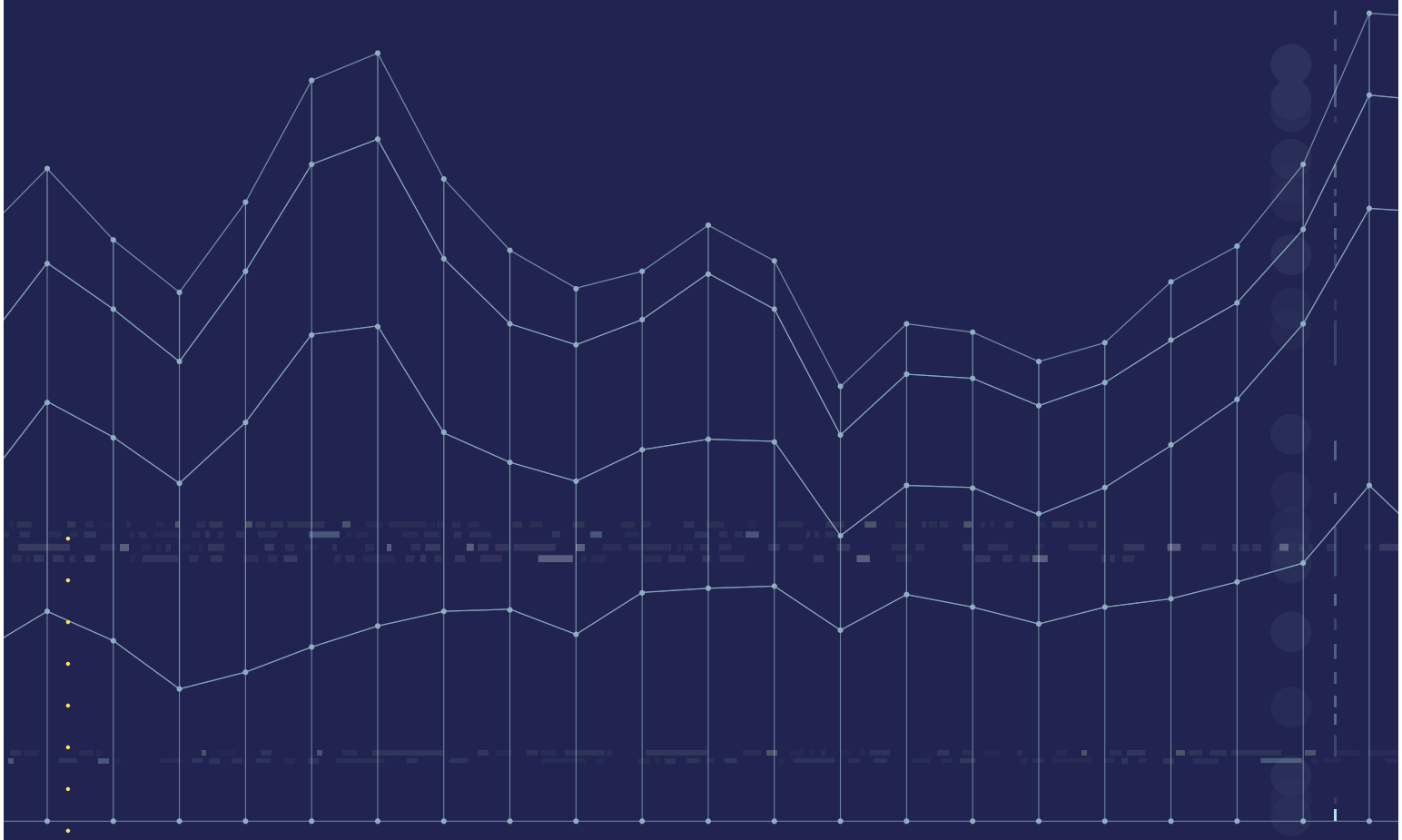
+ a b | e a u

Managing the data prep chaos in your organization

Introducing: Tableau Prep Conductor, for safe and scalable self-service data prep



This paper is intended for IT professionals and IT decision makers to more deeply understand the considerations for self-service data preparation at scale and how Tableau Prep Conductor can help manage a governed data prep environment. This paper assumes the reader has some familiarity with Tableau products, including Tableau Prep Builder, Tableau Server, Tableau Online, and Tableau Desktop.

Contents

Introduction: Why democratize data preparation?

The challenges of self-service data prep in an evolving landscape

Today, self-service analytics and data-driven decision-making are the norm in leading organizations worldwide. Data preparation has historically been an IT function. Few had the ability to prep and bring new data sources into the organization's centralized data warehouse. But the landscape has evolved: the volume and variety of data organizations collect has exploded; data is not always centralized, but often flows through pipelines where use cases determine its ingestion, storage, transformation, and distribution; and analytics platforms are increasingly available to broader audiences to inform decision-making.

Typically, teams and individuals outside of IT (or without access to formal data prep processes) must either wait for another team to prep their data or try to solve their own data problems. Often this means users extract data from systems and prepare their data in spreadsheets. The result is a newly structured data set that serves a singular purpose. Not only can this lead to an abundance of data silos, but departments often duplicate efforts without even knowing it. These individual solutions aren't efficient, scalable, or governed.

With self-service analysis the new norm for data-driven organizations, many people will do what prep they can with the available tools and capabilities to make the data work for them—whether it's cutting and pasting or writing huge calculations that aren't optimal for the server. Even analysts report that the majority of their job is not analysis, but actually cleaning and reshaping data—with an ETL process, in self-service data prep tools, or even in spreadsheet tools like Excel.

Prep like a pro: Understand the wide range of data prep needs

Factors like human error, disparate systems, and changing business requirements can contribute to dirty data, but data prep often necessitates more than simple cleaning steps. Users may need to adjust the granularity of the data or transform it to align and then union or join with other data. This means the data ready for analysis often looks very different from the original data source. Cleaning, shaping, and enriching data may involve these steps and more:

Pivot — Switch a field from columns to rows or vice-versa.

Join — Add more fields to your data source, expanding the number of fields that you can analyze.

Union — Append two data sets together, keeping the same table structure, but adding more rows.

Filter or remove — Exclude values or fields for analysis.

Assign data roles — Validate fields that represent an email, a URL, or geographic data.

Edit values — Manually change a value or use quick cleaning operations to change text case, remove letters, numbers, punctuation, spaces, and more.

Group and replace — Clean values that may vary by pronunciation, common characters, or spelling.

Split values — Split units of information in a field into multiple fields

Create a calculated field — Create a field for analysis using calculations with other values.

Aggregate data — Return a single value from multiple values—like sum, average, count, or minimum.

Learn more about dirty data and how to solve common data prep challenges

Why we should democratize data prep

More and more people are collaborating with data, breaking down data silos across the organization and discovering new insights that drive impact. At the same time, we are seeing more data that, in its state of collection, is not ready or meant for analysis. In many cases, there is a significant gap between the optimal states of the data for efficient capture vs. effective analysis—such as with transactional data or high-velocity streaming data. Whether it's the structure, format, or lack of business context, it needs cleaning, and sometimes curation—for example, healthcare data that requires business rules or claim types—before analysis.

Many organizations are adopting self-service data preparation solutions to explore and prototype new data sources and analytical use cases. Self-service data preparation tools put the power in hands of the people who know the data the best while reducing the burden on IT to prepare it. But self-service data prep is still a brand new skill set; it needs to be developed and rolled out to enable users to understand and perform prep functions effectively, establish repeatable processes, automate them for efficiency, and ultimately build trust and confidence in the data for wider use.

Is all the prep work worth it? According to a recent data prep study, there are many benefits of data preparation that have significantly exceeded companies' expectations, namely: gaining a single, complete view of relevant data across the organization; reducing analytical siloes; and improved data-driven decision-making.

Get the free BARC report, "Data Preparation — Refining Raw Data into Value"

Meet Tableau Prep Conductor

What Tableau did for visual analysis, we are now doing for data preparation. With Tableau Prep Builder, released in the spring of 2018, data is even more accessible with visual, smart, and direct data prep. Analysts and business users can prepare their own data for analysis, and it's fully integrated with Tableau Desktop to keep users in their analytical flow. Now with Tableau Prep Conductor, we're extending the data prep capabilities of the Tableau platform so you can automate data prep flows to run without manual updates as your data changes, and make your prepped data more discoverable by the organization.

Tableau Prep Conductor empowers you to schedule your flows to run in a centralized, scalable, and reliable server environment so your data is always up to date. It also gives administrators visibility into self-service data preparation across their organization. With Tableau Prep Conductor, you can manage, monitor, and secure flows using your Tableau server environment.

Tableau Prep Conductor is integrated to Tableau Server and Tableau Online, leveraging existing scheduling, tracking, and security functionality. Just like extract refreshes, scheduled flow tasks and on-demand flow runs are queued as background tasks. Seamlessly publish flows from Tableau Prep Builder to Tableau Server or Tableau Online using functionality that is similar to publishing data sources and workbooks with Tableau Desktop today.

Keep data fresh automatically
Schedule your flows to run when you need them. Automate the tasks of running flows and create a repeatable process, so there's consistency in the delivery of prepared data.

Stay informed with alerts and run history
See a historical view of your flows' run history, including successful or failed runs at a glance. Keep track of the quality of your prep flows with out-of-the-box alerts if a flow fails.

Create a governed prep environment
Build rules and permissions around data sharing and refreshes. Leverage existing permissions and infrastructure in Tableau Server or Tableau Online to control who can publish, view, and run flows.

Increase data discoverability
Use simple management capabilities, including keyword tagging, moving flows between projects, and setting user permissions, to help users across the organization find relevant, prepared data.

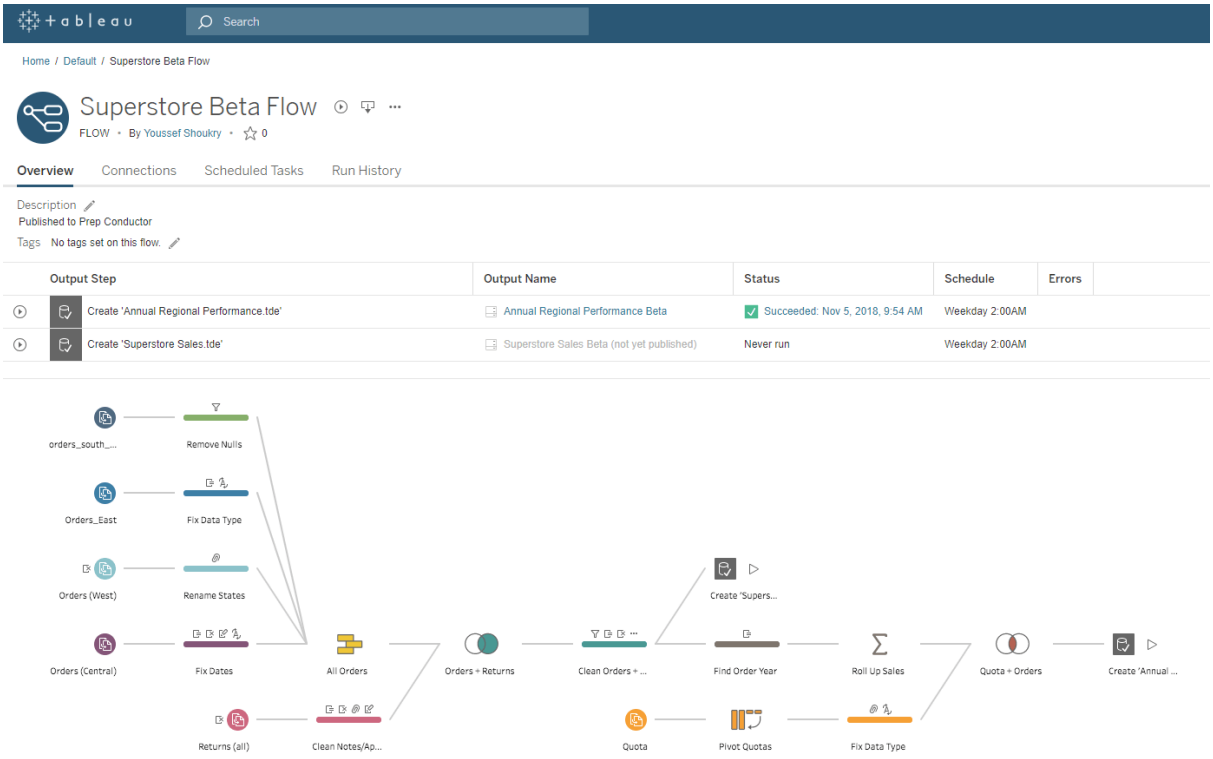


Figure 1 View data prep flows published to Tableau Server with Tableau Prep Conductor.

> Tableau Prep completely streamlined how we clean and prepare marketing data, which is often extraordinarily dirty. It has allowed us to transform and blend data in ways that enable new and different types of analysis. We are thrilled to try Tableau Prep Conductor and expect it will help us automate and scale data prep across the company. It will speed up our ability to report on campaigns for our clients, which will increase the value we're providing and ultimately help our bottom line.
>
> **ANDREW RICHARDSON**
> VP OF ANALYTICS AND MARKETING SCIENCE, ELITE SEM

Thinking about data prep in your organization

To grasp the full context of your data, it's important to not only understand who is using it, but who prepared it, where it came from, where it's been made available for analysis, and how end users will benefit from it. This understanding is foundational for scalable self-service data prep.

Who is doing prep? Who should be?

In many organizations, analysts prepare data for other roles to consume as often as they prepare data for their own analysis. Advanced data preparation tools can be complex, which means this capability is often restricted to a select number of power users. But even if analysts and business users don't have access to data preparation tools, it doesn't mean that they aren't already performing these tasks in other applications.

Self-service business intelligence tools have opened up data analysis capabilities to every level of user, but in order to get insights into their data, these users still need to rely on IT for well-structured data. As you consider self-service data prep at scale, it may be helpful to think of the roles that currently exist and which responsibilities need to be shared or added.

A data steward role may work with more traditional roles like a database administrator. The data steward is more responsible today in seeing that the data being distributed across the organization—to analysts, to business users, and more—is trusted, so users can get value out of the data. Database administrators and data engineers typically prioritize how data is stored and accessed. Columns may be added that are strictly for databases to leverage, not humans. When an engineer builds a data warehouse specifically for analysis, they prioritize the core business metrics that answer the majority of questions. If the information that data analysts need isn't already in the data set, they may need to adjust aggregations or bring in outside sources, which can risk inaccuracies in the data or lead to silos.

As you scale self-service data prep, think about who should be involved in establishing and enforcing agreed-upon governance practices, who is using which tools for prep, what training is required, and how you'll measure success as an organization.

Prep like a pro: Staff the critical data roles within your organization

IT/BI professional roles

The **Database Administrator (DBA)** is responsible for the administration, monitoring, maintenance, and security of databases in the organization. Coordinating with data engineers and data stewards, the DBA will provide data access and assist with modeling, structuring, and optimizing sources of data that connect to Tableau products.

The **Systems Administrator** installs, configures, manages, and maintains the hardware and operating system on which the Tableau Server is installed in the data center or the cloud, while enforcing company policies in compliance with business and technology strategy.

The **Client Administrator** configures client software, including installing database drivers and Tableau products, and enabling Tableau Prep Conductor within Tableau Server or Tableau Online.

Tableau Administrator roles

The **Server Administrator** has full access to Tableau Server settings, all sites on the server, users and groups, and all content assets, such as projects, data sources, and workbooks to monitor and maintain overall server health.

Tableau Site Administrators create and manage the site's users and groups, create projects to organize content on the site, and assign permissions to allow users (groups) to access the content. They also promote and certify content and measure the use of analytics within their site.

Content Creator roles

Data Stewards (with a Tableau Creator license) understand the business domain and the interaction of business processes with analytics. A data steward ensures that there are documented procedures and guidelines for data access and use and will work with the DBA and/or data engineers to plan and execute an enterprise-wide data governance and compliance policy. A data steward may publish prep flows and/or data sources.

Content Authors (with a Tableau Creator license) create and publish dashboards, prep flows, and/or data sources.

Where did the data come from? Where does it live now?

For many organizations today, there is a surprising lack of visibility into how data prep is done. Without a standard, governed approach, ad hoc prep work and analyses can lead to duplicated efforts, manual work without repeatable processes, and inconsistencies in data sources. A big factor in remediating these issues is understanding where the data comes from, and once it's cleaned, where it will be available—essentially, the links between the person who is prepping the data and the person using it for analysis.

- How is data secured, and who needs the right permissions to access and shape it?
- Which users are accessing raw data sources vs. cleaned data?
- Will users require combined data sources—or perhaps external data—to get to the root of questions or make for more robust analysis?
- How are prepped data sources shared with others for analysis?

It's not uncommon to export data into a CSV or other spreadsheet file for cleaning and ad hoc analysis. But this can present security concerns, as files may not be securely shared. For organizations using Tableau Server or Tableau Online as their repository for data sources and workbooks, Tableau Prep Conductor will easily provide visibility into the data prep process with published flows to Server or Online. This will not only centralize where users can find and access prep flows, but to see the integrity of flows and provide the opportunity to learn best practices for self-service data prep.

How do the results benefit your organization?

For many organizations today, there is a surprising lack of visibility into how data prep is done. Without a standard, governed approach, ad hoc prep work and analyses can lead to duplicated efforts, manual work without repeatable processes, and inconsistencies in data sources. A big factor in remediating these issues is understanding where the data comes from, and once it's cleaned, where it will be available—essentially, the links between the person who is prepping the data and the person using it for analysis.

- How are requirements for data sources and reports being gathered?
- What types of questions need to be asked or answered?
- What are the strategic business priorities of the users accessing the data?
- Are you balancing the need to serve immediate answers to known questions, and allow for further exploration?
- What processes are in place to ensure the quality of data flows and published data sources (e.g. quality assurance, certification)?

Managing the data prep chaos with Tableau Prep Conductor

For self-service data prep to succeed at scale, people and technology should be brought together under a governance framework that balances IT control with the flexibility and agility the business demands. IT can focus on enablement through operationalizing flows with automation and monitoring usage, performance, and access to ensure data prep practices scale efficiently and effectively.

How are you operationalizing data prep flows?

In certain cases, as with ad hoc data exploration of small, simple, or already clean data sets, the basic data prep functions of Tableau Desktop (like pivoting or hiding columns) may be sufficient. However, with larger, complex data sets, or flows that feed critical dashboards, the data sources may need to remain fresh for trusted decision-making. Depending on the use case, you may need to go beyond scheduled extract refreshes with Tableau Server. Rather than updating the extract, automating prep flow runs will apply the necessary cleaning steps to the data and produce a data source that can then be used for analysis.

If there are people in your organization using Tableau Prep Builder to clean data, Tableau Prep Conductor allows you to take advantage of their work automatically. Schedule flows to run at a specific time or on a recurring basis with scheduled tasks in Tableau Server, or create flow tasks to run on pre-defined schedules in Tableau Online. Rerunning flows can be done during non-work hours without the need of a human to initiate it each time, saving time and overhead. This also helps users take advantage of a stable server environment, instead of relying on their own desktop resources to run flows.

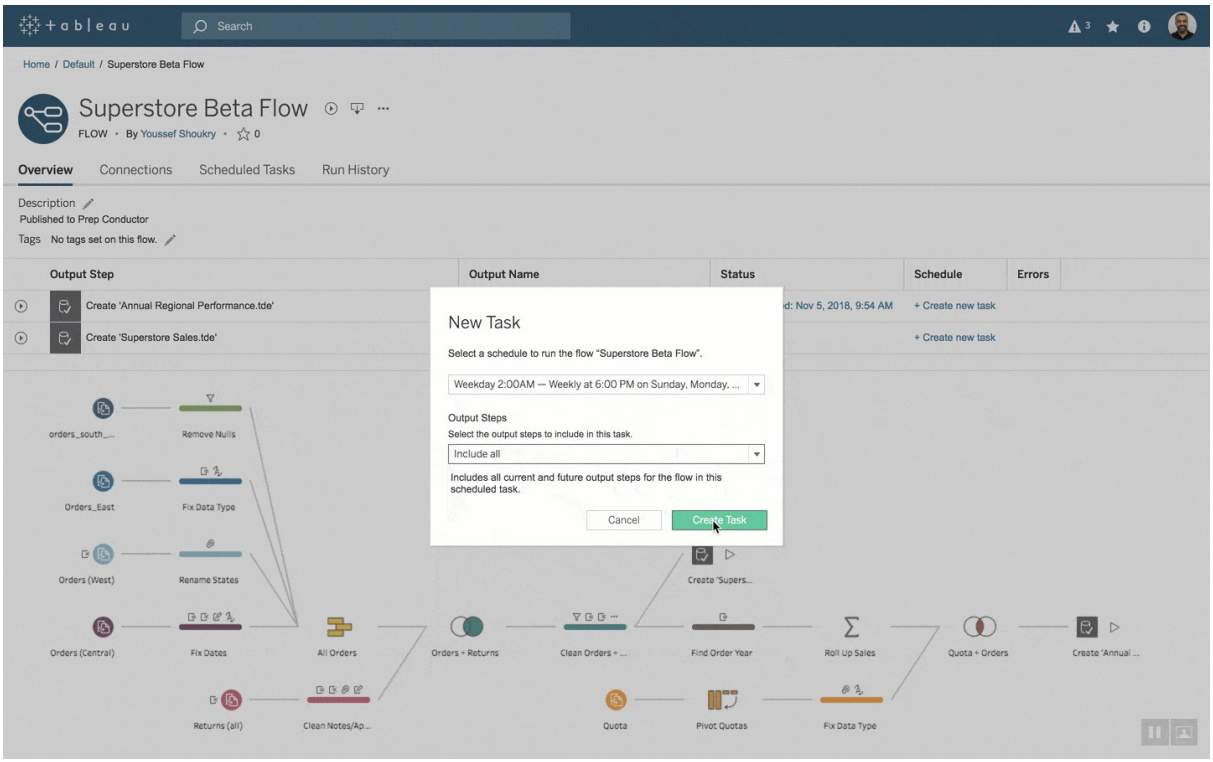


Figure 2 Create a flow task to regularly run prep flows on a set schedule.

Prep like a pro: Automate with the REST API

Build workflows with third party systems using the REST API and connect your data pipeline in powerful ways: Publish, schedule, download, and query flows; update flow connections; run flows and flow tasks on demand; manage permissions, and more.

Learn more about the REST API capabilities with Tableau Prep Conductor

How are you monitoring prep flows?

Monitoring is an area in which IT can strongly enable the organization as data prep practices scale. Performance and usage monitoring, automation, and alerting can help ensure the organization's data remains fresh and secure as its prepared with efficient flows.

With Tableau Prep Conductor, administrators can monitor flows with the same tools available on Tableau Server today. This includes Tableau Services Manager (TSM), the "Status" page, and out-of-the-box Administrative Views. Integrated as a part of Tableau Server and Tableau Online, these views can help answer important questions about your organization's data prep practices. Please note: Not all of the views in Tableau Server are available in or relevant to Tableau Online.

- Performance of flow runs — Understand what flow tasks are currently scheduled, currently running, the duration of flow tasks, which flow tasks run most frequently, and which flows were ad hoc vs. scheduled.
- Actions by all users, by specific users, or by recent users — The latter can be especially useful if you need to do maintenance on the server and want to know what users are doing on the server and how maintenance will affect them.
- Stats for space usage — Identify which flow outputs are taking up most disk space on the server.
- Backgrounder task delays — Use this view to help identify places to improve server performance by optimizing tasks and distributing task schedules.

Prep like a pro: Build custom reports with the Tableau Server Repository

In addition to the pre-built administrative views, you can use Tableau Desktop to query and build your own analyses of server activity. To do this, you can connect to and query views in the Tableau Server Repository, a PostreSQL database.

Learn more about building custom views for Tableau Server using Windows or Linux.

How are issues detected?

Whether it's connection issues or errors within a flow, you need to be aware of and fix problems that can stall your data prep flows. To minimize the risks of using stale data, Tableau Prep Conductor not only informs you of any issues that occur as your flows run, but also helps you with suggested fixes for any errors that occur.

- **Run history** — Users can view the historical refreshes of a flow to see successful or failed runs at a glance. This helps to understand the quality of all your flows and have greater confidence in the accuracy of your data.
- **Alerting** — Timely alerts will inform you if your flows are operating correctly. If a flow does encounter errors, users receive an email notification, as well as an alert within the Server interface. Links within the alert allow analysts and data stewards to quickly take action and examine errors, fix them with suggested remedies, and get back to their prep or analysis.
- **Status pages** — The Tableau Server status page and Tableau Services Manager status page include Tableau Server processes, along with links to troubleshooting documentation if a process is not running as expected. If you hover your mouse pointer over the status indicator for a process, a tooltip shows the node name and the port the process is running on.

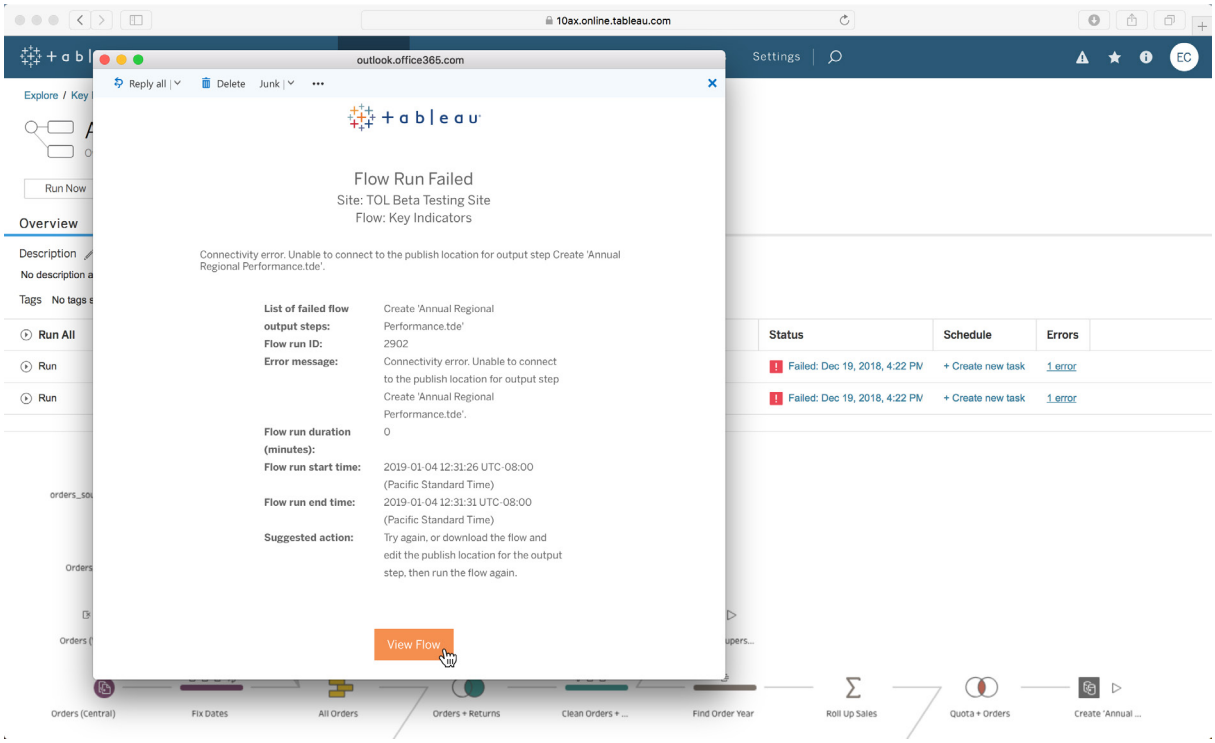


Figure 3 Receive alerts for a failed flow run—seen here in an email and in Tableau Online.

> A data prep tool should equip the one-off questions from the analysts and also be repeatable.
>
> **GORDON STRODEL**
> INFORMATION MANAGEMENT AND ANALYTICS CONSULTANT, SLALOM

How do you utilize access controls and keep data safe at scale?

A critical aspect of monitoring and governing flows is ensuring the right users have access to data prep flows. IT administrators can save time and effort with consolidated permissions controls in Tableau Server or Tableau Online. For prep flows managed by Tableau Prep Conductor, you can set permissions when publishing the flow—who can see it, edit it, run the flow, and more. If a flow connects to databases, you can specify the authentication type and set credentials to access the data.

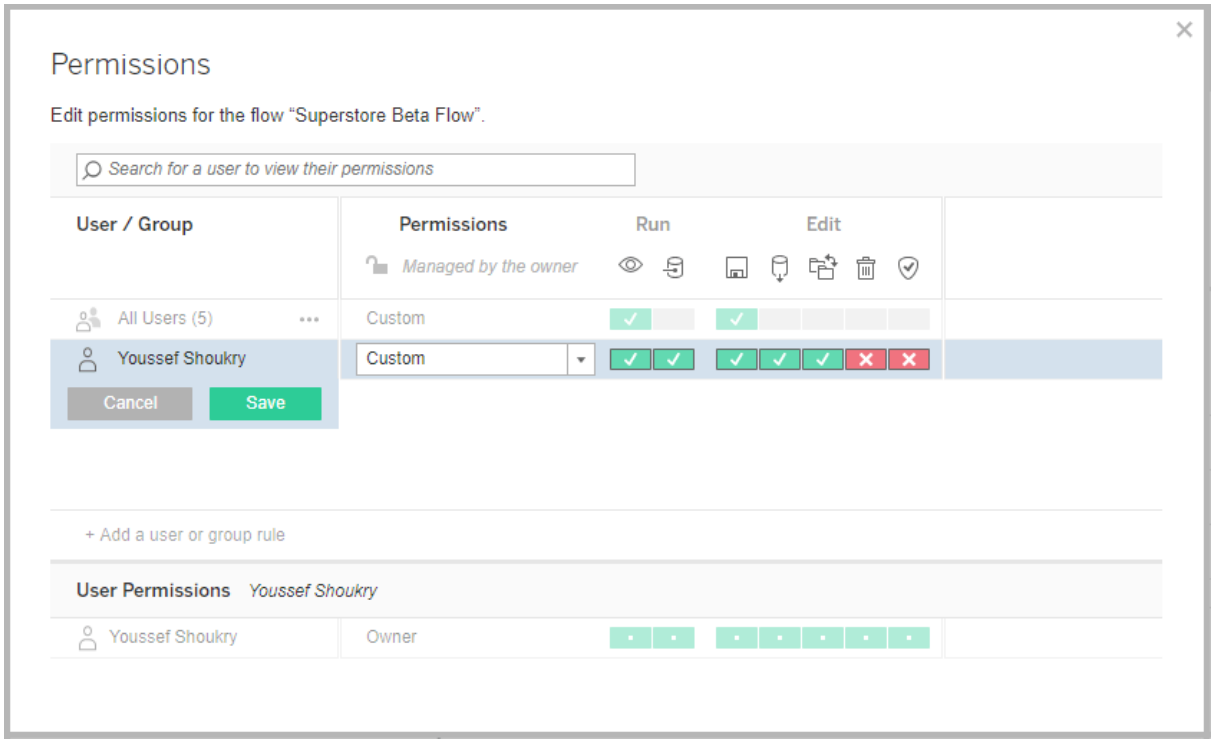


Figure 4 Manage and customize flow permissions with a simple and familiar interface.

Additionally, the Administrative Views provide reports to see permissions and actions by a specific user or by all users, so you can get a deep understanding of who is doing what with each flow.

Discovering the value of your data

The organization will get more value from prepped data if they can discover it, understand its relevance, and know to trust it—especially as data practices scale and data is more widely available to more users. This where self-service data prep can fuel and further the benefits of self-service analysis in a governed environment, like Tableau Server or Tableau Online.

How can your organization discover prepped data sources?

Many of the management capabilities of Tableau Prep Conductor will help users to better discover relevant and trusted data when they search, filter, and find flows and data sources published to Tableau Server or Tableau Online.

Tagging — Apply keywords to flows help users find, filter, and categorize content—just like with workbooks. You can add tags to an individual flow or a selection of multiple flows at once.

Organization — Move flows between projects to keep flows organized with relevant data sources, workbooks, and other content. Flow owners can do this by default, but others can be granted the Move permissions by the appropriate administrators.

Lifecycle management — Additional management capabilities will help to ensure data prep flows are relevant and organized, including permissions to save, rename, and delete flows as necessary. Administrators, flow owners, and project owners can also reassign ownership of flows.

The ability to search for data flows that already exist in Tableau Server or Tableau Online will also help to avoid redundancies and duplicate data prep operations. If a user can find a data source and see that the prep flow suits their needs for analysis, they won't need to spend the time recreating the prep steps or running the flow. Or, they may find an existing flow that they can downloaded and modify to meet a similar use case—without needing to build the prep flow from scratch.

How do you help the organization understand the data available to them?

Data literacy is a necessary investment for any organization scaling self-service data prep and analysis. In many organizations, IT and analytics champions collaborate on a Center of Excellence, including resources and support, an internal user group, and an investment in training and development to build analytical skills. These needs should be assessed in your organization based on the types and volume of users who will utilize tools like Tableau Prep Builder and Tableau Prep Conductor for self-service data prep.

Analysts and business users gain trust in your data by quickly knowing how a Tableau data source was created. With Tableau Prep Conductor, any user can see the origin of a data source created by a flow and navigate directly to it to see how it was constructed. With a base-level understanding of the data and ability to "read" the flow steps, the user should be able to determine whether or not the data source is valuable to them.

To enable Creators and Explorers (the Tableau license types that allow users to connect to data sources and author new content or explore and customize existing content), IT and data stewards should have a process in place to certify data sources that leverage consistently run prep flows. Certified data sources tell the organization that the data is trusted and ready for analysis. Certified data sources will also surface in searches and filtering with a higher priority in Tableau Server or Tableau Online.

Visual data prep allows people to see the full end-to-end process, seeing potential flags earlier on—like misspellings in the data, extra spaces, or incorrect join clauses. It also increases confidence in the final analysis.

JASON HARMER CONSULTANT, AT NATIONWIDE INSURANCE

Considering governance at scale

Every organization has specific needs and there is no "one-size-fits-all" approach to data preparation. However, when selecting a self-service data preparation tool, organizations should consider how the tool will evolve processes towards an iterative, agile approach instead of creating new barriers to entry. People will have a greater desire to prepare and understand their data if they can see the impact of their data prep steps.

Proper collaboration and governance are key—individuals may attempt to solve their own data problems, but IT is a critical player in solving organizational data problems. Governance practices will help the right people access the right data, ensure that the data driving your users' decisions is accurate, and maintain compliance with internal policies or external regulations.

A shift in governance is not asking IT to relinquish control so much as allowing the business to be more self-reliant within a trusted and centralized environment. Analysts and business users become the first line of defense in identifying data issues or irregularities within a governance model that IT and the business agree upon together.

Like the paradigm shift of self-service analytics, broadening how the business participates in governance to democratize data prep has its challenges—including process and technology change management, security risks to be mitigate, and skills gaps for users. But it's important to remember that with an iterative, agile approach to deployment, and a collaborative approach to governance, the benefits of bringing data prep to more people can exceed expectations.

> Before Tableau Prep our team would spend hours and hours making sure that our data sources were clean and organized, just to make sure that our analysis was accurate and effective. We've been able to save hours of work with Tableau Prep, completely reinventing the way we look at our data, and dramatically shortening the time between data collection and actionable insights.
>
> **GESSICA BRIGGS-SULLIVAN** TABLEAU ADMINISTRATOR, CHARLES SCHWAB, INC.

About Tableau

Tableau is a complete, easy-to-use, enterprise-ready visual business intelligence platform that helps people see and understand data through rapid-fire, self-service analytics at scale. Whether on-premises or in the cloud, on Windows or Linux, Tableau leverages your existing technology investments and scales with you as your data environment shifts and grows. Unleash the power of your most valuable assets: your data and your people.

Additional resources

Learn more: Data preparation with Tableau

Learn more: Data management with Tableau

Online Help: Tableau Prep Conductor

Whitepaper: Dirty data is costing you (How to solve common data prep issues)

Whitepaper: Best practices for tidy data

BARC research study: Data preparation - refining raw data into value

Tableau for the enterprise: Analytics powered by IT