perspecta LABS

Arroyo data management suite

Efficient, easy-to-use tools for data manipulation, investigation, analysis and management

The challenge

We live in a data-saturated environment in which organizations, agencies and enterprises are easily overwhelmed with large sets of messy, diverse, complex—and potentially quite useful—data. Leveraging the ever-increasing collections of data to extract insights and information requires efficient, easy-to-use tools for data manipulation, investigation, analysis and management. Across diverse domains from health care, finance and engineering to compliance, operations and intelligence, the challenges are the same. Analysts need to reconcile, correct, validate and integrate data sets. They need both interactive capabilities to investigate and explore and high-performance processing at scale.

The Arroyo solution

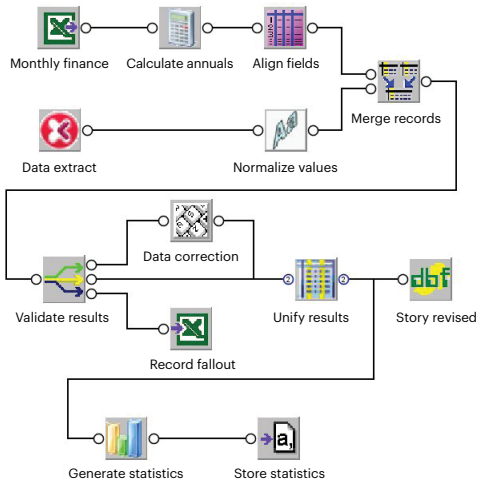
Perspecta Labs' Arroyo data management suite provides an interactive, graphical environment for fast, code-free development of solutions to extract, transform, validate, explore and analyze diverse data. Arroyo efficiently supports the full-spectrum of data management tasks from ingestion, normalization, combination and reconciliation to geospatial analyses, visualizations and sophisticated analytics and machine learning. Arroyo is a high-performance, all-purpose tool—scaling to meet demanding transaction volumes involving terabytes of data and billions of records. It reads, processes and writes structured and unstructured data in diverse types, repositories and formats.

The Arroyo suite of tools enables rapid data exploration, correction, modification and management at scale and for large sets of varied data via:

- Easy-to-use graphical interface used to define, debug, execute and visualize complex data manipulations
- Nearly 100 reusable, configurable building blocks (filters) in the Arroyo suite that provide customizable, built-in functionality to ingest, transform, validate, analyze and output data
- Simple drag and drop construction so that the development and debugging of complicated data management solutions (flows) is an agile configuration task—not a strenuous coding process
- Fast review, update and verification of flow behavior with real-time views and analysis of data flowing in and out of each filter within a solution flow
- Flexible, high-performance operations and bulk data processing via Arroyo's execution engine which can schedule and execute flows based on time, presence of data and other factors
- Easily extensible with full-featured scripting language to incorporate additional filters, enhanced visualizations and new processing and analytic techniques in the users language of choice

The simple Arroyo flow shown below demonstrates the construction of a basic data management process to:

- Extract, align and normalize data from two disparate sources
- Merge the data and perform data validation, reconciliation, error detection and correction
- Unify and store the cleansed data
- Create statistical summaries of the data and data processing, including uncorrectable errors, for use in reports and dashboards



Using Arroyo's drag and drop interface and its wealth of configurable filters, users create solution flows to read, modify, evaluate and write data without programming. Arroyo flows enable quick, interactive data discovery as well as design and testing of data processing solutions for advanced analytics and bulk data management.

Arroyo runs on Windows and Unix and is flexible and extensible:

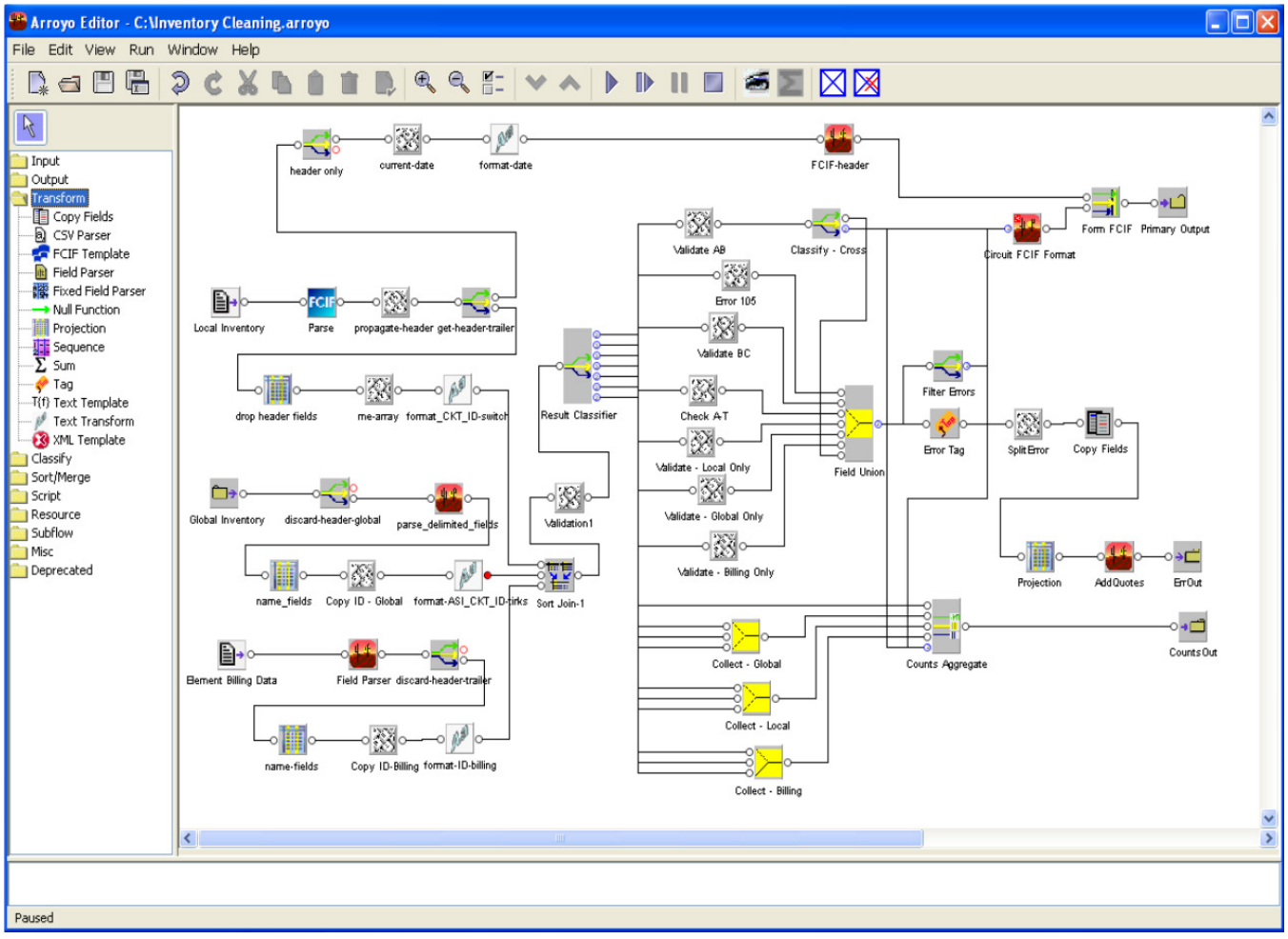
- More than 30 configurable input/output filters to support read/write operations from local files, remote (web-based) sources, local or remote databases and distributed Hadoop file systems
- Easy processing of diverse data structured as flat files (raw text, delimited data, fixed-width data), structured files (HTML tables, XML, JSON, spreadsheets), databases (JDBC, XBase file) and messages (JMS)
- Customizable filters for geo-spatial operations and visualizations, including 2D, 3D and maps
- Comprehensive suite of text analytics capabilities enabled via seamless Python integration
- More than 40 configurable processing filters for efficient sorting, classification, comparison, cleansing, modifying and unification of data
- Simple, comprehensive scripting capability so users can develop and add new features and functionality using choice of languages
- Full set of machine learning capabilities available via the Python scikit-learn suite, including a wide range of machine learning algorithms and preprocessing capabilities used to generate train/test splitting and perform n-way cross-validation and results scoring
- Flexible output filters that produce analytic results, plots and statistical results in standard formats for use in reporting and dashboards

Arroyo, as shown below, enables the rapid development, debugging and validation of complex data processing solutions (flows) without programming. The user-friendly graphical interface makes solution development straightforward via drag-and-drop construction. The resulting Arroyo data processing and management flows can be run interactively for investigation and analysis and executed at scale using Arroyo's high-performance execution engine.

The Arroyo advantage

Automation is critical to cost effectively resolve data quality and transformation issues. Our Arroyo data management suite automates the data extract, transform and load activities allowing visual exploration of data for pattern detection. With Arroyo—you can efficiently support the full spectrum of data management tasks.

Learn more about the Arroyo suite of tools at **perspectalabs.com**.



Arroyo case studies

Case study #1

Health care data cleansing and integration

Health care data arises in a wide variety of diverse forms with different formats, semantics and data structures. Cleansing and integrating multiple sources requires flexible analytics to identify potential errors and replicates and automates data processing steps to correct, harmonize and de-duplicate the resulting data. Arroyo's comprehensive set of filters and interactive flow development capability have been successfully used to develop solutions for cleansing and integration, including de-duplication of health care data involving medical tests/procedures, medications, interventions and patient visits.

Case study #2

Data migration for regulatory compliance

A major telecommunications firm faced the challenge of migrating, validating and consolidating large data sets across five disparate systems to meet new requirements in a changing regulatory environment. The data processing operation involved cleanup, merging, transformation and migration to a target system. Arroyo was successfully used to design, develop, deploy and execute a high-performance migration solution. The migration solution ingested and reconciled data across the systems, and met challenging requirements for flow-through, processing speed and accuracy.

Case study #3

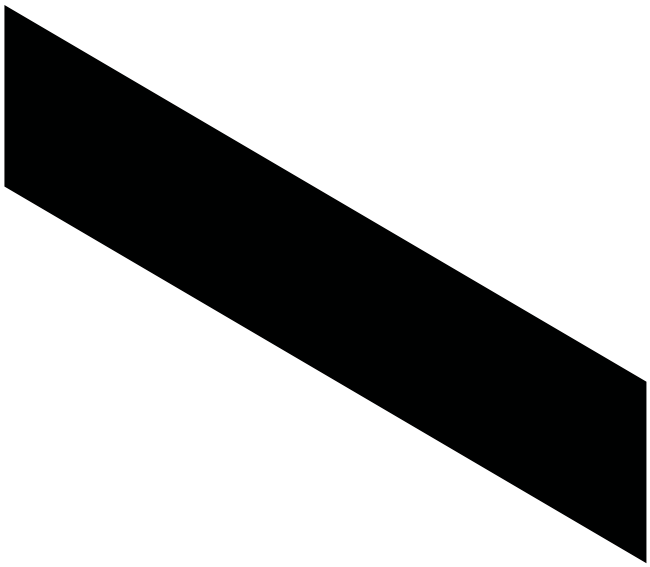
Textual and geospatial data extraction and modeling

The U.S. Army Corps of Engineers was looking to leverage Automated Identification System (AIS) ship tracking information to understand the flow of vessels between ports. Arroyo was used to map detailed ship paths to registered ports and construct a social network model of port interconnectivity. Based on that model, analyses in Arroyo identified patterns and trends in the relationships, such as common paths and seasonal variations. The results provided insights into both waterway usage and the direct benefits of Army Corps waterway maintenance activities.

Case study #4

Enterprise efficiency improvement via analytics

A large U.S. enterprise was struggling to improve a complex order handling process which was not achieving required levels of flow-through operation. Analysis of orders falling out of the system identified more than 50 different types of discrepancies, omissions and inaccuracies that caused order processing failures. Arroyo was used to create a system for near-real-time failure assessment. Triggered by an order-handling error, the solution gathers relevant data across component systems and validates against element specifications and the data model to identify and correct the root cause.



Learn more at perspectalabs.com

perspectalabs.com | info@perspectalabs.com
150 Mount Airy Road
Basking Ridge, NJ 07920