IRI Voracity
An Insatiable Appetite for Data

Stock, Govern & Fish the Data Lake in a Bigger, Better Boat

Whether on-premise or in the cloud, data lakes all present the same challenges: how to most efficiently populate them, govern them for quality and security, and best leverage them for analytic value. The IRI Voracity data management platform—powered by CoSort or Hadoop engines and built on Eclipse—provides a uniquely fast, versatile and affordable environment for doing it all.

STOCK (POPULATE)

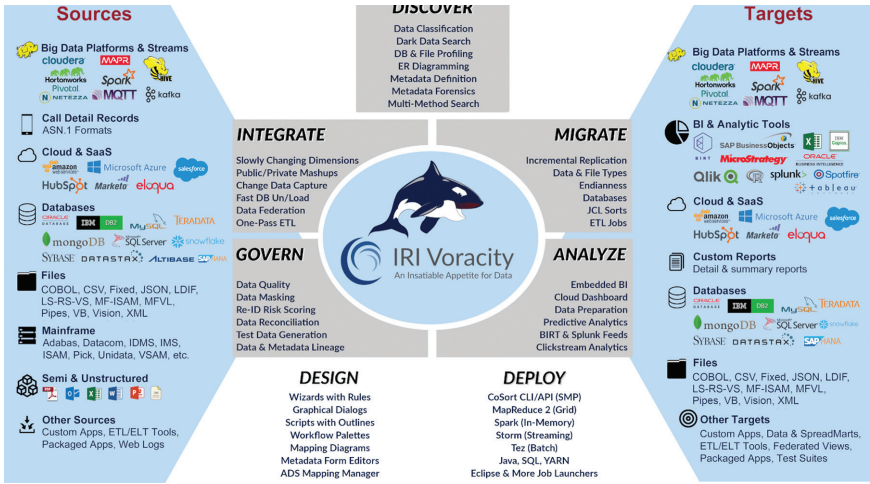
Getting the right data into your data lake will affect its storage footprint, downstream analytic potential, and how data owners perceive your intentions.

Thorough profiling processes that involve source and target stakeholders, and tools that stratify, classify, and search through data, you can determine the nature and fitness of data for purpose, and inspire confidence. These activities are also necessary prerequisites for metadata management, data integration, quality, security, and analytic activities.

To replace or mimic a data warehouse, consolidated high-speed ETL operations can populate a data lake with the right raw, and refined, data. Diagrams and metadata from these jobs can also lend structure to what can otherwise turn into a forgotten "data dumping ground."

To speed both discovery and integration, use fast acquisition and transformation engines like those in Voracity: IRI Fast Extract and IRI CoSort or Hadoop which do not tax databases and applications, or require Java steps that legacy ETL tools see crawling or crashing in volume. Ensure your data mapping engine also handles legacy and semi/unstructured data, S3 and HDFS, URLs and brokered data streams.

GOVERN (CLEAN, PROTECT, DOCUMENT)



Without the structure developed from prior profiling and ETL steps, there is no understanding. And without understanding, there can be no trust in the quality or security of data in the lake, either.

It makes sense therefore to use a common layout format—and metadata management infrastructure—in the data lake. Voracity uses the same simple, shared 4GL to define and document data cleansing and masking jobs as it does for ETL operations.

Voracity can also combine all these operations in the same job script and I/O pass, and produce visual workflows and transform mapping diagrams to reflect everything you do. So, add to your ETL jobs data cleansing functions that:

- Filter and de-duplicate
- Validate and replace
- Synthesize and enrich
- Unify and standardize

And to those, add the simultaneous or separate ability to mask PII in structured, semi-structured, and unstructured sources. Voracity includes deterministic, secure, and reversible (or not) functions to:

- Encrypt, hash, and/or tokenize
- Pseudonymize
- Redact or randomize
- Blur or bucket to anonymize
- Encode or scramble

Voracity can also score the risk of re-identification from quasi-identifiers, and create compliance logs.

FISH (ANALYZE)

Once data is wrangled through the separate or combined processes of ETL, cleansing, and masking, it is ready to analyze through any models you need and visualize in any dashboarding system you prefer, including reporting and statistical analysis that can be performed during the aforementioned preparatory work. Such consolidation led Dr. Barry Devlin in 2018 to label Voracity a "Production Analytic Platform."

Voracity supports these analytic options for data lake users:

- Embedded BI—report and analyze *while* blending
- BIRT & KNIME—use Voracity data from APIs in the same Eclipse IDE
- Data Wrangling—Give display-ready subsets to Power BI, Qlik, R, Spotfire, Tableau, et al.

LEARN MORE AT
iri.com/voracity