EBOOK

GUIDE TO SELF-SERVICE DATA PREPARATION

Whether you call it data preparation, mining, extracting, cleaning, joining, blending, or masking, it's all data transformation. Professionals can spend hours every week working with their data, trying to join data from disparate sources, re-keying info from static reports or PDFs, or formatting data for accurate reporting. Self-service data preparation tools empower users to import, transform, and export data efficiently and with vastly fewer errors than may arise in manual workflows. The right tools also enable them to automate processes to ensure consistently repeatable results.

This guide will help you assess your requirements and develop an implementation plan to increase efficiency and reduce errors in your data transformation processes.

CONTENTS

SIX KEY REQUIREMENTS OF SELF-SERVICE DATA PREPARATION

There are numerous books, courses, and websites designed to help define requirements that will drive every stage of an IT project and a general discussion of a requirements gathering process is beyond the scope of this guide. However, data preparation presents specific challenges that must be considered when selecting technology partners and deciding on the best approach to implementing data transformation workflows. In particular, make sure your full list of requirements addresses these six areas:

1. Access to Semi-Structured and Structured Data Sources

Often the most important information you need is locked away in semi-structured documents and sources that seem impossible to access without rekeying the data. The most common semi-structured data sources are PDF, text, and printer reports (PRN files). The most common structured data sources are Microsoft® Excel®, Microsoft® Access®, delimited files, and Microsoft® SQL Server® or Oracle® database tables. You may also need to access sources like web pages and text files. The ability to access sources like this adds tremendous value to your operations.

Start by making an inventory of all the types of files you would like to be able to incorporate into your analytics processes. Then collect a range of typical examples of all the file types you expect to encounter, including examples that contain known errors or are structured in particularly odd ways. Test the data transformation solutions you are considering using this collection.

2. Data Masking

As data volume grows exponentially, businesses are faced with the challenge of scaling their protection of sensitive information to meet government standards and to defend against theft. Data masking has emerged as the leading solution to accomplish this across all industries and verticals. Data masking is the process of creating a new version of an original data set that maintains the data formatting but generates new data values. Hiding the original data in this way allows it to still be usable for analytics, training, or testing, while ensuring that the underlying data is only visible to approved users.

When evaluating data transformation systems, test their data masking capabilities using your collection of real-world example files.

3. Automated Processes

Automating data prep processes can obviously save time but remember that automated processes also reduce the chances of introducing errors. Document all the data transformation processes your organization must do regularly, including those that may occur only once a year. Identify the most business-critical processes and be sure to test the automation capabilities of your proposed technology solutions.

4. Reduced Risk with Improved Data Governance

The move to self-service is all about speed and agility for the business user. But giving up control can increase the risk of unwelcome data breaches, errors, and compliance issues. You must introduce streamlined data governance that takes into consideration that data preparation often involves non-managed data sources like CSV extracts, PDF reports, or third-party data.

Look for an enterprise solution that can control access to prepared data sets, reusable models, visualizations, and dashboards.

5. Ease of Use

The keys to user acceptance of new data transformation software are a short learning curve and overall ease of use. As you look at different solutions, get a variety of qualified people involved in the evaluation process. Can they learn the basics of using the new tool quickly or do they struggle? Are prebuilt models included with the software that users can access and make sense of? Can power users create and save manipulation functions that connect to data sources, add calculated fields, and mask data without learning a new script language or complex flow diagrams? Can business users get to those models and use them effectively without excessive support?

6. Interoperability

After you have accessed, cleansed, and organized your data, you must be able to make fast effective use of it within the context of your analytics infrastructure. All data transformation solutions can produce flat files in CSV or similar formats, but the most efficient implementations will work directly with your other commonly used tools. A system that requires "export and then import" steps is generally not sufficient. Look for software that offers native data connectors for the data visualization and business intelligence applications you use to guide your decision-making processes.

BRIDGE THE GAP BETWEEN SELF-SERVICE DATA PREPARATION AND ENTERPRISE GOVERNANCE

More organizations are recognizing that data transformation tools are a necessary component for all their data discovery and advanced analytics implementations, and many have developed strategies to govern data that resides in managed systems like enterprise applications or data warehouses.

One of the biggest benefits of self-service analytics lies in its ability to combine and analyze data from a variety of disparate sources. However, this creates a serious governance challenge, especially in cases in which more than half of the data originates from sources not typically managed by the IT department. For example, analysts may wish to pull data from CSV or text files extracted from transaction processing systems, personal spreadsheets, third party reports, and similar semi-structured and unmanaged sources. Issues can arise around version control, data breaches, reconciliation, auditing, and more. An enterprise data preparation platform must address these governance risks without introducing undue delays or cumbersome workflows that erode efficiency or create pushback from users.

Governing “Non-Managed” Data

Most organizations have well-formed strategies to govern data that resides in managed systems like enterprise applications or data warehouses. However, there is often a need to also create a content repository; this lays the foundation for governing “non-managed” data. The repository must support these capabilities:

- **Data Retention:** Document version control is a must-have to ensure consistency. Note that your organization may also need to persist and archive all source data/documents to meet regulatory or business requirements. Document all applicable policies, regulations, and informal business requirements that may affect the configuration and retention capabilities for your repository.

- **Data Masking:** Employees and contractors are often involved in data breaches which may be intentional or accidental. Data discovery tools are great ways for users to build and share information, but in many cases the underlying data is not protected and includes personally identifiable data (like Social Security numbers), personal sensitive data (like medical procedures), or commercially sensitive data. Industry rules and government regulations may require an organization to pay substantial fines in the aftermath of a breach which will be in addition to the legal costs and reputational damage an organization will suffer when a breach occurs.

  When selecting data preparation tools, test their ability to mask sensitive data and maintain that masking throughout your workflows. Be certain that your teams can put effective masking in place and ensure that it cannot be removed or altered easily either by accident or by unauthorized users.

- **Data Lineage:** Your systems must provide a complete data lineage that allows authorized users to drill down to the cell level of source documents. Auditors must be able to identify the transformations that were applied to your data from the source through to the final output of your systems, and trace back the origination point for all data. This is critical for effective audits and data reconciliation processes. A well-implemented system will provide your organization with the ability to trace data back to its ultimate source, understand exactly which sub-systems were used to extract data, and the source files and functions used to generate calculated values.

- **Role-Based Access and Data Curation:** Your systems must allow multiple individuals and teams to share frequently used data sources and automated data transformation routines, and this control must be based on well-defined and enforced user roles. Those systems must also allow appropriate team members to assign and control which user groups everyone belongs to and manage access to subsets of data based on user roles. These capabilities instill confidence among users and managers and helps ensure consistency in the data used to make business decisions.

While it is possible to create a data prep infrastructure that functions like a “black box,” this model will not serve you well if someone questions the data. When your team is asked to substantiate a particular output value, you must be able to respond with complete information, including timestamps that indicate when each process was run. The best data prep tools support efficient operations and ensure that decision-makers can trust your systems to generate accurate information.

When selecting data preparation software, make sure that its auditing functions cannot be disabled.

USE AUTOMATION TO EXPAND CAPACITY AND INCREASE DELIVERY SPEED

Data prep software should make accessing, preparing, and combining data from a variety of sources relatively easy. Timely delivery of clean, governed, and properly prepared data is equally important.

While it is not an absolute necessity for small-scale deployments, the ability to automate data acquisition, transformation, and delivery is a must-have capability at the enterprise level. Look for systems that provide comprehensive frameworks that support alerting, distribution, and multi-input processing and that do not require any coding or scripting. This will keep your operational costs low and expand the scale at which you can deploy data transformation processes. In addition, the centralized management inherent in an automated system provides a single, consistent, secure “source of truth” for data and transformation models.

If automation is a requirement for your operations, make sure any systems you deploy can support the following capabilities.

Automated Data Access and Distribution

The system should be able to access source data, whether it be in the form of PDFs, flat files, spreadsheets, traditional relational databases, or web pages, in a consistent, reliable, and repeatable way. After processing, the system should be able to transfer the data reliably to a variety of destination systems and formats. For example, in some applications you may require the output of your data transformation workflow to be a set of CSV files on a Microsoft® Sharepoint® server, while in other workflows you need the data to be exported into a SQL database. In other cases, you may need to have the data loaded directly into a third-party enterprise system like Salesforce.

Scripting and API Functionality

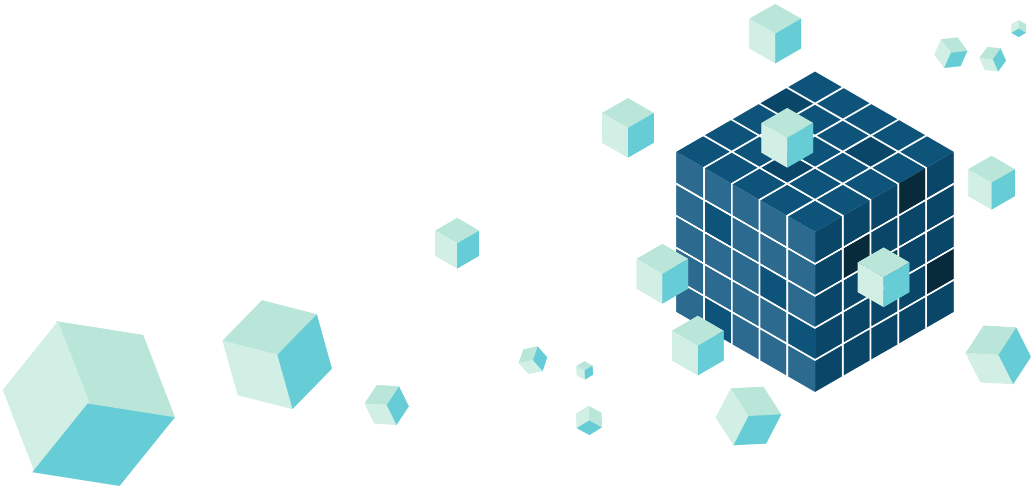
Future demands are impossible to predict, so your system must offer a rich application programming interface (API) that will enable you to invoke processes and workflows without using the product’s standard user interface. A rich API permits users to create powerful scripts or leverage third-party tools to invoke, monitor, or customize processes as part of larger, more complex solutions.

Scheduling and Monitoring

Automated systems must be capable of performing data access, transformation, and export processes based on a schedule and be able to perform operations on an event-driven basis. For instance, you may need to have certain processes run once a day after the close of business, while others must be triggered to start when new data files arrive in a specified location.

Scalability

Once reliable and trusted data prep workflows are in place, it is inevitable that the demand for such services will expand. Regardless of whether you need such capabilities now, the ability to scale up is essential. Look for systems that can support multi-server deployments that enable failover, process engine pooling, and active/passive recovery models.



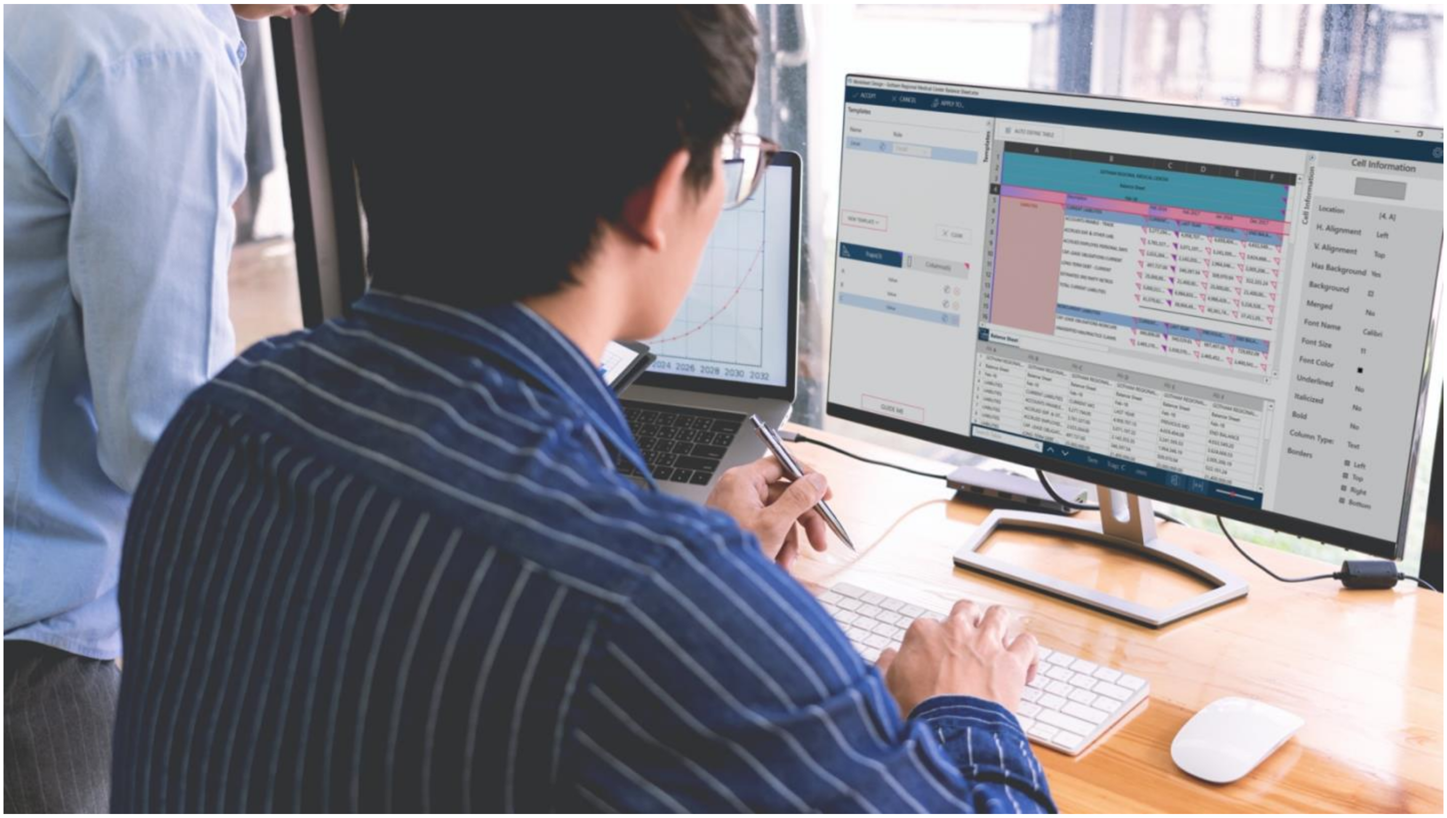
THE VALUE OF SELF-SERVICE DATA PREPARATION

Self-service means that analysts and business users — people without specialized training in databases, data science, or programming — can set up, run, and maintain data preparation processes without the need for IT intervention. This implies that your systems offer intuitive user interfaces that allow people to capture every step of the process in reusable and human-readable workflows they can execute on demand or in automated routines.

Managers must make business-critical decisions based on information generated by data preparation processes, which must be transparent to everyone involved should questions arise. If someone asks, “Where did that outlier come from?”, the system should be able to report on the complete data lineage and even allow users to drill down to the underlying data sources, which for example might be a PDF of a customer invoice.

Self-service preparation systems can deliver trusted datasets to data warehouses or departmental data marts using industry-standard database drivers and allow frequently used data sources or automated data preparation routines to be shared among authorized users. This adds significantly to the efficiencies to be gained from implementing a self-service platform since users do not have to reinvent the wheel with each new project. People can simply reuse existing workflows or modify them as needed to accommodate new requirements.

With self-serving data preparation, business users can achieve excellent results quickly with fewer people involved. They can reduce or eliminate the delivery of outdated, inaccurate information to decision-makers while ensuring that the data is complete and clean. In addition, self-service capabilities foster employee empowerment and help the firm attain strong ROI for its transformation technology investments.



DATA ANALYTICS SOLUTIONS FROM ALTAIR

Insight doesn't come from what people can see on the surface of a data set. It comes from the hundreds or thousands of dimensions hidden in complex data. People need the right tools to easily access these hidden dimensions. Altair empowers business users to collaborate efficiently to access meaningful data, generate insight from this data, and share their finding throughout the enterprise. Altair enables people of different skill sets to easily build complete analytics applications to support fully informed, insightful decision-making.

Data Preparation

Altair's data preparation software enables business professionals to discover, build, share, and collaborate on secure, governed, and trustworthy data sets and models. It can access, cleanse, and format data from a wide variety of sources (including Excel, CSV, PDF, TXT, JSON, XML, HTML, SQL databases, big data like Hadoop, and more) without any manual data entry or coding. Dozens of pre-built data preparation functions make combining disparate but related data sets easy to do quickly. This simple approach to data cleansing eliminates the need to code, script, or create pivot tables or vLookups in Excel. Clients can deploy these tools on the desktop, with on-premises servers, or in the cloud.

AI and Machine Learning

Altair's open, flexible machine learning platform is designed for data scientists and business analysts alike. Its industry-leading visual approach to analytic modeling enables data science teams to create high quality machine learning and artificial intelligence (AI) models. Our collaborative approach to machine learning enables your data scientists and business users to minimize repetitive takes related to creating curated and governed data sets, share knowledge across the enterprise, and reuse steps within connected model workflows for faster analysis and sharing of insight. Altair's code-optional development environment enables data science teams to build models using combinations of SAS language, Python, R, and SQL code.

Data Visualization

Altair's visual analytics tools are optimized to handle time-critical data, including data that may be changing very quickly. Business users can connect to data sources, build, and publish sophisticated real-time dashboards. The platform's filtering tools enable users to zoom in and out on the timeline, remove false positives from the screen, and focus on exceptions. Users can solve difficult problems quickly, understand complex relationships in seconds, and identify issues requiring further investigation with just a few clicks.

Stream Processing

Altair's stream processing engine event processing engine connects directly to a wide range of real-time streaming and historic time series data sources, including MQTT, Kafka, Solace, and many others. Users can build complex stream processing applications with a fully drag-and-drop interface, without writing any code. Applications may combine streaming data with historic data, calculate performance metrics using a wide variety of statistical and mathematical functions, aggregate, conflate, and compare data sets, and automatically highlight anomalies against user-defined thresholds.

ABOUT ALTAIR

Altair is a global leader in computational science and artificial intelligence (AI) that provides software and cloud solutions in simulation, high-performance computing (HPC), data analytics, and AI. Altair enables organizations across all industries to compete more effectively and drive smarter decisions in an increasingly connected world – all while creating a greener, more sustainable future.
For information visit www.altair.com.

Contact Us

Try Us! Click here to obtain a free 30-day trial of Monarch,
our market leading self-service data preparation tool.