



# **What is Data Preparation? An In-Depth Guide to Data Prep**

## In this guide:

[What is data preparation?](#)  
[An in-depth guide to data prep](#)

[Purposes of data preparation](#)

[What are the benefits of data preparation?](#)

[Steps in the data preparation process](#)

[What are the challenges of data preparation?](#)

[Data preparation tools and the self-service data prep market](#)

[Data preparation trends](#)

[How to get started on data preparation](#)

This 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 and vendors, best practices and common challenges faced in preparing data. Throughout the guide, there are hyperlinks to related articles that cover the topics in more depth.

## In this guide:

[What is data preparation?](#)  
[An in-depth guide to data prep](#)

[Purposes of data preparation](#)

[What are the benefits of data preparation?](#)

[Steps in the data preparation process](#)

[What are the challenges of data preparation?](#)

[Data preparation tools and the self-service data prep market](#)

[Data preparation trends](#)

[How to get started on data preparation](#)

# What is data preparation? An in-depth guide to data prep

CRAIG STEDMAN, INDUSTRY EDITOR | ED BURNS, EXECUTIVE EDITOR | MARY K. PRATT

Data preparation is the process of gathering, combining, structuring and organizing data so it can be used in business intelligence ([BI](#)), analytics and [data visualization](#) applications. The components of data preparation include data preprocessing, profiling, cleansing, validation and transformation; it often also involves pulling together data from different internal systems and external sources.

Data preparation work is done by information technology (IT), BI and data management teams as they integrate data sets to load into a data warehouse, NoSQL database or data lake repository, and then when new analytics applications are developed with those data sets. In addition, data scientists, data engineers, other 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*. It's also known as *data wrangling*, although some practitioners use that term in a narrower sense to refer to cleansing, structuring and transforming data; that usage distinguishes data wrangling from the [data preprocessing](#) stage.

## In this guide:

[What is data preparation?](#)  
[An in-depth guide to data prep](#)

[Purposes of data preparation](#)

[What are the benefits of data preparation?](#)

[Steps in the data preparation process](#)

[What are the challenges of data preparation?](#)

[Data preparation tools and the self-service data prep market](#)

[Data preparation trends](#)

[How to get started on data preparation](#)

This 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 and vendors, best practices and common challenges faced in preparing data. Throughout the guide, there are hyperlinks to related articles that cover the topics in more depth.

## PURPOSES OF DATA PREPARATION

One of the primary purposes of data preparation is to ensure that raw data being readied for processing and analysis is accurate and consistent so the results of [BI and analytics applications](#) will be valid. Data is commonly created with missing values, inaccuracies or other errors, and separate data sets often have different formats that need to be reconciled when they're combined. Correcting data errors, validating data quality and consolidating data sets are big parts of data preparation projects.

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 this guide:

[What is data preparation?](#)  
[An in-depth guide to data prep](#)

[Purposes of data preparation](#)

[What are the benefits of data preparation?](#)

[Steps in the data preparation process](#)

[What are the challenges of data preparation?](#)

[Data preparation tools and the self-service data prep market](#)

[Data preparation trends](#)

[How to get started on data preparation](#)

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 most of their time gathering, cleansing and structuring data instead of analyzing it. 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:

- 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; and
- get a higher ROI from BI and analytics initiatives.

## In this guide:

[What is data preparation?](#)  
[An in-depth guide to data prep](#)

[Purposes of data preparation](#)

[What are the benefits of data preparation?](#)

[Steps in the data preparation process](#)

[What are the challenges of data preparation?](#)

[Data preparation tools and the self-service data prep market](#)

[Data preparation trends](#)

[How to get started on data preparation](#)

Effective data preparation is particularly beneficial in [big data](#) environments that store a combination of structured, semistructured and unstructured data, often in raw form until it's needed for specific analytics uses. Those uses include predictive analytics, machine learning (ML) and other forms of advanced analytics that typically involve large amounts of data to prepare. For example, in an article on [preparing data for machine learning](#), Felix Wick, corporate vice president of data science at supply chain software vendor Blue Yonder, is quoted as saying that data preparation "is at the heart of ML."

## 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 this step, data scientists, members of the BI team, other data professionals and end users who [collect data](#) should confirm that it's a good fit for the objectives of the planned analytics applications.
2. **Data discovery and profiling.** The next step is to explore 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 patterns, relationships and other attributes in the

## In this guide:

[What is data preparation?](#)  
[An in-depth guide to data prep](#)

[Purposes of data preparation](#)

[What are the benefits of data preparation?](#)

[Steps in the data preparation process](#)

[What are the challenges of data preparation?](#)

[Data preparation tools and the self-service data prep market](#)

[Data preparation trends](#)

[How to get started on data preparation](#)

data, as well as inconsistencies, anomalies, missing values and other issues so they can be addressed.

3. **Data cleansing.** Next, the identified data errors and issues are corrected to create complete and accurate data sets. For example, as part of [cleansing data sets](#), 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 the analytics requirements. For example, data stored in comma-separated values (CSV) files or other file formats has to 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](#) may 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 routines are run against the data to [validate its consistency, completeness and accuracy](#). The prepared data is then stored in a data warehouse, a data lake or another repository and either used directly 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 and oversees ready-to-use data sets for BI and analytics. 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

## In this guide:

[What is data preparation?](#)  
[An in-depth guide to data prep](#)

[Purposes of data preparation](#)

[What are the benefits of data preparation?](#)

[Steps in the data preparation process](#)

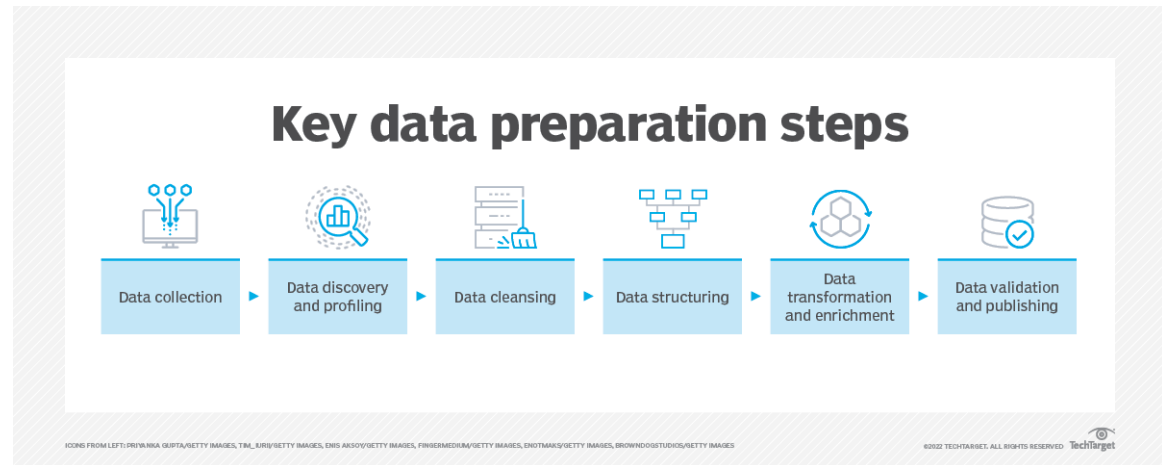
[What are the challenges of data preparation?](#)

[Data preparation tools and the self-service data prep market](#)

[Data preparation trends](#)

[How to get started on data preparation](#)

others, data may be curated by data stewards, data engineers, database administrators or data scientists and business users themselves.



## WHAT ARE THE CHALLENGES OF DATA PREPARATION?

Data preparation is inherently complicated. Data sets pulled together from different source systems are highly 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, it's a time-consuming process: The 80/20 rule [is often applied](#) to analytics applications, with about 80% of the work said to be devoted to collecting and preparing data and only 20% to analyzing it.



## In this guide:

[What is data preparation?](#)  
[An in-depth guide to data prep](#)

[Purposes of data preparation](#)

[What are the benefits of data preparation?](#)

[Steps in the data preparation process](#)

[What are the challenges of data preparation?](#)

[Data preparation tools and the self-service data prep market](#)

[Data preparation trends](#)

[How to get started on data preparation](#)

In an article on [common data preparation challenges](#), Rick Sherman, managing partner of consulting firm Athena IT Solutions, detailed the following seven challenges along with advice on how to overcome each of them:

- **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 may be inconsistent in data from different systems, with variations that can affect views of customers and other entities.
- **Inconsistent data across enterprise systems.** Other inconsistencies in data sets drawn from multiple source systems, such as different terminology and unique identifiers, are also a pervasive issue in data preparation efforts.
- **Data enrichment.** 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.

## In this guide:

[What is data preparation?](#)  
[An in-depth guide to data prep](#)

[Purposes of data preparation](#)

[What are the benefits of data preparation?](#)

[Steps in the data preparation process](#)

[What are the challenges of data preparation?](#)

[Data preparation tools and the self-service data prep market](#)

[Data preparation trends](#)

[How to get started on data preparation](#)



## DATA PREPARATION TOOLS AND THE SELF-SERVICE DATA PREP MARKET

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, various software vendors have introduced self-service tools that [automate data preparation methods](#), enabling both data professionals and business users to get data ready for analysis in a streamlined and interactive way.

The self-service data preparation tools run data sets through a workflow to apply the operations and functions outlined in the previous section. They also feature graphical user

## In this guide:

[What is data preparation?](#)  
[An in-depth guide to data prep](#)

[Purposes of data preparation](#)

[What are the benefits of data preparation?](#)

[Steps in the data preparation process](#)

[What are the challenges of data preparation?](#)

[Data preparation tools and the self-service data prep market](#)

[Data preparation trends](#)

[How to get started on data preparation](#)

interfaces (GUIs) designed to further simplify those steps. As Donald Farmer, principal at consultancy TreeHive Strategy, wrote in an article on self-service data preparation ([linked to above](#)), people outside of IT can use the self-service software "to do the work of sourcing data, shaping it and cleaning it up, frequently from simple-to-use desktop or cloud applications."

In a July 2021 report on emerging data management technologies, consulting firm Gartner gave data preparation tools a "High" rating on benefits for users but said they're still in the "early mainstream" stage of maturity. On the plus side, the tools can reduce the time it takes to start analyzing data and help drive increased data sharing, user collaboration and data science experimentation, Gartner said.

But, it added, some tools lack the ability to scale from individual self-service projects to enterprise-level ones or to exchange metadata with other data management technologies, such as data quality software. Gartner recommended that organizations evaluate products partly on those features. It also cautioned against looking at data preparation software as a replacement for traditional [data integration](#) technologies, particularly extract, transform and load (ETL) tools.

Several vendors that focused on self-service data preparation have now been acquired by other companies; Trifacta, the last of the best-known data prep specialists, agreed to be

## In this guide:

[What is data preparation?](#)  
[An in-depth guide to data prep](#)

[Purposes of data preparation](#)

[What are the benefits of data preparation?](#)

[Steps in the data preparation process](#)

[What are the challenges of data preparation?](#)

[Data preparation tools and the self-service data prep market](#)

[Data preparation trends](#)

[How to get started on data preparation](#)

bought by analytics and data management software provider Alteryx in early 2022. Alteryx itself already supports data preparation in its software platform. Other prominent BI, analytics and data management vendors that offer data preparation tools or capabilities include the following:

- Altair
- Boomi
- Datameer
- DataRobot
- IBM
- Informatica
- Microsoft
- Precisely
- SAP
- SAS
- Tableau
- Talend
- Tamr
- Tibco Software

## In this guide:

[What is data preparation?](#)  
[An in-depth guide to data prep](#)

[Purposes of data preparation](#)

[What are the benefits of data preparation?](#)

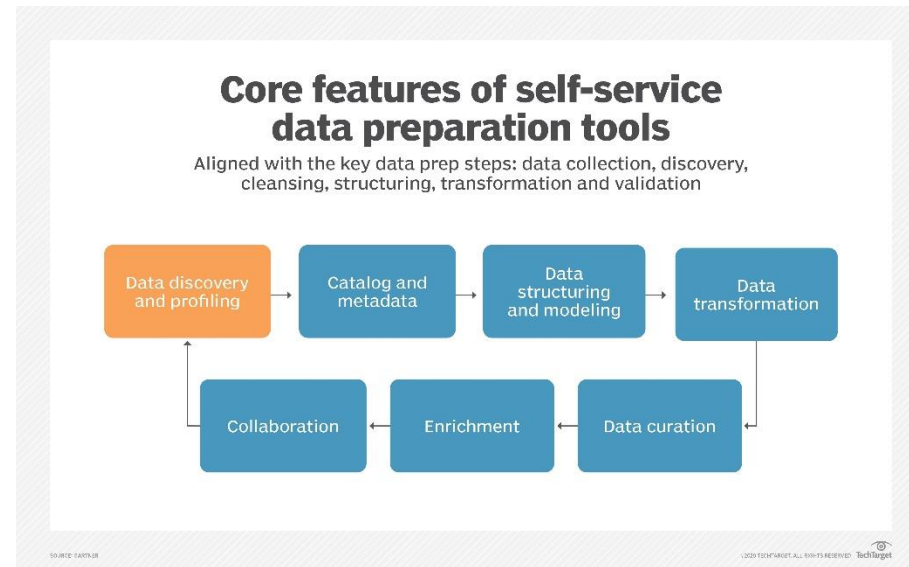
[Steps in the data preparation process](#)

[What are the challenges of data preparation?](#)

[Data preparation tools and the self-service data prep market](#)

[Data preparation trends](#)

[How to get started on data preparation](#)



## DATA PREPARATION TRENDS

While effective data preparation is crucial in machine learning applications, machine learning algorithms are also increasingly being used to help prepare data. Gartner said in its July 2021 report that automating data preparation work "is frequently cited as one of the major investment areas for data and analytics teams," and that data prep tools with embedded algorithms can automate various tasks.

For example, tools with [augmented data preparation capabilities](#) can automatically profile data, fix errors and recommend other data cleansing, transformation and enrichment

## In this guide:

[What is data preparation?](#)  
[An in-depth guide to data prep](#)

[Purposes of data preparation](#)

[What are the benefits of data preparation?](#)

[Steps in the data preparation process](#)

[What are the challenges of data preparation?](#)

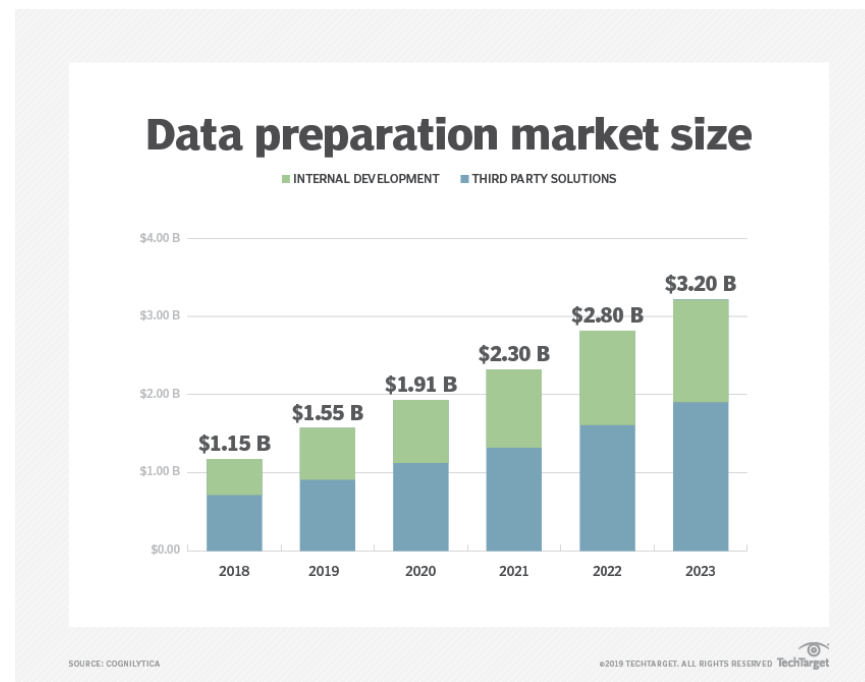
[Data preparation tools and the self-service data prep market](#)

[Data preparation trends](#)

[How to get started on data preparation](#)

measures. 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.

There's also a growing focus on cloud-based data preparation, as more vendors 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.



## In this guide:

[What is data preparation?](#)  
[An in-depth guide to data prep](#)

[Purposes of data preparation](#)

[What are the benefits of data preparation?](#)

[Steps in the data preparation process](#)

[What are the challenges of data preparation?](#)

[Data preparation tools and the self-service data prep market](#)

[Data preparation trends](#)

[How to get started on data preparation](#)

## HOW TO GET STARTED ON DATA PREPARATION

In an article on [data preparation best practices to adopt](#), Donald Farmer of TreeHive Strategy listed the following six items as starting points for successful data prep initiatives:

1. **Think of data preparation as part of data analysis.** Data preparation and analysis are "two sides of the same coin," Farmer wrote. Data, he said, 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, 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 your 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, processing and storage resources, and the people involved in preparing data should be watched closely to ensure that they don't get out of hand.

