



DATA PREPARATION CHALLENGES FACING EVERY ENTERPRISE

- 1 Q&A: Data Preparation Challenges Facing Every Enterprise
- 5 Why Data Preparation Matters
- 8 The Personas of Data Prep
- 11 About SAS

Sponsored by:





Q&A: DATA PREPARATION CHALLENGES FACING EVERY ENTERPRISE

Time spent cleaning data is eating away at the time available for analysis. What steps can your enterprise take to get analytics back on track?

Pure, clean, complete, and accurate data is the hallmark of data quality, but it's not the only factor enterprises must pay attention to if their analysts and data scientists are to get the best results. Today's data is found in many places (both internal and external to the organization) in dozens of formats.

We spoke with Fiona McNeill, product marketing manager at SAS, about the challenges enterprises face with data preparation and what best practices she can recommend to make the process easier.

TDWI: Estimates of the time data scientists spend preparing data (instead of analyzing it) range from 50 to 80 percent. Why is this figure so high?

Fiona McNeill: Substantial time is spent in data preparation because data is typically not ready-made for business discovery and analysis. Data is typically collected from systems, applications, and channels that were designed to serve a function, such as tracking supply chain activity or providing a forum for Web commentary. Such designs focused on the elements needed to make that system or application work, a very different set of engineering

principles than needed for defining data and discovering analytics insights.

Data prep becomes an even bigger issue when you consider big data collected from a host of sources. Once you've identified the question you want to answer, you must curate your data (it will likely be in different formats and collected at different rates). In addition, if your analysis includes a variety of data types, you will need to address the nuances of each one. For example, text data is complicated by the subtleties, abbreviations, and misspellings in human language. Each type of data also needs to be linked somehow with other types, such as streaming data from sensors or relational data tables. Defining your question is only the beginning; preparing the data to answer that question can be an extensive—but critical—task.

Aren't ETL and other data munging processes supposed to clean up and prepare data before it is sent to a data scientist or business data analyst? What's gone wrong? What are the kinds of things traditional ETL doesn't handle that a data prep process must?

Nothing has gone wrong. What we are seeing is an evolution in capabilities that bridge traditional enterprise ETL and self-service data preparation.

Traditional ETL is focused on integrating and synchronizing data, often provisioned from a variety of operational systems and structured into an organizational standard. This vetted and trusted source of information then becomes the source for other databases, reports, and business visualization and analytics tools. Within the realm of ETL, developers and data engineers who are well versed in data integration, data quality, and information governance set the corporate standard for business-critical information.

Self-service data preparation is designed for end users to explore data, often curating data from data lakes (data there may have little or no structure), corporate sources, flat files, and more—all to address their questions. Because the data needed varies with the question at hand, substantial data reconciliation, cleansing, and standardization are

typically required before analysis can begin. As users explore the data, they may bring additional sources into the analysis, further refining it and improving the conclusions and actions that will be based on it.

We used to think that data prep was about adding missing values (“fixing” data) and finding outliers (identifying skewed data), thereby improving data quality. Today the term encompasses a wider range of tasks. What kinds of things now fall under the banner of data prep?

The fundamental tasks for successful analytical data preparation include blending, profiling, cleansing, filtering, transforming, and standardizing data for analytics. These activities aren't something done in batch. They are done in a responsive environment, so the blending and shaping of data are immediately visible to the practitioners who are exploring and creating data for their analysis.

Data preparation activities need to consider how representative your data is of the total population, the question being asked, and the ability to change business processes or activities based on your findings. For some analytics, redundant data can lead to ill-conditioned models. Other methods get more accurate the more times they analyze the same data, and many require complete rows of data. Questions practitioners need to ask during data preparation include:

- **Data completeness:** Is there enough data available to represent the pattern of interest? Is additional data from different sources required? Should random, systematic, or no sampling be done as part of training data construction? Are the outliers significant?
- **Data correctness:** Is the data accurate? Plausibility checks for accuracy, misspellings, parsing, standardization to specific locales, and the like are identified in profiles and cleansed in accordance with the analysis to be performed.
- **Data quantity:** Is there enough? Assess the number of events required per observation period and determine

how much data is needed to represent segment variability or to simply come to a successful conclusion. An easy method for matching data from alternate sources is key to ensuring sufficient data.

- **Data usability:** How does the data need to be aligned for this analysis and the actions taken from it? Assessments often involve cardinality, binning, correlations, derivations of new values, gender or identity analysis, and other methods to prepare data at the needed level of granularity.

Why is this becoming more important to organizations now?

The need to prepare data for business intelligence, discovery, and analytics has always existed. Now, however, we see a shift from overburdened IT to the end users, enabling them with self-service tools that don't require any coding and empowering them to prepare data themselves. The shift is driven by an increasing business demand for data; businesses require more data to be more responsive to customers, more dynamic in response to markets, more efficient in operations, and more innovative with new solutions. Democratizing analytics and giving anyone within the organization the tools to be data informed means users now need to access, explore, and prepare their own data.

How have enterprises traditionally tackled data preparation and who has generally been responsible for preparing the data? How is that evolving?

Generally, there have been two sources for preparing data. One was IT professionals, who constructed multidimensional tables of data that end users could explore and analyze from many different perspectives using BI tools. The other source was individual coders found in pockets throughout the organization. Given their knowledge of data sources, coders would construct analytics base tables upon request.

These roles are now evolving. Analytics practitioners can access data easily through a Web browser from both vetted, trusted IT sources and programs defined by coding experts, in addition to other curated sources. Instead of serving data

to a limited few, data access can now be granted to many more users. Analytics practitioners now have the flexibility to consider any authorized input source in their data preparation. It's no longer a trade-off between IT rigor and end-user flexibility. Data preparation tools that are centrally managed include controls for sources and even individual data elements. These tools can automatically generate code (for inclusion in IT scheduling to ensure current data for a particular user as part of regular updates) and the data preparation recipes can be shared as assets to use individually or in collaboration with others.

Obviously, the better the data, the more productive the data scientists, the more accurate the analysis, and the better the final decisions. What other benefits do enterprises enjoy when they have good data prep processes in place?

The self-sufficiency of data preparation benefits many parts of any data-driven organization. Analytics practitioners (specifically business analysts and citizen data scientists) are more productive because they have the access and the features to cleanse and prepare data for their reporting and analytics in a familiar environment that doesn't require coding.

For data scientists, data preparation can provide faster results than they are used to—at least if they are using data preparation technology that's designed to take advantage of distributed, parallel processing environments.

Enablers to the business, such as ETL developers and data engineers, can further improve practitioner effectiveness without sacrificing governance controls when data preparation tools are provided from a centralized, managed environment.

Finally, sponsors of data preparation adoption are helping to break down internal barriers to data, and insights from data, achieving better balance in the rigor of IT with end-user experimentation and innovation.

What are three best practices for getting the most out of your data prep process?

First, and perhaps most important, consider data preparation as an integrated part of modernizing your data environment. Data preparation is undoubtedly empowering for the end user. When you consider what that end user wants to do with data (such as access and blend a variety of sources, standardize features and formats, create desirable reports, or explore and discover new insights), recognize that such activities occur in a constantly changing data landscape. Distilling data into something useful is an important part of understanding both the question you want to answer and its domain.

Data preparation may not be glamorous, but it is necessary, and no group or individual will be able to keep up with the growing demand alone. Organizations need to recognize that the productivity and efficiency gains of data preparation extend beyond specialists and are now necessary for any part of the organization charged with examining data and driving insights.

Second, as a best practice, adopt data preparation that is managed by IT. This may seem contradictory, with practitioner self-service emanating from IT increasingly unable to fulfill the breadth of use cases and needs, but it isn't. This refers to the management of this self-service environment. Centrally managed technology ensures that organizational rigor is protected, system-of-record data is consistent, sensitive data is masked, authorization to source inputs is controlled, governance is retained, currency is maintained, and lineage from source to result is traceable.

It also means that the processing of tasks and computations can be optimized to available hardware and system memory resources. For the data prep practitioner, centrally managed data preparation provides a location that can schedule defined plans for customized reuse as part of naturally occurring updates.

Third, identify how data preparation is used as part of analytics workflow. Analytics is an iterative process, data relationships between potential inputs are merely guessed at, and quality from some sources is uncertain. Until data is

sourced, profiled, and explored, it's unknown how complete, correct, and usable it is. Once initial triage is complete, more or different data is often needed, and data preparation begins again.

Keeping in mind that the promise of data preparation is productivity, bouncing in and out of different interfaces to prepare, explore, and analyze data is often a frustrating waste of time. Copying code, even if automatically generated from an interface (or worse, copying data for analysis by a different tool) is inefficient. Ask questions: Can data simply be accessed without movement? Can data preparation and analytics computations be pushed to the data and run there or held in memory for multiple users to examine simultaneously?

Look for efficiencies from the perspective of the entire analytics life cycle, not just one piece of it.



WHY DATA PREPARATION MATTERS

Technology modernization rarely occurs in a vacuum. It usually supports or is driven by business modernization.

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.

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.

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.

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.

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 selfservice business and governance demands for faster development of integrated views of data.

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.

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.

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.

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.

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 byproduct 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.



By Brian J. Dooley

THE PERSONAS OF DATA PREP

Different data-preparation use cases demand features and functionality for different types of roles.

As new data processing solutions such as cloud, big data, and machine learning become more prevalent, the basics of database and analytics processing must be transformed. One of the most significant areas of change is in data preparation. Many vendors have been looking at this area recently, providing new mechanisms to ingest, manage, and orchestrate data sourced from areas that might include structured, semistructured, and completely unstructured data. ETL must be adapted to handle many of these possibilities; at the same time, the new data types need to be entered seamlessly into the data storage environment and they must be verified and secured as had been done by the ETL and data-cleansing routines of the past.

However, data prep isn't just about ensuring data is pre-defined for one, common use. Today, it must cater to a much wider range of needs. As analysis has grown in popularity and shifted beyond data analysts to specialists and subject matter experts in other fields, nontechnical users have become a principal focus of data prep. Although these individuals are not schooled in SQL and code development, data is central to their daily tasks, and they need to work with it across a wide variety of use cases—though they

may not be able to write a JOIN statement or expand their technical capabilities beyond point-and-click data access.

Data Prep to the Rescue

Among the companies prominent in data prep is SAS. “There are a number of developing areas that are important to data prep,” says Ron Agresta, product manager director for data management at the company.

“It is important that it should be easy to gain access to data and find data regardless of where it might be; this should be straightforward. Another key issue is that data prep must support the notion of collaboration. If someone wishes to get a colleague to help, they must be able to share work artifacts with them and share the outcome.”

Collaboration, Agresta points out, “needs to be both within the nontechnical community and with data experts across the organization. Companies want data prep to empower nontechnical users, but they also want to have some control over ultimately what they’re using and how they are using it.”

Within this context, security is also important. A solution such as SAS Data Preparation can put constraints on data use availability to ensure that people can only see certain rows or columns of data or hide some items such as a Social Security number or government ID. These issues are important for nontechnical users, who might need to be blocked from access to things that they shouldn’t see.

Organizations are also moving to automate routine processes. This introduces additional issues in preparation of data for self-service. According to Agresta, “In a lot of data preparation applications, we can profile and analyze the data to find problems and opportunities within it before sending it to analytics. For example, it is possible to ensure that you are using the correct variables or the correct pairs of variables or relationships—not just in a single data set but across data sets—that would enable you to easily merge the data.”

This is important in predictive analytics, forecasting, and other processing that requires the ability to join multiple data sources in a smart way—including a need to add missing values and extrapolate values not available in the data. Agresta points out that “the easier we can make this

happen using SAS deep analytics in our data management capabilities, the better. It lets the data scientist get to their job more quickly rather than spending time trying to find the right data joined in just the right way.”

Businesses need to pre-process their data in new ways even before they use it for analytics. “We apply a number of metrics to data and analysis and profiling ahead of any transformation process,” Agresta says. “Things like central tendency, percentiles, and cardinality help to determine whether data is suitable for a selected activity, whether it’s analytics, model building, testing of models, or overall performance. They also help make sure the data you think you have aligns with the expected result.”

Addressing Changing Demands

Although batch processes are still being used to do the bulk of the heavy lifting in most organizations, some of that has been moving to real time. This requires data crunching at the point of data capture, placing an additional burden on data prep.

“This is something that’s really important to SAS,” Agresta explains. “We have an event stream processing engine that lets you build your data quality and cleansing rules in the data stream itself. This means that you don’t have to move the data to one engine and then send it back to some other source. Event stream processing can just listen to the data stream and make corrections on the fly as data is moving through it. You can work on the data much closer to the point of capture, and you can also filter the data at the same time, reducing the processing workload.”

An ongoing trend over the past several decades has been the move toward self-service for many aspects of business intelligence and analytics. The user interface has become extremely important. Within self-service data prep, interfaces are changing to adapt to multiple uses of data sets, and access by different types of users, or what Agresta calls roles. It is important to satisfy both the nontechnical data worker (who requires simplicity) and the data developer or engineer (who requires advanced access and capabilities suitable for coding as well as data manipulation).

“Self-service can go in several ways,” says Agresta. “There can be a coding interface for very technical users; this may include capability to write SAS code or Python code to gain access to deeper data quality algorithms. We also need more interactive or visual-based ways to deal with data. For nontechnical users, we need to provide the right level of capabilities; they need to work through the data using an easy path and a more interactive or visual way to deal with data.”

Another area of growing importance is metadata. The new SAS Data Preparation solution allows enterprises to share metadata, providing a critical aid to automation and future data handling.

“Metadata is important for two reasons,” says Agresta. “Internally, across SAS applications or SAS components, we don’t want people to do the same thing twice. If they do some transformations on a data set such as labeling columns or renaming the table, and want to move that data to a report, we don’t want them to go back to the reporting application to transform the data again. It’s really important to have a seamless experience. Externally, we want to make sure that SAS can import and export metadata in external applications and vice versa to interoperate in a larger IT landscape.”

Meeting the Needs of Different Users

The new SAS Data Preparation product helps enterprises prepare data for different use cases, with new enhancements expected in 2018. It serves several roles within the organization: traditional data analysts or data scientists, anybody for whom working with data is the primary job—whether or not they are experts in manipulating data—plus evolving new roles of data developers or engineers as well as data stewards. Data developers have the traditional ETL skill set and need additional capabilities. Data stewards are tasked with making sure data is fit to purpose, and meets policies that ensure data is only available to certain users by permission.

“We have started to introduce capabilities for these two groups of end-users as well as for the data analyst,” Agresta notes.

“The driving force is that we want these different users to collaborate. If I’m a data analyst building a report and I did

an ad hoc transformation, I may want to share it with, say, a data engineer who could alter it for performance. At the same time, data stewards need to be able to understand that the data coming out of the system is of high quality and fit for use.

“Ultimately SAS Data Preparation will support those three different users with different perspectives and different access to properties of the system, with product capabilities specific to their needs expanded in 2018.”



sas.com

SAS is a leader in business analytics software and services, helping organizations across the globe transform their data into well-defined insights—insights that give a fresh perspective on your business, helping you identify what's working, fix what isn't, and innovate in ways that keep you ahead of the competition. SAS develops software that turns large amounts of data into intelligence you can act on and automate, all from a single code base.

Preparing data can consume 80 percent of an analyst's time, leaving just 20 percent for unlocking insights with analytics. SAS Data Preparation helps reverse those numbers, improving efficiency for both business users and IT by letting analysts quickly prepare data for analytics in a self-service, point-and-click environment.

The intuitive interface empowers business analysts and citizen data scientists to access, blend, shape, and cleanse data without coding or help from IT, and it automatically integrates with downstream analytics and reporting tasks. Once defined, templates can be scheduled as part of data-refresh cycles and shared with others. With simplified data preparation tasks seamlessly defined as part of the activities involved in analytics processing, users can now spend more time analyzing data and less time preparing it.

Learn more at sas.com/dataprep.

Try a free 14 day trial of SAS Data Preparation here:
sas.com/trydataprep



tdwi.org

TDWI is your source for in-depth education and research on all things data. For 20 years, TDWI has been helping data professionals get smarter so the companies they work for can innovate and grow faster. TDWI provides individuals and teams with comprehensive business and technical education and research that allow them to acquire the knowledge and skills they need, when and where they need them.

TDWI advances the art and science of realizing business value from data by providing an objective forum where industry experts, solution providers, and practitioners can explore and enhance data competencies, practices, and technologies.

TDWI offers four major conferences, topical seminars, onsite education, a worldwide membership program, business intelligence certification, live webinars, resource-filled publications, industry news, an in-depth research program, and a comprehensive website at tdwi.org.

© 2017 by TDWI, a division of 1105 Media, Inc. All rights reserved.
Reproductions in whole or in part are prohibited except by written permission.
Email requests or feedback to info@tdwi.org.

Product and company names mentioned herein may be trademarks and/or registered trademarks of their respective companies.