1 Introduction

The Datawatch Data Preparation Buyer’s Guide is your manual as you navigate the self-service data prep technology landscape and provides information on assessing your organization’s needs. It focuses on questions to ask when considering new analytics technology and illustrates how to fluidly transition into the phases of purchase and implementation. It should serve as a resource to help you evolve towards a data-driven organization by enabling self-service across your business analysts and information workers.

This guide is designed to take the pain and frustration out of choosing a self-service data preparation solution, and to help you select the solution that’s right for your business. The information in the guide comes directly from the experiences of our customers who, like you, converted manual operational processes into automated data access and preparation. While every organization is different, this guide represents most of their basic criteria for successfully screening and selecting a solution.
Data warehouses are inflexible. Once the data model has been defined and the data has been loaded into the warehouse, the paths for analyzing the data get frozen into place, limiting the number of potential insights that can be derived from it. Data warehouses are difficult to access if you are an analyst, and requests for data are time consuming.

Data warehouses emphasize reporting over ad hoc exploration. Data warehouses are architected to support scheduled reports or real-time dashboards created by the IT team. The rigidly structured ways in which they store data lend themselves to reports and dashboards that track pre-defined KPIs, but aren’t as well suited to exploratory analysis. For instance, the particular data attributes in which a business analyst is interested in order to answer a one-time inquiry from the CEO may not have been considered when the data was transformed into a normalized format and loaded into the warehouse.

Only IT can prepare this data. ETL tools aren’t designed for business users, even business analysts. Moreover, since one of the points of the data warehouse is to centralize and normalize the organization’s data in standardized formats, it makes organizational sense to task a single unit with preparing it.
What is Self Service Data Prep?

Data Preparation is the iterative, agile process of exploring, collecting, and manipulating data into a form suitable for analysis (reporting or processing) by cleaning and often combining or consolidating data into one file or data table. Data preparation includes transforming raw data into curated datasets for operational processes, data science, data visualization and BI/ analytics, and is most often used when business analysts are challenged with:

- Limited access to data sources, dependency on IT for access to datasets
- Trying to combine data from multiple sources
- Manual data entry into spreadsheets, reporting on error-prone data
- Dealing with data that was pulled from an unstructured source, such as PDF documents, enterprise application reports or web pages

Data Prep can be broken down into three simple and fundamental steps:

1. **Data Acquisition:**
   - Identifying and obtaining access to the data within your sources

2. **Data Cleansing:**
   - Manipulating and preparing data into a usable, functional format and correcting or removing any bad data

3. **Data Blending:**
   - Combining and enriching data with other datasets for detailed analysis or process improvements

A self-service data prep tool enables non-IT users to repeat this three-step process as many times as necessary to add or subtract data sources as needed. In addition, users can extract and blend a variety of data from disparate data sources they typically wouldn’t have access to. Data preparation is a critical component of both operational process efficiency and enabling self-service analytics.
Why is self-service data preparation important?

Self-service data prep improves agility, productivity and data accuracy

The most prominent environments where data preparation (access, blending, transforming, enriching) is utilized are within finance, operations, sales/marketing, HR and site/merchandising operations. Research has found that analysts spend up to 80 percent of their time manually preparing data, which doesn’t leave many hours in the day for analysis.

Only 12 percent of enterprise data is used for information

Most organizations have well-formed strategies to govern data that reside in managed systems like enterprise applications or data warehouses. One of the biggest benefits of self-service analytics lies in the ability to rapidly combine and analyze data from a variety of sources. However this approach also provides a serious governance challenge in that it is estimated that half of this data comes from sources not typically managed by IT. For example, analysts will pull from sources including CSV or text extracts from transactional systems; personal spreadsheets; third party reports; semi-structured content; and more. Issues then arise around version control, data breaches, reconciliation, auditing and more.

What is required is an enterprise data preparation platform that can address these governance risks but do so in a frictionless manner so the business user has the speed and agility they require. A self-service data preparation solution enables business analysts and IT workers to unlock and access more data, including “dark data” that often goes unanalyzed.

Self-service data prep provides better and faster decisions

Business analysts strive to reveal a “deeper intelligence” within their analytics. Enriching data with additional detail thanks to easier access to time-sensitive and previously inaccessible “dark data” in unstructured formats such as PDFs, web pages, machine data, log files and more, allows additional time to be focused on the analysis needed to support business-critical decisions. This enhanced data also instills greater confidence and trust when it is shared with others or used for audit purposes.

Analysts also aim for better, faster decision-making in their organization. When it comes to analytics, dark data in particular, can provide immense value and is a critical requirement in ensuring an accurate and a holistic view of the business. In turn, individuals are empowered to guide the business toward promising opportunities and benefits.
Why is self-service data preparation important?

Data Governance ensures integrity, security, repeatability and compliance

BI managers must be able to share data throughout their organization. Instead of creating a series of one-off requests, imagine if you could create a data model and could post it to your server where it is accessible to anyone who has permission to access it. Now imagine if that model could update the data on a regular basis, whenever it needs to be refreshed. With a self-service data prep tool, that defined data access or data set can be accessible to anyone who has permission. Data is made available in a common way to a common set of consumers on a regular basis. If certain steps need to be taken to prep that data, those steps can be saved and automated.

As businesses strive to evolve their analytics, BI managers are oftentimes at the mercy of IT personnel who are often challenged with being responsive to their requests and mitigating the risk of financial penalties, regulation violations and security breaches while ensuring data integrity. Among the issues at hand:

1. Excel data is often ungoverned
2. Data may contain sensitive information
3. Data breaches can be extremely costly
4. Data sources and manipulation tasks must be tracked and auditable
5. Data integrity requires consistent models and processes for reuse

The adoption of a self-service data preparation solution can help BI managers leverage features such as data masking to protect sensitive information such as patient names or social security numbers to meet regulatory requirements. In addition, a self-service data preparation tool provides IT with data governance and control while enabling self-service analytics. BI managers in turn, can:

- Accelerate the creation of analytics prototypes to be controlled and governed
- Consistently track data lineage
- Provide audit logs for all manipulation
- Reduce dependency on manual data input and blending in Excel
- Share automation processes and source content for reuse
- Maintain compliance by redacting and masking sensitive data
3. Getting started

Your first step is to assess the current state of your operational or analytics processes. By identifying your current manual processes and dependencies you will pinpoint labor-intensive tasks and areas of risk due to potential human error, along with other redundancies and inefficiencies.

### Questions to describe your current process

<table>
<thead>
<tr>
<th>Are you creating reports?</th>
<th>Who manages the process and who else is involved?</th>
<th>Is the process centralized or do different locations also perform the same or similar tasks?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reconciling data?</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Creating dashboards/visualizations?</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Other?</td>
<td></td>
<td></td>
</tr>
<tr>
<td>What system/tool is used to prepare your data today?</td>
<td>How often is the process done?</td>
<td>How long does each step take to complete?</td>
</tr>
<tr>
<td>Is all the data readily available for each task?</td>
<td>What systems are accessed?</td>
<td>If data access must be requested, how long does that take (on average)?</td>
</tr>
<tr>
<td>Is manual data entry required?</td>
<td>Once data is prepared, how is it used/where is it sent?</td>
<td>What are the limitations of your current process?</td>
</tr>
</tbody>
</table>
4. Set your goals / define product requirements

Gartner describes self-service data preparation as a key component for “ad hoc or recurring analysis”, particularly when faced with “multiple and frequently changing sources”.

Adding self-service data preparation to your analytics and operational processes will help business analysts to produce data-driven content while leveraging enterprise data and governance practices.

The next step is to define your specific data preparation goals. These goals will guide your choice of products, services, and vendors. You will evaluate vendors’ proposals based on their ability to help you meet your goals. Goals might be stated in terms of system capabilities and features or in terms of benefits your company expects to derive from the new system. Here are some typical goals that may be relevant to your organization.

- Comply with internal and/or government regulations
- Improve analyst productivity
- Reduce time to complete processes
- Improve IT responsiveness
- Enable self-service analytics
- Improve data quality and trust
- Eliminate manual processes and redundant data entry
- Automate or reuse common process tasks
- Leverage existing IT & BI investments
- Produce timely, accurate reports & dashboards
- Improve data governance
- Enable more business analysts and information workers to produce data-driven information
5. Cost / ROI

Return on Investment (ROI) is one of the most important factors to take into consideration when investing in technological solutions for your company. No one wants to spend a lot of money on a tool that provides little, or no value to the organization. Therefore, there are a few key questions to ask when considering data preparation tools:

Here are a few questions you can ask to evaluate your current cost of data preparation in your organization:

1. What is the average analyst salary at your organization?
2. How many analysts are currently on your payroll?
3. How many hours per day does your average analyst work each year?
4. What percentage of your analysts’ time is usually spent on preparing data?

Asking these questions will help you to identify how much money is currently being spent on data preparation, compared to actual analysis. More often than not, analysts are spending at least 80% of their time preparing data and only 20% of their time doing analysis. A study done by Blue Hill Research calculated that, on average, organizations are wasting $22,000 per analyst, per year when they do not introduce data preparation tools. The study also found that introducing a data preparation solution enables analysts to gain back at least 50% of the time previously spent on preparing data, if not more.

Another important factor for calculating ROI is the initial set-up costs associated with purchasing a data preparation solution. Surprisingly, the cost of data preparation tools varies widely across the industry. Therefore, it is important do some research on:

5. How much do licenses cost?
6. How many licenses are you going to need?
7. How do you plan on managing multiple licenses across your organization?
8. What are the initial start-up costs associated with a roll out?
9. Will training and/or Professional Services be necessary?
10. Do you need data preparation for a one-time project, or is it something that is mission-critical for your daily operations?
11. How are you planning on disseminating data across your organization?
12. Would a centralized location for all data be helpful?

Perhaps most importantly, decision makers should consider the potential benefits of introducing a data preparation solution. Some common benefits of data preparation include:

13. Better analytics and visualizations
14. More productive, less fatigued analysts
15. Cleaner, more up-to-date data, analysis and reporting, which leads to better business decisions
16. Reducing opportunity for human error drastically reduces possibility of incurring fines for non-compliant official reporting
17. Ability to shift focus to other priorities, such as customer satisfaction
6. Researching and Selecting your Provider

The first step in the provider selection phase is identifying and reviewing existing, relevant providers. There are many effective ways that leaders identify providers. Some leaders ask colleagues who work in the same profession, some run exhaustive searches online, others hire consultants, and still others use analyst reports to help narrow down the number of solution providers to assess. Whatever method you choose, research will have to be done to verify capabilities.

Many companies procure software solutions by assessing vendors through a detailed Request for Proposal (RFP). By preparing an RFP, you can solicit feedback on your specific requirements in a common format, making it easier to compare different vendor solutions.

Whether or not you send out RFPs, you should prepare a comprehensive list of the requirements for your data preparation solution. This exercise will help you select a system with the right capabilities and features for your needs.

Most companies list their detailed requirements as a series of capability questions and then indicate if each vendor:

- Currently has the feature in its standard product
- Will have the feature by a specified date
- Can make the feature available on a custom basis for an additional cost

You can now utilize your product requirement and RFP responses to evaluate and choose the most appropriate vendor offering by considering deployment model, end-to-end capabilities, pricing, support for data sources, end-user roles and any integration with existing vendor platforms in which you have already invested.

When creating your RFP, include several key items:

- An overview of your organization and your software needs
- System functionality and IT specifications
- Architecture and technical requirements, including security, accessibility, compatibility with specific databases, Web technology, reporting, server/desktop requirements, etc.
- Required or desired system features
- Request for the product roadmap for the next several years
- Company financial viability, market share, strategic partnership, industry leadership, number of employees, years in the business, and support model
- Software costs (Ask that the provider includes all fees associated with SaaS or On Premise software, any support fees, training, customer support, and implementation support fees)
- Integration requirements from both a personnel and price perspective
- Information on the number of customers the provider serves and the retention rate of those customers. Also ask for information on whether they support other organizations in your industry.
- Integration capabilities with existing enterprise applications, BI platforms, visualization tools or other systems. Also is this solution part of a broader analytics platform
- Customer references

DATAWATCH
## Sample Self-Service Data Preparation RFP Questions:

### Business & Functional Requirements

<table>
<thead>
<tr>
<th></th>
<th>Data access</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Provides the ability to visually access and import structured and semi-structured data from a wide variety of sources.</td>
<td>a. Relational</td>
<td>b. Big data</td>
<td>c. Enterprise applications</td>
<td>d. Documents</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Data exploration and profiling</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>A visual environment that enables users to interactively prepare, search, sample, profile, catalog and inventory data assets, as well as tag and annotate data for future exploration.</td>
<td>a. Searching</td>
<td>b. Sampling</td>
<td>c. Data profiling</td>
<td>d. Cataloging/inventorying data assets</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Collaboration</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>Facilitates the sharing, commenting, curation and promotion of datasets, recipes and exports.</td>
<td>a. Sharing queries</td>
<td>b. Sharing datasets</td>
<td>c. Publishing, sharing and promoting models</td>
<td>d. User rating of datasets</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Data transformation, blending and modeling</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>Supports data enrichment, data mashup and blending, data cleansing, filtering, and user-defined calculations, groups and hierarchies. This includes agile data modeling/structuring that allows users to specify data types and relationships.</td>
<td>a. Data enrichment</td>
<td>b. Data mashup/data blending</td>
<td>c. Data cleansing</td>
<td>d. Data masking</td>
</tr>
</tbody>
</table>
6. Researching and Selecting your Provider

5. **Data curation and governance**
   Supports workflow for data stewardship and capabilities for data encryption, user permissions and data lineage. This also includes security features that enable governance, such as data masking, platform authentication and security filtering at the user/group/role level, as well as through integration with corporate LDAP and/or Activity Directory systems, SSO, source system security inheritance, row/column-level security, and logging and monitoring of data usage and assets.

6. **Metadata repository and cataloging**
   Supports creating and searching metadata; cataloging of data sources, transformations, and user activity against the data source, data source attributes, and data lineage and relationships; and APIs to enable access to the metadata catalog for auditing or other uses.

7. **Use of machine learning**
   Use of machine learning and artificial intelligence (AI) to automate and improve the self-service data preparation process.

8. **How well do the candidate Data Prep tools support functional expectations?**
   For Example:
   a. **Interface Quality**
      ease of use, complexity of development tools, ramp time to productivity.
   b. **Performance**
      speed of data processing work/jobs, multi-threading, job prioritization.
   c. **Sophistication**
      ability to handle complex job streams, availability of advanced statistical functions for transformation and quality rules.
   d. **Reliability**
      error handling, restart-ability, alert/notifications, job monitoring for end users

9. **Architectural Fit:**
   How well does the Data Prep tool align with existing environment and architectural standards for, data movement, security, infrastructure, etc.?
Datawatch invented self-service data preparation. Twenty years ago our Datawatch Monarch product was introduced to solve the vexing data challenge of the time – unlocking mainframe print spool data so it could be used for other analytic tasks. And over the years, although the types of “dark data” formats have evolved, our core mission has never wavered – allow ordinary people to achieve extraordinary results with any data.

With the advent of new self-service analytics tools designed for business users, the need for a complementary self-service data preparation solution is becoming widely recognized. Overcoming organizational data challenges is critical to achieving business success:

- **80%** of an analyst’s time is spent on data preparation
- **12%** of enterprise data is used for decision making
- Data will grow **800%** over the next 5 years and **90%** will be multi-structured

Datawatch products are unique in how they empower customers to solve some of the most difficult data challenges with simple drag-and-drop, or mouse clicks. No other solution can unlock data as quickly and powerfully as Datawatch. We know that business agility is best achieved when users can interact directly with all their data. And IT becomes the hero when they choose solutions that business users love and can be deployed at any scale when required.

Over **40,000** global customers of every size, including **431** of The Fortune 500 rely on Datawatch for self-service data preparation.