# Basic 101

## Course Overview

The value proposition of Dataiku revolves around its ability to democratize AI and streamline the process of AI project development and deployment. Here are the key aspects of Dataiku's value proposition based on the provided information:

**1. Democratizing AI with Dataiku**

Dataiku aims to make AI accessible and usable for everyone in an organization, not just data scientists or IT professionals. It enables organizations to integrate AI into everyday business processes, thereby enhancing decision-making and driving exceptional business outcomes.

* **Systemized Approach**: Dataiku provides a structured approach that balances rationality with creativity, individual efforts with collective collaboration, and routine tasks with ambitious projects. This approach helps organizations harness the full potential of AI while managing complexity effectively.

**2. Centralizing AI Initiatives**

Dataiku serves as a centralized platform for AI initiatives, spanning from initial data exploration to the deployment of AI models in production.

* **Unifying Teams**: It brings together business, data, and IT teams, allowing them to collaborate seamlessly on AI projects. Each team member can contribute their expertise, ensuring a holistic approach to AI development.
* **Upskilling**: Dataiku empowers business analysts and domain experts with visual tools that simplify data handling and AI model development. This democratizes the creation of AI applications across the organization.
* **Single Platform**: Projects can be developed, tested, and deployed in a unified environment, streamlining the entire AI lifecycle.

**3. Streamlining Path to Production AI**

Dataiku facilitates the transition of AI projects from design to production with efficiency and reliability.

* **Balancing Customization and Reusability**: It allows teams to balance custom code with reusable components and standard frameworks, ensuring that AI projects are production-ready.
* **Automation**: Dataiku automates both the design and production phases, reducing repetitive tasks and enhancing project quality and reliability.
* **Deployment**: Projects can be deployed to production environments quickly and efficiently, leveraging elastic cloud infrastructure for scalability.

**4. Managing & Governing Projects at Scale**

Dataiku provides robust capabilities for managing and governing AI projects across the organization.

* **Built-in Governance**: It ensures transparency and portability across teams by documenting project goals, critical decisions, and models. This helps in maintaining consistency and compliance with organizational standards.
* **Lifecycle Management**: Dataiku supports the entire lifecycle of data and AI projects, from development to production, even with limited resources.
* **Operational Risk Management**: It helps organizations manage operational risks associated with AI projects, ensuring compliance with legal and regulatory requirements.

## Discover the Interface

### Concept | Dataiku Cloud Launchpad

The Dataiku Launchpad serves as the administrative hub for users accessing Dataiku, particularly in Dataiku Cloud environments. Here’s a detailed overview and key concepts related to the Launchpad:

**Overview of Dataiku Launchpad**

1. **Access and Administration**:
   * **Access**: Dataiku Cloud users access Dataiku via the Launchpad (<https://launchpad-dku.app.dataiku.io/>). Users of self-managed instances manage functions through the Administration page.
   * **Administrative Functions**: Admins manage configurations and settings through the Launchpad, ensuring smooth operation and updates.
2. **Spaces in Dataiku**:
   * **Definition**: Spaces are configurable instances within Dataiku, each running a specific version of the software. They are independent, allowing actions in one space (e.g., project creation, job running) without affecting others.
   * **Management**: Users manage spaces through the Launchpad’s Overview panel, ensuring each space is activated and configured appropriately.
   * **Design Node**: Spaces include a Design Node instance, the starting point for AI lifecycle tasks like data preparation, modeling, and deployment.
3. **Space Add-ons**:
   * **Extensions**: Spaces may have optional extensions like Solutions (accelerators for business use cases), platform enhancements (Automation, API nodes), integrations (Spark, Git, R), scenario reporters (Email, Slack), and plugins (reusable components).
   * **Connections and Code Environments**: Spaces also feature connections to data sources/storage and custom code environments (Python/R packages).
4. **Monitoring and Administration**:
   * **Audit Trail**: Tracks user access within a space, enhancing security and compliance.
   * **Usage & Monitoring**: Provides real-time visualization of tasks and computational resources, optimizing efficiency.
   * **Administrative Settings**: Manage users, profiles, groups, permissions, and space settings (rename, support access, maintenance).
5. **Next Steps**:
   * **Exploring Dataiku**: After familiarizing with Launchpad, users proceed to the Design home to utilize Dataiku’s features fully, starting with the Design Node.

Spaces

In Dataiku, "Spaces" refer to organizational units within the platform where users can organize and manage projects, datasets, models, and other assets. Spaces provide a structured way to partition and manage resources based on teams, departments, or specific projects within an organization. Here’s a detailed explanation of what Spaces are and how they function within Dataiku:

**Key Features of Spaces in Dataiku:**

1. **Organizational Structure**:
   * **Hierarchical Organization**: Spaces allow for a hierarchical organization of projects and resources. You can create multiple levels of Spaces to reflect different organizational units or project categories.
2. **Access Control and Permissions**:
   * **Granular Permissions**: Each Space can have its own set of access control and permissions. This means you can define who can view, edit, or administer resources within each Space, ensuring data security and governance.
3. **Resource Management**:
   * **Centralized Management**: Within each Space, you can manage projects, datasets, connections, recipes (data preparation steps), models, notebooks (for code development), and other assets specific to that Space.
4. **Collaboration**:
   * **Team Collaboration**: Spaces facilitate collaboration among teams working on different projects or within different departments. Users within the same Space can share and collaborate on resources seamlessly.
5. **Isolation and Segregation**:
   * **Data Isolation**: Spaces provide a level of isolation between different projects or departments, allowing organizations to manage sensitive data separately while maintaining a unified platform.
6. **Project Visibility**:
   * **Visibility and Organization**: Spaces help organize and categorize projects, making it easier for users to locate and access relevant resources quickly. This organizational structure improves efficiency and reduces confusion.

**Practical Use Cases:**

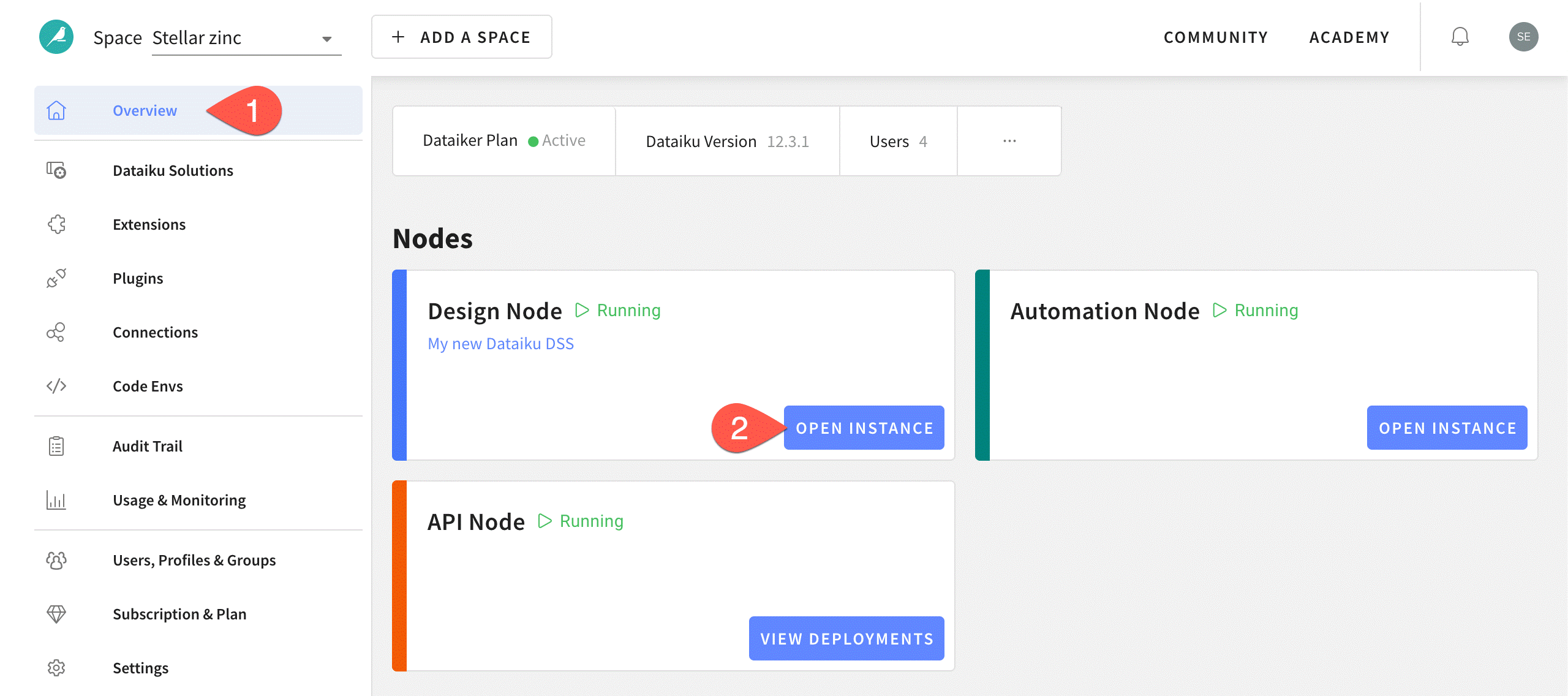
* **Departmental Spaces**: Organizations can create Spaces for each department (e.g., Marketing, Finance, IT) to manage projects and data specific to their functions.
* **Project-Based Spaces**: Spaces can also be created for specific projects or initiatives, allowing teams to collaborate effectively within dedicated environments.
* **Security and Compliance**: Spaces help enforce data access and compliance policies by restricting access to sensitive data based on Space-level permissions.

**Conclusion:**

Spaces in Dataiku provide a flexible and scalable way to organize, manage, and collaborate on data science and AI projects within an organization. They enhance efficiency, security, and collaboration by structuring resources into logical units that align with organizational goals and workflows. By leveraging Spaces, organizations can effectively manage their data and AI initiatives while maintaining governance and compliance standards.

To get started:

1. Make sure the space is turned on from the **Overview** panel.
2. Once the space is active, you can open the **Design Node** instance.

[](https://academy-content.dataiku.com/latest/_images/launchpad.png)

Tip

Dataiku offers multiple nodes for different kinds of tasks in the AI lifecycle, but it all starts in the Design node!

**Space add-ons**

A space is more than just a specific version of the core Dataiku platform. A space may also have several optional extensions or features installed on top. These could be:

| **Add-on** | **Description** |
| --- | --- |
| Solutions | [Solutions](https://academy-content.dataiku.com/latest/cloud-add-ons/solutions/index.html?opals=true) are installable accelerators that leverage Dataiku to answer business problems and use cases. |
| Extensions | Extensions include a variety of platform enhancements such as additional nodes for production tasks (Automation, API, Govern), integrations for Spark, Git and R, scenario reporters (Email, Slack, Microsoft Teams), etc. |
| Plugins | [Plugins](https://academy-content.dataiku.com/latest/cloud-add-ons/plugins/index.html?opals=true) are packages of reusable components that extend the functionality of Dataiku. |
| Connections | [Connections](https://academy-content.dataiku.com/latest/cloud-connections/data-connections/index.html?opals=true) are read-only data sources and read/write data storage locations. |
| Code environments | [Code environments](https://doc.dataiku.com/dss/latest/code-envs/index.html) are custom lists of Python or R packages available to the space. |

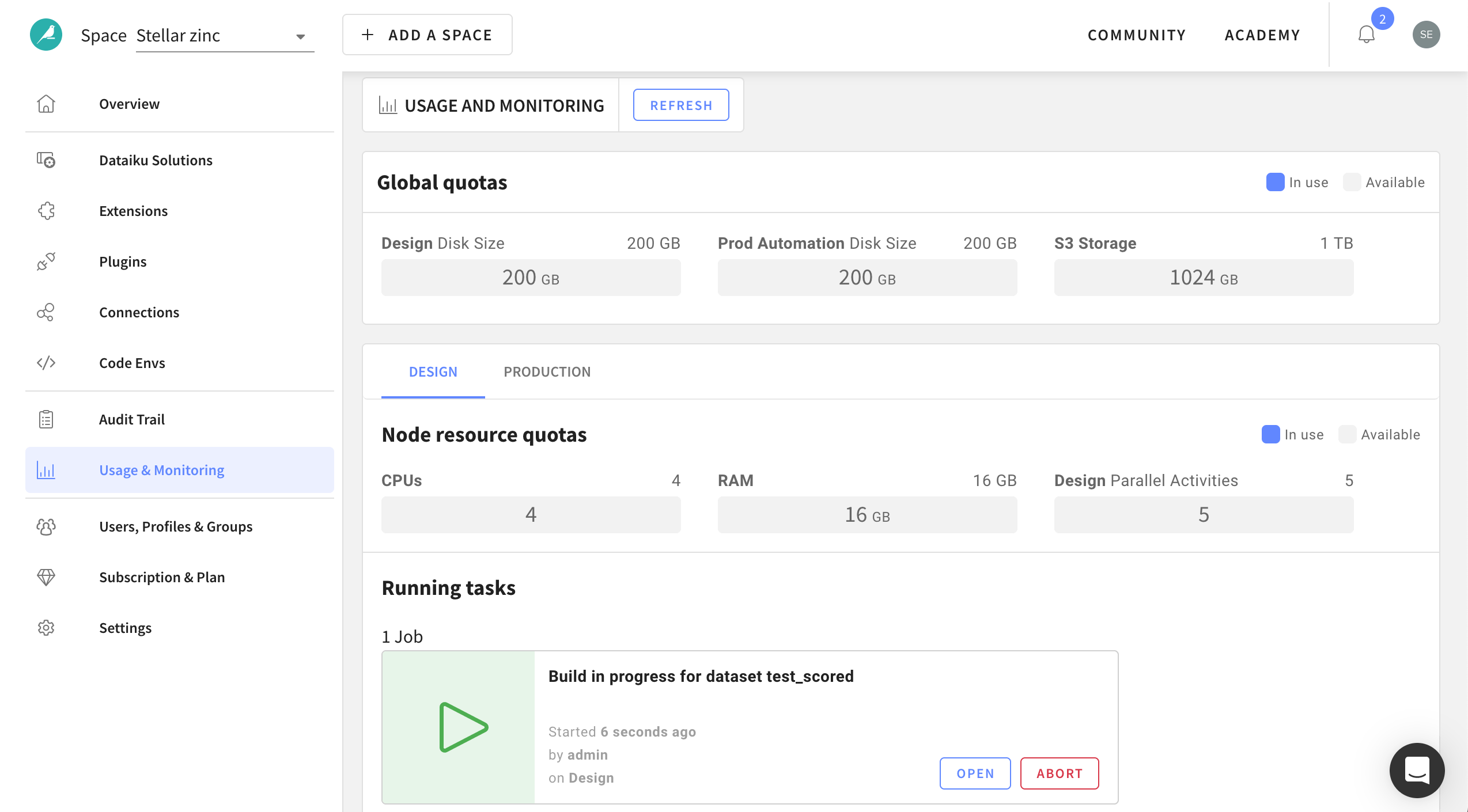
Tip

Recall that spaces are independent from each other. Accordingly, add-ons installed on one space are only available to that space.

**Space monitoring**

The homepage for a space also includes references for monitoring user activity:

| **Audit tool** | **Description** |
| --- | --- |
| Audit Trail | A record of who is accessing the space, when, and from where. |
| Usage & Monitoring | Visualizations of real-time running tasks and available computational resources. |

[](https://academy-content.dataiku.com/latest/_images/launchpad-usage1.png)

Note

For more information, see our resources on [Compute and Resource Quotas on Dataiku Cloud](https://academy-content.dataiku.com/latest/cloud-connections/quotas-compute/index.html?opals=true).

**Space administration**

You can also find basic administrative settings governing the space:

| **Administrative setting** | **Description** |
| --- | --- |
| Users, Profiles & Groups | Invite new users to share the space and manage permissions to govern collaboration.  For more information, see the [Users, Profiles & Groups](https://academy-content.dataiku.com/latest/cloud-space-management/users-roles-groups/index.html?opals=true) article. |
| Subscription & Plan | Details on your current Dataiku plan. |
| Settings | Rename your space; grant access to support; and perform space maintenance. |

**What’s next?**

Once you’ve taken a tour of the Launchpad, move on to the Design home to dive into what Dataiku has to offer.

#### Concept | Dataiku Design homepage

-> The Dataiku Design homepage is like the default page of a website. It’s where you land:

* When you sign in to your Dataiku design instance via your web browser, if you’re using a self-managed installation.
* When you open an instance from the Launchpad, if you’re using Dataiku Cloud.

It’s where you will be able to see shared projects if you are collaborating with colleagues or other users, and create new ones.

-> By default, projects appear at the top. And below, you will see additional sections, including:

* Workspaces
* Applications
* Project folders
* Wikis
* Dashboards

-> You can search for items using the Filter items box at the top of the screen. Or find recent and favorite items on the right sidebar.

-> Homepage customization

You can customize your homepage to fit the way you want to work. For example, you can configure your homepage so that project folders or dashboards display first or to make workspaces your default homepage.

To configure your homepage:

1. Visit your profile and settings in the upper right.
2. In Design homepage layout, arrange the sections in any order you choose.

Configurable Dataiku homepage profile settings.

-> Applications menu

The Applications menu is where you’ll find:

* Access to your workspaces, projects, applications, feature stores, wikis, etc.
* A searchable data catalog which indexes and details all project artifacts including datasets.
* Access to administration settings at the instance level for on-premise installation, including the management of the license, connections, security, code environments, etc.

Note

On Dataiku Cloud, the administration settings are accessible in the launchpad. There is no Administration menu in the Applications menu.

* A plugin library and possibility for developers to create custom plugins.

## Concept | Searching in Dataiku

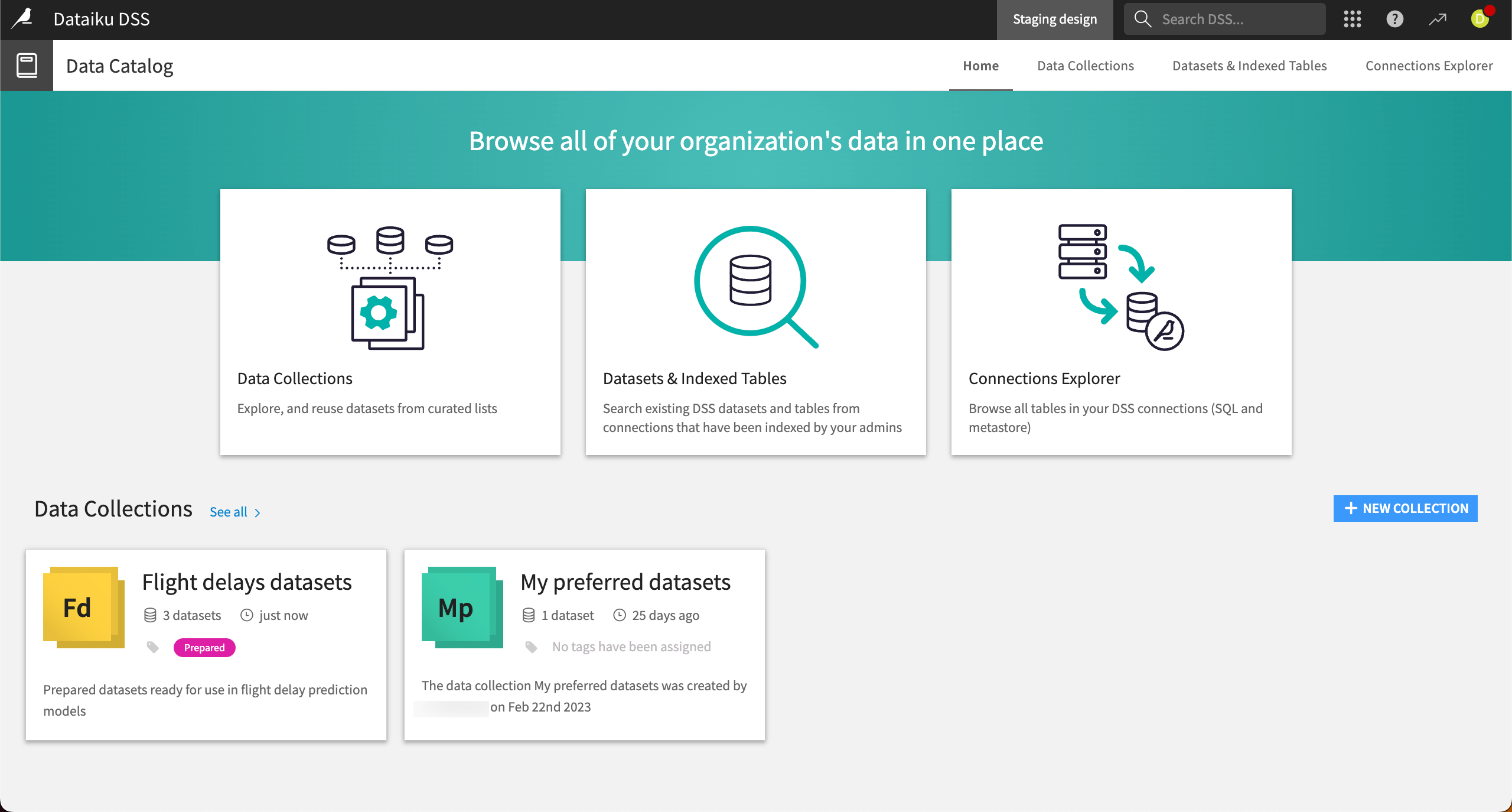
Dataiku provides multiple ways to search for online help, datasets, recipes, projects, and other Dataiku items.

**Data Catalog**[#](https://academy-content.dataiku.com/latest/getting-started/dataiku-ui/concept-search-in-dataiku.html?opals=true#data-catalog)

The **Data Catalog**, accessible from the Applications menu, is the recommended way to search for datasets in Dataiku.

In it, you can:

* Find curated data collections.
* View information about datasets.
* Reuse them in your own projects.
* Import tables from external connections.

[](https://academy-content.dataiku.com/latest/_images/data-catalog1.png)

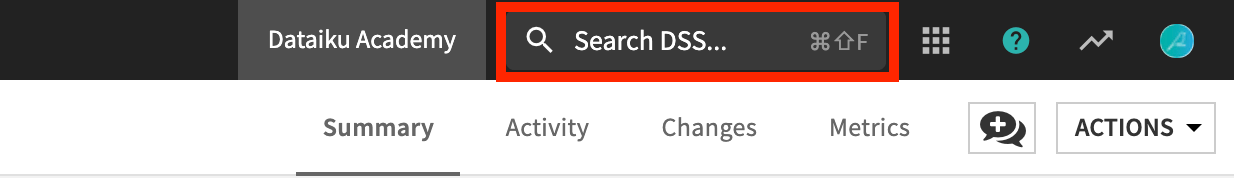
See also

For more information, see:

* The [Concept | Data Catalog](https://academy-content.dataiku.com/latest/collaboration/sharing-projects-assets/concept-data-catalog.html?opals=true) article in the Knowledge Base.
* The [Data Catalog](https://doc.dataiku.com/dss/latest/data-catalog/index.html) article in the reference documentation.

**-> Global Search**

On the top right of your Dataiku instance, the **Global Search** bar enables you to quickly find and navigate to many different types of Dataiku elements. It searches across several sources including help pages and Dataiku items like datasets and recipes, screens and settings within the product.

[](https://academy-content.dataiku.com/latest/_images/kb-global-search-ui.png)

**Using autocompletion**

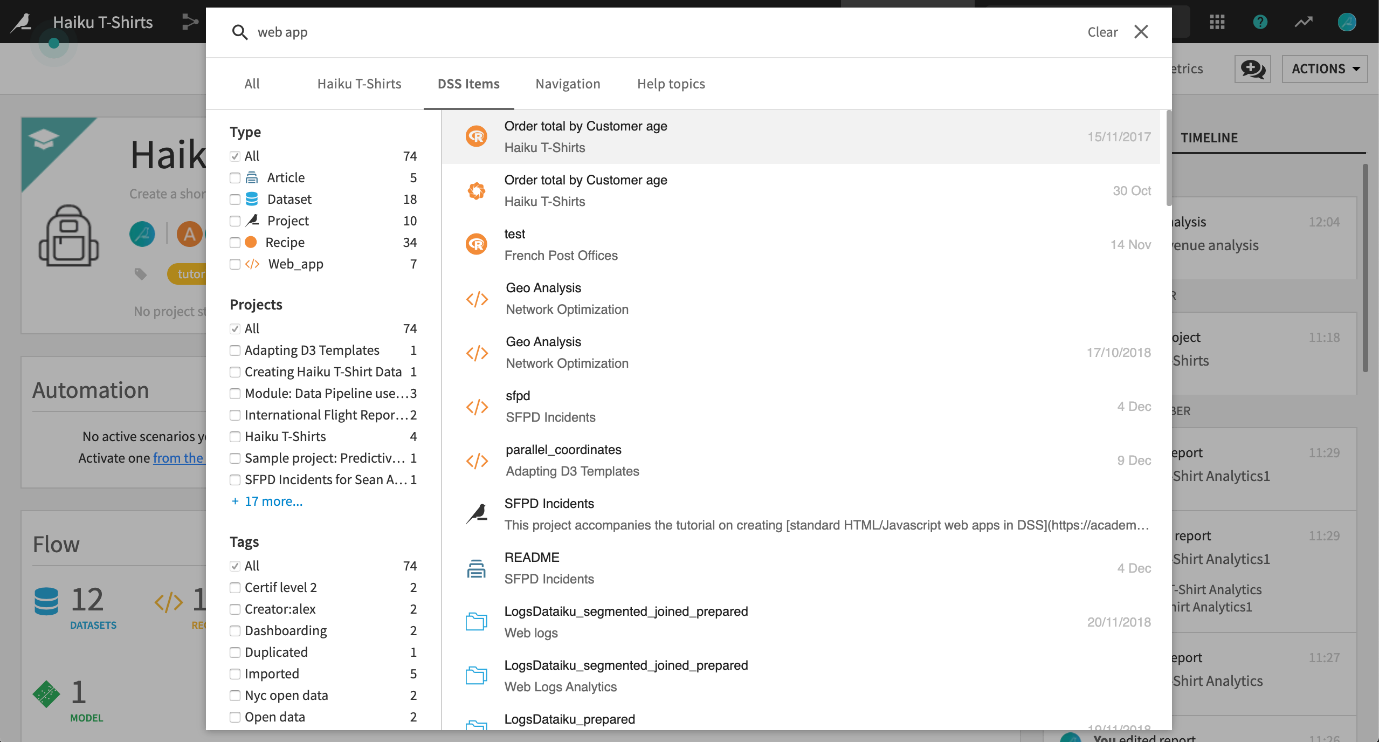
Start typing your search term, and Dataiku begins searching for matches across several sources, including:

* The data catalog.
* The documentation (reference documentation, knowledge base, developer guide, Dataiku blog).
* Community posts.
* The history of recently accessed items.
* An index of screens and settings.

**Filtering a global search**

After entering your search keywords, you can view and filter results in three categories.

| **Tab name** | **Description** |
| --- | --- |
| DSS items | Within **DSS items**, you can filter by type of items — such as dataset, insights, projects, recipes, and statistics worksheets — or by projects, tags, or users. |
| Navigation | In **Navigation**, you can view administrative settings such as plugins. |
| Help topics | In **Help topics**, you can find and filter reference documentation, tutorials, and other resources. |

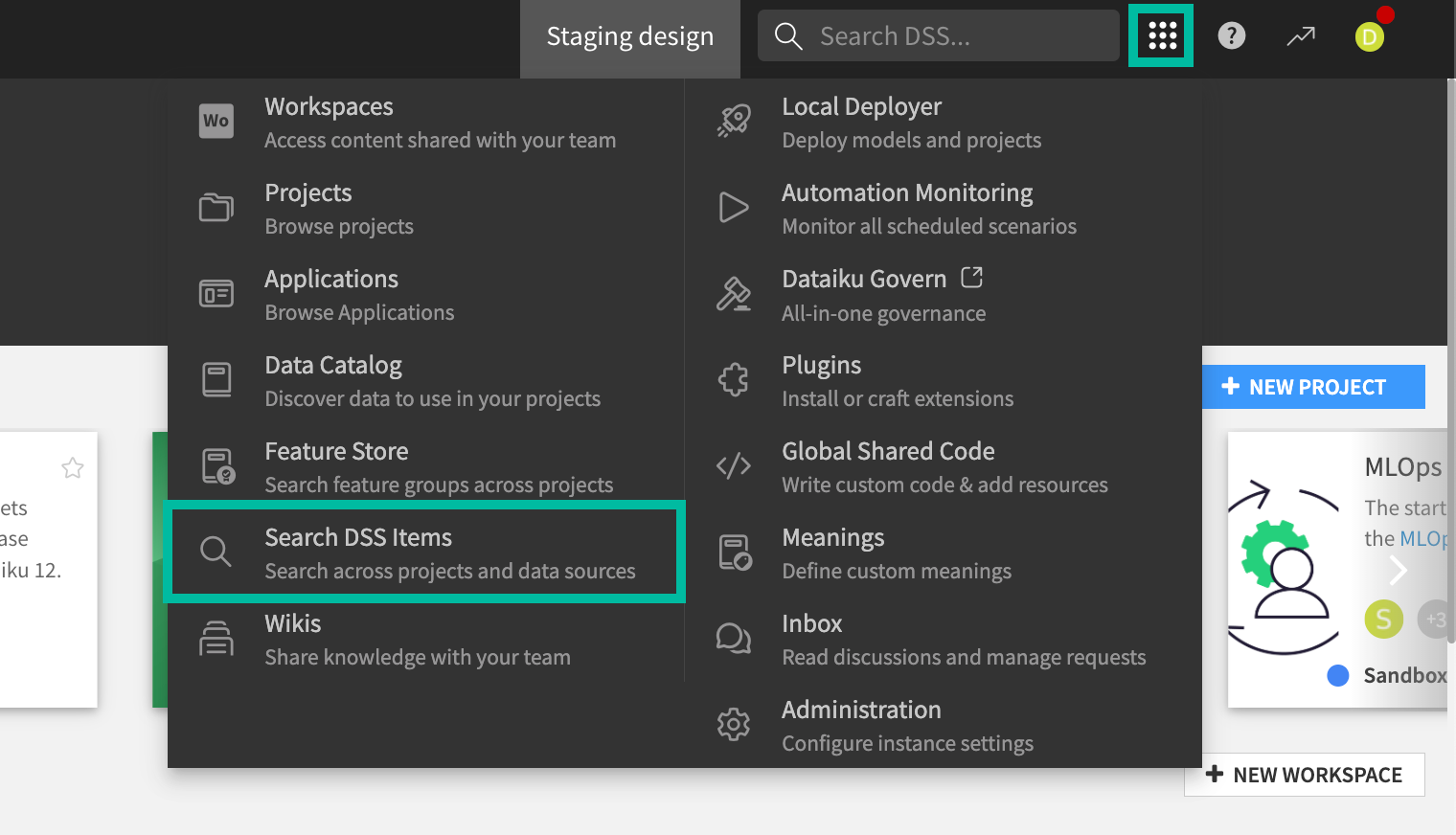
[](https://academy-content.dataiku.com/latest/_images/kb-global-search-tabs.png)

*Searching for “webapp” returns Dataiku items from the current project or elsewhere on the instance, along with relevant help resources.*

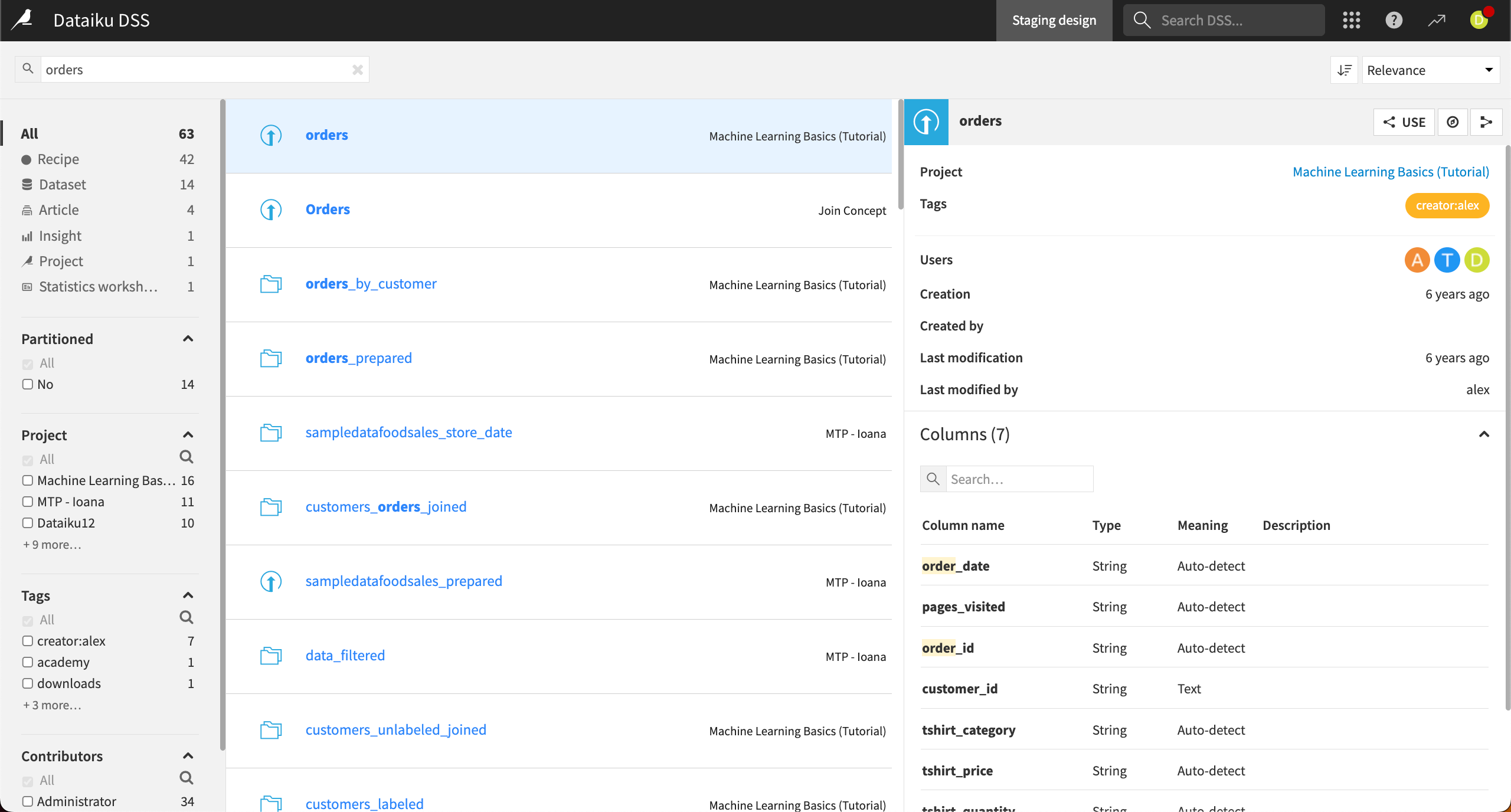
Clicking on a result takes you directly to it.

**-> Search DSS Items**

**Search DSS Items**, found in the Applications menu, is another way to search across projects and data sources in Dataiku. This page allows you to explore DSS items in further detail than you can through the Global Search bar.

[](https://academy-content.dataiku.com/latest/_images/search-dss-items-menu.png)

To manage search results, you can sort or apply filters by type of item, associated projects, tags, and contributors. Select any item to view details such as projects, tags, and schema. Click on an item name to go directly to that item.

[](https://academy-content.dataiku.com/latest/_images/search-dss-items.png)

Create the Project

### Concept | Project

The project is your command center. It contains all your work on a specific activity. For example, a project can include datasets, recipes, models, discussions, and dashboards.

Projects are created from the homepage and can be organized into project folders. You can do things like:

* Check the project’s overall status.
* View recent user activity.
* View contributors.
* Check off items on a to-do list.

Project commands include duplicating, exporting, and even deleting a project (if your access level permits it).

### Tutorial | Create your first project

In this tutorial, we’ll cover much of what you need to familiarize yourself with Dataiku projects

-> Objectives

In this tutorial, you will:

* Create a Dataiku project.
* Explore the project homepage.
* Prerequisites

To complete this tutorial, you’ll need the following:

A Dataiku instance (version 9.0 and above).

Note

If you do not already have access, there are two ways to get started. From the Dataiku website, you can:

1. Start a 14 day free trial of Dataiku Cloud.
2. Install the free edition.

Create your project

Let’s get started! The first step is to create a new Dataiku project.

1. From the Dataiku homepage, click +New Project > DSS tutorials > Core Designer > Basics 101. This action creates the starter project for you with a project ID (e.g. “DKU\_TUTORIAL\_BASICS\_101”) unique to your instance.
2. Explore the project homepage to discover the different tabs and features available on it.

Important

You cannot change the project ID once it is created in Dataiku. If you want a project with a different ID, you can duplicate the project and specify a new ID for the duplicate project.

What’s next?

You now have a project, but no data in it. Acquiring some data will be covered in the next tutorial!

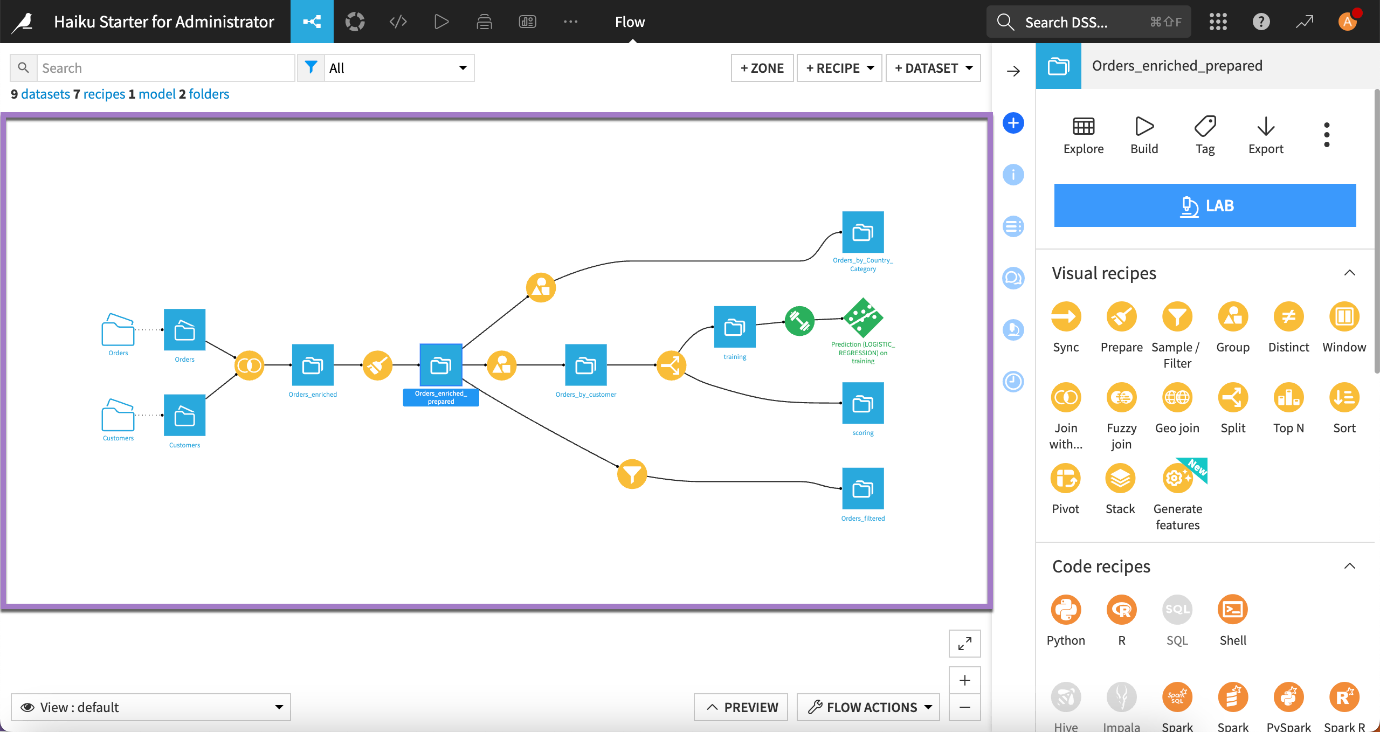
### Concept | Flow

In Dataiku, the **Flow** is the visual representation of how data, recipes, and models work together to move data through an analytical pipeline.

Note

Dataiku positions the different items dynamically, in an optimized way. You cannot change the layout manually.

From the initial data to the final output, the Flow in Dataiku allows you to trace the dependencies among the different items and becomes a visual narrative of your data’s journey.

[](https://academy-content.dataiku.com/latest/_images/flow1.png)

**-**

**> Improving the Flow readability**

Sometimes, the Flow can be quite complex, which can impact readability.

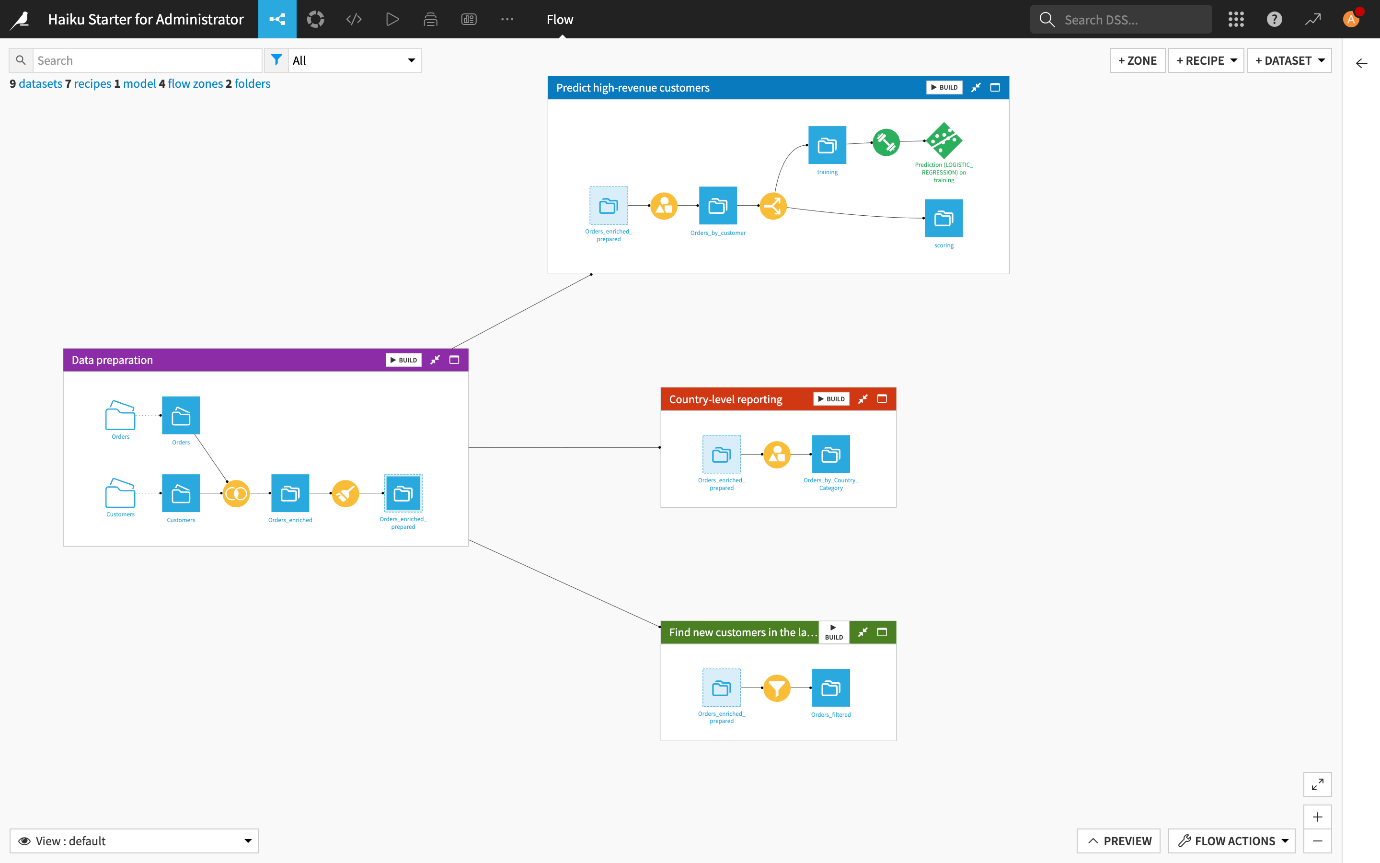
In such cases, for more clarity, you can use Flow zones, tags and filters.

**-> Using Flow zones**

By default, a Flow is displayed in a single zone.

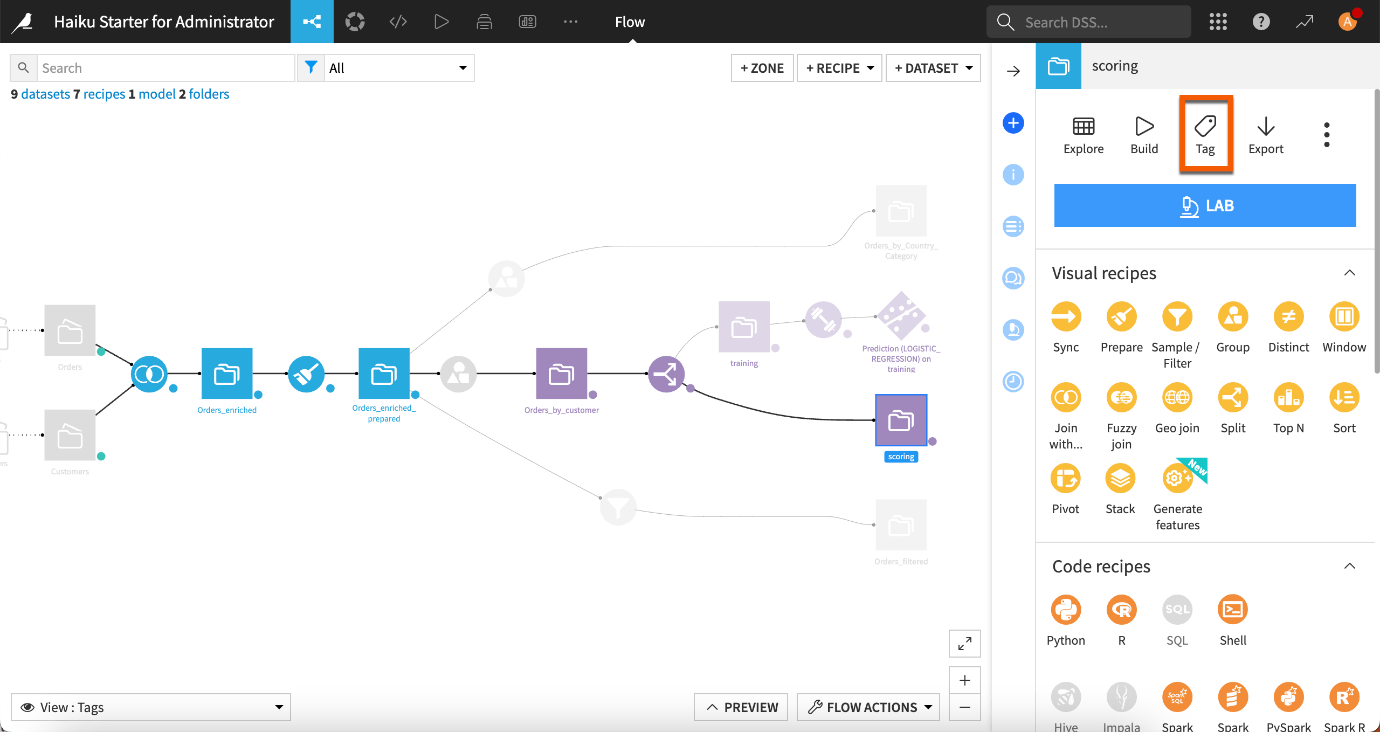
At any time, clicking the **+ Zone** button at the top right of the Flow allows you to add zones to the Flow in order to organize the objects.

For more information, see the [Tutorial | Flow zones](https://academy-content.dataiku.com/latest/getting-started/dataiku-ui/tutorial-flow-zones.html?opals=true) article.

[](https://academy-content.dataiku.com/latest/_images/flow-zones-02.png)

* **Tagging the Flow items**

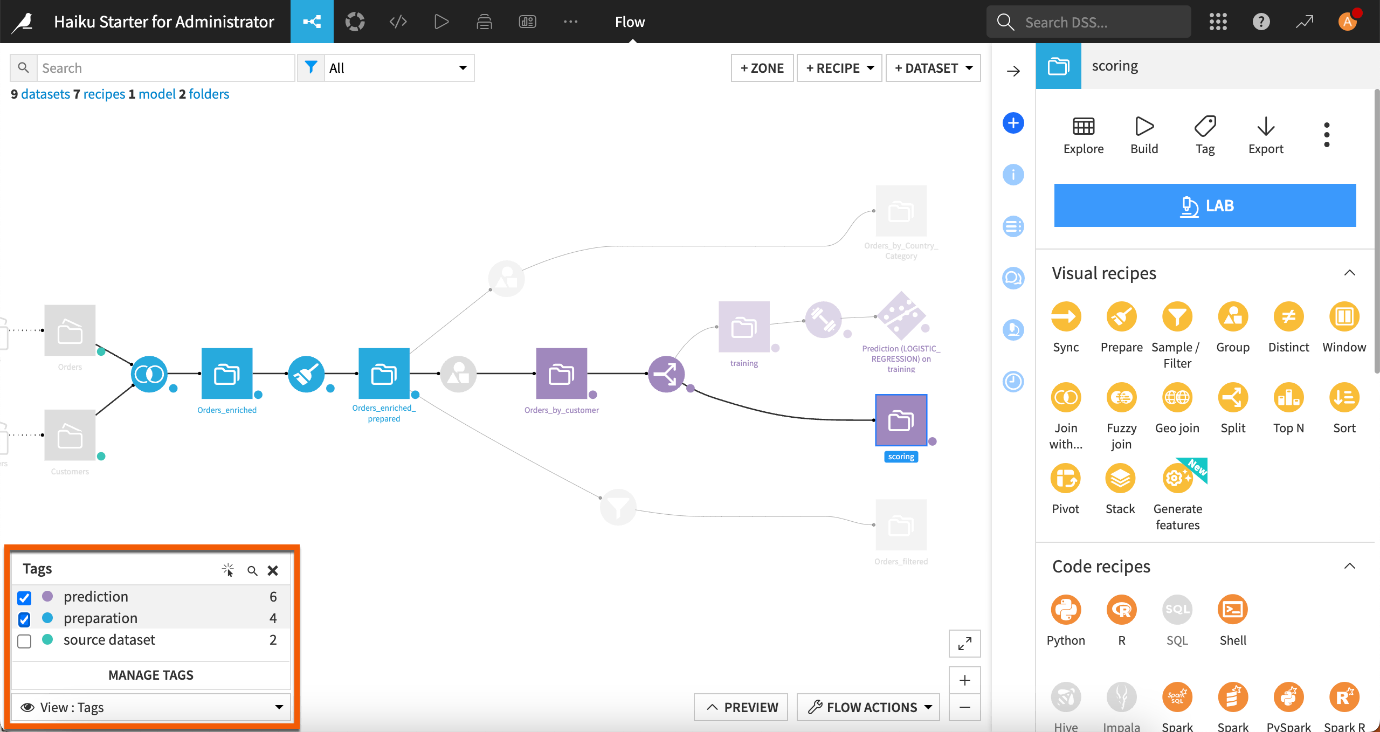
When there are too many objects on the screen, you can add tags to the different items of the Flow then use these tags to select which parts of the Flow to view. Tags can be based on attributes such as creator, purpose, status and so on.

[](https://academy-content.dataiku.com/latest/_images/tags-in-the-flow.png)

* **Filtering the Flow**

The View menu from the bottom left corner of your Flow lets you filter the Flow based on different elements such as Flow zones, tags, connections, recipe engines, the last modification date, etc.

For example, you can show or hide parts of the Flow by selecting or unselecting some tags to decrease the overall number of objects that appear on the screen.

[](https://academy-content.dataiku.com/latest/_images/filters-on-tags.png)

For more information, see the [Concept | Flow views](https://academy-content.dataiku.com/latest/getting-started/dataiku-ui/concept-flow-views.html?opals=true).

* **Building the Flow**

The **Flow Actions** dropdown at the bottom right of the Flow includes a **Build all** menu to build the entire Flow.

Yet, you can also right click on any item in the Flow and select the **Build** menu from there.

Keep in mind that the Flow in Dataiku has an awareness of the relationships and dependencies between datasets in your project. For example, if you make changes to a dataset or recipe, you may choose to dynamically rebuild dependent items upstream or downstream in your Flow to reflect the update.

For more information, see the [Concept | Build modes](https://academy-content.dataiku.com/latest/data-preparation/pipelines/concept-dataset-build-modes.html?opals=true) article.

Key features of Flow in Dataiku include:

1. **Visual Representation**: An intuitive graphical interface that shows the sequence of steps in data preparation and analysis.
2. **Drag-and-Drop Functionality**: Users can easily add, remove, and adjust steps in the data pipeline using drag-and-drop.
3. **Execution and Monitoring**: The ability to run the entire data pipeline and monitor the status and performance of each step.
4. **Integration with Tools and Data Sources**: Flow integrates with various databases, file systems, cloud platforms, and data access and analysis tools.
5. **Version Control and History**: It tracks changes in the data pipeline over time and allows users to revert to previous versions if needed.

ChatGPT

In Dataiku, "Flow Zones" refer to a feature that allows users to organize and manage their data pipelines and workflows more effectively within the Flow interface. Here's a detailed explanation:

**Purpose of Flow Zones**

Flow Zones are designed to:

1. **Organize Workflows**: They provide a way to visually group related datasets, recipes (data preparation steps), models, and other components within the Flow. This helps in organizing complex workflows and projects into logical sections.
2. **Improve Visibility**: By dividing the Flow into zones, users can improve visibility and reduce clutter, especially in large and complex projects. Each zone can represent a specific stage or aspect of the data science or analytics workflow.
3. **Enhance Collaboration**: Flow Zones facilitate collaboration among team members by providing a structured way to divide responsibilities or focus areas. Different team members can work on separate zones of the Flow simultaneously without interference.

**Key Features and Functionality**

* **Creation and Management**: Users can create multiple zones within the Flow and assign components (like datasets, recipes, models) to these zones.
* **Customization**: Zones can be customized with labels and colors to visually distinguish different parts of the workflow.
* **Navigation**: Users can easily navigate between zones to focus on specific parts of the project or workflow.
* **Permissions**: Dataiku allows administrators to set permissions at the zone level, controlling who can access and modify components within each zone. This ensures security and data governance.

**Benefits**

* **Organizational Structure**: Provides a structured and organized view of complex workflows, making it easier to understand and manage.
* **Efficiency**: Enhances productivity by allowing team members to work independently on different parts of the project within their assigned zones.
* **Scalability**: Supports scalability by accommodating larger projects and diverse teams working on multiple aspects of data science and analytics.

Create the Dataset

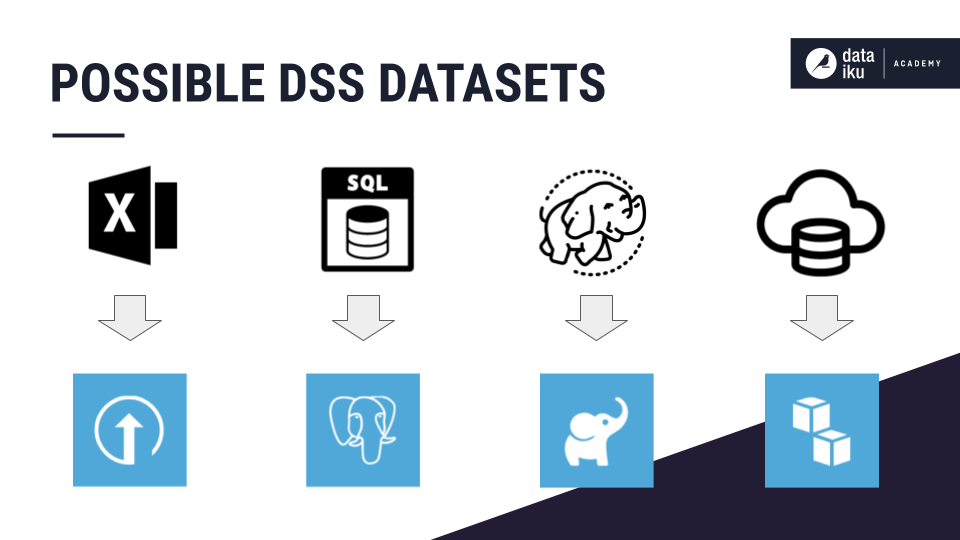
## [Concept | Dataiku datasets](https://play.vidyard.com/cin5EsHqVYo2hCkz1gBxjz?v=3.1.1&type=inline&hubspot_id=a568ca4b5a2b8d791ca438d8e8c2671a&referring_url=https%253A%252F%252Facademy.dataiku.com%252Fbasics-101%252F497926&)

[A dataset in Dataiku can be any piece of data in a tabular format. Examples of possible Dataiku datasets include:](https://play.vidyard.com/cin5EsHqVYo2hCkz1gBxjz?v=3.1.1&type=inline&hubspot_id=a568ca4b5a2b8d791ca438d8e8c2671a&referring_url=https%253A%252F%252Facademy.dataiku.com%252Fbasics-101%252F497926&)

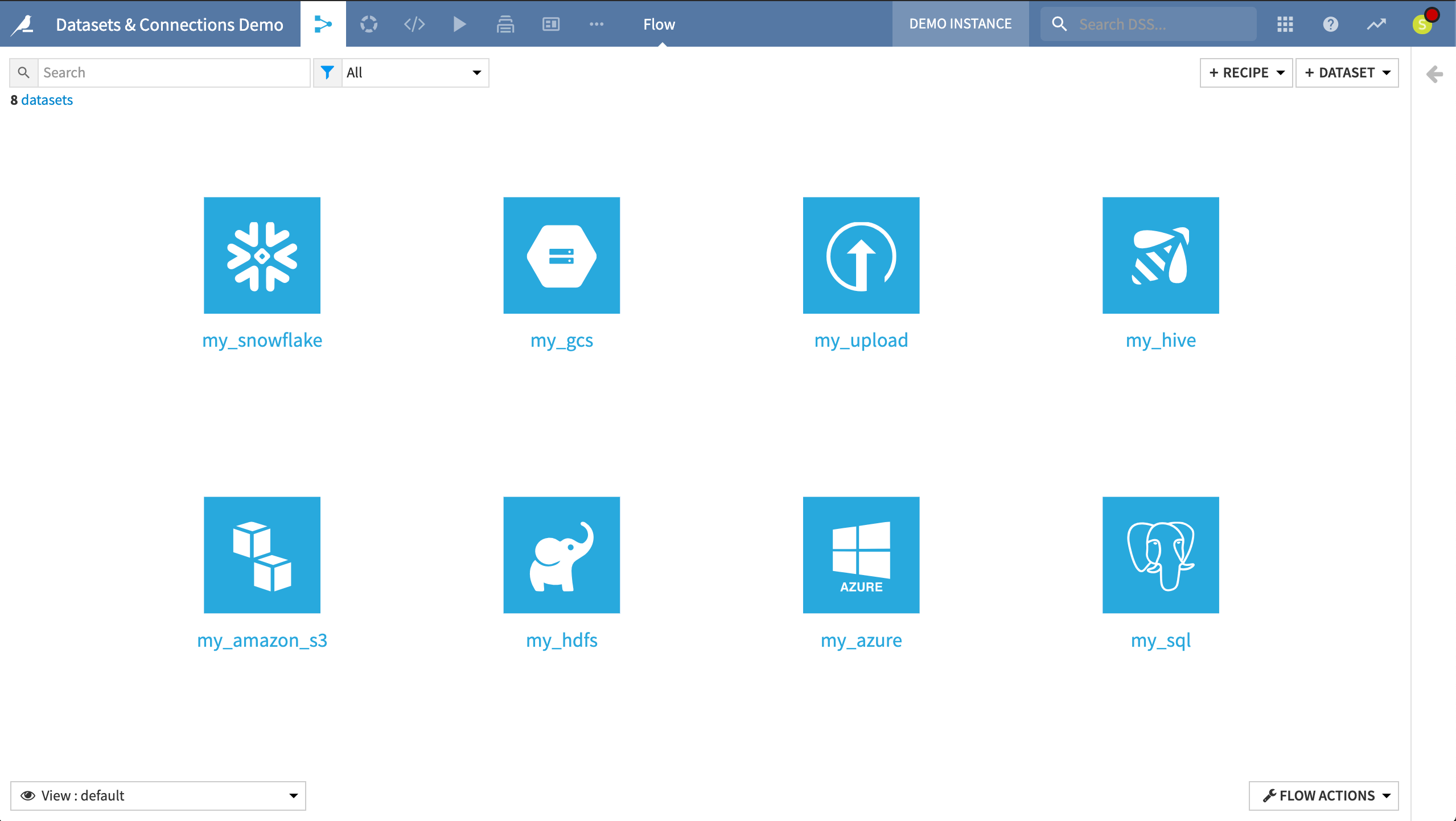
* [An uploaded Excel spreadsheet](https://play.vidyard.com/cin5EsHqVYo2hCkz1gBxjz?v=3.1.1&type=inline&hubspot_id=a568ca4b5a2b8d791ca438d8e8c2671a&referring_url=https%253A%252F%252Facademy.dataiku.com%252Fbasics-101%252F497926&)
* [An SQL table](https://play.vidyard.com/cin5EsHqVYo2hCkz1gBxjz?v=3.1.1&type=inline&hubspot_id=a568ca4b5a2b8d791ca438d8e8c2671a&referring_url=https%253A%252F%252Facademy.dataiku.com%252Fbasics-101%252F497926&)
* [A folder of data files on a Hadoop cluster](https://play.vidyard.com/cin5EsHqVYo2hCkz1gBxjz?v=3.1.1&type=inline&hubspot_id=a568ca4b5a2b8d791ca438d8e8c2671a&referring_url=https%253A%252F%252Facademy.dataiku.com%252Fbasics-101%252F497926&)
* [A CSV file in the cloud, such as an Amazon S3 bucket](https://play.vidyard.com/cin5EsHqVYo2hCkz1gBxjz?v=3.1.1&type=inline&hubspot_id=a568ca4b5a2b8d791ca438d8e8c2671a&referring_url=https%253A%252F%252Facademy.dataiku.com%252Fbasics-101%252F497926&)

**[Dataset representation in the Flow](https://play.vidyard.com/cin5EsHqVYo2hCkz1gBxjz?v=3.1.1&type=inline&hubspot_id=a568ca4b5a2b8d791ca438d8e8c2671a&referring_url=https%253A%252F%252Facademy.dataiku.com%252Fbasics-101%252F497926&)**

[Dataiku represents all the datasets in the Flow of a project with a blue square with the icon matching the type of the source dataset.](https://play.vidyard.com/cin5EsHqVYo2hCkz1gBxjz?v=3.1.1&type=inline&hubspot_id=a568ca4b5a2b8d791ca438d8e8c2671a&referring_url=https%253A%252F%252Facademy.dataiku.com%252Fbasics-101%252F497926&)

[[](https://academy-content.dataiku.com/latest/_images/dataset-examples.png)](https://play.vidyard.com/cin5EsHqVYo2hCkz1gBxjz?v=3.1.1&type=inline&hubspot_id=a568ca4b5a2b8d791ca438d8e8c2671a&referring_url=https%253A%252F%252Facademy.dataiku.com%252Fbasics-101%252F497926&)

[The following example Flow includes different types of datasets, such as an uploaded file, a table in a SQL database, and cloud storage datasets.](https://play.vidyard.com/cin5EsHqVYo2hCkz1gBxjz?v=3.1.1&type=inline&hubspot_id=a568ca4b5a2b8d791ca438d8e8c2671a&referring_url=https%253A%252F%252Facademy.dataiku.com%252Fbasics-101%252F497926&)

[[](https://academy-content.dataiku.com/latest/_images/dataset-flow.png)](https://play.vidyard.com/cin5EsHqVYo2hCkz1gBxjz?v=3.1.1&type=inline&hubspot_id=a568ca4b5a2b8d791ca438d8e8c2671a&referring_url=https%253A%252F%252Facademy.dataiku.com%252Fbasics-101%252F497926&)

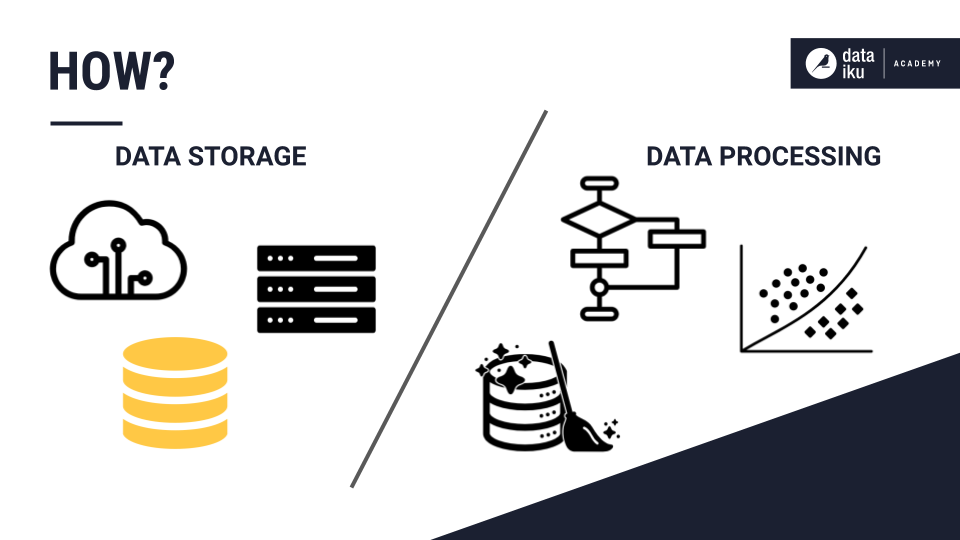
**[Interactions with datasets](https://play.vidyard.com/cin5EsHqVYo2hCkz1gBxjz?v=3.1.1&type=inline&hubspot_id=a568ca4b5a2b8d791ca438d8e8c2671a&referring_url=https%253A%252F%252Facademy.dataiku.com%252Fbasics-101%252F497926&)**

[Regardless of the origins of the source dataset, the methods for interacting with any Dataiku dataset are the same. You can read, write, visualize, and manipulate datasets within Dataiku using the same methods.](https://play.vidyard.com/cin5EsHqVYo2hCkz1gBxjz?v=3.1.1&type=inline&hubspot_id=a568ca4b5a2b8d791ca438d8e8c2671a&referring_url=https%253A%252F%252Facademy.dataiku.com%252Fbasics-101%252F497926&)

[Indeed, the dataset interface includes:](https://play.vidyard.com/cin5EsHqVYo2hCkz1gBxjz?v=3.1.1&type=inline&hubspot_id=a568ca4b5a2b8d791ca438d8e8c2671a&referring_url=https%253A%252F%252Facademy.dataiku.com%252Fbasics-101%252F497926&)

* [An Explore tab for investigating the dataset](https://play.vidyard.com/cin5EsHqVYo2hCkz1gBxjz?v=3.1.1&type=inline&hubspot_id=a568ca4b5a2b8d791ca438d8e8c2671a&referring_url=https%253A%252F%252Facademy.dataiku.com%252Fbasics-101%252F497926&)
* [A Charts tab for visualization](https://play.vidyard.com/cin5EsHqVYo2hCkz1gBxjz?v=3.1.1&type=inline&hubspot_id=a568ca4b5a2b8d791ca438d8e8c2671a&referring_url=https%253A%252F%252Facademy.dataiku.com%252Fbasics-101%252F497926&)
* [A Statistics tab for in-depth statistical reports](https://play.vidyard.com/cin5EsHqVYo2hCkz1gBxjz?v=3.1.1&type=inline&hubspot_id=a568ca4b5a2b8d791ca438d8e8c2671a&referring_url=https%253A%252F%252Facademy.dataiku.com%252Fbasics-101%252F497926&)
* [A Data Quality tab for establishing rules](https://play.vidyard.com/cin5EsHqVYo2hCkz1gBxjz?v=3.1.1&type=inline&hubspot_id=a568ca4b5a2b8d791ca438d8e8c2671a&referring_url=https%253A%252F%252Facademy.dataiku.com%252Fbasics-101%252F497926&)
* [A Metrics tab for tracking important measurements](https://play.vidyard.com/cin5EsHqVYo2hCkz1gBxjz?v=3.1.1&type=inline&hubspot_id=a568ca4b5a2b8d791ca438d8e8c2671a&referring_url=https%253A%252F%252Facademy.dataiku.com%252Fbasics-101%252F497926&)
* [A History tab for following the dataset history (creation date, commits, etc.)](https://play.vidyard.com/cin5EsHqVYo2hCkz1gBxjz?v=3.1.1&type=inline&hubspot_id=a568ca4b5a2b8d791ca438d8e8c2671a&referring_url=https%253A%252F%252Facademy.dataiku.com%252Fbasics-101%252F497926&)
* [A Settings tab including details about the source of the dataset, either the underlying connection or the original files that were uploaded](https://play.vidyard.com/cin5EsHqVYo2hCkz1gBxjz?v=3.1.1&type=inline&hubspot_id=a568ca4b5a2b8d791ca438d8e8c2671a&referring_url=https%253A%252F%252Facademy.dataiku.com%252Fbasics-101%252F497926&)
* [The same sets of visual, code, and plugin recipes](https://play.vidyard.com/cin5EsHqVYo2hCkz1gBxjz?v=3.1.1&type=inline&hubspot_id=a568ca4b5a2b8d791ca438d8e8c2671a&referring_url=https%253A%252F%252Facademy.dataiku.com%252Fbasics-101%252F497926&)

[This is possible because Dataiku decouples data processing logic (such as recipes in the Flow) from the underlying storage infrastructure of a dataset.](https://play.vidyard.com/cin5EsHqVYo2hCkz1gBxjz?v=3.1.1&type=inline&hubspot_id=a568ca4b5a2b8d791ca438d8e8c2671a&referring_url=https%253A%252F%252Facademy.dataiku.com%252Fbasics-101%252F497926&)

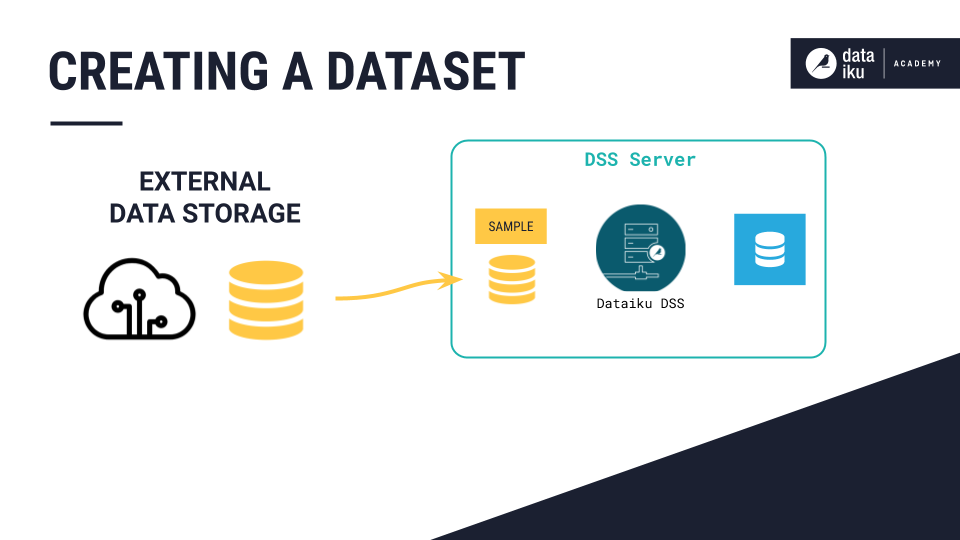
[[](https://academy-content.dataiku.com/latest/_images/dataset-how.png)](https://play.vidyard.com/cin5EsHqVYo2hCkz1gBxjz?v=3.1.1&type=inline&hubspot_id=a568ca4b5a2b8d791ca438d8e8c2671a&referring_url=https%253A%252F%252Facademy.dataiku.com%252Fbasics-101%252F497926&)

**[Connections to the data](https://play.vidyard.com/cin5EsHqVYo2hCkz1gBxjz?v=3.1.1&type=inline&hubspot_id=a568ca4b5a2b8d791ca438d8e8c2671a&referring_url=https%253A%252F%252Facademy.dataiku.com%252Fbasics-101%252F497926&)**

[With the exception of directly uploading files to Dataiku, the Dataiku server does not need to ingest the entire dataset to create its representation in Dataiku.](https://play.vidyard.com/cin5EsHqVYo2hCkz1gBxjz?v=3.1.1&type=inline&hubspot_id=a568ca4b5a2b8d791ca438d8e8c2671a&referring_url=https%253A%252F%252Facademy.dataiku.com%252Fbasics-101%252F497926&)

[Generally, creating a dataset in Dataiku means that the user merely informs Dataiku of how it can access the data from a particular connection.](https://play.vidyard.com/cin5EsHqVYo2hCkz1gBxjz?v=3.1.1&type=inline&hubspot_id=a568ca4b5a2b8d791ca438d8e8c2671a&referring_url=https%253A%252F%252Facademy.dataiku.com%252Fbasics-101%252F497926&)

[Dataiku remembers the location of the original external or source datasets. The data is not copied into Dataiku. Rather, the dataset in Dataiku is a view of the data in the original system. Only a sample of the data, as configured by the user, is transferred via the browser.](https://play.vidyard.com/cin5EsHqVYo2hCkz1gBxjz?v=3.1.1&type=inline&hubspot_id=a568ca4b5a2b8d791ca438d8e8c2671a&referring_url=https%253A%252F%252Facademy.dataiku.com%252Fbasics-101%252F497926&)

[[](https://academy-content.dataiku.com/latest/_images/dataset-sample.png)](https://play.vidyard.com/cin5EsHqVYo2hCkz1gBxjz?v=3.1.1&type=inline&hubspot_id=a568ca4b5a2b8d791ca438d8e8c2671a&referring_url=https%253A%252F%252Facademy.dataiku.com%252Fbasics-101%252F497926&)

## Concept | Data connections

Dataiku lets you change, shape, and analyze your data through a variety of actions. It also enables you to perform these manipulations on externally stored datasets through [connections](https://doc.dataiku.com/dss/latest/connecting/index.html).

See also

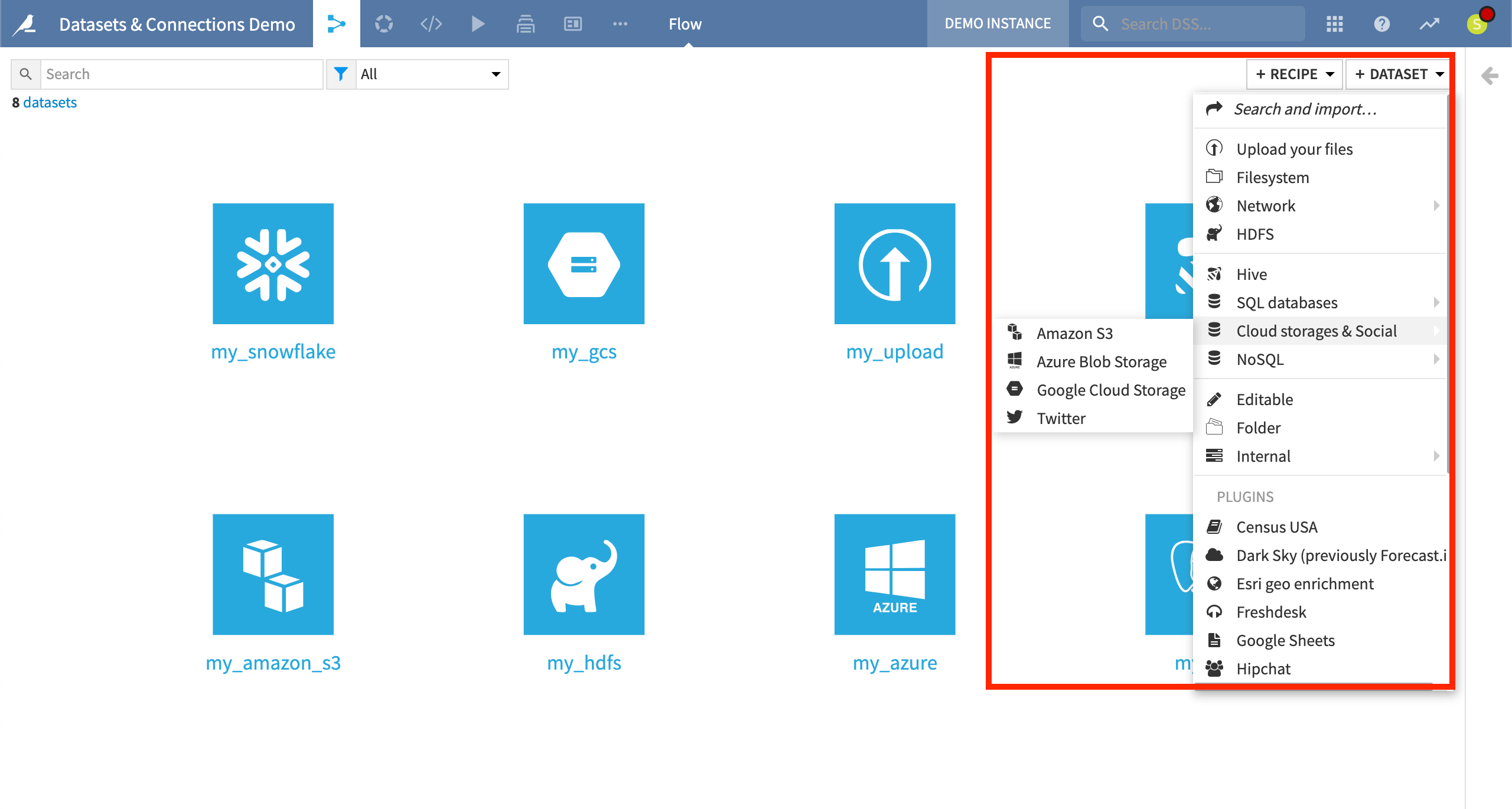
As a developer, for more information on connections, see the following articles in the developer guide:

* [Connections](https://developer.dataiku.com/latest/concepts-and-examples/connections.html) (Concept and examples)
* [Connections](https://developer.dataiku.com/latest/api-reference/python/connections.html) (Python API reference)

Let’s see how Dataiku manages connections to make this possible

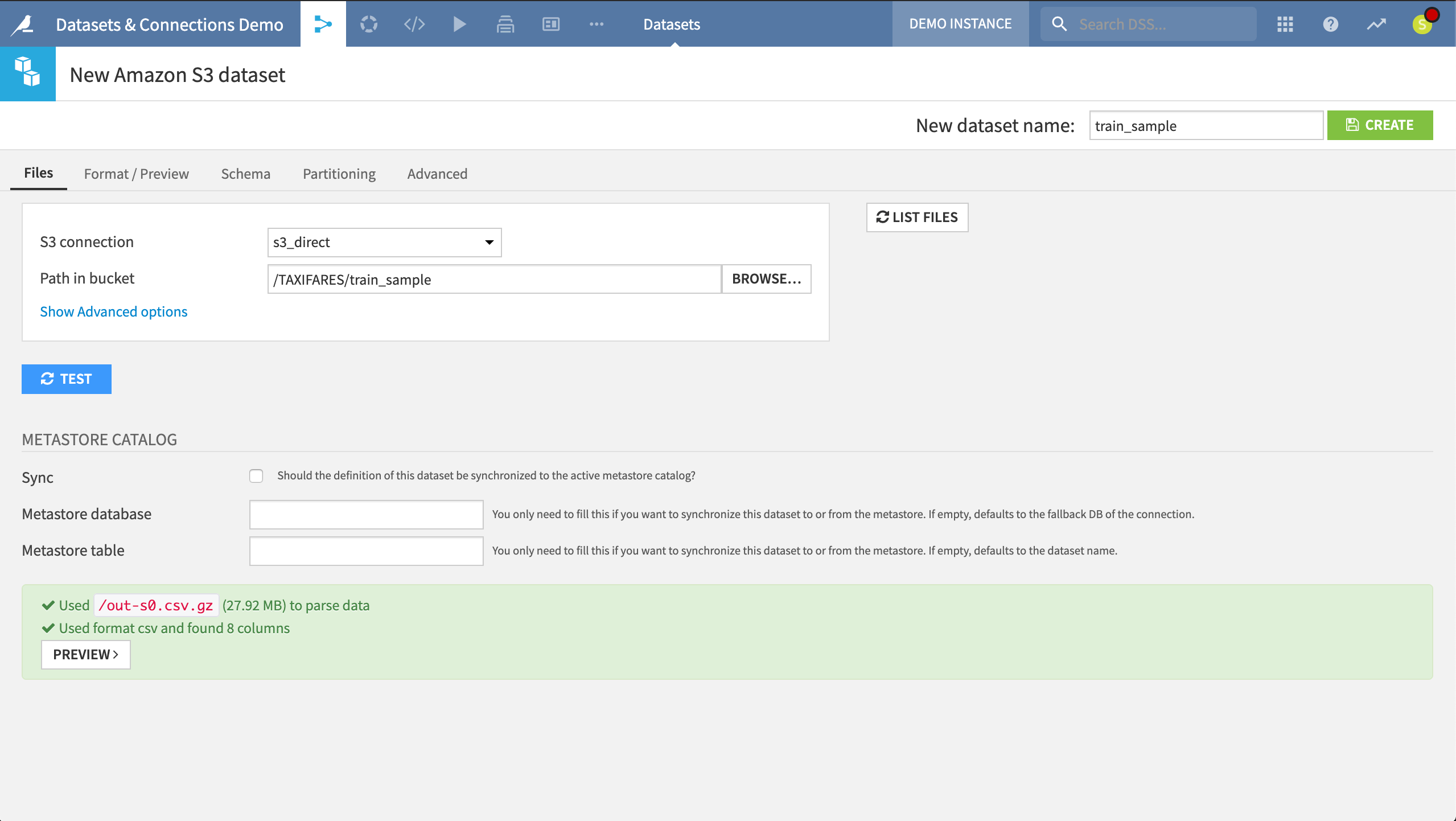
**Importing a new dataset**

You can import a new dataset in the Flow by uploading your own files or accessing data through any previously established connections, such as SQL databases, cloud storage, or NoSQL sources. You might also have plugins allowing you to import data from other non-native sources.

[](https://academy-content.dataiku.com/latest/_images/connections-import.png)

While importing a dataset, you can browse connections and available file paths, and preview the dataset and its schema.

Once you have done that, the user interface for exploring, visualizing, and preparing the data is the same for all kinds of datasets. This is because the processing logic that acts upon a dataset is decoupled from its underlying storage infrastructure.

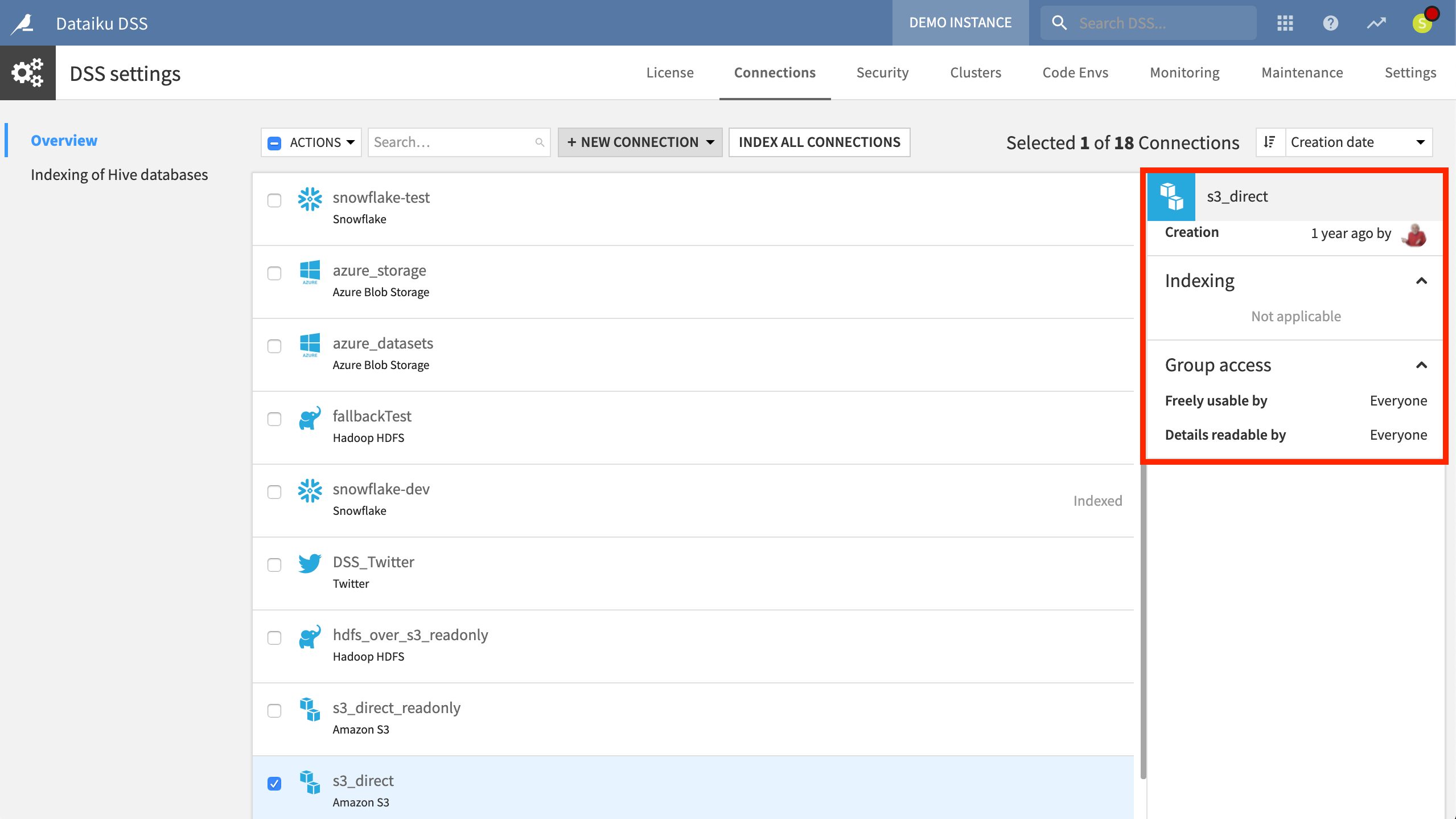
[](https://academy-content.dataiku.com/latest/_images/connections-new.png)

**Managing connections**

Instance administrators can configure connections in:

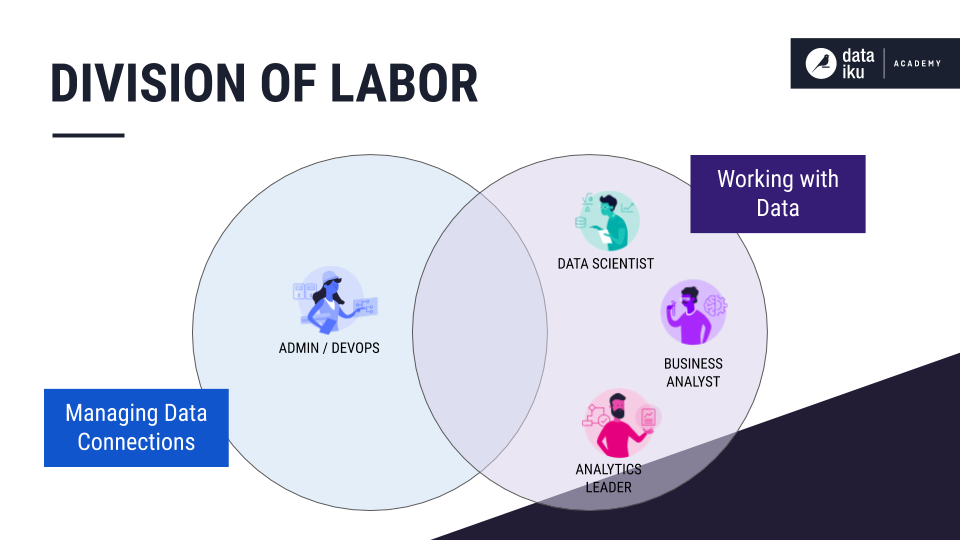
* The **Connections** menu from the [Launchpad](https://academy-content.dataiku.com/latest/cloud-space-management/launchpad/concept-launchpad.html?opals=true) of Dataiku Cloud.
* The **Applications > Administration > Connections** menu from the top navigation bar in a self-managed Dataiku installation.

From here, they can control settings such as credentials, security settings, naming rules, and usage parameters. Admins can also establish new connections to SQL and NoSQL databases, cloud storage, and other sources. Many additional connection types are available in the plugin store for any non-native connections.

[](https://academy-content.dataiku.com/latest/_images/connections-mgmt.png)

One benefit of this system is a clearer division of labor between those who manage data connections and those who work with data.

While understanding a dataset’s storage is often beneficial (particularly with large datasets), those working with data do not always necessarily need expertise in how their organization warehouses its data.

[](https://academy-content.dataiku.com/latest/_images/connections-labor-div.png)

# Create the dataset

Once you’ve created a project, let’s upload our first dataset!

## Objectives

In this section, you will:

* Upload a local file to Dataiku.

## Prerequisites

To complete this tutorial, you must download the [orders CSV file](https://downloads.dataiku.com/public/website-additional-assets/data/orders.csv).

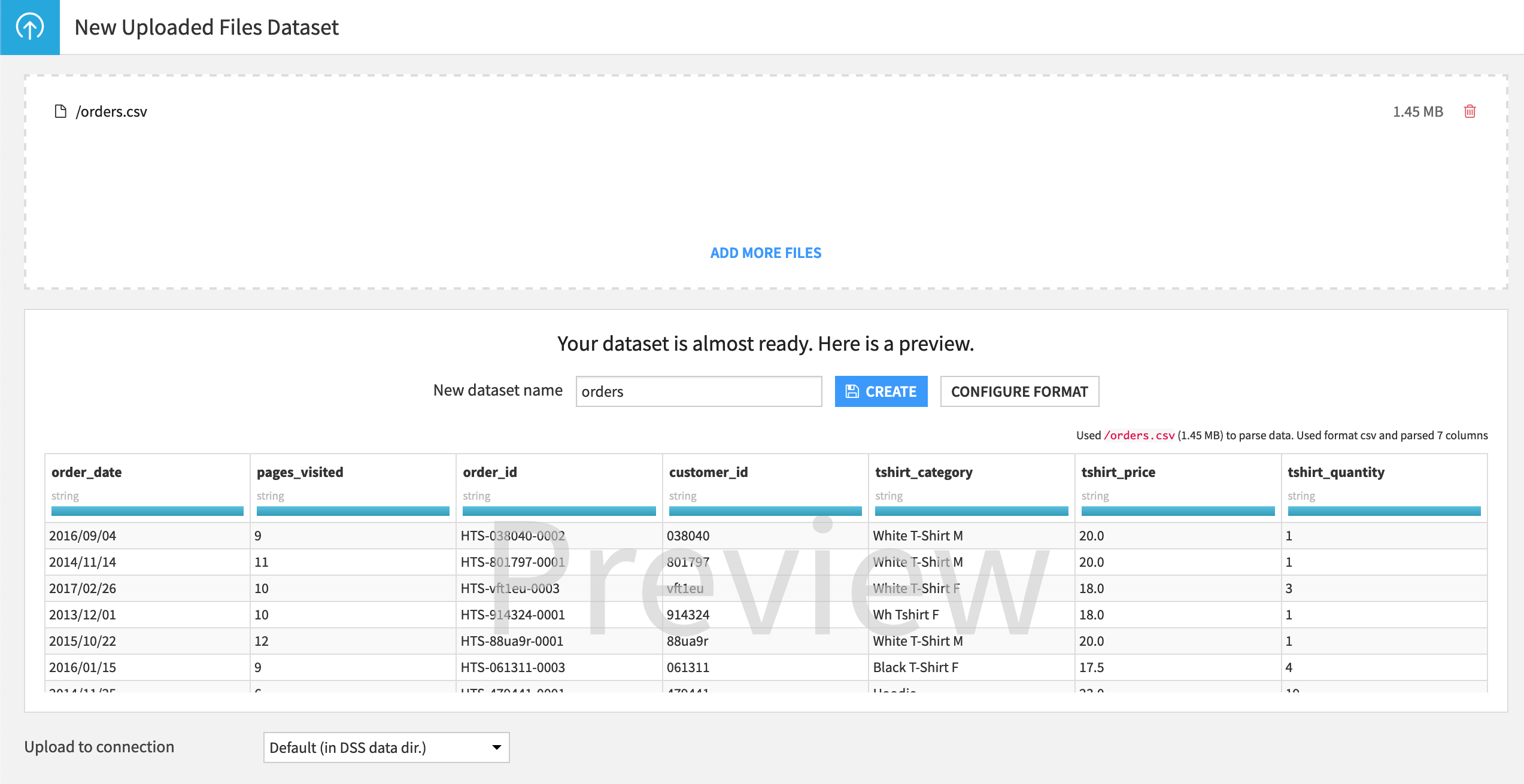
## Import data

Dataiku lets you connect to a wide variety of data sources, but for this tutorial, let’s start by uploading a local file.

1. From the project homepage or the Flow (G+F), click the blue button **+ Import Your First Dataset**.
2. Click on **Upload your files**.

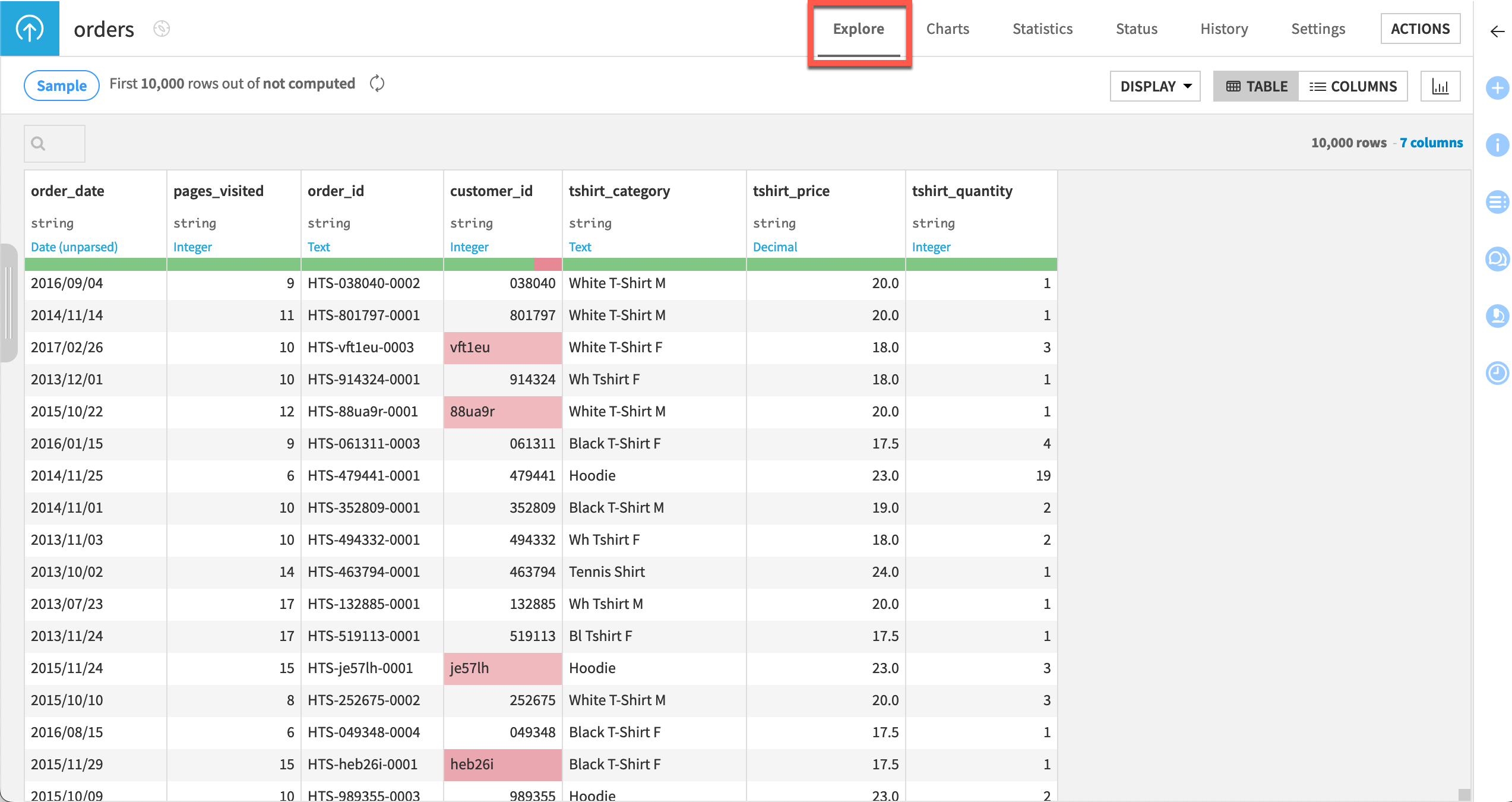
The **New Uploaded Files Dataset** page opens.

1. Click on **Select Files**, and choose the [orders.csv](https://downloads.dataiku.com/public/website-additional-assets/data/orders.csv) file.
2. Review the dataset preview to make sure Dataiku detected the CSV format correctly. As you can see, the data is in a tabular format, with columns (features) and rows (records or observations). In this case, Dataiku has correctly formatted the data and set a default dataset name *orders* based on the file name.

[](https://academy-content.dataiku.com/latest/_images/tshirt-orders-preview.png)

1. Since that’s OK for us, finish importing the dataset by either hitting the **Create** button or using the shortcut **Cmd/Ctrl+S**.

This procedure creates the new dataset and lands you on the **Explore** tab of the *orders* dataset.

[](https://academy-content.dataiku.com/latest/_images/orders-explore-tab.png)

**What’s next?**

Congratulations! Now that you have imported your first dataset, the next step in our sequence of tutorials is to begin visually exploring it.

Explore Your Data

### Concept | Dataset Characteristics

[In Dataiku, columns in datasets are characterized by two main labels: storage type and meaning. Here’s the distinction between these two concepts:](https://academy.dataiku.com/basics-101/497933)

**[Storage Type](https://academy.dataiku.com/basics-101/497933)**

**[Definition:](https://academy.dataiku.com/basics-101/497933)** [The storage type of a column specifies how the dataset backend should store the data in that column. It determines the format and memory allocation for the values.](https://academy.dataiku.com/basics-101/497933)

**[Types:](https://academy.dataiku.com/basics-101/497933)** [Common storage types include:](https://academy.dataiku.com/basics-101/497933)

* **[String:](https://academy.dataiku.com/basics-101/497933)** [Textual data.](https://academy.dataiku.com/basics-101/497933)
* **[Integer:](https://academy.dataiku.com/basics-101/497933)** [Whole numbers.](https://academy.dataiku.com/basics-101/497933)
* **[Float:](https://academy.dataiku.com/basics-101/497933)** [Numbers with decimal points.](https://academy.dataiku.com/basics-101/497933)
* **[Boolean:](https://academy.dataiku.com/basics-101/497933)** [True/False values.](https://academy.dataiku.com/basics-101/497933)
* **[Date:](https://academy.dataiku.com/basics-101/497933)** [Date values.](https://academy.dataiku.com/basics-101/497933)

**[Importance:](https://academy.dataiku.com/basics-101/497933)**

* **[Data Transformations:](https://academy.dataiku.com/basics-101/497933)** [The storage type influences how you can manipulate the data, such as during joins where the key columns must match in storage type.](https://academy.dataiku.com/basics-101/497933)
* **[Formula Language:](https://academy.dataiku.com/basics-101/497933)** [Certain operations may be restricted based on the storage type (e.g., date operations).](https://academy.dataiku.com/basics-101/497933)

**[Note:](https://academy.dataiku.com/basics-101/497933)** [Storage types are typically defined automatically when importing data from sources like SQL tables, and they should generally not be changed unless necessary.](https://academy.dataiku.com/basics-101/497933)

**[Meaning](https://academy.dataiku.com/basics-101/497933)**

**[Definition:](https://academy.dataiku.com/basics-101/497933)** [The meaning of a column provides a semantic label that describes the type of data or its intended use. It's inferred by Dataiku based on the values within the column.](https://academy.dataiku.com/basics-101/497933)

**[Examples:](https://academy.dataiku.com/basics-101/497933)** [Meanings can include descriptors such as country, email address, text, array, temperature, etc.](https://academy.dataiku.com/basics-101/497933)

**[Usage:](https://academy.dataiku.com/basics-101/497933)**

* **[Column Transformations:](https://academy.dataiku.com/basics-101/497933)** [Meanings can suggest transformations that might be applicable to the data. For instance, a column labeled with "Date" might offer options to compute time differences or filter based on dates.](https://academy.dataiku.com/basics-101/497933)
* **[Data Quality:](https://academy.dataiku.com/basics-101/497933)** [Dataiku can assess data quality based on the meaning, determining if values fit the expected semantics (e.g., valid email addresses).](https://academy.dataiku.com/basics-101/497933)
* **[User Interface:](https://academy.dataiku.com/basics-101/497933)** [Meanings also affect the context-sensitive menus in Dataiku, providing relevant transformation options under different recipes like Prepare.](https://academy.dataiku.com/basics-101/497933)

**[Flexibility:](https://academy.dataiku.com/basics-101/497933)** [Meanings are more flexible than storage types and can be adjusted or manually set to better reflect the actual content of the column. For example, a column initially labeled as "Integer" might have some text values, prompting a change to "Text" to encompass all data types present.](https://academy.dataiku.com/basics-101/497933)

**[Dataset Schema](https://academy.dataiku.com/basics-101/497933)**

**[Definition:](https://academy.dataiku.com/basics-101/497933)** [The schema of a dataset lists all its column names along with their associated storage types. It provides a structured view of the dataset's structure.](https://academy.dataiku.com/basics-101/497933)

**[Management:](https://academy.dataiku.com/basics-101/497933)** [You can view and modify the schema within Dataiku, either during dataset creation or later in your workflow. Schema changes affect how downstream processes interpret and handle data.](https://academy.dataiku.com/basics-101/497933)

**[Note:](https://academy.dataiku.com/basics-101/497933)** [Changes made to the schema within Dataiku (e.g., adding or modifying columns, changing types) require running subsequent recipes or processes to propagate these changes downstream.](https://academy.dataiku.com/basics-101/497933)

[In summary, while storage types define how data is physically stored and processed, meanings provide a semantic layer that enhances data understanding and usability within Dataiku's environment. Both concepts are crucial for effective data management, transformation, and analysis in Dataiku.](https://academy.dataiku.com/basics-101/497933)

# [Concept | Sampling on datasets](https://academy.dataiku.com/basics-101/497933)

Sampling in Dataiku serves a crucial role in exploring and preparing datasets, especially when dealing with large volumes of data. Here’s an overview of sampling benefits, methods, and considerations within Dataiku:

**Sampling Benefits**

1. **Efficiency:** Sampling allows you to work with a subset of data, making operations like sorting, filtering, and visualization much faster. This is particularly useful when dealing with datasets that are too large to handle in their entirety efficiently.
2. **Immediate Feedback:** By displaying a sample, Dataiku provides immediate visual feedback on the dataset’s characteristics and distributions. This helps in quickly gaining insights and making initial data exploration easier.
3. **Resource Conservation:** Sampling conserves computational resources and reduces processing time compared to working with the entire dataset. It optimizes performance, especially in interactive data exploration scenarios.

**Sampling Methods**

**Default Sampling:**

* Dataiku defaults to sampling the first 10,000 rows of a dataset for exploration and preparation tasks. This method is quick but may introduce bias if the dataset's initial rows do not represent the overall distribution.

**Other Sampling Methods:**

* **Random Sampling:** Selects rows randomly from the dataset, ensuring a more representative sample than the default method.
* **Stratified Sampling:** Ensures that the sample maintains the proportion of key variables (strata) as in the full dataset, useful for maintaining representation across categories.
* **Class Rebalancing:** Particularly useful in imbalanced datasets (e.g., fraud detection), where the sample can be adjusted to reflect a better balance of classes.

**Adjusting Sampling:**

* You can customize the sampling method and sample size in Dataiku by clicking on the "Sample" button on the Explore tab of a dataset. This opens the Sampling settings panel where you can specify the method (random, stratified, etc.) and increase the sample size if needed.

**Considerations**

* **Bias vs. Representativeness:** While default sampling is quick, it may not be representative of the entire dataset. Choosing a sampling method should balance between speed and the need for a representative sample.
* **Performance Impact:** More complex sampling methods (e.g., stratified) may require additional processing time as Dataiku might need to pass through the dataset multiple times to ensure accuracy.
* **Sampling Scope:** Sampling impacts not only data exploration but also affects subsequent data preparation, modeling, and analysis steps. Ensure that the sample size and method chosen align with the goals of your analysis.

In conclusion, sampling in Dataiku optimizes data exploration by providing quick insights into large datasets while managing computational resources effectively. Choosing the right sampling method depends on the dataset's characteristics and analysis objectives, balancing between speed and representativeness.

# Concepts | Analyze Window

In Dataiku's Explore tab, the Analyze window provides powerful tools to assess data quality and gain insights into individual columns within your dataset. Here’s an overview of how you can analyze column data using the Analyze window:

**Accessing the Analyze Window**

1. **Navigate to the Explore Tab:** Start by selecting your dataset in Dataiku and navigating to the Explore tab.
2. **Select a Column:** Click on the header of any column you wish to analyze. This will bring up a context menu where you can choose "Analyze."
3. **Analyze Window Overview:** The Analyze window opens with several key features:

**Elements of the Analyze Window**

1. **Navigation Arrows:** Located at the top left corner, these arrows allow you to quickly move between adjacent columns to view similar metrics calculated for other columns.
2. **Sample Management:** By default, Dataiku calculates statistics based on the current sample configured in the Explore tab. However, you have the option to compute these statistics on the entire dataset for more comprehensive insights.
3. **Tabs:**
   * **Categorical Tab:** For categorical columns, this tab displays a bar chart sorted by the most frequent observations. It helps visualize the distribution and frequency of different categories.
   * **Numerical Tab:** For numeric columns, this tab provides a histogram and boxplot of the distribution. It also includes summary statistics such as mean, median, min, max, and standard deviation. Detected outliers are highlighted as well.
4. **Summary Section:** Regardless of the tab (Categorical or Numerical), a Summary section is available. This section allows you to review the data quality of the selected column:
   * **Valid Values:** Number of valid entries that conform to the expected data type and format.
   * **Invalid Values:** Number of entries that do not meet the expected format or are outliers.
   * **Empty Values:** Number of entries that are blank or null.
   * **Unique Values:** Count of distinct values present in the column.
   * **Singletons:** Values that appear only once in the dataset.
5. **Values Clustering:** This feature allows you to identify groups of similar values, particularly useful for standardizing text fields with variations or discrepancies.

**Using the Analyze Window**

* **Data Exploration:** Quickly gain insights into the distribution, quality, and characteristics of each column.
* **Data Quality Assessment:** Identify potential data issues such as missing values, outliers, and unusual distributions.
* **Visualization:** Utilize charts (bar charts, histograms, and boxplots) to visualize data distributions and patterns.

**Conclusion**

The Analyze window in Dataiku’s Explore tab is a comprehensive tool for exploring and evaluating individual columns within datasets. It provides essential metrics, visualizations, and summaries that facilitate thorough data quality assessment and exploration, aiding in data preparation and decision-making processes.

# Concept | Charts

In Dataiku, the Charts tab provides a robust platform for visualizing and exploring data through an intuitive drag-and-drop interface. Here's a detailed overview of how you can create, customize, and refine charts within Dataiku:

**Creating a Chart**

1. **Select Chart Type:** Begin by choosing a chart type such as bar chart, line graph, pivot table, or scatter plot from the available options.
2. **Drag Variables:** Drag variables from the Data panel on the left onto the desired axis (X-axis, Y-axis, etc.) to define how data will be represented in the chart.

**Customizing a Chart**

1. **Configuring Variables:**
   * **Aggregation:** Choose the appropriate aggregation method for each variable, such as sum, count, average, etc. This determines how data values are summarized.
   * **Sorting:** Adjust the sorting order of values within the chart to highlight specific trends or patterns.
   * **Formatting:** Customize the appearance of values and labels, affecting tooltips and legends for clarity and context.

**Advanced Features in Chart Builder**

* **Time Series Features:** For time-based data, utilize features like zooming into different periods, adjusting aggregated date intervals, exploring multiple series simultaneously, comparing them in subcharts, or creating basic animations.
* **Filtering Data:** Manage large categorical datasets by selecting specific values for display or grouping less prevalent categories into an "Others" bucket. Drill down into data subsets by clicking on filter icons within tooltips, applying changes directly from the dropdown menu.

**Sampling and Execution Engine Configuration**

* **Sampling & Engine Panel:** Located on the left sidebar, this panel allows you to configure:
  + **Sampling:** By default, charts use the same sample as configured in the Explore tab for consistency. You can adjust the sampling method or sample size as needed.
  + **Execution Engine:** Dataiku defaults to its own processing engine (DSS engine), but you can opt for other engines, especially beneficial for data sources that support SQL queries for enhanced performance.

**Conclusion**

Dataiku's Charts tab offers a versatile environment for visual exploration and analysis of datasets. It empowers users to create various types of charts effortlessly, customize them extensively, and leverage advanced features such as time series analysis and data filtering. With options for sampling and execution engine selection, Dataiku ensures flexibility and efficiency in data visualization, supporting comprehensive data-driven insights and decision-making processes.

# Tutorial | Explore Your Data

**1. Dataset Sampling**

* Purpose: Dataiku displays only a sample of large datasets for interactive work.
* Default Sampling: Shows the first 10,000 records by default.
* Viewing Total Row Count: Click on "Compute row count" (cyclic arrows icon).
* Changing Sample Settings: Click on the "Sample" badge to adjust settings.

**2. Storage Type and Column Meaning**

* Displayed Information:
  + Storage type (in black).
  + Meaning (in blue), indicating the inferred type (e.g., integer, text).
* Adjusting Column Meaning:
  + Click on the meaning to update (e.g., from integer to text).
  + Helps manage data quality, flagged by color (red for inconsistencies, green when resolved).
* Handling Missing Values: Gray indicates missing values if present.

**3. Creating Charts**

* Purpose: Visualize dataset insights.
* Accessing Charts: Click on the "Charts" tab or use shortcut G+V.
* Creating a Chart:
  + Drag and drop variables:
    - Y variable: Count of records.
    - X variable: Column to categorize (e.g., tshirt\_category).
  + Displays a chart (e.g., column chart) showing distribution based on the sample data.
* Data Consistency: Note inconsistencies in data entry (e.g., "Black" vs. "Bl" for tshirt categories).

**Next Steps**

* Congratulations on completing initial dataset import and exploration!
* Further learning in Basics 102 includes preparing datasets with recipes to address issues such as inconsistent data entries.

**References**

* Detailed guidance available in Dataiku's documentation on sampling, column meanings, and chart creation.

Top of Form

Bottom of Form