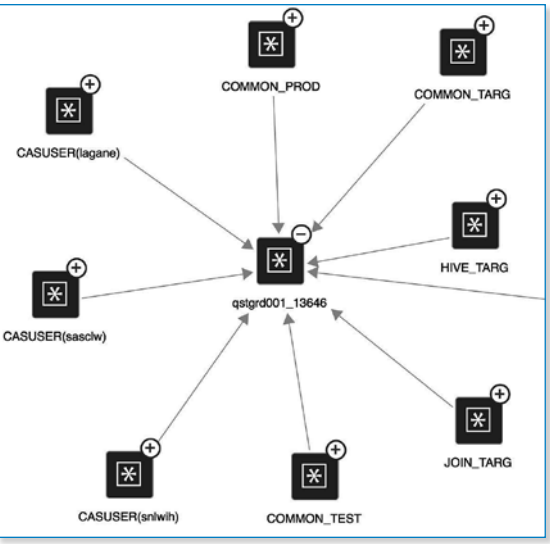
SAS® Data Preparation

Empower analysts to quickly prepare data for analytics in a self-service, point-and-click environment



What does SAS® Data Preparation do?

SAS Data Preparation provides an interactive, self-service environment for users who need to access, blend, shape and cleanse data to prepare it for reporting or analytics.

Why is SAS® Data Preparation important?

SAS Data Preparation saves time on the preliminary tasks done to prepare data for reporting and analytics. Its intuitive interface provides point-and-click actions for critical functions – no coding or SQL skills required. With simplified data preparation tasks seamlessly defined as part of the activities involved in analytics processing, users can spend more time analyzing data and less time preparing it.

For whom is SAS® Data Preparation designed?

It's designed for business analysts, citizen data scientists and other nontechnical users. Data scientists and IT can use the same interface to prepare reusable plans for business analysts.

To answer time-sensitive business questions, organizations need fast access to consistent, trusted data that they can use for analytics. Without it, they may not be able to respond quickly enough to market and customer requirements. But most organizations have massive volumes of data spread across silos. This raw data often contains errors or is duplicated, outdated or lacks identifiers needed to merge sources. Preparing it for analytics can consume up to 80 percent of an analyst's time.

It's a frustrating issue for business and IT. Nontechnical users lack the skills to move and transform data to make it ready for analytics. Alternatives require extensive coding, SQL or scripting knowledge, and training in data engineering for extract, transform, load (ETL) tools. In most cases, IT has to provision data for business users when they could have focused on more strategic activities. And business users have to wait in line for IT to create their data sources before they can get data in the right form for analytics.

Many organizations want to give business users direct access to data to free IT from never-ending custom data requests and improve everyone's productivity. Through its self-service tools, SAS Data Preparation empowers business users to take vetted data from IT and customize it for any report or analysis they need.

Built on SAS® Viya®, the intuitive, visual interface of SAS Data Preparation[1] makes it easy for business users to quickly prepare data without coding or help from IT. The software runs in a fast, in-memory distributed environment. This frees IT from the mundane task of provisioning data, and business analysts and data scientists get relevant results that drive faster business insights. The interface automatically generates code that can be scheduled to ensure currency with source system refreshes. Templates can be defined and reused, promoting sharing and collaboration.

[1] SAS Visual Analytics (sold separately) is a required product for SAS Data Preparation.

Key Benefits

Boost productivity through self-service data preparation. No specialized skills or coding are required to access, merge and shape data, and data preparation tasks are defined within the same visual experience – automatically integrated with downstream analytics and reporting tasks.

Gain efficiency through reusability, collaboration. Automatically generated code and defined transformations can be shared with IT and scheduled to run with each source code update. Data preparation tasks can be saved in projects, then shared and reused by others.

Empower analytics users with fast results. Prebuilt transformations and data cleansing functions assist users as they explore data, refine it and explore some more. And with in-memory distributed processing and parallel I/O, responses can be delivered in near-real time.

Reduce total cost of ownership. Make the most of your existing resources by giving them a visual, interactive interface that guides them through routine reporting and analytics data preparation tasks, with software that requires very little training.

Product Overview

With the volume and variety of data available today, business analysts need to curate data to answer specific questions. This requires different views of the data, which often needs to be examined in different ways, multiple times a day. Even when IT has prepared and cleansed the data for them, analysts still need to iteratively examine and prepare it further for their particular needs.

SAS Data Preparation provides the type of ad hoc environment today's analytics professionals crave. With its simple, interactive user interface designed for self-service data preparation, nontechnical users have flexibility to integrate data from virtually any source they need, cleanse it and prepare it for analysis quickly and easily. Data can be loaded in memory so multiple users will share the same view simultaneously. Users' data preparation tasks are fully integrated with downstream reporting and analytics processing – all from the same intuitive interface. Market-validated data integration and data quality capabilities are prebuilt for quick data vetting and correction. And the seamless, consistent user experience extends across the entire analytics life cycle.

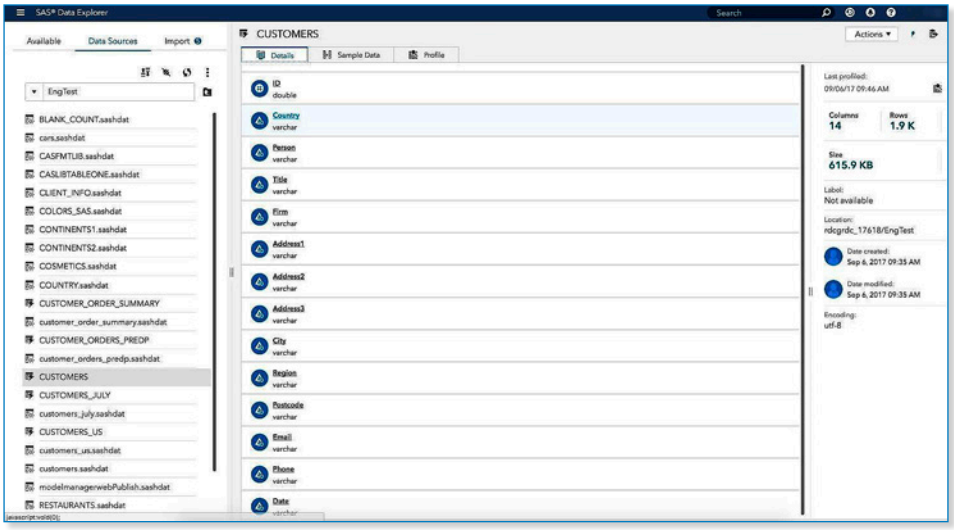


Figure 1. Explore data accessed from multiple sources.

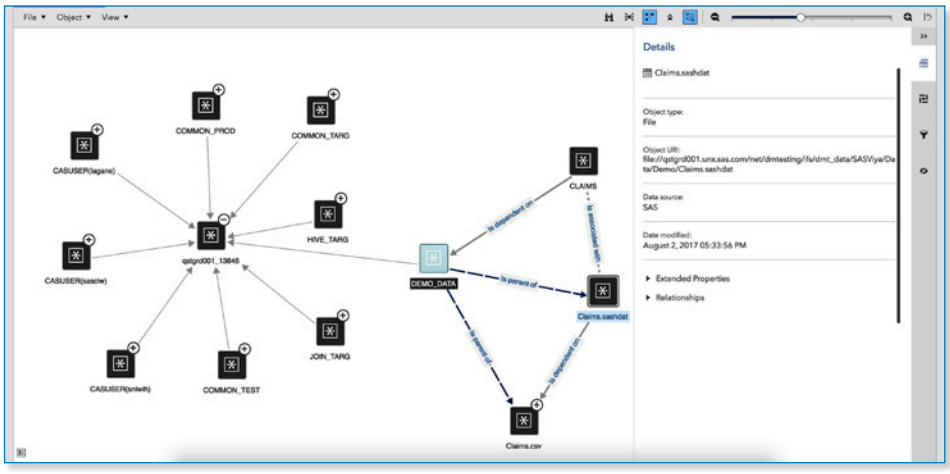


Figure 2. Object lineage shows the relationships between different objects.

Easy-to-use capabilities

With SAS Data Preparation, it's easy to access, integrate, browse and cleanse data. Visually explore external data sources, and big data stores like Hadoop and data in SAS Viya. Create connections to external data sources on the fly – curate what you need, when you need it. And get fast insight into the data by profiling physical metadata information – column names, data types, encoding, column and row counts.

You can access data from flat files, relational data sources, social media sources, SAS data sets, Apache Hadoop, Teradata, CSV files, text files and other sources. Technical users who prefer to code can access the SAS Data Quality routines from SAS code or from third-party coding languages, like Python.

Speed and scalability

High-performance, high-quality data fuels high-performing results. With SAS Data Preparation, users can interactively blend and shape data in near-real time, without having to wait on batch processes. Data preparation functions can be loaded in parallel and processed in memory. For some sources, processing can be pushed to run where the data resides – speeding execution of SAS code, minimizing data movement and delivering rapid responses.

Visual interface for self-service data preparation

Business analysts and data scientists can use the wizard-based interface to access, integrate, view, filter, join, transform, cleanse and query data. Each transformation is designed to guide users through the data orchestration process so they can easily understand the impact of how any single data preparation task affected results.

Variety of prebuilt transformations

Several types of prebuilt transformations are included in SAS Data Preparation – column-based, row-based, code-based and multiple-input-based transformations. These prebuilt transformations assist with filtering, blending, shaping, remediating and standardizing data.

Built-in data quality

Out of the box, SAS Data Preparation includes SAS Data Quality functions to help create analytics-ready data. Functions include profiling, casing, standardizing, parsing, identification analysis and more. Users can generate column-based and table-based basic and advanced profile metrics to uncover data quality issues and get insights into the data itself. Data quality and other data preparation tasks are accessible from coding interfaces other than SAS, including Python.[2]

Data governance and lineage

SAS Data Preparation lets users explore the relationships between data sources, data objects and actions taken on the data – so it's easy to trace pipeline activity.

Collaboration, reuse and automation

With SAS Data Preparation, users can prepare data for their specific analysis, then save and share transformations so they can be reused later. Templates can be defined from a point-and-click interface – or from a coding environment – defining best practices for others to use. Template code can also be scheduled as part of IT processing to keep prepared data current with refreshes.

Key Features

Data and metadata access

- Use any authorized internal source, accessible external data sources and data held in memory in SAS Viya.
  - View a sample of a table or file loaded in the in-memory engine of SAS Viya, or from data sources registered with SAS/ACCESS®, to see data you want to work with.
  - Quickly create connections to and between external data sources.
  - Access physical metadata information like column names, data types, encoding, column count and row count to gain further insight into the data.
- Data sources and types include:
  - Access to more than 20 data sources and types, including relational databases, social sources, etc.

Data provisioning

- Parallel load data from supported data sources into memory simply by selecting them – no need to write code or have experience with an ETL tool.*[3]
- Reduce the amount of data being copied by performing row filtering or column filtering before the data is provisioned.

Guided, interactive data preparation

- Transform, blend, shape, cleanse and standardize data in an interactive, visual environment.
- Easily understand how a transformation affected results, getting visual feedback in near-real time through the distributed, in-memory processing of SAS Viya.
- Quickly extract document content and perform text identification and extraction using batch text analysis.

Column-based transformations

- Save data plans for quick data preparation jobs (through support for wide tables).
- Use column-based transformations to standardize, remediate and shape data without configuring:
  - Change case, convert column, rename, remove, split, trim whitespace, custom calculations.

Row-based transformations

- Use row-based transformations to filter and shape data.
- Create analytical-based tables using the transpose transformation to prepare the data for analytics and reporting tasks.
- Create simple or complex filters to remove unnecessary data.

Code-based transformations

- Write custom code to transform, shape, blend, remediate and standardize data.
- Write simple expressions to create calculated columns, write advanced code or reuse code snippets for greater transformational flexibility.
- Import custom code defined by others, sharing best practices and collaborative productivity.

Multiple-input-based transformations

- Use multiple-input-based transformations to blend and shape data.
- Blend or shape one or more sets of data together using the guided interface – there's no requirement to know SQL or SAS.

Data profiling

- Profile data to generate column-based and table-based basic and advanced profile metrics.
- Use the table-level profile metrics to uncover data quality issues and get further insight into the data itself.
- Drill into column-level profile metrics and see visual graphs of pattern distribution and frequency distribution results.
- Use a variety of data types/sources (listed previously) to profile data from Twitter, Facebook, Google Analytics or YouTube.

[2] Such third-party interfaces to SAS are available for download from GitHub.

3 Data cannot be sent back to the following data sources: Twitter, YouTube, Facebook, Google Analytics, Esri; it can only be sourced from these sites.

Key Features (continued)

Data quality processing[4]

Data cleansing

- Find like records and group together logically.
- Use locale- and context-specific parsing and field extraction definitions to reshape data and uncover additional insights.
- Use the extraction transformation to identify and extract information (e.g., name, gender, field, pattern, identify, email and phone number) in a specified column.
- Generate match codes on data that can be used to perform fuzzy matching.
- Standardize data with locale- and context-specific definitions to transform data into a common format, like casing.

Identity definition

- Create a unique identity for each row with the unique ID generator.
- Analyze column data using locale-specific rules to determine gender or context.
- Identify, find and sort data by tagging columns and tables.

- Use identification analysis to analyze the data and determine its context, and to identify the subject data in each column.
- Use gender analysis to determine the gender of a name using locale-specific rules.

System and job monitoring

- Use integrated monitoring capabilities for system- and job-level processes.
- Understand how many processes are running, how long they're taking and who is running them.
- Easily filter through all system jobs based on job status (running, successful, failed, pending and cancelled).
- Access job error logs to help with root-cause analysis and troubleshooting.

Data import and data preparation job scheduling

- Create a data import job from automatically generated code to perform a data refresh using the integrated scheduler.
- Schedule data explorer imports as jobs so they will become an automatic, repeatable process.
- Specify a time, date, frequency and/or interval for the jobs.

Data lineage

- Create multiple views with different tabs, and save the organization of those views.
- Explore relationships between accessible data sources, data objects and jobs.
- Use the relationship graph to visually show the relationships that exist between objects.

Plan templates and project collaboration

- Use data preparation plans (templates), with one or more sources of data, to improve productivity.
- Reuse the templates by applying them to different sets of data to ensure that data is transformed consistently to adhere to enterprise data standards and policies.
- Rely on team-based collaboration through a project hub used with SAS Viya projects.

Cloud data exchange

- Move data from on-premises locations to SAS Viya running in a private or public cloud.
- Preprocess data locally to reduce the amount of data that needs to be moved to remote locations.
- Use a command line input (CLI) interface for administration and control.
- Use cloud data exchange to securely and responsibly negotiate your on-site firewall.

TO LEARN MORE »

To learn more about SAS Data Preparation system requirements, download white papers, view screenshots and see other related material, please visit: sas.com/data-preparation.

[4] Supported data quality transformations rely on SAS Quality Knowledge Base Locales, a locale-specific library of data quality functions available in over 30 locales, included with SAS Data Preparation.