Supports base64 & hex
- Reversible

Encryption / Decryption

- 3DES EBC & SSL
- AES-128 & -256 CBC
- AES-256 Format-Preserving
- GPG (PGP-compatible)
- FIPS-compliant OpenSSL
- Custom
Static Data Masking Functions (4-6 of 15)

- **Pseudonymization**
  - Provides realistic names
  - Reversible lookup values
  - Non-reversible selection

- **Redaction / Obfuscation**
  - Partial/full-field masking
  - Conditional omission
  - Non-reversible

- **Randomization**
  - Random data generation
  - Random data selection
  - Non-reversible
Static Data Masking Functions (7-15 of 15)

Hashing
- SHA-1 & 2 cryptographic
- Returns hash of fieldstring
- Use for integrity checking

Expression Logic
- Mathematical operations
- PCRE logic
- Custom blurring

String Manipulations
- Find, replace, and add
- Reposition and trim
- Use INSTR information

Blurring & Bucketing
Add random “noise” (perturbate) to ages/dates, and generalize (anonymize) quasi-identifiers

Tokenization
DB-value substitute for PCI DSS

Deletion & Suppression
Erasure for GDPR Right to Be Forgotten

Custom Functions
User’s field-level call
Query-Ready XML Audit Log

```
<?xml version="1.0" encoding="UTF-8" standalone="yes"?>
<Query-Trail>
  <AuditTrail>
    <AuditRecord>
      <Product>CoSort</Product>
      <Version>9.5.3</Version>
      <VersionTag>R081406122-1600</VersionTag>
      <Serial>999999</Serial>
      <OperatingSystem>Windows 7</OperatingSystem>
      <User>IRI</User>
      <ProcessId>12345</ProcessId>
      <Terminal>console</Terminal>
      <Program>sort</Program>
      <Command>/spec-patient_enc.scl</Command>
      <StartTime>2014-02-20 13:31:49</StartTime>
      <EndTime>2014-02-20 13:31:54</EndTime>
      <RunTime>00:00:05</RunTime>
      <ReturnCode>0</ReturnCode>
      <ErrorMessage>normal return</ErrorMessage>
    </AuditRecord>
    <RecordsProcessed>1000</RecordsProcessed>
  </AuditRecord>
</Query-Trail>
```
Re-ID Risk Determination

US HIPAA and FERPA regulations require that patient and student data sets used in research or marketing have a statistically certified “very small” chance of being re-identifiable.

- IRI risk scoring wizard produces re-ID probability scores in 3 modes
- Analyzes quasi-identifiers with multiple, peer-reviewed functions
- Detail and graphed scoring reports
<table>
<thead>
<tr>
<th>President</th>
<th>Party</th>
<th>State</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adams, John</td>
<td>D-R</td>
<td>MA</td>
</tr>
<tr>
<td>Adams, John Quincy</td>
<td>D-R</td>
<td>MA</td>
</tr>
<tr>
<td>Anderson, Chester A.</td>
<td>REP</td>
<td>CT</td>
</tr>
<tr>
<td>Buchanan, James</td>
<td>REP</td>
<td>TX</td>
</tr>
<tr>
<td>Bush, George H.W.</td>
<td>REP</td>
<td>FL</td>
</tr>
<tr>
<td>Bush, George W.</td>
<td>REP</td>
<td>TX</td>
</tr>
<tr>
<td>Carter, James E.</td>
<td>REP</td>
<td>GA</td>
</tr>
<tr>
<td>Cleveland, Grover</td>
<td>REP</td>
<td>OH</td>
</tr>
<tr>
<td>Cleaveland, Grover</td>
<td>REP</td>
<td>OH</td>
</tr>
<tr>
<td>Clinton, William J.</td>
<td>DEM</td>
<td>NY</td>
</tr>
<tr>
<td>Clinton, William J.</td>
<td>DEM</td>
<td>NY</td>
</tr>
<tr>
<td>Coolidge, Calvin</td>
<td>REP</td>
<td>MA</td>
</tr>
<tr>
<td>Coolidge, Calvin</td>
<td>REP</td>
<td>MA</td>
</tr>
<tr>
<td>Eisenhower, Dwight D.</td>
<td>REP</td>
<td>TX</td>
</tr>
<tr>
<td>Eisenhower, Dwight D.</td>
<td>REP</td>
<td>TX</td>
</tr>
<tr>
<td>Fillmore, Millard</td>
<td>REP</td>
<td>NY</td>
</tr>
<tr>
<td>Ford, Gerald R.</td>
<td>REP</td>
<td>MI</td>
</tr>
<tr>
<td>Ford, Gerald R.</td>
<td>REP</td>
<td>MI</td>
</tr>
<tr>
<td>Garfield, James A.</td>
<td>REP</td>
<td>OH</td>
</tr>
<tr>
<td>Garfield, James A.</td>
<td>REP</td>
<td>OH</td>
</tr>
<tr>
<td>Grant, Ulysses S.</td>
<td>REP</td>
<td>OH</td>
</tr>
<tr>
<td>Harding, Warren G.</td>
<td>REP</td>
<td>OH</td>
</tr>
<tr>
<td>Hendricks, Benjamin</td>
<td>REP</td>
<td>IN</td>
</tr>
<tr>
<td>Harrison, Benjamin</td>
<td>REP</td>
<td>OH</td>
</tr>
<tr>
<td>Harrison, William Henry</td>
<td>REP</td>
<td>OH</td>
</tr>
<tr>
<td>Hayes, Rutherford B.</td>
<td>REP</td>
<td>OH</td>
</tr>
<tr>
<td>Hayes, Rutherford B.</td>
<td>REP</td>
<td>OH</td>
</tr>
<tr>
<td>Hoover, Herbert C.</td>
<td>REP</td>
<td>OH</td>
</tr>
<tr>
<td>Jackson, Andrew</td>
<td>REP</td>
<td>OH</td>
</tr>
</tbody>
</table>

This image demonstrates the process of masking data within a MongoDB database using IRI FieldShield. The left side shows the original data before masking, while the right side displays the masked data. The blue arrow indicates the original data, and the red arrow shows the masked data.
IRI FieldShield in Voracity

Masking et al in Hadoop, too

Map once, deploy anywhere
## Dynamic Data Masking Options

<table>
<thead>
<tr>
<th>Method</th>
<th>Operation</th>
</tr>
</thead>
<tbody>
<tr>
<td>ODBC Select / Update</td>
<td>Apply FieldShield column masks to target (view) tables for specific users/rows</td>
</tr>
<tr>
<td>DB App Invocation</td>
<td>Use .NET or Java SDK library functions or system-call job scripts on the fly</td>
</tr>
<tr>
<td>In-Situ Redaction</td>
<td>User and SQL-specific full and partial column masking on query</td>
</tr>
<tr>
<td>Custom I/O Procedures</td>
<td>Drive real-time application data directly to/from FieldShield jobs in memory</td>
</tr>
<tr>
<td>Real-Time Processing</td>
<td>Hadoop Spark and Storm processing of dynamic input streams (via Voracity VGrid)</td>
</tr>
<tr>
<td>Governance Mode</td>
<td>New runtime facility tied to RBAC/IAM infrastructure masks fields for some users</td>
</tr>
</tbody>
</table>

## Encryption Key Management Options

1. Passphrase (key string) embedded in script
2. String as environment variable
3. String in (securable) key file
4. Multi-factor authentication via Azure Key Vault or Townsend Security Alliance Key Manager (which also features VM and HSM supports)
This method has FieldShield masking of structured data flattened into CSV subsets by the Sonra Flexter pre-parsing tool.

Alternatively, just use IRI DarkShield.
User Profiles

- Vertical industries and governmental agencies storing, processing, or outsourcing applications with sensitive data, such as:
  - Banks
  - Census / Tax
  - Defense
  - Health Care
  - Insurance
  - Schools

- Application, DB, and DW users handling sensitive data
- CISOs, compliance teams, consultants, IT managers, and solution architects
Use Cases

**Tesco Bank/RBS UK**
- Decrypt and re-encrypt fields in credit card migration and test files
- Generate and manage encryption and user ID keys
- Other projects protect 38,265 records per minute on Windows

**Accenture Singapore**
- Design and run encryption and masking jobs on Linux servers
- Secure PHI for the Ministry of Health Holdings (MOHH)’s Oracle DB
- Row sequencing and job audits

**Medicx Media Solutions USA**
- Encryption and hashing functions to PII and PHI in geo-medical consumer health databases
- Exceeds HIPAA requirements in provisioning mScores™ data to digital and direct marketers
Key Differentiators

**Developer Support**
- Version controls
- Master data definition
- Secure key management
- Git project management (teaming)
- SDK supports .NET and Java calls
- Data profiling and metadata discovery
- XML (and soon JSON) job logs, IAM

**One-Stop-Shop**
- Integrated data classification & search
- Includes re-ID risk scoring for HIPAA
- Use w/Voracity ETL, migrate, cleanse
- Metadata-compatible with RowGen TDM
- Used in DB subsetting wizard
- Also works in Voracity BI & KNIME jobs
- Runs w/Actifio DB clones, Splunk ES, etc.

**Price Performance**
- The data-centric security tool with:
  - The most sources
  - The most protection functions
  - The most target file formats
- Fastest standalone protection software

**Ease-of-Use**
- Familiar Eclipse GUI
- Self-documenting 4GL syntax
- Easy management and modification of jobs/metadata
Competitive Advantages

vs. IBM
○ FieldShield scripts simpler than Optim interoperability model and Javascript options
○ Seamless integration with more sources
○ More functions
○ Lower cost

vs. CA (Grid Tools)
○ Built-in CoSort engine makes FieldShield faster than GT Fast Data Masking
○ Tight integration with data profiling, ETL, data quality, and BI operations
○ Multi-target/format options
○ Lower cost
○ Built-in re-ID risk determination wizard

vs. Informatica
○ FieldShield DDM inclusive with product (compared to Informatica’s upgrade)
○ More SDM protection functions
○ Integration with Eclipse and Excel
○ Access to 4GL scripts
○ Lower cost

vs. Imperva (Camoflauge)
○ FieldShield has more masking and encryption functions
○ Hash, decode, and pseudonymize functions
○ Faster and more extensible in the IRI Workbench IDE
○ Lower cost

vs. Oracle (click)
What CellShield EE Does

- Discovers, reports, and masks PII and perform audit actions in Excel 2010 & later
- Searches and secures PII in spreadsheets on one PC or throughout an SMB LAN
- Provides common and allow new search pattern definitions for PII formats
- Searches for strings in a dictionary, and find/fix PII *floating* in cells
- Supports reuse and sharing of patterns in project or cloud repositories
- Generates a report of all patterns found and open it for action in a worksheet
- Opens applicable worksheets and highlights the located ranges for protection
- Encrypts, redacts, or pseudonymizes in one-pass with chosen functions and options
- Reveals data with the decryption key, or if reversible pseudonymization was used
- Overlays results directly into the affected cells, or in another worksheet
- Moves between, or bulk-remediates all, identified worksheets and ranges
- Auto-inserts protection details into an un-editable audit column in the report
CellShield PII Discovery

The dark data profiling wizard in the IRI Workbench searches network-wide for sensitive data in spreadsheets based on user-specified (plus popular and saved) Java regular expressions (patterns):
CellShield Reporting

The report produced by the profiling wizard opens in a dynamic worksheet supported by an action dialog for protection and auditing activities:

![CellShield Spreadsheet Selector](image)

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td>Include File</td>
<td>File Path</td>
<td>File Name</td>
<td>Sheet Name</td>
</tr>
<tr>
<td>2</td>
<td></td>
<td>C:\Users\pekarch\Documents\Testing Documents for DDD\NamesNHSN1.xlsx</td>
<td>NamesNHSN1.xlsx</td>
<td>Sheet1</td>
<td>NHS_Number \b[0-9]{1}\s[0-9]{3}\s[0-9]{3}\s[0-9]{4}\b</td>
</tr>
<tr>
<td>3</td>
<td></td>
<td>C:\Users\pekarch\Documents\Testing Documents for DDD\NamesNHSN2.xlsx</td>
<td>NamesNHSN2.xlsx</td>
<td>Sheet1</td>
<td>NHS_Number \b[0-9]{1}\s[0-9]{3}\s[0-9]{3}\s[0-9]{4}\b</td>
</tr>
<tr>
<td>4</td>
<td></td>
<td>C:\Users\pekarch\Documents\Testing Documents for DDD\NamesNHSN3.xlsx</td>
<td>NamesNHSN3.xlsx</td>
<td>Sheet1</td>
<td>NHS_Number \b[0-9]{1}\s[0-9]{3}\s[0-9]{3}\s[0-9]{4}\b</td>
</tr>
<tr>
<td>5</td>
<td></td>
<td>C:\Users\pekarch\Documents\Testing Documents for DDD\NamesNHSN4.xlsx</td>
<td>NamesNHSN4.xlsx</td>
<td>Sheet1</td>
<td>NHS_Number \b[0-9]{1}\s[0-9]{3}\s[0-9]{3}\s[0-9]{4}\b</td>
</tr>
<tr>
<td>6</td>
<td></td>
<td>C:\Users\pekarch\Documents\Testing Documents for DDD\NamesNHSN5.xlsx</td>
<td>NamesNHSN5.xlsx</td>
<td>Sheet1</td>
<td>NHS_Number \b[0-9]{1}\s[0-9]{3}\s[0-9]{3}\s[0-9]{4}\b</td>
</tr>
<tr>
<td>7</td>
<td></td>
<td>C:\Users\pekarch\Documents\Testing Documents for DDD\NamesNINo1.xlsx</td>
<td>NamesNINo1.xlsx</td>
<td>Sheet1</td>
<td>NINo \b[0-9]{1}\s[0-9]{3}\s[0-9]{3}\s[0-9]{4}\b</td>
</tr>
<tr>
<td>8</td>
<td></td>
<td>C:\Users\pekarch\Documents\Testing Documents for DDD\NamesNINo2.xlsx</td>
<td>NamesNINo2.xlsx</td>
<td>Sheet1</td>
<td>NINo \b[0-9]{1}\s[0-9]{3}\s[0-9]{3}\s[0-9]{4}\b</td>
</tr>
<tr>
<td>9</td>
<td></td>
<td>C:\Users\pekarch\Documents\Testing Documents for DDD\NamesNINo3.xlsx</td>
<td>NamesNINo3.xlsx</td>
<td>Sheet1</td>
<td>NINo \b[0-9]{1}\s[0-9]{3}\s[0-9]{3}\s[0-9]{4}\b</td>
</tr>
<tr>
<td>10</td>
<td></td>
<td>C:\Users\pekarch\Documents\Testing Documents for DDD\NamesNINo4.xlsx</td>
<td>NamesNINo4.xlsx</td>
<td>Sheet1</td>
<td>NINo \b[0-9]{1}\s[0-9]{3}\s[0-9]{3}\s[0-9]{4}\b</td>
</tr>
<tr>
<td>11</td>
<td></td>
<td>C:\Users\pekarch\Documents\Testing Documents for DDD\NamesNINo5.xlsx</td>
<td>NamesNINo5.xlsx</td>
<td>Sheet1</td>
<td>NINo \b[0-9]{1}\s[0-9]{3}\s[0-9]{3}\s[0-9]{4}\b</td>
</tr>
</tbody>
</table>
CellShield Protection

Perform point-and-click encryption and decryption, masking (full or partial cell), or pseudonymization (reversible and non-reversible) of the applicable ranges within the spreadsheets in the report.
CellShield Intra-Cell Search & Mask

Feature finds and fixes floating PII, *ad hoc*, or *en masse*

![Screenshot of Excel spreadsheet showing data protection features](image-url)
CellShield Auditing

An uneditable log entry for the protection applied to each pattern identified in the report is automatically appended on each action:
**What’s New in CellShield EE**

**Going into V2 Shortly**
- Faster search/mask in volume
- Improved audit column support

**What’s Planned Later**
- DarkShield-side masking alternative
- Sharepoint access
IRI Data Protector Suite

IRI DarkShield
Unstructured Data Search & Security

IRI Data Protector Suite
What DarkShield Does

- Simultaneously scans, extracts, and de-IDs or deletes PII (and audits actions) in all supported file formats
- Finds faces or defined data classes tied to RegEx patterns, lookup sets, NER models, and/or image regions
- Builds, saves, and re-uses semi-supervised, machine learning models in project or cloud repositories
- Blacks-out PII in images, blurs faces, and applies functions of encryption (including FPE), pseudonymization, hashing, encoding, bit scrambling, redaction, or erasure functions for PII in text files and documents
- Writes masked files atop originals, or to different folders with the same file names and formats
- Shows search, remediation, and model training job status via real-time progress bars
- Shares search methods and masking functions with CellShield EE and FieldShield
- Generates logs of all values found or masked, along with IRI-compatible metadata for BI, queries, etc.
- Creates graphical, interactive displays of search and mask results, or hand-offs log files to Splunk
- Runs in IRI Workbench with other IRI and Eclipse tools, or from the command line
Granular Sourcing/Targeting

Use DarkShield’s dark data discovery wizard to find sensitive data in unstructured files LAN-wide, mask it, and re-target the results.
IRI DarkShield

Apply width-preserving redaction, blackout, deletion, encryption, pseudonymization, and other data masking functions to protect PII and comply with data privacy laws like the GDPR.
Deletion Function

IRI FieldShield, DarkShield & CellShield and other features in Voracity combine to comply with GDPR (and thus CCPA, KVKK, etc.) provisions like:

- Discovery and **De-Identification** of PII and PI
- The right to be **Forgotten** (via erasure like this)
- Data **Portability** (via extraction and reformatting)
- Data **Rectification** (via discovery and cleansing)
Optionally and automatically extract all of the values you searched for (think GDPR data portability), plus the metadata associated with the files containing those values.
at his touch of a certain icy pang along my blood. “Come, sir,” said I.
“You forget that I have not yet the pleasure of your acquaintance. Be
seated, if you please.” And I showed him an example, and sat down.
myself in my customary seat and with as fair an imitation of my or-
dinary manner to a patient, as the lateness of the hour, the nature of
my preoccupations, and the horror I had of my visitor, would suffer
me to muster.
“I beg your pardon, Dr. Lanyon,” he replied civilly enough. “What
you say is very well founded, and my impatience has shown its heels
to my politeness. I come here at the instance of your colleague, Dr.
Henry Jekyll, on a piece of business of some moment, and I under-
stood...” He paused and put his hand to his throat, and I could see,
in spite of his collected manner, that he was wrestling against the
approaches of the hysteria—“I understood, a drawer...”
But here I took pity on my visitor’s suspense, and some perhaps
on my own growing curiosity.
“That it is, sir,” said I, pointing to the drawer, where it lay on the
floor behind a table and still covered with the sheet.
He sprang to it, and then paused, and laid his hand upon his
heart; I could hear his teeth grate with the convulsive motion of his
jaw, and his face was so ghastly to see that I grew alarmed both for
his life and reason.
“Compose yourself” said I.
He turned a dreadful smile to me, and as if with the decision of
despair, plucked away the sheet. At sight of the contents, he uttered
one loud sob of such immense relief that I sat petrified. And the
next moment, in a voice that was already faintly well under control,
“Have you a graduated glass?” he asked.
I rose from my place with something of an effort and gave him
what he asked.
He thanked me with a smiling nod, measured out a few mini-
tims of the red tincture and added one of the powders. The mix-
ture, which was at first of a reddish hue, began, in proportion as the
DarkShield supports both pre-trained OpenNLP Name Finder models or new Named Entity Recognition (NER) models that you can build and train inside its semi-supervised machine learning dialog. This iterative process improves the accuracy of searches for names and other nouns based on their Natural Language Processing (NLP) context in sentences.

Compare this method to other DarkShield search methods, like pattern and lookup matches, path filters, or bounding-box areas (for images).
DarkShield can *detect* faces in any image and blur (all of) them, or just those it *recognizes* from your trained library of faces.
Search via Path Filters

Allows the user to take the structure of a JSON file into account during searches. Additional filters for other formats like XML are being added now. This:

- Ignores fields that do not match the filter
- Increases search speed, and narrows the scope of the search results
Documented CLI

**Command Prompt**

```
C:\Users\cosort>darkshield -h
Usage: darkshield [[Criteria Path [-r] [-t=Tessdata]] ; [-d=Dark Data Path]]
       [-hU] [-b=Binary Directory] [-e=Executable]
Command Line API for the DarkShield search and remediation engine.
   -b, --bin-dir=Binary Directory
       The path to the binary directory where the CoSort executable is found. Defaults to "$COSORT_HOME/bin"
   -e, --executable=Executable
       The CoSort executable to use for the remediation. Defaults to "sortcl."
   -h, --help
   -V, --version
Search Criteria Options
   -r, --remediate
       Remediates the files as they are found by the search.
   -t, --tessdata=Tessdata
       Path to the tessdata folder containing Tesseract training files.
Dark Data Remediation Options
   -d, --darkdata=Dark Data Path
       Path to the dark data file.
Options
C:\Users\cosort>
```
Easily query, analyze, and format the results of search and mask operation through built-in reports and this graphical display.

Or, export DarkShield log data for visualizations your preferred BI tool, or to SIEM environments like Splunk ES, shown here.

It is also then possible to take actions through the Splunk Adaptive Response Framework or a Phantom playbook.
Current Benefits

1. Combines PII discovery, delivery, deletion, and reporting in multiple unstructured source formats into one or more ergonomic operations
2. Allows pattern definition reuse and combination to consolidate searches
3. Consolidates multiple right to be forgotten and data portability requests into the same find/fix operation through literal names or lookup-file matches
4. Supports multiple drives, nodes, and threads for searching and masking work
5. Operates in the same Eclipse job design and metadata environment, IRI Workbench, with related data governance and management activities
6. Features affordable licensing options (standalone, bundled, or free in Voracity)
7. Integration with IRI FieldShield/Voracity data classification and masking functions
8. Parameter serialization and modeling for easy modification and batch execution

Development Roadmap

1. More unstructured format support, including A/V, proprietary apps, cloud silos, etc.
2. Additional ergonomic convergence with structured and embedding sources
3. Plug-in integration with more SIEM tools beyond Splunk ES and Phantom Playbooks which are now supported, like IBM QRadar and SolarWinds
4. Additional logging and application integration options
IRI Data Protector Suite
What is DMaaS?

- IRI Data Masking as a Service (DMaaS) is a professional service engagement
- DMaaS makes use of trusted IRI ‘shield’ software products described above ONLY
- Certified IRI experts classify, discover, and de-identify PII of concern in supported silos
- Also available: HIPAA re-ID risk scoring and anonymization, and ‘fake PII’ for testing
- IRI services are performed under a SoW with NDA, BAA, or other data security terms
- Source data that cannot be sent is accessed via VPN or secure public/private cloud
- Data is only accessed by IRI engineers in the US or certified partners like Capgemini
- All data access, classification, discovery (search) and masking operations are logged
- Billing is hourly or daily, with project rates available; IRI software costs are subsumed
- Customer is responsible for payment of cloud infrastructure of their choice

IRI DMaaS
User Profiles

- DBAs and sysadmins responsible for PAN, PHI, PII or other sensitive information
- Sites needing standard data classification and consistent masking functions
- CISOs without sufficient internal IT resources to do this work internally
- Data governance and C-suite officers subject to compliance audits

Use Cases

RBS / Tesco (PCI DSS)
  - Produced and implemented custom encryption for testing data in M&A

Confidential (HIPAA)
  - Cataloged and de-identified protected health information

University of Adelaide (Privacy Act)
  - Data classification, search, and de-identification of PII in massive PeopleSoft financial, HR, and campus test data schemas in Oracle
Also available with IRI Data Protector or Manager Suites, and the IRI Voracity Platform
What RowGen Does

- Creates synthetic but realistic random and random-real test data simultaneously
- Improves DB prototypes, application quality, benchmarking, and outsourced operations
- Uses standard DB DDL, production file, and custom metadata to define layouts
- Preserves structural and referential integrity of real EDW DBs for testing
- Produces data in any types, structure, volumes, value ranges, and if condition
- Synthesizes composite data values and custom (master) data formats
- Generates computationally valid and invalid NID (Codice Fiscale, etc.) SSNs, CCNs
- Sets and graphs test data value distributions (linear, normal, random, etc.)
- Applies common attribute rules (like lookups) rules for pattern-matched field names
- Filters, transforms, and pre-sorts test data while it’s being generated
- Writes loader metadata and perform direct path loads for test DB populations
- Builds test flat-file and custom/structured detail and summary report targets
- Subsets and masks databases automatically for test purposes
- Provides SDK functions for generating test data in Java apps and Hadoop
Use Existing Data Models and Metadata

Build Test Data for:
- Altibase
- CLF/ELF
- COBOL
- CSV
- DB2
- Hadoop
- Hive
- JSON
- LDIF
- MySQL
- NoSQL DBs
- Oracle
- SQL Server
- Sybase
- Teradata
- XML
DB Subsetting, Masking Optional

Including subsetting and test data generation wizards facilitate DB and EDW prototyping, as well as test data virtualization for DevOps. Masked and referentially-correct copies of production table extracts ensure production data is safe and test data is realistic.
User Profiles

Anyone doing DB testing, app development, stress-testing, or benchmarking, including:

- Developers (programmers)
- DBAs and DW (ETL) architects
- Analysts and consultants

Use Cases

Bank of Montreal
- Generates safe, realistic 20GB Oracles tables with RI for query testing

MasterCard Peru
- Synthesizes PAN and PII in files to support OLTP and app testing

Transitive UK
- Simultaneously creates and transforms data to test cross-OS virtualization
Key Differentiators

1. Big data generation and population performance
   (embedded CoSort pre-sorting engine speeds bulk loads)

2. Synthetic data that’s broader and safer than real data

3. Concurrent test data manipulation and custom report outputs

4. Simple, portable, and modifiable test data generation and auto-built DB loader scripts, all managed visually in Eclipse

5. Metadata compatibility with IRI software, Erwin (AnalytiX DS), and MIMB: to facilitate test data generation for 3rd-party BI, CRM, and ETL tools
# What’s New in RowGen

<table>
<thead>
<tr>
<th>Recently Added</th>
<th>Development Underway</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ability to generate Data Vault test data</td>
<td>Random direct DB column lookups</td>
</tr>
<tr>
<td>New email generator</td>
<td>On-demand TDM via TAF integration</td>
</tr>
<tr>
<td>New credit card number generator</td>
<td>Provisioner for Splunk test data</td>
</tr>
<tr>
<td>New national ID number generator</td>
<td>KNIME node test data integration</td>
</tr>
</tbody>
</table>

IRI offers four methods for producing safe, intelligent test data in referentially correct database, flat-file, semi-structured file, and formatted report targets:

1. Production data **masking/scrambling** in IRI FieldShield or IRI Voracity
2. Database **subsetting** & masking in FieldShield or Voracity
3. Synthetic **test data** creation (via random generation/selection) in RowGen or Voracity
4. A **combination** of the above techniques in Voracity