# 8 simple building blocks for data preparation

An introductory guide to understand how ML can accelerate data preparation to achieve business-ready data

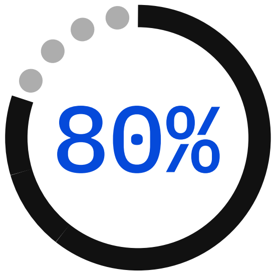
## Introduction

Data preparation accounts for about 80 percent[1] of the work of data citizens. This means time is spent cleaning, organizing and collecting data sets rather than mining curated datasets for analytics and building predictive models.

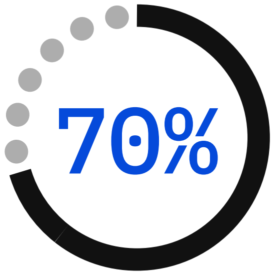
It’s expected that by 2022, automated data preparation will be utilized in more than 70 percent of new data integration projects for analytics and data science.[2]

Modernizing your data preparation can help in reducing the time to insight and data delivery. You can increase your data users productivity by spending more time on delivering insights and creating predictive models with an automated machine-learning (ML) based platform.

These trends are driving the adoption of automated data preparation to scale from self-service models to expanded analytics capabilities.



Data preparation accounts for about 80% of data citizens time.



Automated data preparation will be utilized in 70% of new data integration projects by 2022.

## What is data preparation?

Data preparation is a self-service activity to convert disparate, raw, messy data into a clean and consistent view of your data.

By automating and operationalizing your data preparation with tools like IBM InfoSphere® Advanced Data Preparation, you can create efficiencies of 10x and greater.

# 8 building blocks:

**1. Discovering**

Interactive exploration helps you discover features of your data and quickly determine the value of your data set.

IBM's data type inference, column-level profiles, interactive quality bars and histograms provide immediate visibility into trends and data issues, guiding the transformation process to supply accurate data for ML model development and testing.

**2. Structuring**

Structuring refers to actions that change the form or schema of your data. Splitting columns, un-nesting hierarchies, pivoting rows and deleting fields are all forms of structuring. Structuring needs to happen to provide well-formed tabular datasets to ML models.

Predictive transformation allows you to simply highlight sections of your data to get suggestions of the appropriate transforms based on the data you're working with and the type of interaction you applied to the data.

**3. Cleaning**

During the cleaning stage, users identify data quality issues, such as missing or mismatched values, and apply the appropriate transformation to correct, filter, or delete these values from the data set.

A guided cleaning process is critical to provide accurate data to ML models and achieve the best predictions.

## What is data preparation?

Data preparation is a self-service activity to convert disparate, raw, messy data into a clean and consistent view of your data.

By automating and operationalizing your data preparation with tools like IBM InfoSphere® Advanced Data Preparation, you can create efficiencies of 10x and greater.

# 8 building blocks:

### 4. Enriching

The data required to build, tune, and test ML models can often be spread across multiple data sources. In order to gather all the necessary insights, you need to enrich your various data sets by standardizing, combining, and aggregating multiple data sources.

IBM's data enrichment features allow you to easily execute lookups to data dictionaries or execute joins and unions with disparate data sets. Intelligent join and union inference uses ML to rapidly identify appropriate keys to combine your diverse data sets.

### 5. Build

IBM's ML platform features a massively parallel modeling engine that can scale to hundreds or even thousands of powerful servers to build ML models in one click.

Leveraging the clean data input generated, the intuitive web-based interface makes it easy for you to let automated ML models do all the work. Or you can write your own predictive models for evaluation by the platform.

### 6. Validate

The ML platform automatically searches through millions of combinations of algorithms, data preprocessing steps, transformations, features, and tuning parameters for the best ML model for your data.

Integrated ML speeds up model evaluation and builds a leaderboard so you can see which models perform best with your data. In addition, it also provides the tools you need to explore and compare individual models.

## What is data preparation?

Data preparation is a self-service activity to convert disparate, raw, messy data into a clean and consistent view of your data.

By automating and operationalizing your data preparation with tools like IBM InfoSphere® Advanced Data Preparation, you can create efficiencies of 10x and greater.

# 8 building blocks:

**7. Tune**

While automation and speed usually come at the expense of quality, IBM uniquely delivers on all those fronts. ML automates model tuning, but supports manual tuning so you can tune and adjust machine learning algorithms for even better results.

Each model is unique — fine-tuned for the specific data set and prediction target.

**8. Deploy**

The best predictive models have little to no organizational value unless they are rapidly operationalized within the business. With IBM's automated ML platform, deploying models for predictions can be done with a few mouse-clicks. The built-in ML model publishes a REST API endpoint, making it a breeze to integrate within modern enterprise applications.

You can now derive business value from ML in minutes, instead of waiting months to write scoring code and deal with the underlying infrastructure.

# IBM InfoSphere Advanced Data Preparation

IBM InfoSphere Advanced Data Preparation (ADP) provides self-service access, transformation and automated cleaning of diverse data, while maintaining IT compliance standards.

## Three benefits of IBM InfoSphere Advanced Data Preparation

### Establish self-service

By making data preparation an intuitive, visual process instead of a coding exercise, it increases productivity by spending more time on delivering insights. ADP enables self-service through helping build predictive models within an automated ML based platform.

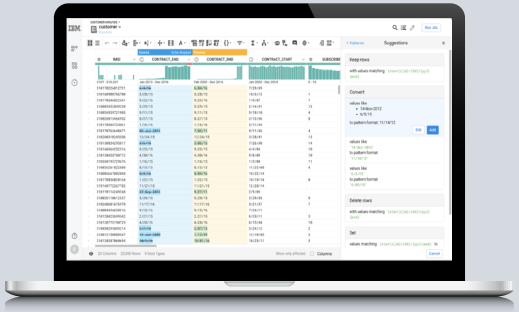
### Maximize investments in BI and analytics

Automated data preparation operationalizes data transformation across the entire organization. It enables your data citizens to prepare any type of data so it can be incorporated into the organizations BI and analytics efforts. With ADP your organization can build once and then reuse across many use cases.

### Accelerate ROI

On average, ADP users reduce the time they spend combining and cleaning disparate data for modeling by 70%[3] compared to their existing approach (such as Excel, SQL or coding). Using data visualization and machine learning helps surface data quality issues and reduce data preparation time.

**Ready to modernize your data preparation?**
Read the blog to learn more.



**Read the IBM InfoSphere Advanced Data Preparation solution brief**

Read solution brief

IBM Corporation
New Orchard Road
Armonk, NY 10504

Produced in the United States of America
June 2019

1. Cleaning Big Data: Most Time-Consuming, Least Enjoyable Data Science Task, Survey Says
2. Market Guide for Data Preparation Tools, Gartner - April 17, 2019
3. Trifacta: Data Wrangling - Prepare Raw and Diverse Data Faster