IRI
Total Data Management

CELEBRATING 40 YEARS 2018
The CoSort Company

Embracing Velocity With Voracity

The data on which businesses now rely not only exists in large and small collections, but it can come from new places at unpredictable speeds and intervals, and in unfamiliar or inconvenient formats. It can also be in different states of quality and security, and it may need to be mashed up with static legacy sources for analytic value.

In this increasingly complex context, what does an ideal technical and commercial solution for managing data and getting value from it look like? And what does ideal mean in terms of data integration and insight speed, governance functionality, and cost (of deployment, maintenance, and change)?

IRI users have asked these questions since 2003 when we started talking about big data and data franchising (now called data wrangling).

A FOUNDATION IN SPEED

Since IRI's beginnings 40 years ago, the company's CoSort utility has brought affordable data processing speed in volume to thousands of users worldwide. CoSort and its spin-offs do their work in state-of-the-art data transformation algorithms running inside I/O- and memory-optimized multi-threaded C code.

Sort, join, aggregate, lookup, cleanse, mask, reformat and other custom mapping and layout functions all happen in the same program at once. And they apply to multiple, multi-gigabyte data sources which become multiple targets in seconds, on one node.

IRI customers have found this combination of discrete task optimization and single-pass consolidation is essential to SLA delivery amid growing data volumes and shrinking production windows. This time-to-value advantage has remained unmatched in the data management market anywhere near IRI price points.

In most other data integration environments (think about the top ETL tools you know), critical sort, join, and aggregation transforms run in separately compiled Java programs that require partitioning and enough memory to hold the entire input set. When their engines fail, they "push down" the problem to an already query-taxed database, a costly appliance, or a complex Hadoop paradigm.

THE POWER OF THE PLATFORM

Beyond handling data rapidly without more hardware or engineers, there is also efficiency to be had by marshalling data centrally. Staging data for BI and analytics in one place allows that data to be prepared faster and consumed consistently. In addition, acquiring and integrating disparate data sources in the same way and UI makes it easier to design data-driven jobs.

Similarly, when disparate data elements like DB columns, CSV fields, and JSON key values are expressed simply—and in a common, open syntax—you can work with them faster. Data class unification and rule application, as well as common symbolic mapping references, are all enabled. So too are data and metadata lineage, because field names correspond to source locations specified in industry-standard syntax.

Many of these good ideas are already leveraged in ETL tools on the market. However, most of them are slow, expensive, and hard to use. They are also limited in functional scope, particularly in data governance and metadata management. For this reason, IRI released a new platform in 2016 called Voracity to address the technical challenges of managing today's large and evolving data landscape, and the commercial challenges of "megavendor" ETL providers.

Beyond affordable speed lies a unique combination of linear scalability in volume, functional versatility, and centralized ergonomics.

MAP ONCE, DEPLOY ANYWHERE

In Voracity, jobs run multi-threaded in the CoSort engine. Many also run interchangeably in Hadoop via MR2, Spark, Spark Stream, Storm, or Tez from the same IDE.

IRI agrees with Informatica on the benefits of this approach (which they call multi-latent abstraction). There is no code to rewrite, and the current state of the data does not matter. Choose the engine based on where the data and available hardware are.

BUT DO MORE AT ONCE

Voracity is an end-to-end data lifecycle management and solution stack. It was built for "big data" lakes and BI/DW users who want to do more at once, and as quickly and affordably as possible.

Voracity enables this goal by supporting and combining discovery, integration, migration, governance, and analytics. It runs many of these tasks in the same job on Unix, Linux, or Windows machines—from a Raspberry Pi all the way up to a Hadoop cluster in EC2.

AND MAKE IT EASIER

Six interlinked graphical design modes in Voracity's Eclipse IDE (IRI Workbench) give users a choice of how to build and modify jobs. Open metadata and teaming plug-ins speed job ramp-up and sharing, as well as data lineage and compliance tracking.

IRI
www.iri.com