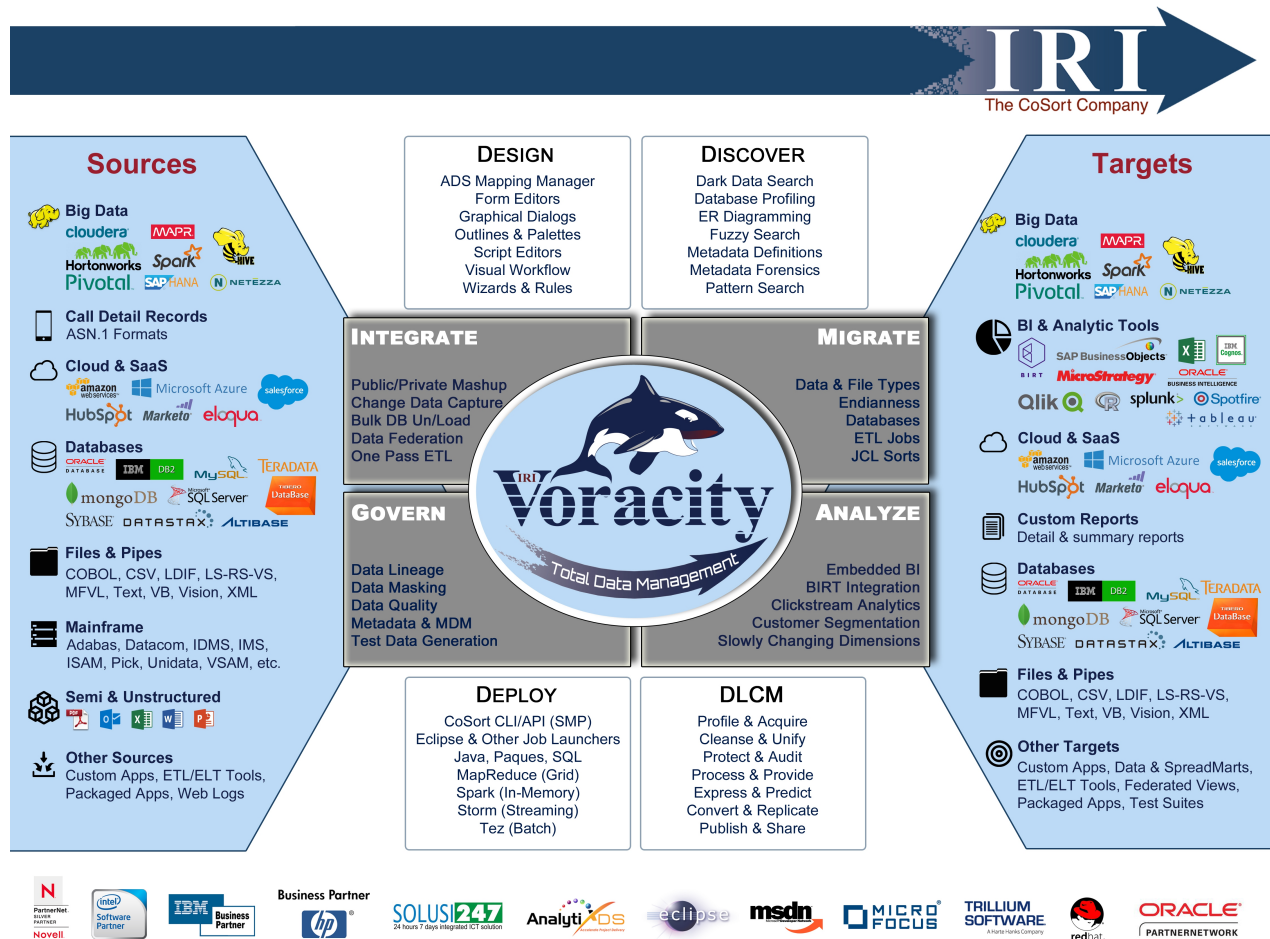# Voracity First Steps

IRI Voracity combines the discovery, integration, migration, governance, and analysis of data in multiple sources ... all in place, and often in one pass. Manipulate, migrate, mask, munge, and map structured, semi-structured, and unstructured data into multiple targets at once.

See the tables below for taking your first steps in the Voracity platform. More technical details are available in the help for applicable wizards and dialogs, or otherwise from the IRI Workbench help menu, blog articles, product manuals or other documentation, on and offline.

The toolbar items in the table below indicate the location of new task or job definition wizards. The wizards automatically build project-ready, visually-integrated data, job, and batch specs in scripts that you can also edit by hand in or outside the GUI. Details on Job Design and Deployment options -- as shown in the above diagram, and some cases available for use in other IRI software products -- are documented in the Welcome section's Voracity Platform Overview.

Note that the Data Integration section steps below are ETL-centric. Additional data warehouse related capabilities like segmentation, change data capture, and slowly changing dimensions are

listed under **Analytics**, while data quality, masking, and test data generation are under **Data Governance**. There are also many other advanced data formatting, expression, statistical, and analytic features available through the CoSort SortCL language which IRI Workbench supports.

IRI engineers, plus trained representatives or consultants worldwide can help you use or optimize the program in the course of normal support, or professional service, engagements.

Data Discovery	
Database profiling	Use the DB profiling wizard from the toolbar to: • Produce and report stats on columns and values • Pattern search using Java RegEx patterns • Fuzzy-search report near matches and probabilities • Compare foreign and primary keys to check integrity
ER diagramming	Use the ER diagram creator to customize views of tables and relationships in any DB connected to DTP via JDBC
Dark data discovery	Run the dark data discovery wizard to find and report on Java RegEx pattern-matching values and metadata from HTML, MS Office, PDF, RTF, TXT and XML documents
Dark data structuring	Use the dark data discovery wizard to move the extracted values into flat files with pre-built data definition files (DDF)
Big data discovery	NoSQL, Hadoop, Cloud, and SaaS platform data search and classification (soon, from the DB profiling wizard above)
Structured data discovery	Use the metadata discovery wizard to preview sequential file/DBs layouts and define IRI DDF metadata automatically
Data Classification	Use the data classification wizard to classify your data and apply transformation and protection rules to data classes in multiple sources at once

Data Integration	
High-performance, native ETL	Design, preview, debug, share, and execute: extract, transform, and load (ETL) jobs on schedule or externally by pipe-lining IRI FACT or ODBC* (E), CoSort/SortCL or Hadoop (T), and bulk DB loaders or ODBC* (L) steps. Build Voracity ETL jobs in: • New ETL Job Wizard • Visual workflow and transform mapping diagrams • FACT (E/L) and CoSort (T/L) wizards • Syntax-aware script editors w/outline & dialogs; and/or • AnalytixDS Mapping Manager
Accelerate another ETL tool	Build faster extract, sort, join, aggregation, or pre-sorted load scripts, and run them as system or command-line calls (e.g. sortcl /spec=VoracityXform.scl) with/in another ETL tool: • IRI FACT (E) jobs: use the new FACT job wizard or write a FACT .ini

Data Integration				
	or .xml config script. • IRI CoSort (T) jobs: launch a new CoSort sort or join job wizard. Filters are source and target options, while aggregations, functions, PII masks, etc. are defined in the target field specs. • DB load utility config files using pre-CoSort'ed input files (L): select the DB loader from the load dialog menu in a CoSort new sort or join job wizard, the Voracity ETL wizard, the FACT offline DB reorg wizard, or RowGen DB test data wizard. See this to "pushdown" and optimize transforms in Informatica, and this for examples of Pentaho augmentation.			
Switch from another ETL tool	Auto-convert ETL project metadata to IRI Voracity ETL from: • IBM InfoSphere DataStage 7.5 and up • Informatica 9.0 and up • Microsoft SQL Server SSIS 2008 and greater • Talend • Oracle Data Integrator (ODI) • SAP BODS • Pentaho • Oracle Warehouse Builder • Actian/Pervasive Data Integrator • Ab Initio Get a license for the AnalytixDS (ADS) Mapping Manager and the CatFX template for your ETL tool to auto-generate Voracity ETL XML flow files. Contact your IRI or ADS representative.			
Fast Extract via IRI FACT	See the FACT Platform Overview and First Steps in Welcome.			
Multi-threaded transformations	Define and combine filter, sort, join, aggregate, pivot, reformat, surrogate, lookup, and many other transforms in SortCL jobs built with any of the 5 ETL design options above, and run them with CoSort or Hadoop engines. See Custom Transforms and review the *Custom Transforms* chapter of the CoSort manual for defining complex transformations.			
Sorted, auto-configured loads	Feed and speed your RDB load utility with pre-sorted flat files. Create standalone sort and loader control files from the CoSort sort or join job wizard launched from the CoSort toolbar menu, or embed pre-sort specs in ETL or offline reorg jobs built from their wizards in the ETL and FACT menus.			
Data federation	Integrate and report in the same SortCL (.scl) or NextForm (.ncl) job script with a virtual target (specify /OUTFILE=stdout) to see, not persist, purpose-built mashups.			
Data replication	Move, migrate, manipulate and/or copy data from heterogeneous sources so you can use regular, fresh datasets for testing, reporting, or other applications. Specify your source and target layouts in a NextForm			

Data Integration	
	migration wizard, or, if you also need to sort, join, or aggregate, use the CoSort new or sort or join job wizard. For complex needs, use ETL. To refresh, see change data capture (CDC) under Analytics below.
Data pivoting	Use the Pivot wizard to convert (pivot) rows to columns when dealing with denormalized tables. Unpivot converts columns to rows and can be used to create reports based on normalized tables. This wizard also works on flat files.

Data Migration	

Data type translation	(Re)define the data type of any output (or pre-action layour / inrec) field in your CoSort SortCL, NextForm, FieldShield or RowGen script; or, via GUI:
	  1. Open the Data Targets screen from any of the new job wizards in the toolbar menus, or right-click in the script editing window and click **IRI > Edit Targets**
	  2. Click **Target Field Layout**
	  3. Select the output field from the named target tab below and right-click **Edit**
	  4. In the Field Editor dialog, select from the Data Type menu in the Layout tab.

File format conversion	Switch between popular flat-file and legacy COBOL or index file format through /PROCESS type declarations in the Edit Target Options dialog in any of the above wizards, or from any job script per the process above.
Endian conversion	Support and change between big and little endian conditions directly in infile, inrec or outfile /FIELDs in job scripts, or through the Target Options dialog.
Database migration	Launch a new Voracity ETL, CoSort, or NextForm Multi-Table Migration job wizard to define source and target layouts on a multi-table basis, and apply cross-table mapping rules. Run the auto-created batch file on the command line or from a workload automation (scheduling) tool with inserted checkpoints and previews of the transform mapping blocks. To migrate from a relational DB to a data warehouse star schema, see this example.
Metadata translation	Use the Import File Metadata or Import Table Metadata wizards available in the IRI menu, and/or make use of external "To IRI" metadata conversion technology available from AnalytixDS or Meta Integration.

Data Governance	
Static data masking	Find and protect personally identifiable information (PII) in flat-file and DTP-connected data sources in the Data Source Explorer pane. Use the Data Discovery tools to identify the precise locations of PII.

Data Governance		
	Use FieldShield, CoSort, or Voracity ETL wizards to build mapping jobs with ad hoc or reusable field-level protection functions. Use the FieldShield New Multi-Table Protect Job job wizard to apply common masking rules across tables to save design time and post-protection referential integrity.
Dynamic data masking	Consult the IRI FieldShield reference manual or Java / .NET SDK documentation for calling encryption, hashing, encoding, and redaction libraries into applications moving or accessing plaintext.
Synthetic test data generation	Use IRI RowGen wizards to generate and populate structurally and referentially correct database prototypes, test files, and report simulations using production metadata, not data. See the test data management blog. When you are ready to switch from test data generation to real data processing, change the /PROCESS=RANDOM declaration to the real input format in the /INFILE section of the script, or with the Edit Source Options dialog.
Metadata management	Use the metadata wizards in the IRI menu, and Eclipse-compatible source code control systems like EGit, to build, search, share, secure, and track the lineage of IRI and other EMM assets, including Voracity-supported BAT, DDF, INI, JAR, SCL, SQL, and XML files.
Master data management (MDM)	Use the new MDM wizard in the Voracity toolbar menu to search for, report on, and unify redundant values into master data repositories.
Data quality	Find bad data through fuzzy logic searches in the Data Discovery and MDM wizards. Use the MDM wizard to unify and make master values from source values. Use CoSort SortCL scripts to cleanse: scrub, validate, enrich, bucket values, and template formats. Embed third-party API functions as custom transforms in SortCL to leverage advanced standardization libraries.
Forensic metadata and job audits	Use Data Discovery wizards to expose discrete values and metadata in those sources. Name an XML audit file in SortCL, NextForm, FieldShield, and RowGen scripts, or in the CoSort Resource Control File (Run menu) dialog's Performance tab to save job details.
DAM/DAP, DLP, and Redaction (Pending)	Integrated data-centric audit and protection along with database activity monitoring and policy setting, plus innovative data loss prevention and atomic data remediation technologies will be available from ISVs using IRI's "Gulfstream" API for Voracity metadata exchange.

Analytics	
Detail/summary reports (embedded BI)	Use the file and record formatting features in the Target Option and Target Field dialogs, or refer to existing reporting scripts like this. Include transformations and field-level masking to save job steps and I/O passes in producing information.
Bulk/batch change data capture (CDC)	Use the CoSort New Join Job wizard or build a join task with an ETL designer (above) that reflects the logic in this script to trap changes in any two sources off-line, without DB logs. Faster in volume, this method also supports cross-calcs and reformatting. Specify /query or /include filters to limit the analyzed data, and set conditions in workload automation tools like Oracle Jobs Scheduler to refresh/update the deltas in near real-time.
Clickstream analytics (data webhousing)	Use the ETL and reporting design options above to build data integration and reporting jobs from CLF and ELF web log sources. See below regarding preparing data for hand-offs to third-party visual BI and analytic platforms.
Customer data integration (CDI) & segmentation	Use the ETL and reporting design options of New Join Job wizards in the CoSort or Voracity ETL menu to build ETL flows from DB and file sources of customer and transaction data. Apply custom include and omit logic in scripts or through the Record Filters option in the Data Sources or Data Targets dialogs to direct new value groups to different targets. Define custom PII field masking functions and record layouts for each target based on data privacy and business requirements in the Target Field Layout Editor dialog.
Slowly changing dimension (SCD) & trend reports	Refer to existing SCD job script samples in the SortCL chapter of the CoSort manual at section 24.3.2, Slowly Changing Dimension Reporting. IRI plans to front end SCD applications in Voracity menu wizards in the next release to generate these scripts automatically.
BIRT integration	Follow these steps to install BIRT into the Workbench, and then design and launch CoSort SortCL data preparation jobs that get run when BIRT data presentations (reports) do. BIRT will use the IRI Data Source visible in the ODA folders of the Data Source Explorer. This hook enables real-time transfer of SortCL's stdout records and field definitions into BIRT reports.
Data preparation (franchising) for BI/analytic tools	Use an ETL or new CoSort job wizard to gather, filter, cleanse, sort, join, aggregate, de-normalize, mask and otherwise transform disparate data sources into ODBC, CSV, or XML file subsets that BI and analytic tools can more readily ingest.

INNOVATIVE ROUTINES INTERNATIONAL (IRI), INC.
Suite 303, Atlantis Center
2194 Highway A1A
Melbourne, FL 32937-4932 USA
Phone +1 321-777-8889
http://www.iri.com