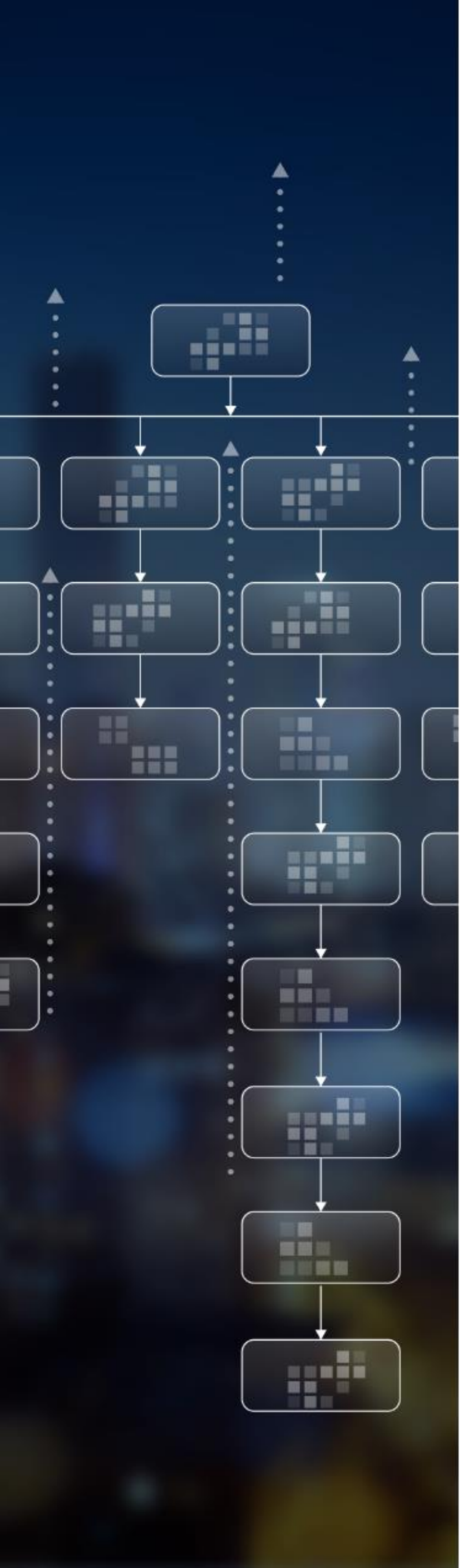


What is data preparation? An in-depth guide

June 2024

In this guide:

Data preparation is the process of gathering, combining, structuring and organizing data for use in business intelligence, analytics and data science applications. This comprehensive guide to data preparation further explains what it is, how to do it and the benefits it provides in organizations. You'll also find information on data preparation tools, best practices and common challenges faced in preparing data. Throughout the guide, hyperlinks point to related articles that provide more information on the covered topics.

In this guide:

What is data preparation? An in-depth guide

CRAIG STEDMAN, INDUSTRY EDITOR

Data preparation is the process of gathering, combining, structuring and organizing data for use in business intelligence, analytics and data science applications. It's done in stages that include data preprocessing, profiling, cleansing, transformation and validation. Data preparation often also involves pulling together data from both an organization's internal systems and external sources.

IT, BI and data management teams do data preparation work as they integrate data sets to load into data warehouses, data lakes or other repositories. They then refine the prepared data sets as needed when new analytics applications are developed. In addition, data scientists, data engineers, data analysts and business users increasingly use self-service data preparation tools to collect and prepare data themselves.

Data preparation is often referred to informally as *data prep*. Alternatively, it's also known as *data wrangling*. But some practitioners use the latter term in a narrower sense to refer to cleansing, structuring and transforming data, which distinguishes data wrangling from the data preprocessing stage.

In this guide:

This comprehensive guide to data preparation further explains what it is, how to do it and the benefits it provides in organizations. You'll also find information on data preparation tools, best practices and common challenges faced in preparing data. Throughout the guide, hyperlinks point to related articles that provide more information on the covered topics.

WHY IS DATA PREPARATION IMPORTANT?

One of the main purposes of data preparation is to ensure that raw data being processed for analytics uses is accurate and consistent. Data is commonly created with missing values, inaccuracies or other errors. Also, separate data sets often have different formats that must be reconciled when they're combined. Correcting data errors, improving data quality and consolidating data sets are big parts of data preparation projects that help generate valid analytics results.

Data preparation also involves finding relevant data to ensure that analytics applications deliver meaningful information and actionable insights for business decision-making. The data often is enriched and optimized to make it more informative and useful -- for example, by blending internal and external data sets, creating new data fields, eliminating outlier values and addressing imbalanced data sets that could skew analytics results.

In addition, BI and data management teams use the data preparation process to curate data sets for business users to analyze. Doing so helps streamline and guide self-service BI applications for business analysts, executives and workers.

WHAT ARE THE BENEFITS OF DATA PREPARATION?

Data scientists often complain that they spend much of their time gathering, cleansing and structuring data. A big benefit of an effective data preparation process is that they and other end users can focus more on data mining and data analysis -- the parts of their job that generate business value. For example, data preparation can be done more quickly, and prepared data can automatically be fed to users for recurring analytics applications.

Done properly, data preparation also helps an organization do the following to gain business benefits:

- Ensure the data used in analytics applications produces reliable results.
- Identify and fix data issues that otherwise might not be detected.
- Enable more informed decision-making by business executives and operational workers.
- Reduce data management and analytics costs.

- Avoid duplication of effort in preparing data for use in multiple applications.
- Get a higher ROI from BI and data science initiatives.

Effective data preparation is particularly beneficial in big data environments that store a combination of structured, semistructured and unstructured data to support machine learning (ML), predictive analytics and other forms of advanced analytics. Those applications typically involve large amounts of data, which is often stored in raw form in a data lake until it's needed for specific analytics uses. As a result, preparing data for machine learning can be more time-consuming than creating the ML algorithms to run against the data -- a situation that a well-managed data prep process helps rectify.

STEPS IN THE DATA PREPARATION PROCESS

Data preparation is done in a series of steps. There's some variation in the data preparation steps listed by different data professionals and software vendors, but the process typically involves the following tasks:

1. DATA COLLECTION

Relevant data is gathered from operational systems, data warehouses, data lakes and other data sources. During the data collection step, data scientists, data engineers, BI team members, other data professionals and end users should confirm that the data they're gathering is a good fit for the objectives of planned analytics applications.

2. DATA DISCOVERY AND PROFILING

The next step is exploring the collected data to better understand what it contains and what needs to be done to prepare it for the intended uses. To help with that, data profiling identifies relationships, connections and other attributes in data sets. It also finds inconsistencies, anomalies, missing values and other data quality issues. While they sound somewhat similar, profiling differs from data mining, which is a separate process for identifying patterns and correlations in data sets as part of analytics applications.

3. DATA CLEANSING

Next, the identified data errors and issues are corrected to create complete and accurate data sets. For example, as part of data cleansing work, faulty data is removed or fixed, missing values are filled in and inconsistent entries are harmonized.

4. DATA STRUCTURING

At this point, the data needs to be modeled and organized to meet analytics requirements. For example, data stored in comma-separated values files or other file formats must be converted into tables to make it accessible to BI and analytics tools.

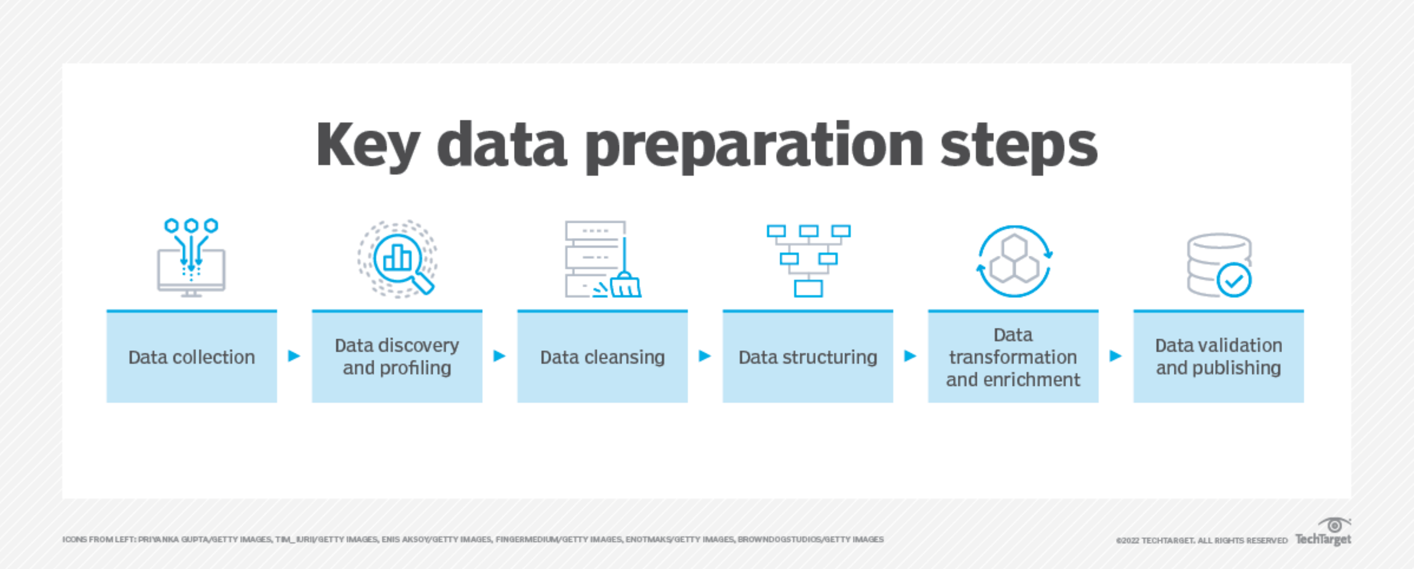
5. DATA TRANSFORMATION AND ENRICHMENT

In addition to being structured, the data typically must be transformed into a unified and usable format. For example, data transformation might involve creating new fields or columns that aggregate values from existing ones. Data enrichment further enhances and optimizes data sets as needed, through measures such as augmenting and adding data.

6. DATA VALIDATION AND PUBLISHING

In this last step, automated data validation routines are run against the data to check its consistency, completeness and accuracy. The prepared data is then stored in a data warehouse, a data lake or another repository, where it's either used by whoever prepared it or made available for other users to access.

Data preparation can also incorporate or feed into data curation work that creates ready-to-use data sets for BI and analytics applications. Data curation involves tasks such as indexing, cataloging and maintaining data sets and their associated metadata to help users find and access the data. In some organizations, data curator is a formal role that works collaboratively with data scientists, business analysts, other users and the IT and data management teams. In others, data might be curated by data stewards, data engineers, database administrators or data scientists and business users themselves.



Key data preparation steps

Data collection ► Data discovery and profiling ► Data cleansing ► Data structuring ► Data transformation and enrichment ► Data validation and publishing

ICONS FROM LEFT: PRIYANKA GUPTA/GETTY IMAGES, TIM_IURII/GETTY IMAGES, ENIS AKSOY/GETTY IMAGES, FINGERMEDIUM/GETTY IMAGES, ENOTMAKS/GETTY IMAGES, BROWNDOGSTUDIOS/GETTY IMAGES

WHAT ARE THE CHALLENGES OF DATA PREPARATION?

Data preparation is inherently complicated. Data sets pulled together from different source systems are likely to have numerous data quality, accuracy and consistency issues to resolve. The data also must be manipulated to make it usable, and irrelevant data needs to be weeded out.

As noted above, doing so is often a lengthy process: In the past, a common maxim was that data scientists spent about 80% of their time collecting and preparing data and only 20% analyzing it. That might not be the case now, partly due to the

In this guide:

increased availability of data preparation tools. But in the 2023 edition of an annual survey conducted by data science platform vendor Anaconda, 1,071 data science practitioners ranked data preparation and data cleansing as the two most time-consuming tasks in analytics applications.

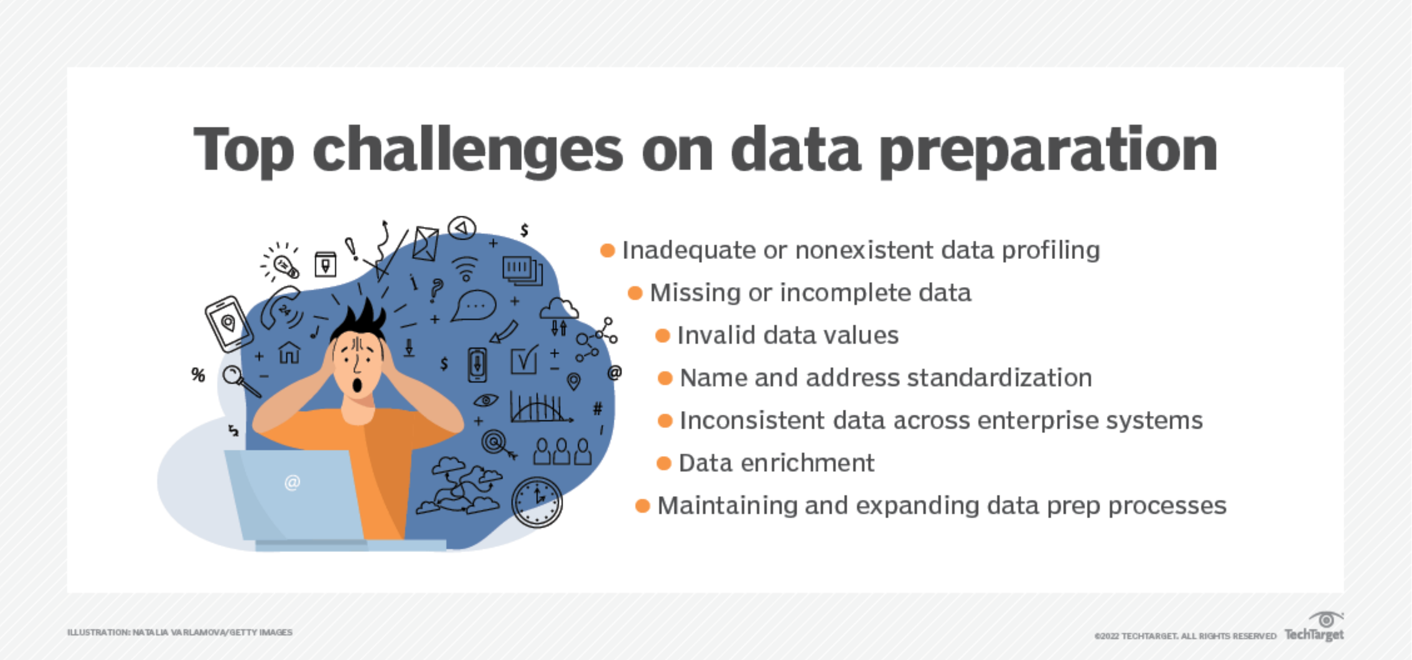
The following are seven common data preparation challenges faced by data scientists and others involved in the process:

- **Inadequate or nonexistent data profiling.** If data isn't properly profiled, errors, anomalies and other problems might not be identified, which can result in flawed analytics.
- **Missing or incomplete data.** Data sets often have missing values and other forms of incomplete data; such issues need to be assessed as possible errors and addressed if so.
- **Invalid data values.** Misspellings, other typos and wrong numbers are examples of invalid entries that frequently occur in data and must be fixed to ensure analytics accuracy.
- **Name and address standardization.** Names and addresses might be inconsistent in different systems, with variations that can affect views of customers and other entities if the data isn't standardized.
- **Inconsistent data across enterprise systems.** Other inconsistencies in data sets drawn from multiple source systems, such as different terminology and

In this guide:

unique identifiers, are also pervasive issues to contend with in data preparation efforts.

- **Data enrichment issues.** Deciding how to enrich a data set -- for example, what to add to it -- is a complex task that requires a strong understanding of business needs and analytics goals.
- **Maintaining and expanding data prep processes.** Data preparation work often becomes a recurring process that needs to be sustained and enhanced on an ongoing basis.



Top challenges on data preparation

- Inadequate or nonexistent data profiling
- Missing or incomplete data
- Invalid data values
- Name and address standardization
- Inconsistent data across enterprise systems
- Data enrichment
- Maintaining and expanding data prep processes

ILLUSTRATION: NATALIA VARLAMOVA/GETTY IMAGES

DATA PREPARATION TOOLS

Data preparation can pull skilled BI, analytics and data management practitioners away from more high-value work, especially as the volume of data used in analytics applications continues to grow. However, the self-service tools now offered by various software vendors automate data preparation methods. That enables both data professionals and business users to get data ready for analysis in a streamlined and interactive way.

The tools run data sets through a workflow that follows the steps of the data preparation process. They also feature GUIs designed to further simplify the required tasks and functions. In addition to speeding up data preparation and leaving more time for related analytics work, the self-service software might help organizations increase the number of BI and data science applications they're able to run, thus opening up new analytics scenarios.

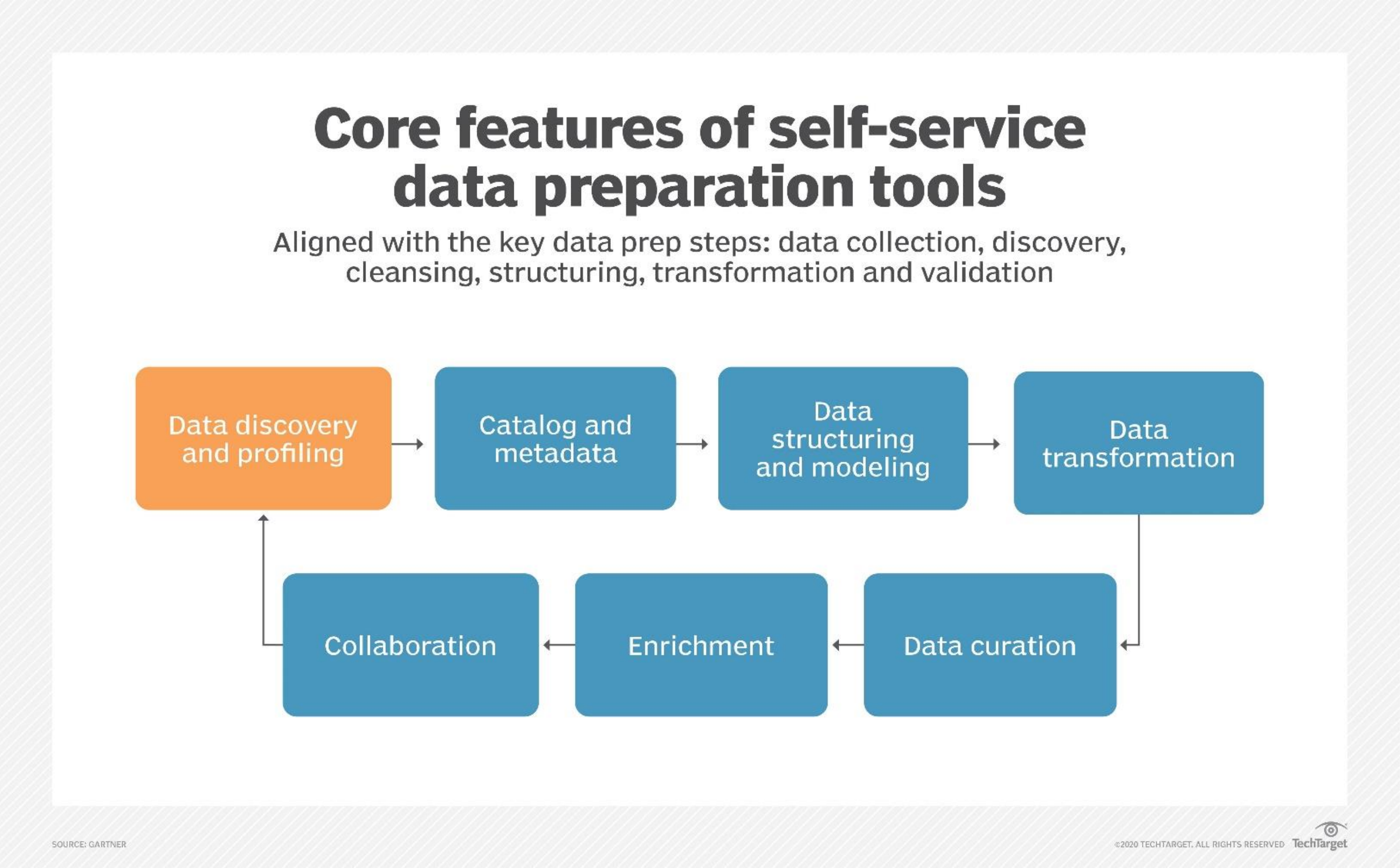
In 2023, consulting firm Gartner removed data preparation tools from its annual Hype Cycle report on emerging data management technologies, saying they had reached full maturity and mainstream adoption. Initially sold as separate products by several vendors that focused on data preparation, the tools have now largely been incorporated into broader data management software suites and BI or data science platforms. For example, Gartner lists data preparation as a core capability of BI platforms.

The following list includes some prominent BI, analytics and data management vendors that offer data preparation tools or capabilities -- it's based on market reports from Gartner and Forrester Research plus additional research by TechTarget editors:

- Altair.
- Alteryx.
- AWS.
- Boomi.
- Datameer.
- IBM.
- Informatica.
- Microsoft.
- Software.
- Oracle.
- Precisely.
- Qlik.
- SAP.
- SAS.
- Tableau.
- Tamr.
- Tibco

One caveat from data management consultants and practitioners: Don't look at self-service data preparation software as a replacement for traditional data integration technologies, particularly extract, transform and load (ETL) tools. While data prep tools enable users to integrate relevant data sets for analytics applications, ETL ones provide heavy-duty capabilities for integrating large amounts of data, transforming it and moving it into a data store. The two technologies often complement one another: A data management team might use ETL software to combine and initially prepare data sets, then data scientists or business analysts can use a self-service tool to do more specific data preparation work.

In this guide:



Core features of self-service data preparation tools

Aligned with the key data prep steps: data collection, discovery, cleansing, structuring, transformation and validation

SOURCE: GARTNER

©2020 TECHTARGET. ALL RIGHTS RESERVED TechTarget

DATA PREPARATION TRENDS

While effective data preparation is crucial in machine learning applications, AI and machine learning algorithms are also increasingly being used to help prepare data. For example, tools with augmented data preparation capabilities based on AI and ML can automatically profile data, fix errors and recommend other data cleansing, transformation and enrichment measures. In its 2023 Hype Cycle report, Gartner said organizations should make such augmented features a must-have item when buying new data management tools.

Automated data prep features are also included in the augmented analytics technologies now offered by many BI vendors. The automation is particularly helpful for self-service BI users and citizen data scientists -- business analysts and other workers who don't have formal data science training but do some advanced analytics work. But it also speeds up data preparation by skilled data scientists and data engineers.

In addition, generative AI (GenAI) tools are starting to be incorporated into data management processes, including data preparation. For example, GenAI offers the potential for conversational interfaces that enable data prep tasks to be performed using natural language. It could also be used to write integration scripts, fix data errors and create data quality rules as part of data preparation work. Conversely, the deployment of GenAI applications further increases data prep workloads in organizations.

There's also a growing focus on cloud-based data preparation, as vendors now commonly offer cloud services for preparing data. Another ongoing trend involves integrating data preparation capabilities into DataOps processes that aim to streamline the creation of data pipelines for BI and analytics.

HOW TO GET STARTED ON DATA PREPARATION

Donald Farmer, principal at consultancy TreeHive Strategy, listed the following six data preparation best practices to adopt as starting points for a successful initiative:

1. **Think of data preparation as part of data analysis.** Data preparation and analysis are "two sides of the same coin," according to Farmer. That means data can't be properly prepared without knowing what analytics use it needs to fit.
2. **Define what data preparation success means.** Desired data accuracy levels and other data quality metrics should be set as goals and then balanced against projected costs to create a data prep plan that's appropriate to each use case.
3. **Prioritize data sources based on the application.** Resolving differences in data from multiple source systems is an important element of data preparation that also should be based on the planned analytics use case.
4. **Use the right tools for the job and your skill level.** Self-service data preparation tools aren't the only option available -- other tools and technologies can also be used, depending on an organization's internal skills and data needs.

5. **Be prepared for failures when preparing data.** Error-handling capabilities need to be built into the data preparation process to prevent it from going awry or getting bogged down when problems occur.
6. **Keep an eye on data preparation costs.** The cost of software licenses, data processing and storage resources, and the people involved in preparing data should be watched closely to ensure that expenses don't get out of hand.

Craig Stedman is an industry editor who creates in-depth packages of content on analytics, data management, cybersecurity and other technology areas for TechTarget Editorial.

Ed Burns, a former executive editor at TechTarget, and freelance journalist Mary K. Pratt contributed to this article.

▼ **CONTINUED READING**

Data preparation best practices for analytics applications

Top data preparation challenges and how to overcome them

Data preparation in machine learning: Key steps