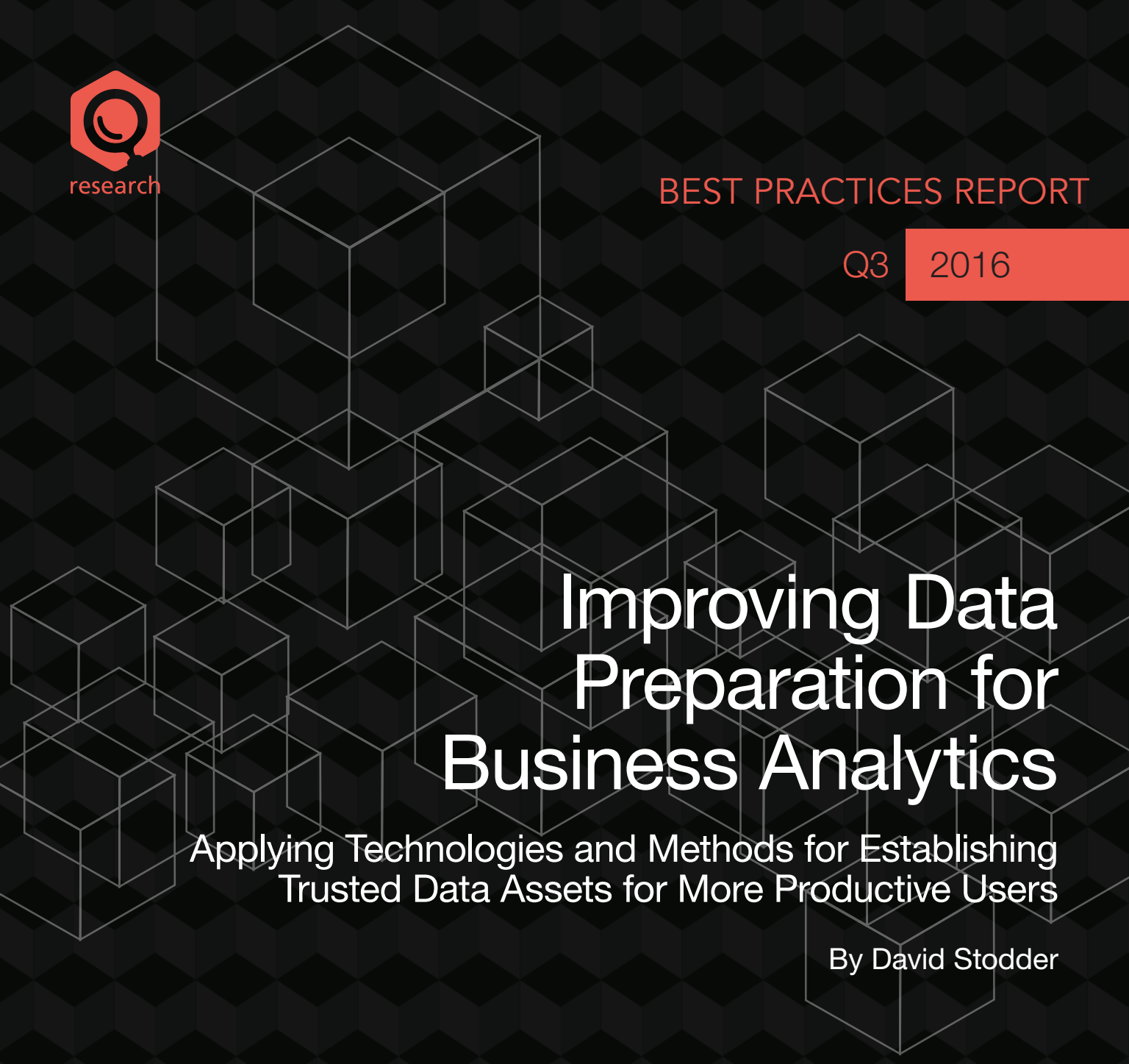


research

BEST PRACTICES REPORT

Q3 2016

Improving Data Preparation for Business Analytics

Applying Technologies and Methods for Establishing Trusted Data Assets for More Productive Users

By David Stodder

Co-sponsored by RedPoint

tdwi Advancing all things data.

Improving Data Preparation for Business Analytics

Applying Technologies and Methods for Establishing Trusted Data Assets for More Productive Users

By David Stodder

Table of Contents

Research Methodology and Demographics 3

Executive Summary 4

Why Data Preparation Matters 5

Better Integration, Business Definition Capture, Self-Service, and Data Governance 6

Sidebar – The Elements of Data Preparation: Definitions and Objectives 8

Find, Collect, and Deliver the Right Data 8

Know the Data and Build a Knowledge Base 8

Improve, Integrate, and Transform the Data 9

Share and Reuse Data and Knowledge about the Data 9

Govern and Steward the Data 9

Satisfaction with the Current State of Data Preparation 10

Spreadsheets Remain a Major Factor in Data Preparation 10

Interest in Improving Data Preparation 11

IT Responsibility, CoEs, and Interest in Self-Service 13

Attributes of Effective Data Preparation 15

Data Catalog: Shared Resource for Making Data Preparation Effective 17

Overcoming Data Preparation Challenges 19

Addressing Time Loss and Inefficiency 21

IT Responsiveness to Data Preparation Requests 22

Data Volume and Variety: Addressing Integration Challenges . . . 22

Disconnect between Data Preparation and Analytics Processes . . 24

Self-Service Data Preparation Objectives 25

Increasing Self-Service Data Preparation 27

Self-Service Data Integration and Catalog Development 28

Governance and Self-Service Data Preparation 30

Vendor Products 32

Recommendations 37

Research Co-sponsor: RedPoint Global 39

About the Author

DAVID STODDER is senior director of TDWI Research for business intelligence (BI). He focuses on providing research-based insight and best practices for organizations implementing BI, analytics, performance management, data discovery, data visualization, data preparation, and related technologies and methods. He is the author of TDWI Best Practices Reports and Checklist Reports on visual analytics, customer analytics, BI/DW agility, mobile BI, and information management. He has chaired TDWI conferences on customer analytics, big data analytics, and other topics. Stodder has provided thought leadership for over two decades. He has served as vice president and research director with Ventana Research, and he was the founding chief editor of *Intelligent Enterprise*. You can reach him by email (dstodder@tdwi.org), on Twitter (@dbstodder), and on LinkedIn.

About TDWI

TDWI, a division of 1105 Media, Inc., is the premier provider of in-depth, high-quality education and research in the business intelligence and data warehousing industry. TDWI is dedicated to educating business and information technology professionals about the best practices, strategies, techniques, and tools required to successfully design, build, maintain, and enhance business intelligence and data warehousing solutions. TDWI also fosters the advancement of business intelligence and data warehousing research and contributes to knowledge transfer and the professional development of its members. TDWI offers a worldwide membership program, five major educational conferences, topical educational seminars, role-based training, onsite courses, certification, solution provider partnerships, an awards program for best practices, live webinars, resourceful publications, an in-depth research program, and a comprehensive website: tdwi.org.

About the TDWI Best Practices Reports Series

This series is designed to educate technical and business professionals about new BI, analytics, data preparation, and data management technologies, concepts, or approaches that address a significant problem or issue. Research for the reports is conducted via interviews with industry experts and user companies and is supplemented by surveys of BI, analytics, and data management professionals.

To support the program, TDWI seeks vendors that collectively wish to evangelize a new approach to solving BI and analytics problems or an emerging technology discipline. By banding together, sponsors can validate a new market niche and educate organizations. Please contact TDWI Research vice president and senior director Fern Halper (fhalper@tdwi.org) and senior directors Philip Russom (prussom@tdwi.org) and David Stodder (dstodder@tdwi.org) to suggest a topic that meets these requirements.

Acknowledgments

TDWI would like to thank many people who contributed to this report. First, we appreciate the many users who responded to our survey, especially those who responded to our requests for phone interviews. Second, our report sponsors, who diligently reviewed outlines, survey questions, and report drafts. Finally, we would like to recognize TDWI colleagues Michael Boyda, Peter Considine, James Haley, Lindsay Stares, Diane Foultz, and Denelle Hanlon.

Sponsors

Alation, Alteryx, Attivio, Datameer, Looker, Paxata, Pentaho (a Hitachi Group Company), RedPoint Global, SAP, SAS, Talend, Trifacta, Trillium Software, and Waterline Data sponsored this report.

Research Methodology and Demographics

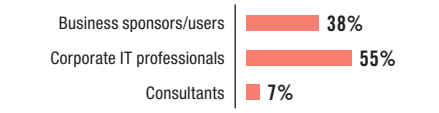
Report Scope. Organizations in pursuit of data-driven goals are seeking to extend and expand business intelligence (BI) and analytics to more users and functions. Users want to tap new data sources, including Hadoop files. However, organizations are feeling pain because as the data becomes more challenging, data preparation processes are getting longer, more complex, and more inefficient. They also demand too much IT involvement. New technology solutions and practices are providing alternatives that increase self-service data preparation, address inefficiencies, and make it easier to work with Hadoop data lakes. This report will examine organizations' challenges with data preparation and discuss technologies and best practices for making improvements.

Survey Methodology. In March and early April 2016, TDWI emailed an invitation to complete an Internet-based survey to business and IT executives; VPs and directors of BI, analytics, and data warehousing (DW); business and data analysts; line-of-business and departmental directors and managers; and other professionals. The invitation was also delivered via websites, newsletters, and publications from TDWI. The survey analysis drew from a total of 411 responses. A total of 271 completed every question. Answers from respondents who answered enough questions for their input to be valuable are included in the results. Thus some questions have different numbers of responses. Students and marketing and sales personnel from technology vendors were excluded.

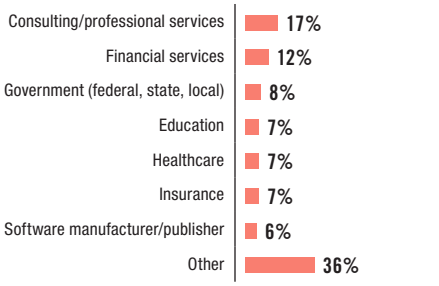
Survey Demographics. The largest percentage of survey respondents is data and IT professionals (55%); included in that group are those with CIO, CDO, CAO, or IT executive management titles (3%); VPs or directors of BI, analytics, or data warehousing (11%); BI system designers and developers (11%); and data architects (10%). The second-largest percentage is business sponsors or users (38%). Along with executive and line-of-business management, this group includes business analysts (9%) and data analysts (5%). Industry representation was varied, with consulting and professional services making up the largest segment (17%); financial services (12%) and government (federal, state, and local; 8%) were the two next highest. Most respondents reside in the U.S. (46%) or Europe (24%), but other regions account for 30%.

Other Research Methods. TDWI conducted telephone interviews with business and IT executives, VPs of BI/DW, business and data analysts, BI directors, and experts in BI and visual analytics. TDWI also received briefings from vendors that offer related products and services.

Position

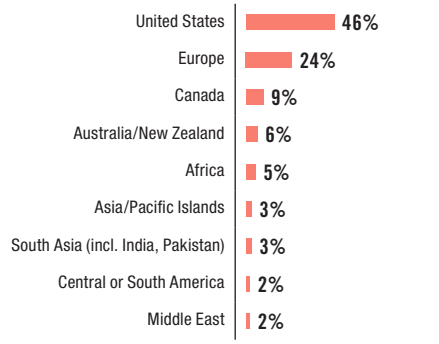


Industry

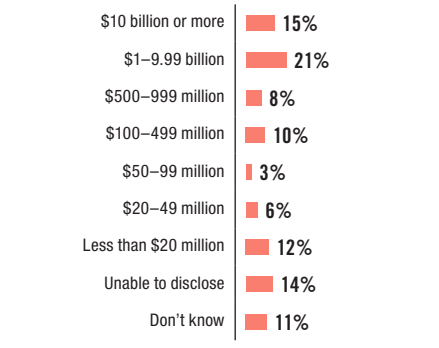


("Other" consists of multiple industries, each represented by 5% or less of respondents.)

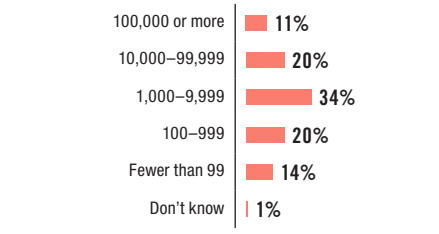
Geography



Company Size by Revenue



Number of Employees



Based on 411 survey respondents.

Executive Summary

Business users want the power of analytics—but analytics can only be as good as the data. To perform data discovery and exploration, use analytics to define desired business outcomes, and derive insights to help attain those outcomes, users need good, relevant data. Executives, managers, and other professionals are reaching for self-service technologies so they can be less reliant on IT and move into advanced analytics formerly limited to data scientists and statisticians. However, the biggest challenge nontechnical users are encountering is the same one that has been a steep challenge for data scientists: slow, difficult, and tedious data preparation.

Data preparation is the focus of innovative software technology methods aimed at accelerating, if not automating, processes necessary to support business analytics.

Data preparation is a hot topic on both business and IT sides of organizations. It is also the focus of innovative software technology and methods aimed at accelerating, if not automating, processes necessary to support business analytics. Preparing, blending, integrating, cleansing, transforming, governing, and defining the metadata of multiple sources of data—including new, raw big data in Hadoop—has been primarily an IT job; however, broadening interest in data science and analytics has drawn non-IT personnel into the execution of these tasks. Non-IT users such as business and data analysts as well as developers are looking for smarter self-service tools that reduce difficulties and make data preparation processes faster. IT, meanwhile, is interested in tools that can streamline data preparation, improve productivity, and enable IT to serve users better.

This TDWI Best Practices Report examines experiences with data preparation, discusses goals and objectives, and looks at important technology trends reshaping data preparation processes. From small organizations using spreadsheets and visual discovery tools to large enterprises trying to improve data quality and delivery for a variety of uses including business intelligence (BI) and advanced visual analytics, data preparation difficulties are a major concern. We find strong interest in improving data preparation and increasing self-service capabilities so that business users and analysts can do more on their own to prepare data without IT hand-holding.

This report examines experiences with data preparation, discusses goals and objectives, and looks at important technology trends reshaping data preparation processes.

In our report, TDWI Research advises organizations to focus improvements on reducing the time it takes to prepare data, which can help users realize insights from data faster. Users are weary of long and repetitive data preparation processes. Rapidly evolving technology is enabling organizations to automate and standardize steps as well as build knowledge about the data for better reuse and sharing, transformations, and analysis. TDWI Research additionally advocates integrating data preparation with governance programs. As organizations increase self-service data use, they need to ensure that users observe good governance and support stewardship of data assets. Clearly documenting data preparation steps can be helpful in making governance more effective and giving users the confidence that they are working with trusted data in BI and analytics projects.

Why Data Preparation Matters

Data preparation may not sound exciting, but it is one of the hottest topics in the technology industry and is important to all organizations making decisions and taking action informed by data. The topic is attracting attention because many of the problems business users and analysts in organizations confront when working with data crop up during data preparation processes. Users across the spectrum deal with data chaos every day. This includes users working with spreadsheets, BI, and visual discovery tools and more technically savvy data scientists developing and applying sophisticated analytics. The BI and data management teams of IT are also burdened by problems with poor, ill-defined data and hand coding of preparation and transformation routines. Better data preparation can remedy these problems and help both business and IT become more productive and effective.

Users deal with data chaos every day; BI and data management teams are also burdened by poor, ill-defined data.

Data preparation covers a range of processes that begin during an organization's initial ingestion of raw, structured, and unstructured data (from one or multiple sources). Data preparation processes focus on determining what the data is and improving its quality and completeness, standardizing how it is defined and structured, collecting and consolidating it, and taking transformation steps to make it useful, particularly for reporting and analysis. The selection and type of preparation processes can differ depending on users' purposes, their data expertise, how they plan to interact with the data, and what kind of questions they want to answer. This TDWI Best Practices Report examines experiences, practices, and technology trends in data preparation for BI, visual discovery, and analytics. (See sidebar, "The Elements of Data Preparation: Definitions and Objectives," on page 8 for more detailed definitions of data preparation processes.)

With improved practices and technologies for data preparation, organizations can better deal with current data troubles and prepare for future challenges arising from new data and user requirements. They can use data preparation to build the value of data assets and manage them more efficiently. Most important, executives, line-of-business managers, and the rest of an organization's personnel will be able to get the necessary data faster and use it with greater confidence for strategic, operational, and financial decisions. This can have a direct and beneficial impact on the business, improving its competitiveness, reputation, and the quality of customer and partner relationships.

Innovation in data preparation could not be coming at a better time. As more of our world becomes digitized, humans and machines are generating data at an astronomical rate—in the quintillions of bytes per day by some measures. Both the creation and consumption of data is occurring in an increasingly mobile and fast-changing environment, which means that important new attributes, such as location, must be taken into consideration for enriching the data. Organizations cannot expect heavily manual data preparation processes to scale, and lack of coordination between processes will also become a greater problem as the big data tsunami hits.

Organizations cannot expect manual data preparation processes to scale, and lack of coordination will become a problem as the big data tsunami hits.

The increasing volume, variety, and velocity of data is putting pressure on organizations to rethink traditional methods of preparing data for reporting, analysis, sharing, and use in the automated execution of business processes. This report will look at how organizations can tap innovations in technology and practices to establish smarter, more scalable, more coordinated, and better-governed data preparation processes for use with traditional platforms, such as data warehouses, as well as data lakes stored in Hadoop or other big data systems.

Better Integration, Business Definition Capture, Self-Service, and Governance

For many business users, integrating views of diverse data is an important mission of data preparation processes. Sales and marketing functions, for example, want to interact with customers and prospects across multiple channels, including social media, and need a steady stream of data flowing from activities in each of these channels. To get a complete perspective and analyze the quality of multichannel marketing campaigns and engagement, organizations need well-integrated data that is transformed appropriately so that sensible insights can be made from it. This is not an easy problem to solve, beginning with the challenge that each channel might define customers or sales differently. Organizations have been implementing traditional, non-self-service data preparation tools to address these needs for some time. Newer tools in the market are enabling organizations to capture business definitions side by side with technical metadata, which is critical to meeting self-service business and governance demands for faster development of integrated views of data.

Organizations need to record tribal wisdom about data assets, including data definitions, best practices about data usage, and the data's applicability for certain metrics and algorithms.

Organizations need data preparation processes that can help them record tribal wisdom about data assets, including data definitions, best practices about data usage, and the data's applicability for certain metrics and algorithms. Understanding the nuances of an individual column in a table can assist discovery of how data elements are actually related. This knowledge is vital when business decision makers and analysts need flexibility—for example, when requirements change due to rapid changes in the business, when users want to do more ad hoc discovery, or when different users want different views of the data and not a single, "one size fits all" report. Technology innovations are enabling smarter, faster, and automated data integration and transformation to support more varied use cases. Through analysis of our survey results, we will look at organizations' experiences with meeting demands for more agile integration and transformation and plans for addressing challenges.

Data transformation can cause major delays and inefficiencies in integrating sources. Extraction, transformation, and loading (ETL) has long been the central activity for converting data values from the original sources to the format of the target source, such as a data warehouse, data mart, or BI report. As organizations seek to extract more meaning from the original data and discover errors or other problems before they get to the target source, data transformation rules can accordingly grow more complex. In addition, data sources are growing more diverse to include unstructured and semistructured data.

Organizations need to rationalize their transformation routines, eliminating those that are no longer valuable.

Most large organizations will have hundreds if not thousands of ETL or other data transformation routines running, which can become a performance and processing burden. Routines are often written and never revised; they will continue to run even though they may not fit users' current purposes. Organizations thus need to rationalize their transformation routines, eliminating those that are no longer valuable so they can incorporate new ones that better fit current BI and analytics requirements. This report will discuss research findings about data transformation and consider best practices.

> **USER STORY BANK GAINS VISIBILITY INTO ITS DATA LAKE WITH HELP FROM A CATALOG**
>
> Data preparation plays a key role in enabling HSBC, one of the world's largest banking and financial services organizations, to gain visibility into its data. The bank is using Attivio Data Source Discovery to automatically catalog all the information stored in the firm's Hadoop data lakes. The bank uses the catalog to develop applications for reducing risks and costs.

Self-service capabilities are on the rise. The popularity of self-service visual analytics and discovery tools is revolutionizing user experiences with data by democratizing BI and data exploration. These tools require less IT attention and offer users easier means of personalizing their experiences, particularly through data visualization. Now self-service functionality is coming to data preparation and transformation. User-driven integration and preparation technologies—going by terms such as *data blending*, *wrangling*, and *munging*—are maturing, enabling nontechnical users to explore data and choose data sets that fit their BI and visual analytics processes. Many of the technologies use machine learning, natural language processing, and other advanced techniques to suggest data sets and guide users to work with the data so they can avoid coding and work at a higher level. This report will examine the self-service data preparation trend.

Governance and stewardship are vital. Finally, data preparation processes need to address data governance, especially as user self-service becomes more prevalent and risks adding data chaos. Data governance is often regarded as being primarily about protecting sensitive data and adhering to regulations; indeed, data preparation processes are vital to meeting those priorities. However, data governance is expanding to include stewardship of data quality, data models, and content such as visualizations that users create and share. Governance committees can ensure that the quality of data, models, and content meets the organization's standards. Reusability is also an objective of data governance; organizations want to improve efficiency and collaboration through sharing and reuse of transformation routines and other work. IT has an important role to play in driving strong but agile data governance. This report will discuss the emerging intersection of data preparation and governance.

Data preparation processes need to address data governance, especially as user self-service becomes more prevalent and risks adding data chaos.

For all organizations, regardless of size, the stakes for managing data and using it effectively for business analytics are getting higher. Organizations can no longer treat data as a mere by-product of business processes. Customers, patients, business partners, regulators, and others expect organizations they deal with to place a high value on data as part of the currency of their relationships. They expect frontline personnel to have quality information at their fingertips and customers to receive relevant marketing messages and recommendations. Improving data preparation steps is vital for organizations to meet these expectations and ensure that business analytics and decision making are based on the best data possible.

USER STORY GLOBALLY DISTRIBUTED FIRM INTEGRATES GOVERNED DATA FOR ANALYTICS

Global businesses have people and machine assets spread across the globe, creating challenges in integrating data from widely distributed sources and in delivering comprehensive views to equally far-flung personnel. A large industrial machinery manufacturer is using Pentaho Data Integration to blend sensor data from a variety of maritime vessels with geospatial data and customer information, applying advanced algorithms to predict equipment failure. The company has also used Pentaho's platform to build interactive dashboards and extend them to customers as a service in order to help them optimize maintenance cycles and fuel consumption.

The Elements of Data Preparation: Definitions and Objectives

Data preparation is not easy to define because it involves multiple steps. The process generally begins with data sourcing and ingestion, moves through making data suitable for use through transformation and enrichment, and then integrates with governance and stewardship for monitoring and improving how data is used for BI and analytics. The steps are interdependent and overlap, and they don't always run sequentially. Some data preparation tools provide workflow (sometimes called a pipeline) to guide processes and steps so they can be easily repeated. Workflow is also important for governance and for documenting the data lineage behind the analytics—that is, how analytical conclusions were reached based on particular data.

As a series of ongoing processes, data preparation is not a onetime event. Business and IT should collaborate and iterate on data set preparation. The purpose of the data matters with data preparation—and purposes often change.

In larger organizations, data preparation may be part of a broader enterprise information management strategy, or the context could be more humble. For example, it could be a business department trying to improve data for spreadsheet and visual discovery tool use. Finally, data preparation intersects with (and should support) data governance.

There are five major objectives common to most data preparation processes. Listed under each objective below are goals, processes, and steps that many organizations apply to reach each objective. Again, different organizations will mix the items differently depending on their requirements.

Find, Collect, and Deliver the Right Data

- Find, discover, explore, and search for data; iterate to settle on the right data
- Source, access, collect, and extract data across multiple platforms
- Determine fitness for purpose (e.g., discovery, reporting, monitoring); filter data to select a subject that meets rules and conditions
- Know where the data is so you can source it again
- Deliver frequently used data sets through automated processes

Know the Data and Build a Knowledge Base

- Understand the data; learn data and file formats
- Use visualization capabilities to examine the current state of the data and spot anomalies and inconsistencies
- Profile and prepare the data; use profiling to generate data quality metrics and statistical analysis of the data
- Sample the data; run profiles; learn primary and foreign key relationships; apply knowledge to transformations
- Discover and learn data relationships within and across sources; find out how the data fits together; use analytics to discover patterns
- Define the data; collaborate with business users to define shared rules, business policies, and ownership
- Build knowledge with a catalog, glossary, or metadata repository
- Gain high-level insights; get the big picture of the data and its context

Improve, Integrate, and Transform the Data

- Assess data quality and accuracy; consider how it fits with governance rules and policies
- Determine the right level of quality for the purpose of the data
- Cleanse and deduplicate the data; note missing data; perform identity resolution
- Refine and merge-purge the data
- Validate the data; validate new sources
- Normalize the data; divide data into groups as necessary; structure the data
- Determine the need for data staging location
- Map data elements from sources to destination; document transformation rules
- Apply transformation functions to structured and unstructured data
- Generate code (potentially automatically) to transform data and attributes
- Add or change attributes (e.g., account names to a customer table)
- Enrich the data; complete the data
- Integrate and blend data with data from other sources; perform joins and merges
- Determine levels of aggregation needed to answer user questions; use data connectors
- Restructure the data according to needed format for BI, integration and blending, and analysis; transpose the data
- Use filter, sort, and join functionality (potentially self-service) to tailor data for reports or analysis
- Incorporate formulas for manipulation requirements
- Model the data for BI and analytics
- Apply statistical and advanced models to data

Share and Reuse Data and Knowledge about the Data

- Operationalize data preparations so they can be rerun on a schedule or on demand
- Automate data preparation steps; automate data flows
- Share and reuse data sets
- Share and reuse data definitions, metadata, and master data
- Deliver prepared data to analytics tools and business processes
- Share and reuse analysis and calculations
- Share knowledge of the data's usefulness for its purpose (e.g., BI, analytics)
- Promote individual, "tribal" data knowledge into best practices that can be shared

Govern and Steward the Data

- Apply policies and rules to protect sensitive data
- Implement data security rules and policies; implement regulatory policies
- Monitor and track data usage
- Track lineage of where the data came from and how it was prepared
- Manage access to enterprise data sets, semantics, and shared catalogs
- Promote data sets that are sanctioned by governance for reuse
- Empower authorized users to be data curators
- Create business metadata and document definitions
- Increase confidence and trust in the data
- Analyze the ongoing health of the data

Satisfaction with the Current State of Data Preparation

Users bring different perspectives to their data preparation journeys based not only on roles and requirements but also on levels of experience.

Users bring different perspectives to their data preparation journeys based not only on roles and requirements but also on levels of experience with particular tools and platforms. To begin our survey, we sought to learn about our 411 research participants' technical proficiencies. Offering a list of tools and platforms, we asked respondents to select those they are proficient with. Most are proficient in using a spreadsheet application such as Microsoft Excel (81%). Nearly as many said they are proficient in SQL coding and/or reporting-tool scripting language programming (79%). Most participants said they are also capable of using data visualization and presentation technologies (75%) and specific BI or visual analytics software (72%). Over two-thirds (68%) are proficient in using relational database management and data extraction software, with 63% able to use extract, transform, and load (ETL); extract, load, and transform (ELT); data virtualization; and data integration technologies. The percentages are a bit lower when we filter results to see just the business sponsors and users segment.

Research participants utilize a range of tools and platforms for their BI and analytics projects. To gain a better sense of the technologies used, we asked participants about their reliance on various types of tools and platforms for data access, querying, reporting, analysis, presentation, and sharing. Responses proved to be roughly the same regardless of organization size. After spreadsheets, which are the most commonly used tools, participants are reliant on ETL, ELT, or data integration tools (36% extremely reliant, 30% reliant, and 18% moderately reliant). Next highest was BI/OLAP systems managed by central IT (31% extremely reliant, 28% reliant, and 15% moderately reliant). These results indicate a fair but not universal degree of prevalence in the use of traditional BI and data warehousing architectures.

Research participants' organizations are least reliant on desktop databases and mobile BI/analytics applications (both 33% not very reliant). Only 17% are reliant or extremely reliant on data catalogs, which are important for consolidating and coordinating data definitions and knowledge about the data, and 18% are moderately reliant on catalogs.

Spreadsheets Remain a Major Factor in Data Preparation

Spreadsheets continue to be ubiquitous as an affordable tool for viewing data, doing calculations, creating graphs, and performing other types of data analysis. Small and midsize firms that have little or no IT function and lack enterprise BI and data warehousing are particularly reliant on spreadsheet applications. Thus spreadsheets come up often as a commonly used tool for preparing data and as a source of data that must be prepared for integrated views from within other applications.

Data preparation either for spreadsheet use or to collect data from spreadsheets goes by many names, but most are not pretty.

Data preparation either for spreadsheet use or to collect data from spreadsheets goes by many names, but for most users the names are not pretty. Cleansing and preparing data manually to remove duplicate records, fix errors, and investigate out-of-range values tends to be boring and time-consuming for most business users. Many will clean and prepare data for use in their personal spreadsheets in a haphazard fashion while they are doing other activities, rather than follow clear and repeatable processes. Spreadsheets are also common for performing data transformation, but too often conversions are done to fit a specific need and no set rules are followed. Users can get distracted during preparation steps, lose their place, miss data, and make mistakes. They then have to start over, and in some cases the errors are missed and become internalized in downstream analysis.

Yet spreadsheets are often where users do their data preparation, even in organizations where there is a data warehouse. In interview research, we hear that users tire of waiting for new data to become available in the warehouse and do not want to wait out long IT processes. Instead they will copy and paste data on their own into spreadsheets and try to cleanse and prepare it there for personal or departmental use. Although this choice may save them time initially, it is likely to increase the level of data chaos in the organization and add to the number of "spreadmarts" or disparate data silos. Documentation of data definitions and transformations can be spotty and conflicting. Users find that they have trouble lining up records and their joins no longer work properly. They stop trusting the data in one spreadsheet and create new ones, exacerbating the problem further. Governance concerns also arise because users might share poor or sensitive data. At this point, either internal IT or outside consultants are often called in to rescue users.

Spreadsheet data chaos is a pain point that many data preparation tools try to address. Certain tools can automate steps to replace manual effort and improve standardization. Others will extract data from spreadsheets—numbering in the hundreds or thousands in some organizations—into more centralized databases where the tools can launch programs to profile the data, improve its quality, and perform other tasks as necessary. Some tools are able to return the improved data to spreadsheets so that users can continue to work with their familiar tools but with better data. There are also tools that can build documentation through workflows about the data and analytics calculations and track data lineage, which is helpful for governance. As a best practice, organizations should not ignore spreadsheet data use and preparation as they develop an overall strategy. They should evaluate technologies that could either reduce the need to rely on spreadsheets for data preparation or improve data preparation involving spreadsheets.

As a best practice, organizations should not ignore spreadsheet data use and preparation as they develop an overall strategy.

Interest in Improving Data Preparation

Before delving into data preparation processes specifically, we wanted to get an overall sense of the satisfaction of users in research participants' organizations with how easily they can find relevant data and understand how to use it appropriately for BI and analytics. These two related objectives are core to the mission of data preparation. The biggest percentage said users in their organizations are somewhat satisfied (36%), with 7% very satisfied (see Figure 1). More than a third (37%) indicated dissatisfaction. The results suggest significant room for improvement.

Organizations confront barriers when trying to make upgrades to data preparation, many of which are not about technology implementation. Our study finds that among research participants, an insufficient budget is the most common barrier to improving how data is prepared for users' BI and analytics projects. The second most common barrier is not having a strong enough business case. The results highlight the difficulty many data professionals have in convincing executive management to invest in improving the quality of shared data assets (quality being a cornerstone of well-prepared data).

Organizations confront barriers when trying to make upgrades to data preparation, many of which are not about technology implementation.

Overall, how satisfied are users in your organization with how easily they can find relevant data and understand how to use it appropriately for BI and analytics?

Figure 1. *Based on answers from 411 respondents.*

In conducting interviews for this report, we found that organizations tend to act only when a harmful incident has occurred, such as a new regulatory policy that must be observed or public embarrassment from customer complaints about bad data. Some will budget for solutions if there is a specific project, such as a migration or consolidation of systems after a business merger or acquisition. It can be difficult for data stewards to demonstrate and quantify the relationship between poor data quality and preparation and lost revenue, missed business opportunities, and costly process inefficiencies.

Complicating efforts to make sustained improvement to data preparation across the enterprise is the third most common barrier cited by research participants: lack of skilled personnel or training. Organizations frequently do not have enough internal expertise to work effectively and efficiently with data—or their expertise is spread unevenly and is dependent on few "power users" rather than built up as part of a broader user training program. In many cases, most of the expertise resides in BI teams that are part of the IT function. However, BI teams themselves may lack skills in data-preparation analytics projects if they are trained for more traditional BI reporting requirements.

USER STORY DONORSCHOOSE.ORG AVOIDS DATA COMPLEXITY TO GET FUNDING TO SCHOOL PROJECTS

DonorsChoose.org, an online nonprofit organization founded in 2000, uses the collaborative crowdfunding potential of the Internet to enable people to fund school classroom projects. It has been a huge boon to strapped teachers across America, channeling more than $430 million to 730,000 classroom projects. The other contribution DonorsChoose.org makes is in providing essentially public access to the voluminous data it collects along the way about what school projects need. The data is now an important resource for media, educators, and policy makers. DonorsChoose.org also performs a wide range of analytics across its own sources and third-party data to understand donation patterns, improve marketing, and monitor user satisfaction.

DonorsChoose.org uses a cloud-based system as the platform for its data warehouse, which it operates using Amazon Redshift on Amazon Web Services. "We do zero ETL ourselves," said Vladimir Dubovskiy, lead data scientist. Rather than devote precious engineering resources to data preparation and writing ETL, the organization instead uses Fivetran to stream data from Salesforce, Google Analytics, Zendesk, Postgres data, and CSV sources into Redshift. DonorsChoose.org then implements Looker as the single access point to the data. Looker also handles ETL and preparation for analytics with features such as Persistent Derived Tables.

Users develop dashboards, set up metrics, explore data, and create analytics. The dashboards have been popular with teams not just inside the organization but with vendors such as Amazon. "We have rock stars using Looker in advanced ways, but folks really thrive on the dashboards, including vendors who supply materials to schools," Dubovskiy said. "They use our portal access to see the heartbeat of their inventory and apply the data to their forecasting."

IT Responsibility, CoEs, and Interest in Self-Service

In most of our research participants' organizations, IT is the function primarily responsible for making it easier for users to find relevant data and understand how to use it properly for BI and analytics. TDWI asked participants to indicate which function is responsible and 70% said IT, not surprising given the dominance of IT (participants could select more than one; in many organizations more than one function is responsible). IT has traditionally had responsibility for managing the collection, preparation, and delivery of enterprise data to users, most often through a BI and data warehousing system. However, not all users are satisfied with IT as the gatekeeper of the data or, for various reasons, the data they receive from the BI and data warehousing system, which is a factor driving the self-service trend.

IT is the function primarily responsible for making it easier for users to find relevant data and understand how to use it properly.

Interestingly, 20% of participants said that a center of excellence (CoE) committee had some responsibility. Usually set up as a joint business-IT leadership team, a CoE can provide leadership for BI, analytics, data warehousing, and data lake projects and ensure that they are aligned with the organization's overall strategy. A CoE is helpful in resolving competition for resources between projects, managing costs, developing road maps, and encouraging open discussion about where projects are succeeding or failing. TDWI Research recommends forming a CoE as a best practice to improve leadership of BI and analytics projects, which includes the data preparation processes that serve those projects. Communication through a CoE can help business and IT stakeholders spot weaknesses and gaps in data preparation and identify where new training or technology investment is needed. The CoE can also contribute to data governance, including the development of rules, policies, and standards for data use, along with transformation, data lineage, and the user development of content such as dashboards.

Communication through a CoE can help business and IT spot weaknesses in data preparation and identify where training or technology investment is needed.

TDWI Research drilled down further into participants' view of the extent of IT involvement in data preparation. Figure 2 shows that although the survey results are divided, the largest percentage (28%) said IT is responsible for all of their data preparation processes (note that respondents could choose only one answer that best fit their organization). The second-highest percentage said that once users receive data from IT, they are responsible for further preparation (20%). In this scenario, IT typically delivers a report, snapshot, data extract, data cube, or other sampling to users, who then employ spreadsheets, enterprise BI tools, self-service visual analytics and discovery tools, or other applications. The data may be delivered into a shared resource such as a departmental data mart or data lake.

To what degree is IT primarily responsible for data preparation processes for BI and analytics projects compared to business analysts and users? (Select the answer that is most accurate for your organization.)

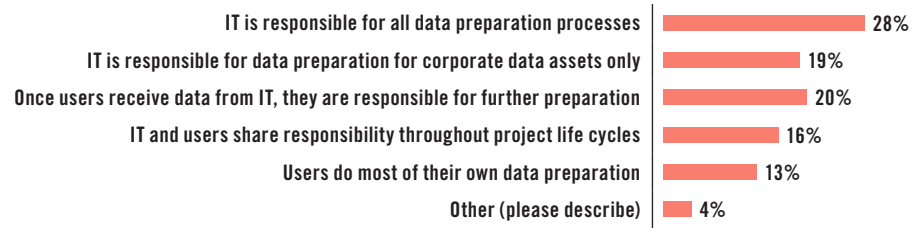


Figure 2. *Based on answers from 369 respondents.*

Nearly the same percentage (19%) said IT is responsible for data preparation for corporate data assets; in other words, users work additionally with other data sources for which they are responsible. Only 13% said users do most of their own data preparation. Some participants wrote in their own answers, which shed light on different ways in which IT or other functions are responsible. For example, one said that their organization's BI team is "responsible for data preparation for corporate data assets and external information and is not part of IT. IT is responsible for the staging previous to the data warehouse but not the analytics." Endorsing the value of a CoE style of committee, another said that an "independent BI competency center (BICC) is responsible for all data preparation processes."

USER STORY DATA CATALOG HELPS INSURER AVOID POTENTIAL DOWNSIDES OF DATA LAKES

"Understanding where your data came from and what it means in context is vital to making a data lake initiative successful and not just another data quagmire—the catalog plays a critical component in this." – Global head of data governance, risk, and standards at an international multiline insurer.

Users are seeking technologies that support ad hoc data integration, transformation, and quality improvement.

Most organizations see self-service data preparation as an important goal. As self-service analytics grows, organizations need to support a wider spectrum of user requirements for accessing and analyzing different types of data. Users are seeking technologies that support ad hoc data integration, transformation, and quality improvement. In addition, organizations that have deployed Hadoop systems need tools that enable users to access data that is frequently not structured and modeled as that in a data warehouse. It is proving difficult for many IT functions to fulfill the number and variety of new user requests. Self-service data preparation is an important trend, giving users more capabilities for working with data on their own, with less IT hand-holding, to suit specific business needs.

Research participants overwhelmingly regard increasing the ability of users to perform self-service data preparation for BI and analytics projects as important.

Research participants overwhelmingly regard increasing the ability of users to perform self-service data preparation for BI and analytics projects as important; 44% said it was very important and 33% said it was somewhat important. When filtered for just the survey segment of business sponsors and users, the "very important" percentage rises to 49%. Technologies are evolving to meet self-service demand by providing easier-to-use graphical interfaces and guidance to help users find the right data and prepare it to fit their requirements. More difficult for many organizations is adjusting their data infrastructure to accommodate self-service data preparation; in most cases, the infrastructure was not built to handle numerous nontechnical users seeking to do their own data preparation. In addition, BI teams will need to shift from owning projects to providing expert help to users with the different phases of data preparation so that they can accomplish doable goals without getting frustrated—and leave more complex projects to IT and BI teams.

USER STORY TRAVEL COMPANY EMPLOYS SELF-SERVICE DATA PREP TO SUPPORT WEBSITE INNOVATION

Few things are more critical to business success than giving customers a great website experience. Being able to innovate in this area depends on analytics that can access large and varied customer-generated data. An online and mobile travel company uses Datameer self-service data integration, preparation, analytics, and visualization tools to provide the firm's business analysts with direct access to big data sources.

Attributes of Effective Data Preparation

The value of data preparation processes can be measured by how well the resulting data meets users' BI and analytics requirements for clean, relevant, and trustworthy data. TDWI Research asked participants how satisfied their organization is with some key attributes of the data being prepared for BI and analytics activities (see Figure 3; attributes are ranked most to least by the combination of "very satisfied" and "somewhat satisfied" percentages). The responses offer an opportunity to discuss these attributes and participants' assessment of how well they are being achieved by their data preparation processes in more detail.

Accuracy, quality, and validity. This group of attributes is at the heart of data preparation and transformation processes and becomes particularly critical as multiple sources are integrated and blended for BI and analytics. To be accurate, data values must be correct and in the right form; otherwise data could be wrong or invalid. Data validation constraints and processes help ensure that clean and correct data is added to data sets. Some data preparation tools are able to move data through a pipeline, or integrated workflow, that improves quality in a series of steps. While in the pipeline, the tools enable organizations to spot problems in the data and address why they are occurring. Some tools let analysts examine source data as it is coming through the pipeline, validate it, and determine how they want to transform it given their requirements. Research participants are the most satisfied with their data's accuracy, quality, and validity compared to the other attributes; 14% are very satisfied and 46% are somewhat satisfied.

Some tools let analysts examine source data as it is coming through the pipeline, validate it, and determine how they want to transform it.

In general, how satisfied is your organization with the following attributes of the data being prepared for BI and analytics activities?

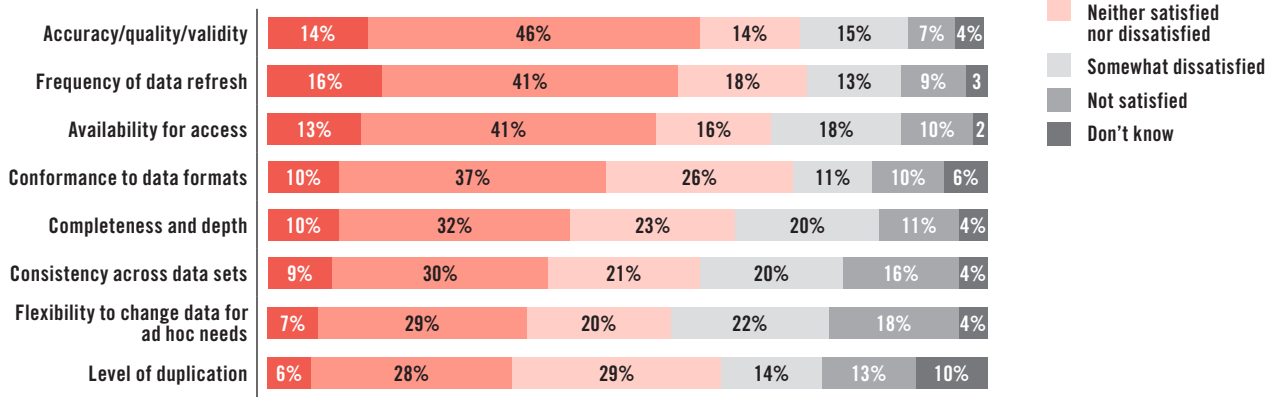


Figure 3. *Based on answers from 384 respondents. Ordered by highest total of "very satisfied" and "somewhat satisfied" combined.*

Frequency of data refresh. Satisfaction with this attribute is similar to that in the previous category; 16% are very satisfied and 41% are somewhat satisfied. With requirements increasing for fresher data to satisfy operational BI and near-real-time analytics, users and applications need more current data refreshes or continuous incremental refreshes that only update changed data or add new data. Sometimes, however, the fresher the data the lower the quality; there may not have been time to run data quality processes. Users may have to work with raw data that has not been fully transformed if they need access to it and do not have time for long transformation processes. Users need to be informed of the data's state, including how recently it was refreshed, its quality, and its cleanliness. Some of the latest tools enable users to learn the structure, quality, and completeness of data files

as they are being ingested or extracted so that users can make decisions about how or whether to transform it. Such tools offer metrics and visualization capabilities for easier evaluation and faster insight into the data.

Hadoop is playing an important role in improving availability, even if consistency and quality in data lakes are uncertain.

Availability for access. Operational BI and analytics are driving the need for higher levels of data availability primarily for access but also for scoring analytics models in operational analytic workflows. Hadoop is playing an important role in availability; a key reason why organizations are creating data lakes is simply to make more data available for analytics, even if consistency and quality are uncertain. Data preparation tools will then work on top of Hadoop data lakes to help users and programs find the right data, improve and transform it, and blend it for specific user requirements. Rather than wait until all the data is loaded, some tools let users examine the data as it is flowing into the data lake; the tools give users a view of the data using a spreadsheet interface or other visualization and provide tools for choosing how to transform it. Only 13% are very satisfied with availability; 41% are somewhat satisfied.

Conformance across data formats. Data's lack of conformity with data models, domain models, or database schema can be difficult and time-consuming to fix, especially manually. Poor conformance creates uncertainty about the data, which slows down BI and analytics and can be a major source of user frustration. Data preparation tools can be helpful in automating checks for conformance to specified formats. Research participants are not highly satisfied with their current preparation processes for achieving this goal; just 10% are very satisfied and 37% are somewhat satisfied.

Completeness and depth. Completeness is important on many levels. Data sets that are missing expected values can hang up ETL and analysis processes and produce erroneous results. Data quality and preparation tools can be helpful in spotting unexpected gaps. Completeness and depth are also major goals of processes for data integration and improving understanding of how data is related within and across sources. For example, organizations can use preparation tools to help build comprehensive views of all data about customers. Research participants clearly see room for improvement; only 10% are very satisfied and 32% are somewhat satisfied with data preparation processes for completeness and depth.

Research participants report that their organizations are less than satisfied with consistency across data sets; only 9% are very satisfied.

Consistency across data sets. This attribute is critical to building trust in the data. Profiling and transformation tools are often aimed at spotting inconsistencies in the data such as spelling variations in names and addresses or outliers that could skew analysis. Transformation steps are critical to taking source data values and converting them so that values are consistent in the destination data systems. For many types of BI reporting, data must be provisioned consistently so that trends and comparisons drawn are valid. For discovery and other types of analytics, perfect consistency may not be critical. Some tools use natural language processing and identity resolution techniques such as fuzzy matching to spot inconsistencies so they can be remedied. Research participants report that their organizations are less than satisfied with consistency across data sets; 9% are very satisfied, 30% are somewhat satisfied, and 36% are either somewhat dissatisfied or not satisfied.

Research participants indicate low satisfaction with their data preparation processes for achieving flexibility to change data for ad hoc needs.

Flexibility to change data for ad hoc needs. Flexibility and agility are critical to supporting self-service BI and visual analytics users, but in many organizations the BI, ETL, and data warehousing infrastructure was built for standardized querying and reporting. More frequently than ever, users will want new data, changes to existing data, and other updates. ETL routines are often difficult to change—if not technically, then difficult to get IT personnel's time to make changes. This is a reason why self-service data preparation and transformation may be helpful in taking some of the load off IT. Research participants here indicate low satisfaction with their data preparation processes for

achieving flexibility to change data for ad hoc needs; 7% are very satisfied and 29% are somewhat satisfied; 40% are either somewhat dissatisfied or not satisfied.

Level of duplication. Data quality, merge-purge, and cleansing processes often aim at reducing or eliminating duplicate data. Automated tools are critical as the number of input files increases. As with improving consistency, fuzzy matching techniques are important to identifying and reducing duplication as well. Some tools are able to take the corrected data and move it back to the sources. The results of our research indicate a need for improvement in processes for addressing data duplication. Just 6% of research participants are very satisfied and 28% are somewhat satisfied.

USER STORY PEPSICO REDUCES TIME AND COMPLEXITY OF DATA PREP FOR SUPPLY PLANNING DASHBOARDS

The data is there for retail companies seeking to improve supply chain performance. The question is whether analysts can prepare and provision the data fast enough for a diverse community of business-to-business (B2B) decision makers to respond optimally to events and trends in the supply chain. Mike Riegling, data analyst on the customer supply planning team at PepsiCo, must bring together data from a variety of sources for about 20 different dashboards so that internal managers and external retail B2B customers can keep an eye on the complete supply chain for the three brands he focuses on: Gatorade, Quaker Oats, and Tropicana.

The data supports two distinct PepsiCo programs with different analytics requirements: collaborative planning, forecasting, and replenishment (CPFR), which enables multiple retail partners to share information with PepsiCo as their buying teams place orders for inventory from PepsiCo; and vendor-managed inventory (VMI), "where we are purchasing our own inventory on behalf of our customers to support their warehouses, and they distribute the products to their stores and get them on the shelves," Riegling said. The goal in both cases is to avoid out-of-stocks but also to ensure that a surfeit of inventory isn't sitting in warehouses where it could spoil or ultimately be returned.

Riegling's team stores data in Hadoop files and uses Trifacta to prepare data for the dashboards. Previously, the team had been using Microsoft Excel and Access, which were "slow, clunky, hard to troubleshoot to fix bad queries or data inaccuracies, limited in capacity for data, and hard to adapt to a changing business environment," said Riegling. "Now, we have no data limit, which is important to our more complex retailers." The team has been able to lower the amount of time analysts need to spend on data preparation and reduce query complexity.

Data Catalog: Shared Resource for Making Data Preparation Effective

An important benefit of more formalized data preparation processes is that individuals or groups in the enterprise can trust the data and analytics shared by other groups and be confident in reusing it for their BI and analysis. Most research participants (46%) are somewhat satisfied with their data preparation for sharing; 16% are very satisfied and 18% are either somewhat dissatisfied or not satisfied.

Most research participants (46%) are somewhat satisfied with their data preparation for sharing; only 16% are very satisfied.

The satisfaction level is higher when we select only research participants who said they were "reliant" on a data catalog. A data catalog is a central repository that typically contains metadata—that is, descriptive information about the data sets, how they are defined, and where to find them. Some data catalogs (or metadata repositories) include information about who produced the data set, its quality, and other important characteristics. Shared resources such as data catalogs, glossaries, and metadata repositories can help users find quality sources and gain better knowledge about how data in multiple sources may be related. However, building these resources manually can be an arduous, never-ending task. Many organizations use spreadsheets or hand-coded SQL programs. In addition, it can be difficult to get different parts of organizations to agree on metadata elements such as higher-

level definitions of customers, sales, and products and definitions of related data sets. Higher-level concepts often get out of sync with the data, rendering many catalogs and repositories somewhat useless for finding data sets unless the organization makes it a priority to keep definitions in sync and up to date.

Tools can automate steps in building catalogs and glossaries and in keeping them up to date.

Fortunately, tools exist that can automate steps in building catalogs and glossaries and in keeping them up to date. The tools can discover metadata from existing data sets to learn details about data, tag data according to higher-level business definitions and rules, and locate and use existing documentation. Users, including analysts and developers, can employ tools to examine data lineage to learn how data has been consumed and transformed by others. Shared resources such as catalogs and glossaries can be essential for governance because they make it easier to locate data and track lineage to oversee adherence to governance rules and policies. Data stewards can use these shared resources to identify where data quality is poor and target where to apply remedies.

USER STORY ANALYSTS EMPLOY DATA CATALOGING TO BROADEN ACCESS AND ENABLE REUSE

Analytics projects are driving business users and analysts to extend their reach to different types of data, including to sources that exist beyond the data warehouse. However, the attempt to access and interact with multiple data sources is often an exercise in frustration because each data source lives in its own silo. Making things even more complicated, BI or data discovery tools typically have their own metadata repositories (or lack thereof) and have their own ways of accessing data sources.

"There is nothing new about this problem," said a data science leader of a 400-person-strong analytics group at a leading pharmaceutical firm. "We have different databases all over the globe. You name it—we have it." His group has analysts with skills and interests ranging from primary market research to statistics and data science. "It's confusing to have to tell folks that 'hey, you're only able to look at data in X system.' What if there's no report—or they can't find the report—that's been generated for the data they need? What if they really just need the raw data or want to do analysis that includes data from a variety of systems, not just one? It would be better to have everyone operating in one environment together."

The data science group leader, who asked not to be named, explained that his group is using Alation Data Catalog to establish such an environment. It provides a form of knowledge management that draws from multiple sources but is tied directly to data elements and queries so that people can reference and learn from the catalog without having to leave their preferred tool for querying the data. The group is building knowledge about the data and analytics to support reproducible analytics. "Let's say there's a great piece of analysis someone has done and wants to build on it," the group leader said. "Usually you have to go find that person, who has probably lost the code or doesn't remember what they did, and you're back to square one. With this system, I am able to trace it all the way back, see the queries, and see that the data sets are real."

Overcoming Data Preparation Challenges

This section will focus on challenges users and organizations face with data preparation. As we discuss the challenges, we will look at how technologies and practices can be advanced to address them. To provide context, let's first consider the broader data-related factors research participants said are their most significant barriers to improving how data is prepared for user BI and analytics projects in their organizations. We asked participants to rank barriers; in Figure 4 we can see that a number of barriers are highly significant. Ranked using a weighted average, we can see that the top barrier is difficulty accessing and integrating data across system or application data silos. The fourth-highest one—existing ETL and data integration lack speed and agility—additionally spotlights data integration woes.

Research finds that difficulty accessing and integrating data across system or application silos is the most significant barrier to improving data preparation.

For years, many organizations have sought to consolidate data from multiple silos into an enterprise data warehouse using ETL routines. Today, some are creating Hadoop data lakes as an operational data store or staging area to collect as much data as possible and load it without having to create or impose a schema as the data comes in. Some are using cloud-based storage and data management for their data lakes rather than expend resources constructing an on-premises system. ETL routines or other analytics programs can then run on data in the lake.

Data preparation tools of the type mentioned earlier can apply machine learning to automatically join disparate data sets and create a metadata catalog or glossary. These capabilities become particularly important as more business users seek to drive data transformation and integration themselves through self-service BI and visual analytics tools. Users want to access a wider variety of data, including that in data lakes, but not have to wait for IT-driven ETL development before they can interact with the data. Smarter and easier-to-use data transformation and integration are critical to overcoming the delays and inflexibility that frustrate users who are dependent on traditional ETL.

Which of the following data-related factors present the most significant barriers to improving how data is prepared for users' BI and analytics projects? (Please rank from most to least significant.)

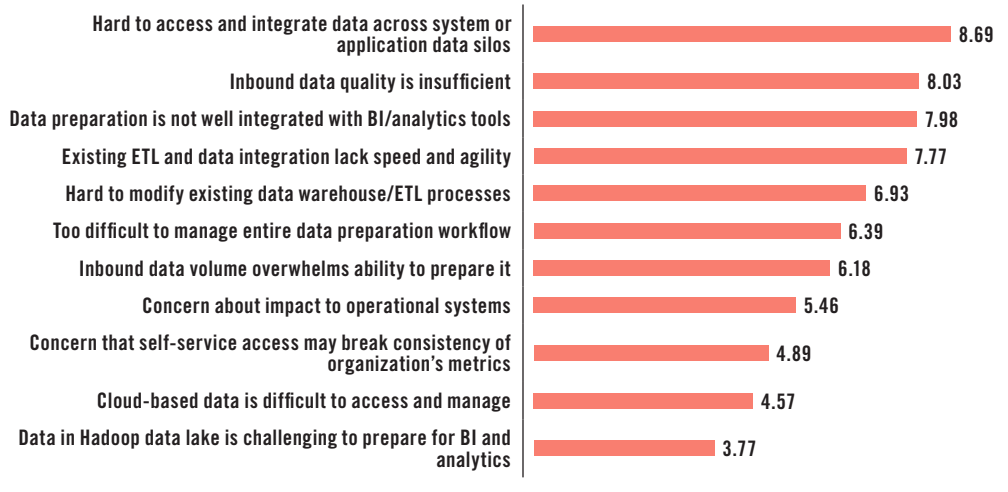


Figure 4. *Based on answers from 311 respondents. Ordered by highest weighted average.*

An alternative to traditional data warehousing for integrating and transforming data is software that can bridge silos by using data federation or virtualization technologies that leave the data in place but are able to send queries and consolidate results. This alternative is becoming useful for organizations that want to provide users with integrated views of data, but some of it is located externally such as on public cloud-based servers or on systems run by third parties such as business partners. No one solution fits all use cases; organizations should determine their BI and analytics requirements carefully before choosing a technology solution.

The second most significant barrier is insufficient inbound data quality. Data preparation tools that have strong data quality analysis and management capabilities can be helpful in enabling organizations to profile, validate, and cleanse data as it arrives or at least before it is loaded into data stores set up for BI and analytics. The tools can help organizations spot recurring problems and improve forms or other data entry interfaces to increase inbound data quality and reduce the cost of fixing problems downstream.

When data preparation is not well integrated with BI and analytics, users are often frustrated as they try to deepen data interaction.

Organizations have difficulty integrating data preparation with BI and analytics tools. The third most common barrier, noted in Figure 4, is that data preparation is not well integrated with BI and analytics tools. Users of self-service BI and visual analytics tools are often frustrated as they try to move beyond exploratory projects and deepen their interaction with more data. Performance will slow down, sometimes because the activity goes beyond the tool's technical limits for manipulating, filtering, blending, and enriching the data. Just as often, however, problems could be due to the organization's data preparation processes, which may not be set up to handle numerous ad hoc questions, particularly from a large number of users. As users grow their implementations of self-service BI and visual analytics tools, organizations should evaluate whether their data preparation processes and technologies are ready for new and different workloads. They should examine technologies that offer improved integration between visual analytics and data preparation, including into a single platform in some cases.

USER STORY VERADATA TACKLES DATA PREP SCALABILITY AND REPEATABILITY TO IMPROVE CLIENT SERVICE

"We only get paid if you improve" is the bold guarantee from VeraData, a marketing services provider based in Fort Myers, Florida. The company develops BI and analytics to optimize client marketing campaigns and data acquisition. Many of VeraData's clients are charities and nonprofits—organizations that are careful with marketing spending given their limited finances. "We go to a client and we say, 'We're going to guarantee you better response rates, acquisition costs, or revenue'—three things we target," said Andris Ezerins, VeraData's executive vice president of operations. "If we don't achieve those through our predictive and prescriptive models, we don't get paid."

VeraData's success is therefore directly related to how efficiently and effectively it can bring data in and prepare it for business analytics. VeraData is using Alteryx for data preparation and blending. Ezerins said scalability is critical because with more accounts come more users and more data. "Some of our biggest clients have data sets of over 100 million records. You can't open 100 million records in a spreadsheet." Part of scalability is therefore making it easier to train new personnel to use the tools. "Our account managers need to be able to answer their own questions about the data and use filtering, sorting, and other capabilities to scroll through large data sets."

Repeatability is also critical to scaling VeraData's data-driven business. "Years ago, when a new client would come in the door, we'd have to rebuild the analytic process from scratch," said Ezerins. "Then, of course, the client would call us back a month later and need five records from the third file, and we'd have to redo the entire

process to find those records." VeraData is now able to build workflows with documentation developed and embedded by the software, which helps the company respond to clients with flexibility.

Addressing Time Loss and Inefficiency

A common result of data preparation problems is lost time and inefficiency. Spreadsheet-based data preparation in particular is often mired in manual effort to combine, cleanse, and transform different types of data using processes that are hard to repeat or share. Users of BI and analytics tools also spend a great deal of time looking for the right data, only to find that the data is flawed or not prepared for their requirements. Users and analysts often cannot share and collaborate effectively; they might, for example, waste a great deal of time developing questions and queries that others have already developed and that could have been used instead.

We asked research participants what percentage of the total time spent on their recent BI and analytics projects was devoted to preparing the data compared to the time spent performing analysis and data interaction (see Figure 5). The largest percentage (37%) said that 61%–80% of the time was spent on data preparation; in total, almost three-quarters (73%) of respondents said that more than 41% of the time was spent on data preparation. We also asked what percentage of the personnel working on projects was involved in data preparation processes. Similarly, the largest percentage of research participants said 61%–80% of their personnel were involved; in total 60% said that 41% or more of their personnel were involved.

Research finds that many participants have to devote the majority of their time to preparing data instead of to analysis and data interaction.

Thinking of your organization's most recent BI and analytics projects, what percentage of the total time was spent preparing the data compared to the time spent performing analysis and data interaction?

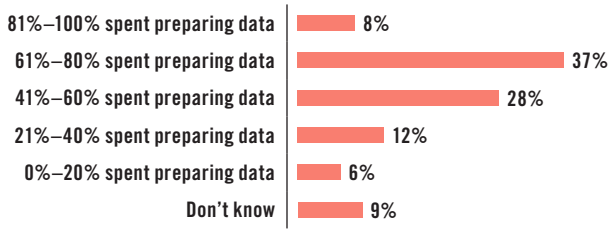


Figure 5. *Based on answers from 372 respondents.*

The results give a sense of the considerable time and personnel devoted to data preparation, which means that steps taken to reduce the time and improve the efficiency of personnel could have a significant impact. In Figure 6, we can see that research participants have a strong interest in reducing the amount of time and resources spent on data preparation processes for BI and analytics. Almost half (48%) said this was a very important objective and 38% said it was somewhat important.

Time spent cleansing and transforming data is about the same. Cleansing data tasks, such as deduplication and removal of dirty data, often appear to take up the largest share of data preparation time. However, our research shows that data transformation—which can include converting data values from source to destination, joining data, enriching data, splitting columns, and shaping and flattening data—may be taking organizations about the same amount of time to execute. When asked about whether they are spending more, the same, or less time cleansing as transforming data, the largest percentage of research participants (28%) said that their organizations are spending about the same amount of time for both, with 19% indicating that they are spending somewhat less time

and 19% indicating that they are spending far less time cleansing than transforming data. Just over a quarter (27%) are spending more time cleansing than transforming.

Overall, how important is it to your organization to reduce the amount of time and resources spent on data preparation processes for BI and analytics?

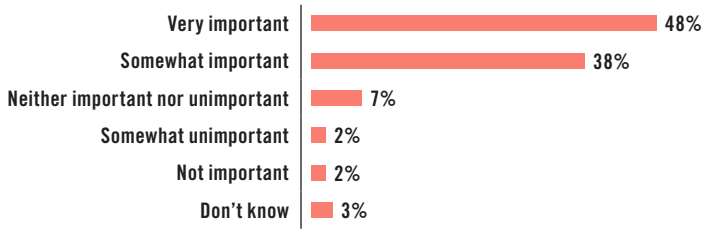


Figure 6. *Based on answers from 369 respondents.*

USER STORY TRANSPORTATION FIRM GETS ON TOP OF CUSTOMER DATA INTEGRATION CHALLENGES

Customer insight is a priority for organizations that want to optimize operations to satisfy customers. Data integration is vital to achieving this objective, but it is often where things founder due to problematic data and difficulties in data preparation steps for verifying and merging customer records. A transportation company in the U.S. Midwest is implementing SAS Data Management to improve customer data integration, in particular to overcome data quality issues associated with manually written invoices that result in variations in names, addresses, and single entities represented by sometimes dozens of records in the database. The transportation company has been able to reconcile records, identify duplicates, and reduce multiple instances into a single master record.

IT Responsiveness to Data Preparation Requests

The largest percentage of research participants said it takes IT two to six days to respond to data preparation requests; the second-largest percentage said it takes one to two weeks.

Given our research showing that IT remains largely responsible for performing data preparation tasks, it continues to fall primarily on IT to service user requests. We asked research participants how much time it typically takes their IT function to respond to data preparation requests, such as cleansing data or adding or changing attributes in existing databases or the data warehouse. The largest percentage (24%) said that IT takes two to six days; 18% said it takes one to two weeks; and the same percentage said it takes three to four weeks. In large organizations with over 10,000 employees, the percentages are about the same, except that the share of participants indicating it takes two to three months rises from 12% to 19%.

Does this mean IT is taking too long to fulfill requests? The answer, as in so many cases, is it depends. Sooner is most always better, particularly when decision makers are waiting on the data to analyze a potentially fleeting business opportunity. Users are left working with either older or lower-quality data. If addressing data preparation requests takes many weeks or months, this can result in business mistakes and poor decisions.

Data Volume and Variety: Addressing Integration Challenges

It's no secret that data volumes and variety are on the rise, no matter what the size of the organization. Traditional structured data is getting more voluminous and now many organizations are focused on tapping big data sources that are semistructured or unstructured, which could include

customer behavior data, machine or sensor data, log file data, geolocation data, and feeds from external sources. Analytics and data discovery projects can demand capabilities for interacting with large volumes of data and iterating through the data more frequently than with traditional BI to explore the data and test and score models.

When asked how large the data volume is for their organization's typical analytics projects, most research participants indicate that volumes are in the millions of rows (52%), with 19% saying they involve billions of rows or more (21% said thousands of rows or less, and 8% said volume varies too much to answer or didn't know). This highlights the need for data preparation processes and technologies capable of scaling not only *up*, within one system, but also *out* so that as data volumes grow, preparation workloads can be distributed dynamically across multiple nodes to improve performance.

The ability to handle data variety becomes important when users want to analyze diverse sources to examine different variables and spot trends, patterns, and correlations. Variety is not just about semistructured and unstructured data types. Even structured data can be varied if it originates from third parties, partners, and websites (such as from governments). However, many organizations are hesitant to encourage users to work with noncorporate or external data because they may be unable to apply their traditional data preparation and transformation processes to it and vet its quality. For the majority of research respondents (63%), only a quarter or less of the data they work with for analytics comes from noncorporate systems.

Organizations are hesitant to encourage users to work with noncorporate or external data because they may be unable to vet its quality.

This is not a surprising result, although the percentage could grow as business users, analysts, and data scientists seek to blend internal and external data views for analytics more frequently. In TDWI's interview research for this report, we talked to companies that are blending hundreds of input files, some of which come from external sources (primarily from partners); at one firm, data preparation processes cleanse the data from hundreds of input sources through matching and deduplication to create comprehensive master record views. They can use the master record to help partners cleanse their input files.

Our research finds that most organizations need to improve integrating noncorporate data. One-third (33%) of research participants are either somewhat dissatisfied or not satisfied with their organization's ability to integrate noncorporate data with corporate data for use in BI and analytics projects. Slightly fewer (30%) are either somewhat satisfied or very satisfied, with 22% neither satisfied nor dissatisfied.

Additionally, research participants are only moderately satisfied with their data preparation processes for enabling their organizations' users to work with new or complex data sources in their analysis. Complex sources are those that do not fit into traditional alphanumeric data field structures, such as geospatial data, social media data, and multimedia. Over one-third (37%) are either somewhat dissatisfied or not satisfied; 35% are either somewhat satisfied or very satisfied; and 23% are neither satisfied nor dissatisfied (6% did not know).

Research participants are moderately satisfied with their data preparation processes for new or complex data sources.

USER STORY ACCOLADE REVAMPS DATA MODEL TO IMPROVE QUALITY, SINGLE VIEWS, AND SCALABILITY

Accolade, a rapidly growing consumer health engagement services company, had to adjust its data strategy. "With our growth, we were bringing in more data, which we had to integrate with additional partners," said Dan Klein, vice president of information management at Accolade. The company provides Accolade Health Assistant, a single human point of contact for health benefits and healthcare needs, to self-insured employers, health plans, and health systems in the U.S. As individuals and families use the solution, interactions generate a considerable

amount of data. "We had to evaluate how we could optimize scaling and have more reusable components so that we could move out of writing custom code and go to more of a configuration-based approach," said Klein.

Accolade headed toward a Hadoop strategy, finding the stack and data technologies powerful, but the need for standardization remained a problem. "Hadoop required a lot of custom code, similar to what we were doing previously but with a more powerful base," Klein noted. Plus, the company needed ETL and data warehousing on top of Hadoop to service data consumption by BI and analytics users. "We wanted to make sure the data was stable before opening it up to those tools." Accolade chose to implement Talend's solutions, including Talend Data Preparation. "The ability to collect, standardize, and connect data from multiple sources is a key component of our business," said Klein. "[We have been able to] accelerate that process, cutting hours off of administrative tasks." As a result, Accolade's users have more time for innovation with analytics.

A big adjustment for Accolade has been the development of a new data model. "We changed the whole way we think about data in the course of a couple of quarters," Klein said. "Before, we just loaded data directly into our transactional system and extracted it for the data warehouse. We have moved to an environment where we have a specific data science phase and added real ETL, master data management, and data quality throughout our process. We've learned that it's very important to give the team enough time to learn the new way and fully adopt it."

Disconnect between Data Preparation and Analytics Processes

About a third of participants are mostly or always moving back and forth between using preparation tools and analysis and discovery tools.

Analytics workflows are often iterative. In many cases, users and analysts build models or scripts, apply them to some data, examine the results, and then repeat the process. Thus it can be frustrating if analytics and data preparation processes are fully separate and require significant waiting time in between. We asked research participants how closely their analytics and data preparations are integrated. In Figure 7, we can see that there is no dominant level of integration between the processes. Only 6% always move in distinct phases from using data preparation tools to using data analysis and discovery tools, with 22% mostly doing so. About a third (34%) are mostly or always working in an iterative fashion, moving back and forth between preparation and analysis and discovery tools. The largest percentage (29%) does both equally.

In your analysis work, do you find that you move in distinct phases from using data preparation tools to using data analysis and discovery tools, or is the process more iterative, where you frequently move back and forth between preparation and analysis/discovery tools?

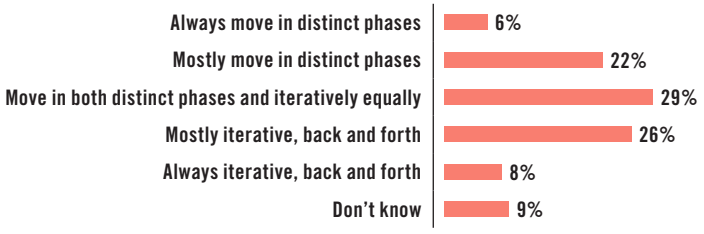


Figure 7. *Based on answers from 298 respondents.*

Analytics processes generally demand greater flexibility than traditional BI processes so that business users and analysts can explore data, pursue what-if inquiries, and evaluate different combinations of variables. Transformation routines geared to standard reporting often apply rules that are too generic for specific types of analytics, such as analyzing social media data. This requires users and analysts to develop transformation rules separately from their analytics workflows, adding time to their work. Some newer transformation tools and capabilities within visual analytics tools increase flexibility by giving users more control to adjust transformation as needed.

Of course, the pursuit of flexibility can have a downside for an enterprise if it creates too much chaos, rework, and processes that are not repeatable or programmed properly to scale. To manage enterprise performance and availability, IT will often try to "productionalize" as much data preparation work as possible. We asked research participants if their organization's data preparation work is largely ad hoc for specific needs or a part of everyday activity that is productionalized, scheduled, and run at scale. The survey results varied; 30% said their work is equally ad hoc and productionalized; 31% said it was mostly productionalized; and 4% said it was all productionalized. On the other side, about a quarter (24%) said their work was mostly ad hoc and 8% indicated that it was all ad hoc (3% did not know).

A best practice in many organizations is to give business users and analysts a source of data—data marts, data lakes, and cloud-based sandboxes are examples—so they can perform ad hoc analysis, transformation, and other data preparation at will. Then IT will consult with the users and analysts to determine which processes should be productionalized so that they can be repeated and applied to new data and new but similar requirements. Technologies available in the market can help organizations increase automation in the data preparation processes that organizations seek to productionalize.

Self-Service Data Preparation Objectives

Self-service data preparation is about enabling users to do more on their own to serve their BI, data discovery, and analytics needs. New technologies for self-service data preparation are automating processes so that users have less need for manual work in finding the right data among a variety of incoming sources, cleansing it, and transforming the data. Self-service data preparation can enable users to be less dependent on IT and data specialists in their organizations, not only for their expertise but for their time. Users can reduce their wait in IT and analytics specialists' backlogs. Self-service data preparation could relieve some IT burdens as long as IT is confident in giving up control of processes. Ideally, users and IT will work together to ensure that self-service enables users to be more productive but not increase data chaos and duplicative, uncontrolled work.

IT dominates ETL or ELT, but this is the only step that more than half of research participants say IT performs alone.

Figure 8 shows which steps in preparing data for BI and analytics are performed by research participants' IT functions, in a self-service fashion by users or analysts, or by both in collaboration. IT dominates the integration of data and performance of ETL or ELT; 63% said IT alone executes these processes. However, this is the only step in which more than half of research participants report IT performing alone. Almost half (45%) said IT only creates and maintains metadata models. When BI and analytics projects expand to enterprise-scale deployment, it is often necessary for IT to manage metadata models so that users are not burdened with the complexity. Some data preparation and cataloging tools are making it possible to automate metadata discovery and model creation so that users and IT can develop the cataloging resources more quickly and efficiently.

As far as steps performed in a self-service fashion by users or analysts, the highest percentage of research participants said that moving or replicating data into spreadsheets is performed in this fashion, which is not surprising. No other activity showed nearly as high a percentage for self-service. Higher percentages of participants indicated that there is collaboration between users and IT on tasks. For example, 42% said IT, users, and analysts collaborate on creating calculated fields, dimensions, or aggregations.

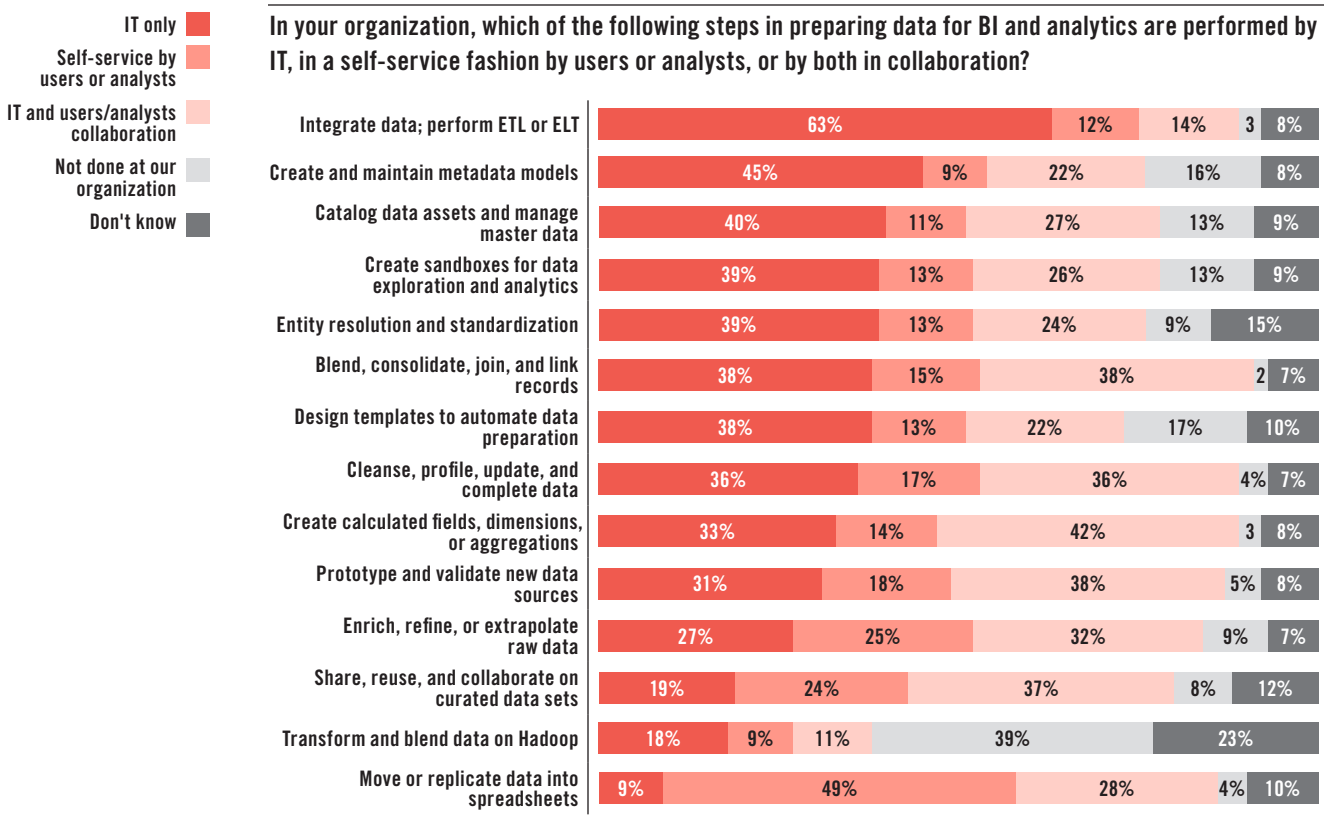


Figure 8. *Based on answers from 295 respondents. Ordered by highest IT only and descending.*

Reducing time to business insight is the most important benefit sought. TDWI Research finds that a high percentage of organizations (81%) are seeking to shorten the time it takes to produce business insight by increasing self-service data preparation. Nearly as many are hoping to increase data-driven decision making (76%). Rounding out the top five most common benefits sought from the selections we provided to research participants were improving the ability to react to changing business conditions (53%), operational efficiency for frontline workers (49%), and gaining a single, complete view of relevant data (43%). Business users and analysts trying to justify new technologies and practices for self-service data preparation would be wise to focus their arguments on how they could deliver benefits in these five areas.

USER STORY SMARTER DATA PREPARATION ENABLES BANKS TO UPGRADE RISK AND COMPLIANCE

Banks function in a highly regulated environment. To manage audits and the risk of compliance failure, they must use data and analytics effectively to catch money laundering, fraud, and trade manipulation. Although many banks have access to varied data sources that they can tap for analyzing risk, "the people who fill compliance roles are not traditionally data savvy," said an expert in risk and compliance analytics who works with many of the biggest banks at a large consulting firm. "It's a good opportunity for us to develop solutions that improve their data expertise and build predictive models so that banks can anticipate events and meet regulatory requirements."

The fundamental component of risk analytics and compliance is data quality. "Before we can do any analytics, we need to make sure that the data is of sufficient quality to be trusted so that when the regulators see the analytics, they have faith in the numbers," said the expert, who asked not to be named for this story. The consulting firm is partnering with Paxata to develop solutions for bank compliance groups to improve data quality for risk and compliance analytics.

The expert said that by implementing Paxata's use of machine learning, banks can move beyond traditional sampling of perhaps one-tenth of financial transactions to look at all the transactions. "Very quickly, based on a few filters and attributes, we can get a much more accurate and complete picture of a bank's risk," said the expert. They can gain more fine-grained knowledge about the parties to a transaction, such as whether they are on sanctions or watch lists. "We can do entity resolution—that is, see if two names spelled slightly differently are really the same person—and apply natural language processing to catch similar practices used by parties that have significant risk."

Increasing Self-Service Data Preparation

What are organizations currently doing to increase self-service data preparation and what are they planning to do? Our research finds that the most common course of action is to keep using existing tools, which is what 76% of research participants said their organizations are doing. The second most common practice is using a spreadsheet application or desktop database for data preparation (63%)—that is, tools that most users already have at their disposal. Most organizations face obstacles to adopting new technologies and will typically try to extend use of their current tools, despite the downsides that have been discussed, including poor repeatability and lack of flexibility. Moving in a new technology direction always demands a strong business and technical case.

Just over a third (35%) are using the data preparation features of their self-service BI, visual analytics, and data discovery tools, with 28% planning to within one year. Similarly, nearly a third (31%) are currently using self-service data preparation features of data integration, ETL, and data quality tools and 19% plan to within one year. These results show that organizations are at least moderately interested in newer, more sophisticated technologies that offer greater integration of data preparation into self-service front-end tools. Participants are showing similar interest in implementing self-service features in their traditionally IT-centric data integration, transformation, and data quality systems. Self-service features are important to enabling users to perform visual analytics and data discovery in a single workflow with data preparation or, at minimum, the ability to iterate back and forth between them from within a single graphical workspace.

Nearly a third of research participants are using self-service data preparation features of data integration, ETL, and data quality tools and 19% plan to within one year.

Self-Service Data Integration and Catalog Development

Many users and analysts would like greater capabilities for integrating or blending disparate data sources together for analysis.

Many users and analysts would like greater capabilities for integrating (or *blending*, to use an industry term) disparate data sources together for analysis. Often the objective is to see data relationships, including correlations and trends that become apparent in an integrated view. For example, structured transaction data could be joined with call center notes to better understand how sentiment is affecting sales; or an organization might want to see all the information they can touch about activities in a customer segment over a defined time period. Some vendors enable integration and blending through a visual drag-and-drop interface supported by machine learning, so that users do not have to work directly with the data.

We asked research participants which sources they are currently integrating or blending using a software solution's self-service data preparation capabilities and which they are planning to integrate or blend (Figure 9). From our list, top sources are not surprising: relational databases (75%) and data warehouses (72%), followed by spreadsheets and desktop databases (68%). More unusual sources are currently integrated or blended by smaller but still significant percentages—for example, demographic and segmentation data (44%); content, documents, and XML and JSON sources (32%); and Web analytics/clickstream data (30%).

Are users in your organization currently integrating or "blending" data from the following sources using a software solution's self-service data preparation capabilities? Is your organization planning to enable users to do so?

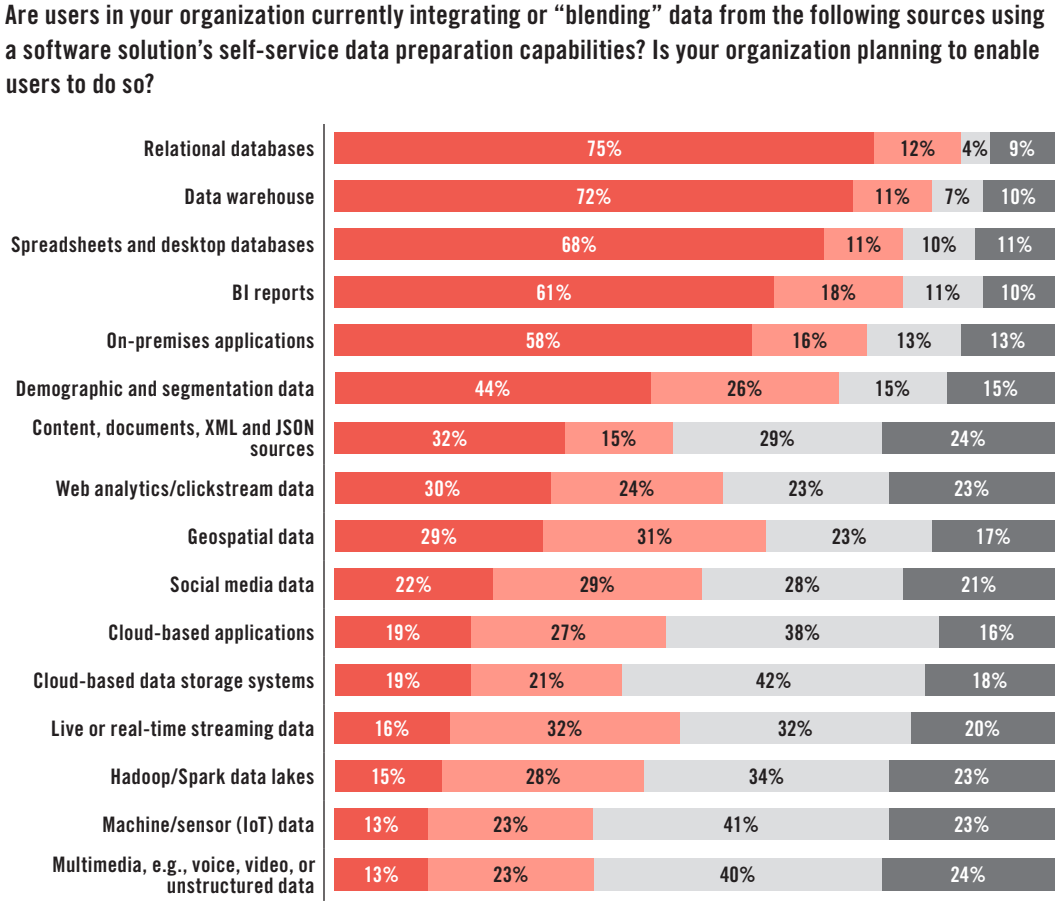


Figure 9. *Based on answers from 276 respondents. Ordered by highest currently doing.*

The largest percentages for sources that participants plan to enable for integration or blending are live or real-time streaming data (32%), geospatial data (31%), and social media data (29%). These choices reflect some organizations' desires to analyze more unusual sources than what typically exist in data warehouses to gain a view of trends and patterns. The sources of least interest for integrating and blending are cloud-based storage systems and machine/sensor (Internet of Things) data (42% and 41% of participants have no plans, respectively). These results suggest current immaturity in working with these sources in most organizations.

Organizations are moderately satisfied with current cataloging capabilities. Gaining an integrated view of multiple data sources can be aided by use of a data catalog, glossary, or similar repository. This resource pulls together information about the data from one or multiple sources. It can help users find data and enable groups to do so in a more repeatable fashion, including for satisfying higher-level searches for information about customers, products, or other objects of interest. Technology trends in data cataloging are moving toward self-service, with smarter systems employing machine learning to enable users to find and understand data without having detailed knowledge about its location.

TDWI Research finds that organizations are moderately satisfied with their capabilities for cataloging data assets from multiple sources and providing a self-service data catalog to support BI, analytics, and data discovery activities. One-quarter (25%) of research participants are somewhat satisfied and 9% are very satisfied. However, 37% are either somewhat dissatisfied or not satisfied (21% are neither satisfied nor dissatisfied and 8% did not know).

USER STORY FARM MARKET ID IMPROVES DATA PREP TO DEVELOP PRODUCTS AND EMBRACE BIG DATA

Big data is generating excitement in the agriculture industry. Farm Market iD, which for over 40 years has been providing data and marketing solutions to firms in the industry, has a potential big data gold mine. It has built up a large, proprietary database of facts about nearly all the farmland in the U.S., including data about farm owners, what is being grown, and geospatial data about field boundaries. As it begins to analyze this data along with massive, externally available streams of data about weather, soil conditions, water tables, and other geospatial data, Farm Market iD could realize one of the agriculture industry's biggest dreams: projecting yield. "If you can start projecting yield, it becomes unbelievably powerful," said Steve Rao, CEO of Farm Market iD. "You can really start to address challenges and disruptions to the food supply."

Yet before embarking on this bright future, Farm Market iD had to upgrade its data preparation to become faster and less repetitive. "We felt like we were running through oatmeal, manually having to touch everything that went through the pipeline and doing a lot of the same fundamental data standardization, deduplication, format checking, and quality steps over again for each client," Rao said. "The speed of the data coming in and the speed with which we needed to push it out to clients were getting faster, but we were slowed using disparate and unintegrated software. At first we didn't perceive it to be a problem, but then we realized that it was going to be hard to grow and to sustain that growth."

The company deployed RedPoint Data Management to automate, standardize, and integrate data preparation tasks. "We went after the low-hanging fruit first, which freed up time to do more development," said Rao. This included creating a Dealer Insights report that takes 60 different processes. According to Damon Horst, Farm Market iD's CTO, "We would not have attempted to productize this integration into a single deliverable the way it was before."

Governance and Self-Service Data Preparation

One danger of increasing self-service data preparation as well as self-service BI and analytics is the potential for additional data chaos. Data governance is important to ensuring that data assets are protected and that quality and consistency are valued and maintained. Data preparation and integration tools can be helpful to organizations' governance by aiding development of catalogs and glossaries through automated tagging and other features plus capabilities for tracking data lineage—that is, how the data has been used and transformed and by whom.

Organizations are planning to employ data preparation to strengthen knowledge of data assets to ensure adherence to policies and security rules.

We asked research participants to rank in importance data governance objectives that their organizations plan to implement as users engage in self-service data preparation (Figure 10; objectives are shown ranked by highest weighted average). The leading selection was creating business metadata and document definitions, followed by implementing data security rules and policies and applying policies to data from ingest to export. The results suggest that organizations are planning to employ key data preparation capabilities to strengthen their knowledge of data assets so that they can ensure successful adherence to policies and security rules.

Not as high priorities currently are some of the objectives for governing implementation of self-service data preparation, including tracking who prepares the data, validating new data sources, and tracking how data was prepared. However, as users mature in their experience with self-service data preparation, these governance objectives are likely to rise in importance.

As users engage in self-service data preparation for BI, analytics, and discovery, what are the most important data governance objectives that your organization plans to implement?

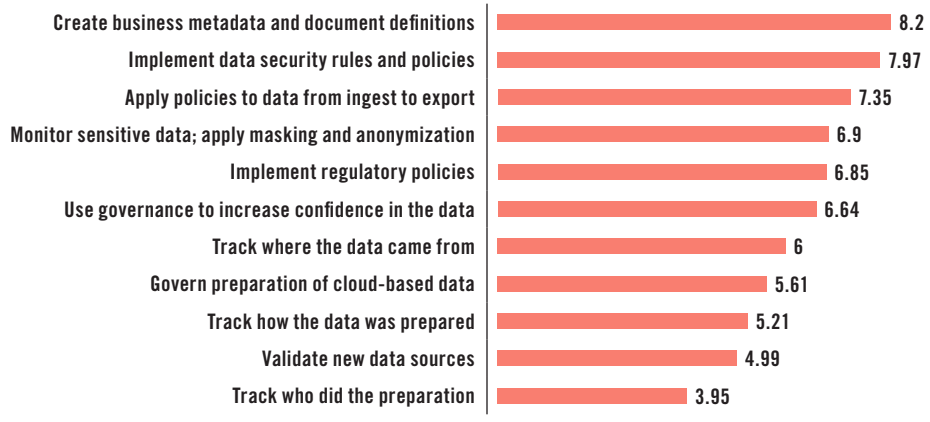


Figure 10. *Based on answers from 274 respondents. Ordered by highest weighted average.*

USER STORY LARGE HEALTHCARE SUPPLY CHAIN RELIES ON HEALTHY, TRUSTED DATA TO OPTIMIZE OPERATIONS

Healthcare providers are increasing their focus on patient care while having to reduce costs, enhance efficiency, and improve quality. Healthcare product and service suppliers must be responsive to changes in the industry by implementing agile, real-time, and optimized supply chain systems. By improving business operations with streamlined and compliant procure-to-pay and order-to-cash processes, they can take advantage of early payment discounts, optimized inventory levels, and reduced day sales outstanding. Data preparation can play a critical role in improving suppliers' business operations. A large provider of healthcare products and services in North America is using SAP Agile Data Preparation for its data conversion processes. It has been able to leverage the data expertise of its business and IT users to save time and improve cost efficiency in data preparation. The firm has been able to institute real-time information management processes that support business agility, improve customer relationships, and increase savings in supply chain operations.

Vendor Products

The firms that sponsored this report are among the leaders and innovators in providing technologies for data preparation, data quality, data integration, data cataloging, and related areas. To get a sense of where the industry as a whole is headed, this section takes a brief look at the portfolios of these vendors. (Note: the vendors and products mentioned here are representative, and the list is not intended to be comprehensive.)

Alation, Inc.

Founded in 2012 and based in Redwood City, California, Alation offers technologies that make business users and analysts productive in finding and querying data and collaborating on their work. By sharing their work, they can engage a business-user-driven approach to data governance and stewardship. The Alation Data Catalog builds a catalog resource by employing machine learning and artificial intelligence techniques to crawl and interpret data and sources of knowledge about the data, including metadata and user behavior. Alation's technology extends the traditional definition of a data catalog by adding to technical metadata key descriptive business information such as the business lexicon associated with a technical data label, aggregate visibility into who has accessed the data, and an understanding of popular joins between data sets. Users can search and query data with their natural language, without needing to know exact data item names or column titles. Within the Alation query application called Compose, Alation SmartSuggest offers recommendations on joins, filters, and relevant data sets. Alation can track data lineage and use the catalog as a growing knowledge base for governance and stewardship.

Alteryx, Inc.

Founded in 2010, Alteryx is headquartered in Irvine, California. The company's solutions aim to overcome data preparation, blending, and analytics challenges that slow business users and analysts, particularly those who have reached the limits of spreadsheet and point applications. Alteryx Designer provides self-service capabilities for blending data from a wide selection of both on-premises and cloud-based applications and databases. Through a graphical drag-and-drop interface, business users and analysts can select data sources and fields to blend with Alteryx providing code and writing the joins behind the scenes. Users can create reusable workflows that integrate preparation and blending activities, such as data cleansing and joining data from multiple sources, with building predictive models and other analytics processes. Alteryx has introduced in-database data blending and analytics, giving organizations the option of utilizing leading database engines for faster processing. Alteryx Server supports centralized deployment and sharing of analytics at scale, which enables organizations to schedule the execution of repeatable workflows to automate future projects on analytics workflow development and enforce IT and data governance controls.

Attivio, Inc.

Founded in 2007 and headquartered in Newton, Massachusetts, Attivio focuses on reducing the time business users need to find, understand, and unify their data. Attivio's perspective on data preparation is that most of the latency is due to difficulty in finding the right data sources upon which to do the data preparation work and in unifying the data for analysis. The company's core technology combines search, text analytics, and SQL database features to enable users to gain a more complete view of their information and find correlations between sources. It can develop a schema-less universal index to structured and unstructured data and support joins of structured and unstructured data. Attivio's search engine spiders data across sources, looking at all elements including schemas, catalogs, tables, and columns. Attivio's content analytics can enrich the data and learn its meaning, ultimately to build a semantic data catalog. Self-service capabilities let users find relevant data on their own, with the technology using its semantic understanding of the data to suggest joins.

Datameer

Founded in 2009 and headquartered in San Francisco, Datameer focuses on integrating data preparation and analytics into a big data analytics solution purpose-built for Hadoop. Users can work with a spreadsheet-like interface and single workflow to visually profile their data and perform other preparation steps as they move forward with transformations and analysis. Preparation processing does not have to be performed separately for big data workloads; as users analyze data, they can view new data coming in or determine whether some data needs more cleansing or other preparation work given the users' chosen analytics scenario. Users can also examine data lineage and track what is being done to the data. The interface lets users work at an abstraction level above the varied data in typical Hadoop systems, which could include geolocation data, social media sentiment data, online behavior data, and transaction data. More than 70 data connectors and schema flexibility ease access from the Datameer interface to structured and unstructured sources.

Looker

Founded in 2011 and headquartered in Santa Cruz, California, Looker offers a lightweight, Web-based platform for enabling business users to go beyond standard reporting to find and explore data in a self-service fashion to accomplish business objectives. It is frequently deployed in the cloud but can be implemented on premises just by running a JAR file. Looker helps organizations overcome lack of data centralization due to data silos by providing users with one data access point while the solution behind the scenes connects to any SQL store, including SQL-on-Hadoop technologies, and relies on the underlying power of the database to run transformations at query time, operating on the data "where it sits." Creating a single access point allows Looker to govern data definitions as set up by analysts so that all users are working off the same set. The in-database approach also enables Looker to avoid the latency of having to set up an intermediate store such as a cube or subset.

Paxata

Founded in 2012 and headquartered in Redwood City, California, Paxata provides an enterprise-grade data preparation platform for a comprehensive approach to data integration, data quality, enrichment, and data mastering. It employs machine learning and natural language processing algorithms to examine semantic content and speed data matching across sources that have multiple or no schemas. Paxata's visual and interactive workspace enables everyone, including nontechnical users, to access, explore, and profile data from a variety of sources. Users go through a series of machine-led suggestions and filters to identify patterns, duplicates, and quality issues (semantic and syntactic) and normalize and merge data. IntelliFusion—Paxata's parallel, in-memory columnar data preparation engine—is built on top of Apache Spark and uses Hadoop and NFS for data storage. Paxata scales to process over one billion rows interactively. Governance, versioning, and sharing of rules are managed in a lightweight Java layer; governance is enhanced by a "visual lineage" capability that can show transformation steps to trace the history of a data set. Annotation and tagging help users collaborate and keep "tribal knowledge" about data preparation projects in one place.

Pentaho, a Hitachi Group Company

Pentaho, a Hitachi Group company, provides an end-to-end analytics platform that incorporates data preparation and integration for analytics processes as well as a fully integrated suite of reports, visualizations, and dashboards for business users. Founded in 2004 and acquired by Hitachi in 2015, Pentaho focuses on data preparation across a variety of traditional and emerging data sources and formats, enabling users to visualize and analyze data at different steps within the data pipelines they create. Pentaho Data Integration (PDI) is at the heart of the Pentaho platform, providing an intuitive visual experience for preparing, blending, and processing data. Pentaho has support for Hadoop and other big data stores, enabling teams to visually create transformation processes that can run in these environments. Users can pass metadata to PDI at runtime to control transformation logic for multiple data ingestion and preparation processes, allowing one template to drive hundreds of autogenerated transformations. Pentaho provides capabilities to operationalize predictive models in data workflows and enables analytics tools and third-party applications to query PDI transformations as virtual blended data sets, with no physical staging required.

RedPoint Global

RedPoint Global, founded in 2006 and headquartered in Wellesley Hills, Massachusetts, offers an integrated portfolio of application and data management platforms dedicated to helping organizations across industries realize value from data for better customer intelligence and engagement. RedPoint brings together data integration, data quality, and master data management; its technologies can feed data to a variety of downstream BI and visual analytics applications as well as its own customer engagement orchestration and execution software. RedPoint provides tools for improving data quality processes, including parsing and matching of name, product, or device data. RedPoint Data Management for Hadoop focuses the company's technology on working effectively in the Hadoop ecosystem, which is growing in use for organizations seeking to use unstructured data such as from social media as part of customer intelligence and engagement. RedPoint technology can do all of its processing directly inside Hadoop clusters and works as a pure YARN application to distribute its code to the nodes where the data lives rather than pulling the data out to another store. RedPoint's tools enable user self-service through easy-to-use, drag-and-drop interfaces.

SAP / Intel

SAP, founded in 1972 and headquartered in Walldorf, Baden-Wurttemberg, Germany (U.S. headquarters in Newtown Square, Pennsylvania), addresses data preparation requirements with a capability that is part of a more comprehensive offering: SAP Enterprise Information Management. SAP Agile Data Preparation aims at enabling self-service, governed data preparation that takes advantage of the SAP HANA in-memory computing framework for performance. It balances the need to empower nontechnical users to discover, prepare, and share data with less IT help with the need to address an organization's governance and stewardship requirements. Users work through a visual interface to prepare data sets, choosing what sources they wish to combine, from CSV files to Hadoop data sets. SAP Agile Data Preparation can import data from on-premises or cloud-based sources. IT can monitor users' data preparation, take steps to improve data quality and trustworthiness, and operationalize data access and usage. The SAP Master Data Governance capability supplies tools for establishing the single-best record from across data domains. SAP Information Steward enables data stewards to develop and manage policies centrally and monitor compliance.

SAS

SAS, founded in 1976, offers market-leading technology products and services that span analytics, data management, and business intelligence. SAS has infused data preparation capabilities throughout its portfolio to improve the productivity of data professionals. Load data in and out of Hadoop without knowing how to write code or burdening IT. Empower business users to prepare data from a visual, easy-to-use interface, and execute data preparation tasks such as profiling, transformation, or data quality inside the database for improved performance.

Talend

Talend, founded in 2006 and headquartered in Redwood City, California, provides a common set of solutions for real time or batch, on premises or cloud, data (including big data) or application integration, data quality, and master data management. Talend is based on a foundation of open source development. Talend Data Preparation, which organizations can begin using as a free desktop application, enables business users to access, discover, profile, cleanse, standardize, enrich, and shape data. As it loads data from spreadsheet files, databases, and a variety of other sources, auto suggestions guide the user on their data discovery journey and preview the effect of potentially relevant preparation functions. Through its subscription version, Talend Data Preparation runs at enterprise scale, leverages Talend's broad connectivity. IT can protect data with role-based access and data masking. Data preparations that need to run regularly or be shared can be promoted into any integration process that the Talend platform can operationalize. Organizations can set up data preparation routines to run not just for BI and analytics but for other use cases such as external data ingestion, data stewardship, real-time integration, and data migration.

Trifacta

Trifacta provides data wrangling applications aimed at improving the productivity of business users, data scientists, and other data professionals, particularly when working with big data in Hadoop. Founded in 2012 and headquartered in San Francisco, Trifacta focuses on the notion of "predictive transformation" as key to giving users an intuitive, guided approach to exploring and preparing diverse data. Trifacta Wrangler is aimed at enabling all types of users to interact with various quantities of data in the flow of analytics processes. Wrangler has free downloadable desktop and commercial enterprise versions. Machine learning and analytics surface suggestions and guidance into visualizations such as histograms for data profiling; users can preview data, make selections, and define transformation rules. Rather than offer just one mode, Trifacta Wrangler can choose the optimal engine for running transformation logic, which could be locally on the desktop or using Spark or MapReduce on a Hadoop cluster. In March 2016, the company announced Photon Compute Framework, which offers a series of enhancements including the embedding of in-memory compute frameworks into the Trifacta interface.

Trillium Software

Trillium Software, a Harte Hanks Company, is headquartered in Burlington, Massachusetts. With its enterprise data quality solutions, Trillium is focused on enabling business and IT leaders to understand and trust their data assets to support analytics initiatives, real-time decision making, data governance, and more. The data quality solutions cover the full range of tasks, including profiling, parsing, standardization, matching, and enrichment. In February 2016, the company introduced Trillium Refine, a solution enabled by a partnership with UNIFi Software. Trillium Refine integrates self-service data preparation with data quality capabilities, with a particular focus on organizations' big data analytics requirements and processing in the Hadoop ecosystem. Native Hadoop processing provides faster execution without requiring manual coding. Trillium Refine can run on premises or in the cloud and features connectivity to any type of data. After Trillium Refine users prepare and cleanse data, they can select to export to Excel, Tableau, or other platforms. Trillium Refine employs machine learning and a recommendation engine to speed identification of data relationships and correlations across sources.

Waterline Data

Waterline Data offers a smart data catalog that empowers business analysts and data scientists to find, understand, and provision trusted data sets needed to do self-service analytics. Waterline Data combines automated data discovery, crowdsourcing of tribal data knowledge, and agile data governance to create and continuously enrich at scale a self-service catalog of curated data assets, including business metadata, sensitive data, and data lineage. Waterline automatically catalogs all the data in the lake and lets analysts find it quickly so they don't need to ask around to find the best data; they can then open it with their preferred data prep or data discovery tool. Waterline provides a unified data catalog that not only knows about all the data assets in the data lake but can also integrate with other data catalogs to ensure you get a complete view of your data assets. Waterline can also integrate with existing business glossaries, ETL tools, and is certified with Hortonworks Atlas and Cloudera Navigator to share metadata, tags, and data lineage.

Recommendations

Make shortening the time to achieving business insight a data preparation improvement priority. TDWI Research finds that the most important benefit sought by organizations we surveyed is reducing time to insight. Data preparation processes are often the primary source of latency—especially in the eyes of those working with data visualizations that make it all look so easy. Make it a priority to apply new technologies and methods that trim out delays in getting users from data to insight.

Research finds that the most important benefit sought from improving data preparation is reducing time to insight.

Focus on reducing how long data preparation takes to deliver valuable data. In many organizations, the biggest overall complaint about data preparation is that it takes too much time. Individual users working with spreadsheets are mired in manual data cleansing tasks. Data transformation routines can be too inflexible, requiring constant development of new routines. At the enterprise level, new data sources and types of data present constant challenges for IT, which is often forced to write specialized code with each new situation. Organizations should evaluate current data preparation procedures to eliminate unnecessary routines and develop strategies for increasing the use of automation and standardizing processes for incorporating and integrating new data. Once prepared, data should be registered in a data catalog for reuse by others.

Use new technologies and methods to achieve higher levels of repeatability. In organizations with uncontrolled data preparation processes being executed individually by every user (and often on spreadsheets), almost every process is a one-off that has to be done over again when there's a new requirement or new data. Transformation processes in particular need to be standardized and recorded for data lineage. Evaluate how you can apply technologies and adjust processes so that you can reuse scripts, workflows, and other elements for the next situations. Encourage sharing of repeatable methods, scripts, and workflows by adopting a collaboration framework.

Develop shared data catalogs, glossaries, and metadata repositories as part of data preparation processes. Shared resources such as these are critical to improving knowledge of the data and applying this knowledge to BI and analytics. Organizations have historically struggled to develop up-to-date and complete metadata repositories, data catalogs and glossaries because it has been mostly a manual effort. Technologies in the marketplace can automate development of these resources to make it easier for users to find data that is relevant, know more about various data sources, and collaborate with others using the same resources.

Evaluate self-service data preparation technologies. TDWI Research finds strong interest among both business and IT research participants in moving data preparation in the direction of more self-service. Organizations should evaluate both new self-service features of their existing BI and data analytics tools and their data integration, transformation, data quality, and metadata management systems. They should look for opportunities to let users test self-service capabilities on selected data samples before they try them on bigger and more complex sources.

There is strong interest among both business and IT research participants in moving data preparation in the direction of more self-service.

Integrate self-service data preparation with self-service BI and visual analytics. Users of popular self-service tools for interacting with data and creating visualizations often run into barriers because they lack technologies and know-how to prepare their own data. New technologies are increasing options for enabling nontechnical users to select, blend, and otherwise prepare data themselves. Evaluate technologies that improve this integration; encourage BI teams and business "power user" experts to share their expertise to help nontechnical users launch self-service data preparation processes to support their use of self-service BI and analytics tools.

Revise data transformation to support flexibility needed for analytics processes. To analyze social media data for sentiment, draw insights from customer behavior data, do predictive analysis of IoT data streams, and more, users and analysts need flexible means of setting transformation rules. Often, they need to explore the data first to develop the rules rather than apply them first to all data sources. Organizations should evaluate the effectiveness of current ETL and other data transformation routines and test newer technologies that offer greater flexibility, including through self-service capabilities.

Improve governance and stewardship over self-service data preparation. Users are excited about using self-service data preparation technologies to increase flexibility and deepen their data interaction as they use self-service BI and visual analytics. However, it is imperative that organizations align data preparation and governance so that the introduction of greater self-service does not add to data chaos. Organizations should emphasize not only the need to adhere to policies and regulations but also the benefits that will accrue to users from participating in processes to achieve higher data quality, more trust in the data they use in analytics, and the ability to apply repeatable processes.

Involve data preparation in data governance objectives. Organizations are under pressure to improve data governance: first to ensure that regulatory policies and rules regarding use of sensitive data are being followed and, second, to support data stewards in overseeing data quality and that content users create meets internal standards. Data preparation processes should be essential to carrying out data governance but not enough organizations are making this link. Use data preparation processes to build a bridge between end user interests and adherence to data governance policies so that in the end data governance becomes more effective.

Innovations are enabling organizations to know more about Hadoop data, find it more easily, and use it to improve insights with analytics.

Use data preparation technologies and methods to increase the value of Hadoop data. Many organizations are amassing data in Hadoop data lakes, which are serving variously as massive data storage, staging areas for ETL and analytics, and operational data stores before some of the data is moved into data warehouses. Innovative technologies in the marketplace are enabling organizations to apply data preparation processes to Hadoop data itself, enabling organizations to know more about this data, find the right data more easily, and improve speed to insight with analytics. Organizations with Hadoop data investments should evaluate data preparation technologies and methods to realize higher value.

Create CoE to help IT learn how to improve responsiveness to users' needs. TDWI Research finds that despite all the attention on the self-service phenomenon, IT still performs the lion's share of data preparation. IT generally prefers regular production data-preparation routines for ETL, data integration, and other steps. Users are unsatisfied if IT cannot address ad hoc or specialized needs and are likely to use their own spreadsheets or other applications in less efficient ways. Center of Excellence and governance committees can be helpful in bringing users and IT together to resolve differences and help IT see where it needs to be more responsive.

RedPoint

www.redpoint.net

Gain Extreme Agility and Performance While Creating Analytics-Ready Data

Business leaders today understand that utilizing big data effectively can lead directly to profits. The simplified calculation goes something like this: collect lots and lots of data, put it in some kind of database, add a data scientist and some software, then sit back and collect money.

We know this isn't true, but the main issue is the large gap between data collection and data analysis. Preparing analytic-ready data is difficult especially when working with structured and unstructured data. So it requires the right tools and processes to achieve the best results.

Today's data analysts have a tremendous amount of responsibility. Not only do they have more data, but that data comes from more structures and is stored in more varied technologies—throwing vast amounts of unstructured data into the mix.

The up-front work alone involves de-normalizing and normalizing data, recoding attributes to the data, data exploration, univariate analysis, and data profiling. On top of all this is the coding of analytics systems. This data prep can take up 80 percent of analysts' time, leaving little to spend in production mode where they can finally tune their algorithm.

However, there is a better way.

RedPoint aims to make the process as fast and as automated as possible so customers have more time to tune their algorithm. This means enabling data analysts and scientists to work directly in Hadoop without requiring them to move data from one place to another with a tremendous amount of data preparation.

Take the recent Hadoop integration benchmark conducted by MCG Services. The analysts found that RedPoint completed workloads very quickly—well within enterprise requirements and faster than they even thought possible. Compared to a similar benchmark conducted by MCG Services with two leading data management competitors, RedPoint ran 5.5 times faster than a product using Spark and 19 times faster than a product using MapReduce.

Using RedPoint Data Management, the same job required no coding. With just 15 minutes of development and no tuning or optimization, it returned the same results in just three minutes.

Not only does RedPoint perform all the functionality inside the data lake, it happens in its raw document format. The system handles the ETL, quality, matching, MDM, de-duping, parsing, merge/purge, address standardization, and master keys. What's more, RedPoint DM can auto structure data instantly.

Empowering organizations to break free from the slog of data preparation is critical in a modern data-driven organization. Setting up a data preparatory environment that will enable data scientists and analysts to spend less time on data preparation and more time on business-critical analytics projects is something that RedPoint believes will help produce better and more accurate analytics, leading to better business outcomes.

TDWI Research provides research and advice for data professionals worldwide. TDWI Research focuses exclusively on business intelligence, data warehousing, and analytics issues and teams up with industry thought leaders and practitioners to deliver both broad and deep understanding of the business and technical challenges surrounding the deployment and use of business intelligence, data warehousing, and analytics solutions. TDWI Research offers in-depth research reports, commentary, inquiry services, and topical conferences as well as strategic planning services to user and vendor organizations.

Advancing all things data.
555 S Renton Village Place, Ste. 700
Renton, WA 98057-3295

T 425.277.9126
F 425.687.2842
E info@tdwi.org

tdwi.org

WP-TDUS0916-02