Quick Start Guide

Quick Start Guide

Overview

The Quick Start Guide is an abbreviated version of the Import User Guide. This document is comprised of high-level information, specific to the Data Operations Center (DOC), which will allow you to begin importing data and navigating this environment with quickness and ease. From adding Organizations and Projects to configuring Flows, you will gain basic insight into essential topics that will help you understand how to deploy data and, ultimately, send required data to a customer's preferred delivery Destination.

You must access the primary User Guide for in-depth information specific to each layer of the DOC environment, as the content here simply provides a high-level glimpse of the platform which will allow you a respectable jump start.

This document addresses:

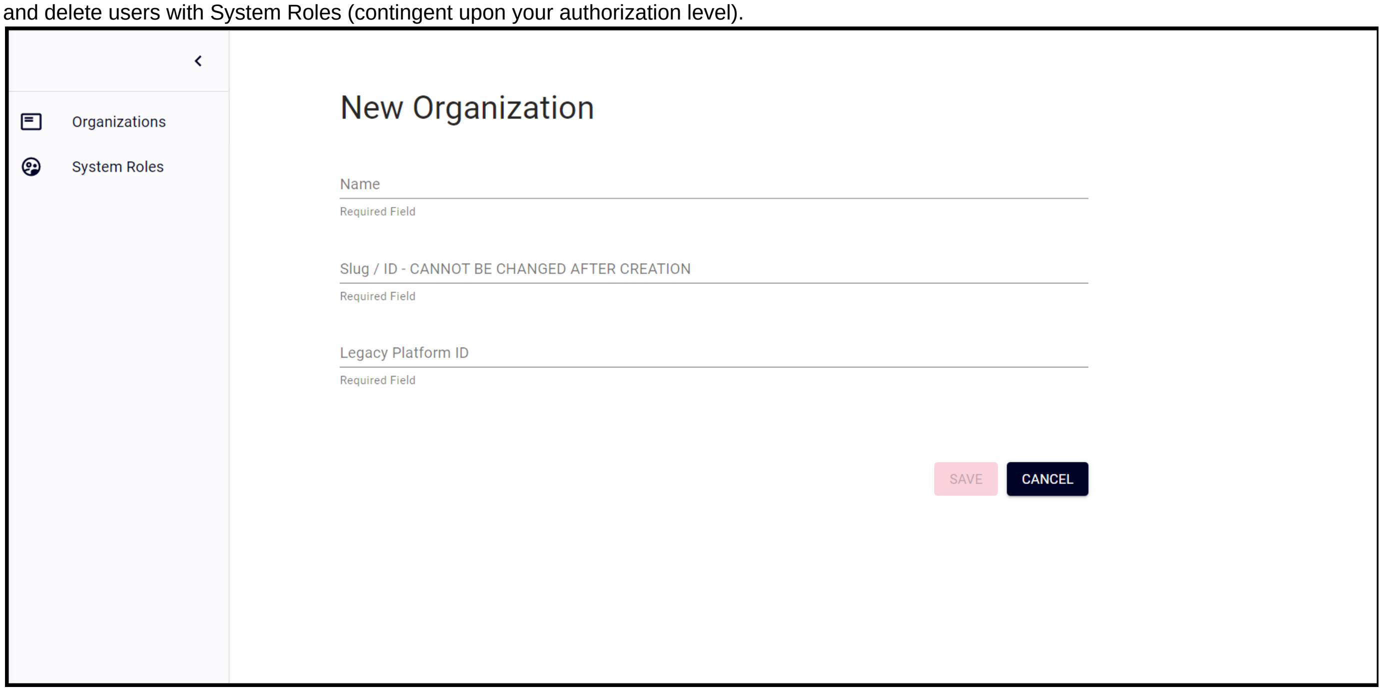
- Organizations
- Projects
- Schemas
- Destinations
- Collections
- Sources
- Snapshots
- Flows
- Deliveries

Organizations

An Organization is the company or entity requesting the data extraction. An Organization may be comprised of multiple users. As you build Extractors and import data into DOC, you first must define the Organization with which to associate the extracted data.

Certain information is defined at the Organization level. These items are global and impact the entire Organization. At this level, Schemas, Destinations, and Integrations – for example – have a global impact and may be used in different Projects and Collections.

This page contains two navigation pane options: **Organizations** and **System Roles**. The Organizations option allows you to add and view Organizations. The System Roles option allows you to add, edit, and delete users with System Roles (contingent upon your authorization level).



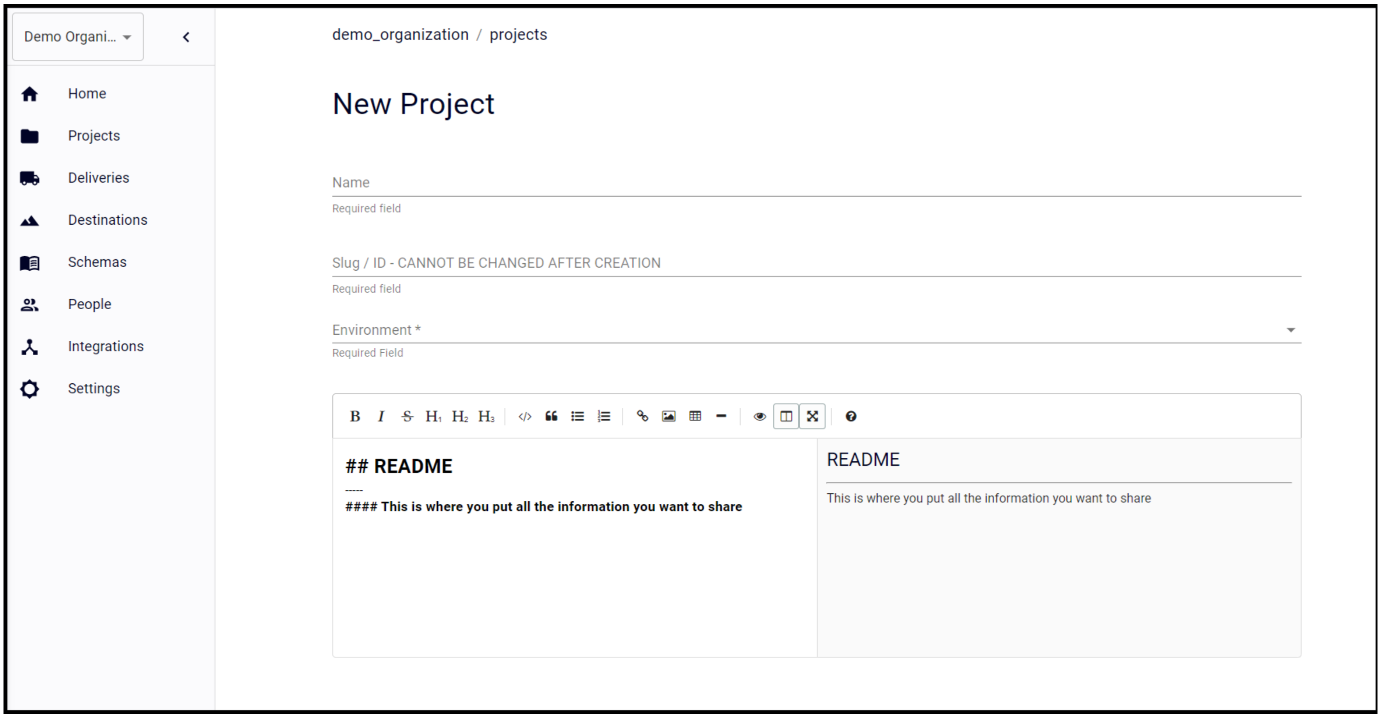
To add an Organization:

1. From the left navigation pane, click **Organizations**.
2. From the top right of the **Organizations** page, click the **Add an Organization** icon or plus (**+**) symbol.
3. From the **New Organization** page, enter text in the **Name** field.

   The name you enter will autofill the **Slug/ID** field. You can associate Slug/IDs with many DOC platform objects, which serve as self-defined identifiers. Slug/IDs can be useful as you reference APIs or create variable names. You cannot change Slug/IDs. That noted, ensure they are meaningful.
4. In the **Legacy Platform ID** field, enter the numeric value that corresponds to the SaaS application account ID.

   You can retrieve this value from the **User ID** field on the **Account Settings** page of the SaaS application, which is accessible via the **Account** icon in the left navigation pane. The value maps to a User ID. This value represents the user account that owns the Extractor for the Organization. Relative to the SaaS application, this value is useful when running Extractors and compiling billing information. The Legacy Platform ID links the SaaS and DOC applications.
5. To store content, click **Save**. To disregard, click **Cancel**.

Projects



A Project represents a group of Collections. These Collections contain Extractors.

As you establish Projects, you can *name* them identically; however, in this case, each Project must be associated with a different Environment (**DEV**, **STAGING**, **PRODUCTION**).

Projects can be locked, restricting write and edit operations under each subsequent DOC layer (such as Collections, Sources, and Snapshots.) You will not have Project access unless you are an ORG_ADMIN, OPS, or assigned to work on a specific layer of this Project. If your role is ORG MEMBER and you are attempting to edit a Source, for example, you cannot make changes unless this Source is assigned to you.

To add a new Project:

1. From the left navigation pane, click **Projects**.
2. From the top right of the **Projects** page, click the **Add a Project** icon or plus (**+**) symbol.
3. From the **New Project** page that now appears, enter text in the **Name** field.

   The name you enter will autofill the **Slug/ID** field. You can associate Slug/IDs with many DOC platform objects, which serve as self-defined identifiers. Slug/IDs can be useful as you reference APIs or create variable names. You cannot change Slug/IDs. That noted, ensure they are meaningful.
4. Use the drop-down arrow to select an **Environment**.

You may choose **DEV**, **STAGING**, or **PRODUCTION**. Two projects can have the same name; however, in this case, their Environments must be different. For now, the Environment references, generally, function as tags or labels.

5. To share Project information with team members, enter **README** text.

   This section, which supports the *markdown* syntax, allows you to provide additional Project context and insight.
6. To store content, click **Save**. To disregard, click **Cancel**.

Schemas

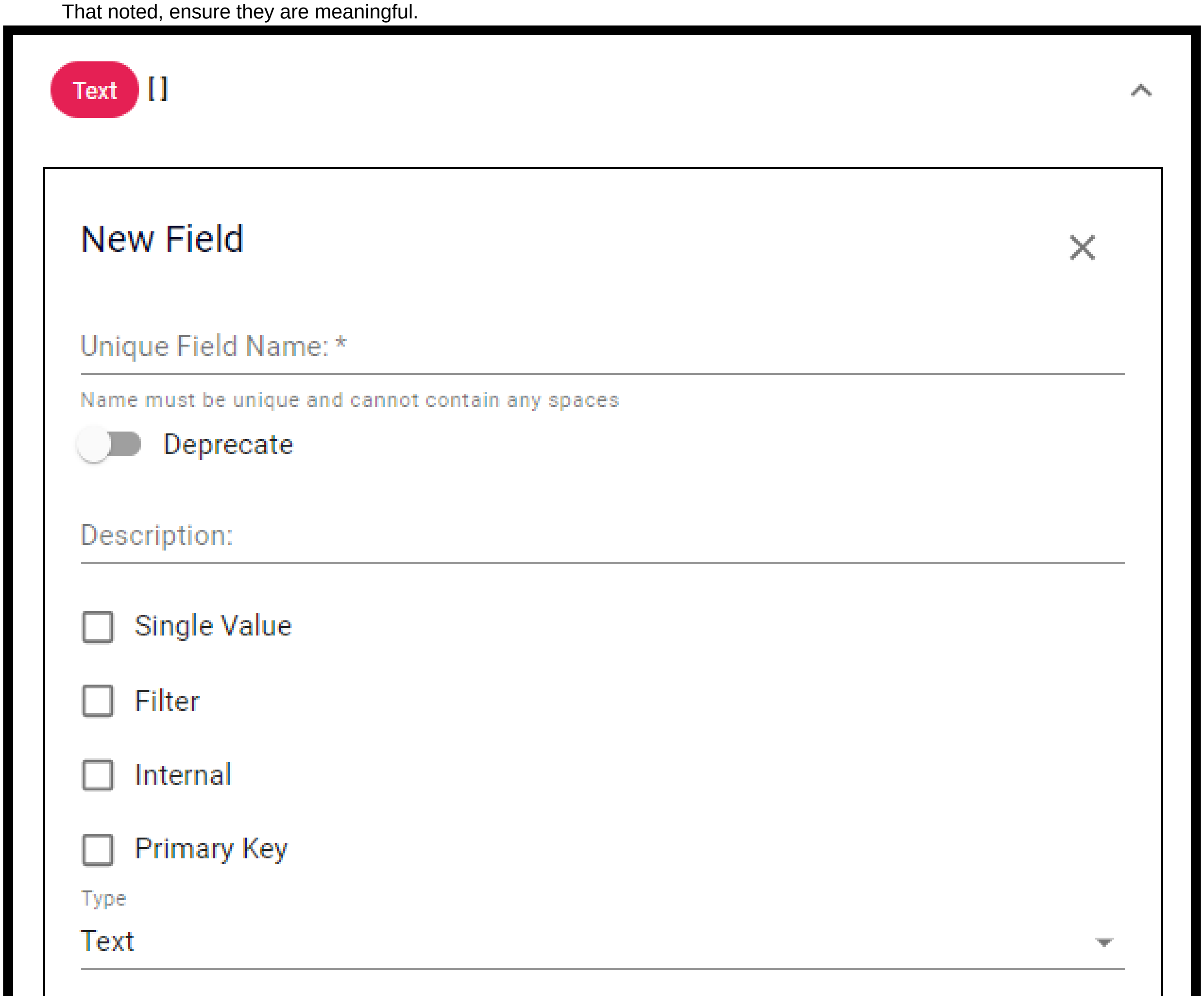
Schemas, essentially, are the Extractor column names. As you define your Schema, the fields that you add and the data types you assign must align with the columns in the related Extractor. The **Schemas** page allows you to construct the shape of the extracted data.

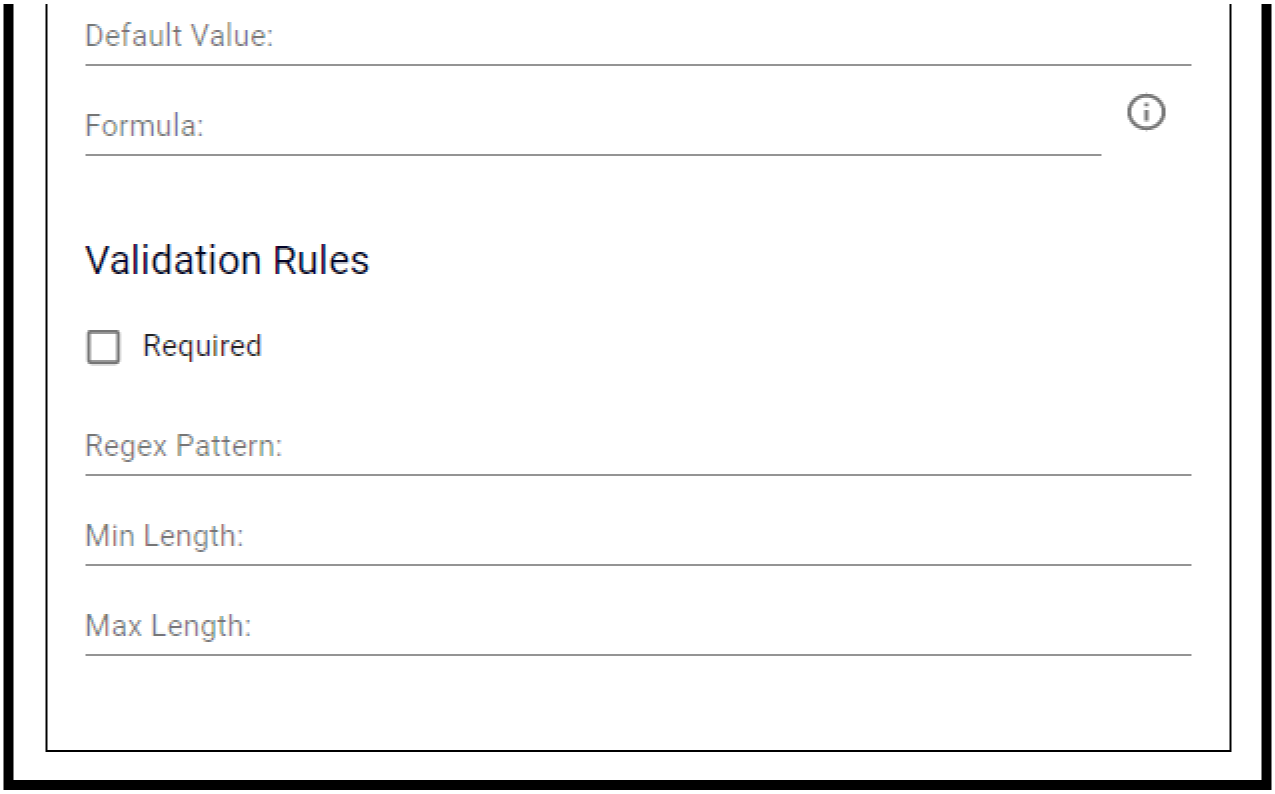
When you initially create a Schema, it is in a **Draft** state. You can make modifications to the Schema while it is in a Draft state. Once you publish the Schema (an act you must perform before data is pushed to the customer Destination), you cannot make any breaking changes. For example, after you publish a Schema, you can add columns; however, you cannot change the data type of a column from text to whole number. This is a breaking change.

To add a new Schema:

1. From the left navigation pane, click **Schemas**.
2. From the top right of the **Schemas** page, click the **Add a Schema** icon or plus (**+**) symbol.
3. From the **New Schema** page that now appears, enter content in the **Name** field.

   The name you enter will autofill the **Slug/ID** field. You can associate Slug/IDs with many DOC platform objects, which serve as self-defined identifiers. Slug/IDs can be useful as you reference APIs or create variable names. You cannot change Slug/IDs. That noted, ensure they are meaningful.





Field	Description
Single Value	Select this checkbox if you must consider whether or not the value is an array.
Filter	If the value is a falsy value (zero, false, blank), mark this row as filtered and do not include it in the data pushed to the customer Destination.
Internal	If this is true, the data is excluded from the data pushed to the customer Destination.
Primary Key	The fields that are part of the composite primary key give the rows the `_id` metadata column – a generated UUID from the hash of the column values. The data pushed to Destinations is deduplicated on this ID.
Type	The type gives the avro/parquet data types and also controls how the system turns extracted text values into typed values. The locale parameter on a Source is used when performing this conversion.
Default Value	A textual default value for the column. Note that this should be in ISO format for date/time and JSON format for numbers and Booleans.
Formula	https://github.com/handsontable/formula.js/tree+ /master/test [Supported functions]. Plus PARAM('name') to get a Source parameter. Plus INPUT('name') to get an input parameter. Plus OUTPUT('name') to get the output value.

Validation Rules

These settings contribute towards the validation error statistics for Snapshots.

Field	Description
Required	Select this checkbox if you want to enact validation rules.

Regex Pattern	RegEX or regular expressions allow you to check a string of characters for matches. You might need to check fields on a form to ensure they meet certain specifications. For example, you might create an expression to check the **Email Address** field to ensure it has an @ symbol and an extension such as .com or .org. In another example, you might create an expression to validate passwords to ensure they satisfy requirements.
Min Length	This field requires that you specify the minimum length of the value associated with the RegEX or regular expression.
Max Length	This field requires that you specify the maximum length of the value associated with the RegEX or regular expression.

WARNING	These rules are soft indicators. They do NOT filter data. You must select the **Filter** checkbox (available in the **Add a Field** view) to invoke this setting.

To add a new Field:

1. Enter a **Field Name**.

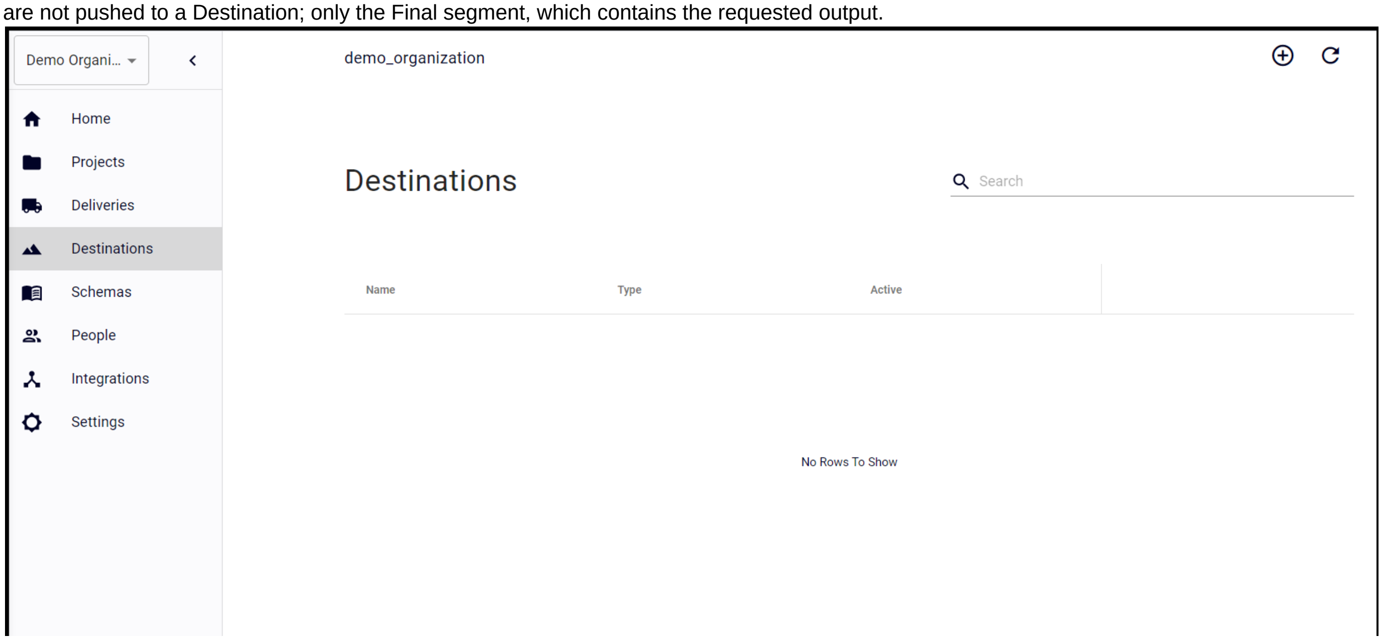
   The name must be unique and cannot contain spaces.
2. Move the **Deprecate** toggle switch to the right or ON position if you want to deem this field inactive.

   The **Deprecate** toggle switch defaults to the left or OFF position, as most fields are active.
3. Enter a **Description**.
4. Select from the **Single Value**, **Filter**, **Internal**, and **Primary Key** checkboxes (as defined above).
5. Enter content in the **Type** field.

   This field denotes a data type. The default entry is **Text**. **Boolean**, **Currency**, **Whole Number**, **Decimal**, **Date**, **Date & Time**, **Image**, **URL**, and **File** represent data types.
6. Enter a **Default Value**.
7. Enter a **Formula** (as defined above).
8. Select the **Required** checkbox if you need to enter **Validation Rules**.
9. Enter a **Regex Pattern** (as defined above).
10. Enter a **Min Length** (as defined above).
11. Enter a **Max Length** (as defined above).

Destinations

A Destination is the location where customer data is delivered. You configure Destinations at the Organization level. While significant and frequent in use, Destinations are not always required. For example, when testing, you might download output as opposed to pushing it to a Destination. In another scenario, with Chained Flows, there are occasions when the first segments are not pushed to a Destination; only the Final segment, which contains the requested output.



To add a Destination:

1. From the left navigation pane at the Organization level, click **Destinations**.
2. From the top right of the **Destinations** page, click the **Add a Destination** icon or plus (**+**) symbol.
3. From the **New Destination** page that now appears, enter text in the **Name** field.
4. Use the drop-down arrow to select a **Type**, choosing from **S3** or **SFTP**.

   An S3 bucket is an AWS service that serves many purposes. In this case, it serves as a storage location. Similarly, the SFTP is another vehicle by which you may transfer and store data.
5. Ignore or clear the **Active** checkbox.

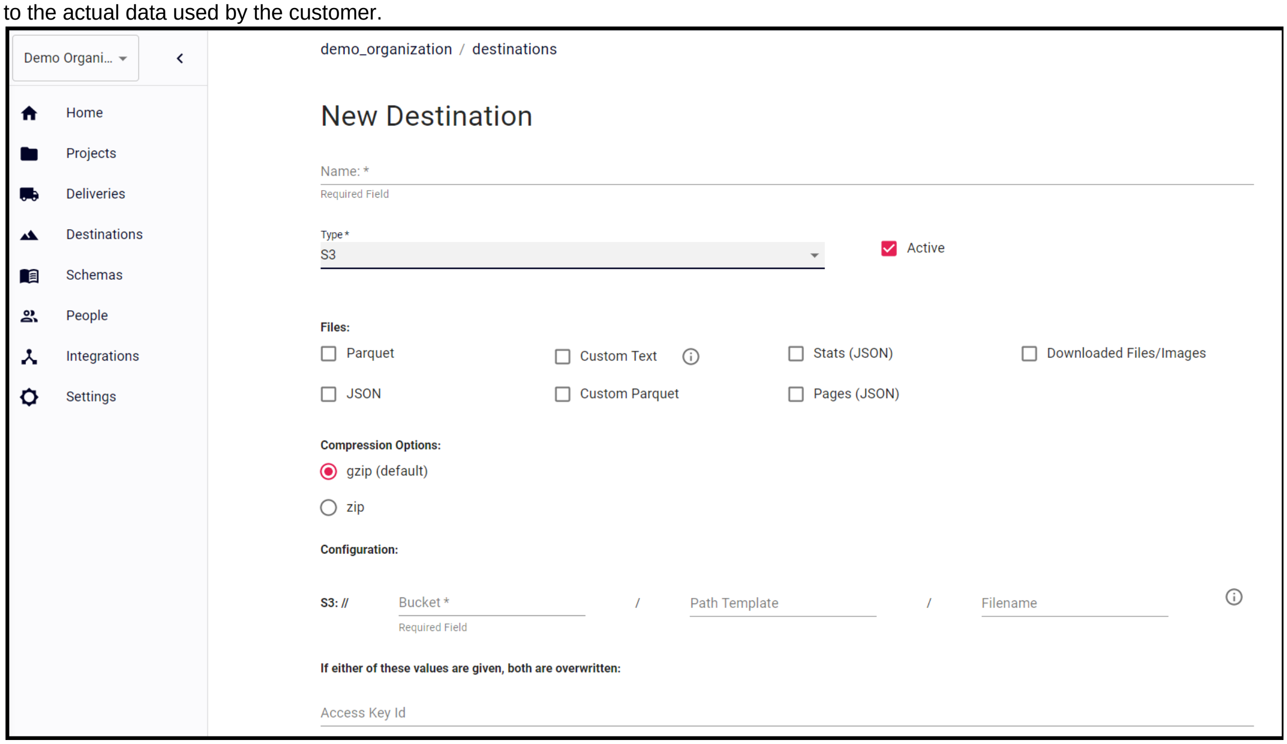
   The **Active** checkbox is selected by default. If you clear this checkbox, no data will be pushed to this Destination. Clearing the **Active** checkbox provides a method to maintain the Destination while *not* pushing data to it. You can render a Destination inactive to avoid any data being pushed to this location.
6. Choose a **File**.

   To further understand the context of Destinations relative to file types, it is important to realize that you can retrieve four kinds of data via an Import.IO web data extraction:

   - The actual dataset
   - Statistics specific to the crawl run and extraction
   - Pages, which represent the web pages that the browser rendered when accessing the websites to extract data.
   - Downloaded files/images (if configured in the Extractor).

     You can configure Import to download actual files that are accessible via a web page; for example, you might extract data from a web page that includes a list of links to PDF documents. You can extract the URL values that point to the PDFs; in addition, you can configure Import to download the actual files (not just extract the URL value of the download link).

     Commonly, customers are only interested in the extracted dataset and, potentially, downloaded files/images (if this is a project requirement). The data specific to *stats* or *pages* is available to monitor the health of the project and used by the developer building the Extractor or the team running the extraction. While useful for development and monitoring of the project, neither stats nor pages are specific to the actual data used by the customer.



Destinations offer different file types for each Snapshot of data. The customer may prefer a specific file type, the Delivery team may make a recommendation, or the two may have a conversation and decide. Generally, a conversation results in a defined or agreed upon file type, which is determined at the start of a project. The file type also might be determined during *Solution Design* by a Solution Architect engaged with the customer.

File Type	Description
Parquet	This is a standard file format and the most common. Parquet has many advantages such as its size, the fact that it includes schema information, the speed at which it will open, and the ability to open partial datasets (subsets) from a parquet file. Data Science teams will often prefer parquet to JSON and our Solution Architects commonly recommend parquet.
Custom Text	This allows you to configure the export of datasets in the .csv and .tsv formats.
JSON	This is a standard file format. JSON is a structured data format, is commonly used by APIs, and is generally the go-to format for structured data for Engineering teams; however, JSON is not favorable for large datasets.
Custom Parquet	This is a standard file format.
Stats (JSON)	This file type allows you to retrieve data about the statistical properties of the extraction to be exported for analysis. While rare for customers, this file type is useful for the internal Managed Services team.
Pages (JSON)	This file type allows you to retrieve data about the HTML pages encountered during the extraction to be exported. While rare for customers, this file type is useful for the internal Development team.
Downloaded Files/Image	This file type allows you to download files/images to be exported. This file type requires that *file download* be configured in the Extractor. This is a relevant user feature.

Compression Options

You can use the radio button to select a **Compression Option**.

gzip, the default selection, is used for single files and has a faster compression than **zip**. In addition, gzip is widely available on *nix machines and is generally preferred; however, some customers only may be able to use zip (which can handle compressing entire directories but is slower). It is the Import practice to send datasets in a compressed format.

Configuration

This section allows you to establish the location where the customer data will be pushed.

S3

You must enter an **S3 Bucket Name**, **Path Template**, and **Filename**. Collectively, this path designates the Destination where the customer data will be pushed or delivered.

1. Enter a **Bucket Name**.

   The Bucket Name represents the AWS S3 cloud storage location. For example, *aws-workbench-assets* represents a bucket name.
2. Enter a **Path Template**.

   The Path Template may represent a folder you created in the AWS S3 environment along with a concatenation of Source Parameter names and template variables. Certain variables require prefixes. For example, *input/:start_YYYY/:source.stage* might represent a Path Template (where **:start_** is the prefix to the *YYYY* date/time variable.)
3. Enter the **Filename**.

The Filename can include template variables and the inferred extension, .ext. For example, *:snapshot_id.:ext* represents the output format and file extension. A combination of the **Bucket Name**, **Path Template**, and **Filename** might appear as follows: workbench-dev-assets **/** input/:start_YYYY/:source.stage **/** :snapshot_id.:ext

4. Enter an **Access Key ID** and a **Secret Access Key**.

   These references, similar to passwords, are your S3 credentials and were likely provided by a member of your IT department. The text you enter will be encrypted.
5. To store content, click **Save**. To disregard, click **Cancel**.

SFTP

The Secure File Transfer Protocol (SFTP) allows you to send and receive Internet files.

1. Enter a **Path Template**.

   The Path Template may represent a folder along with a concatenation of Source Parameter names and template variables.
2. Enter the **Filename**.

   The Filename can include template variables and an inferred extension, .ext. For example, *:snapshot_id.:ext* represents the output format and file extension.
3. Enter a **Host**.

   A Host is a computer device with network accessibility.
4. Enter a **Port**.

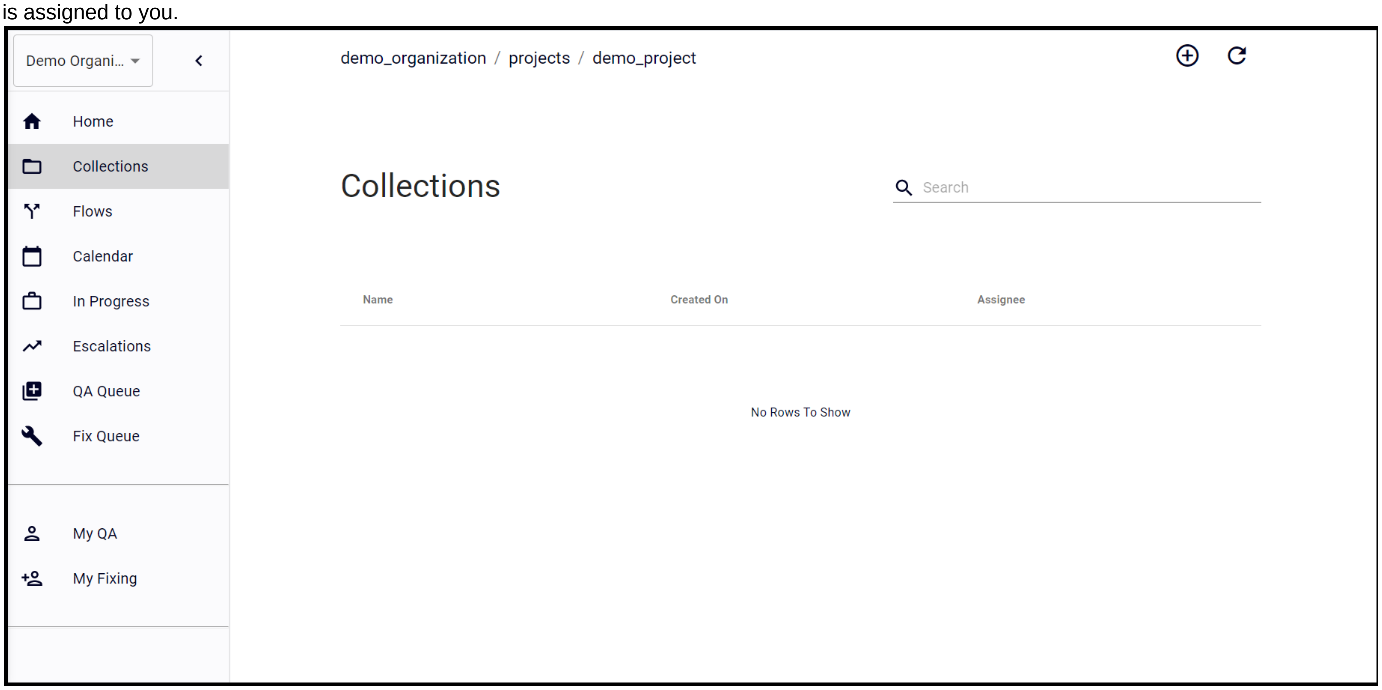
   A Port is a numeric designation that represents a network service or device.
5. Enter a **Username**, **Password**, and an **SSH** Key.

   These references, similar to passwords, are your SFTP credentials.
6. To store content, click **Save**. To disregard, click **Cancel**.

Collections

A Collection is a group of Sources (or Extractors) that adheres to a particular Schema. A Schema represents column names within an Extractor. You cannot create a Collection without a Schema.

Projects can be locked, restricting write and edit operations under each subsequent DOC layer (such as Collections, Sources, and Snapshots.) You will not have Project access unless you are an ORG_ADMIN, OPS, or assigned to work on a specific layer of this Project. If your role is ORG MEMBER and you are attempting to edit a Collection, for example, you cannot make changes unless this Collection is assigned to you.



To add a Collection:

1. From the left navigation pane, click **Collections**.
2. From the top right of the **Collections** page, click the **Add a Collection** icon or plus (**+**) symbol.
3. From the **New Collection** page, enter content in the **Name** field.

   The name you enter will autofill the **Slug/ID** field. You can associate Slug/IDs with many DOC platform objects, which serve as self-defined identifiers. Slug/IDs can be useful as you reference APIs or create variable names. You cannot change Slug/IDs. That noted, ensure they are meaningful.
4. To enter content in the **Schema** field, click the drop-down arrow and select from the list.

   Schemas represent Extractor column names, and you can share them across Organizations. At this point, you have created at least one Schema. It is important to note that when you initially create a Schema, it is in a Draft state. You cannot push data to the customer unless the Schema is in a Published state. Once published, you cannot make breaking changes to the Schema. Adding columns is not a breaking change; however, changing the column type (from text to number, for example) represents a breaking change.
5. Enter **Parameters**.

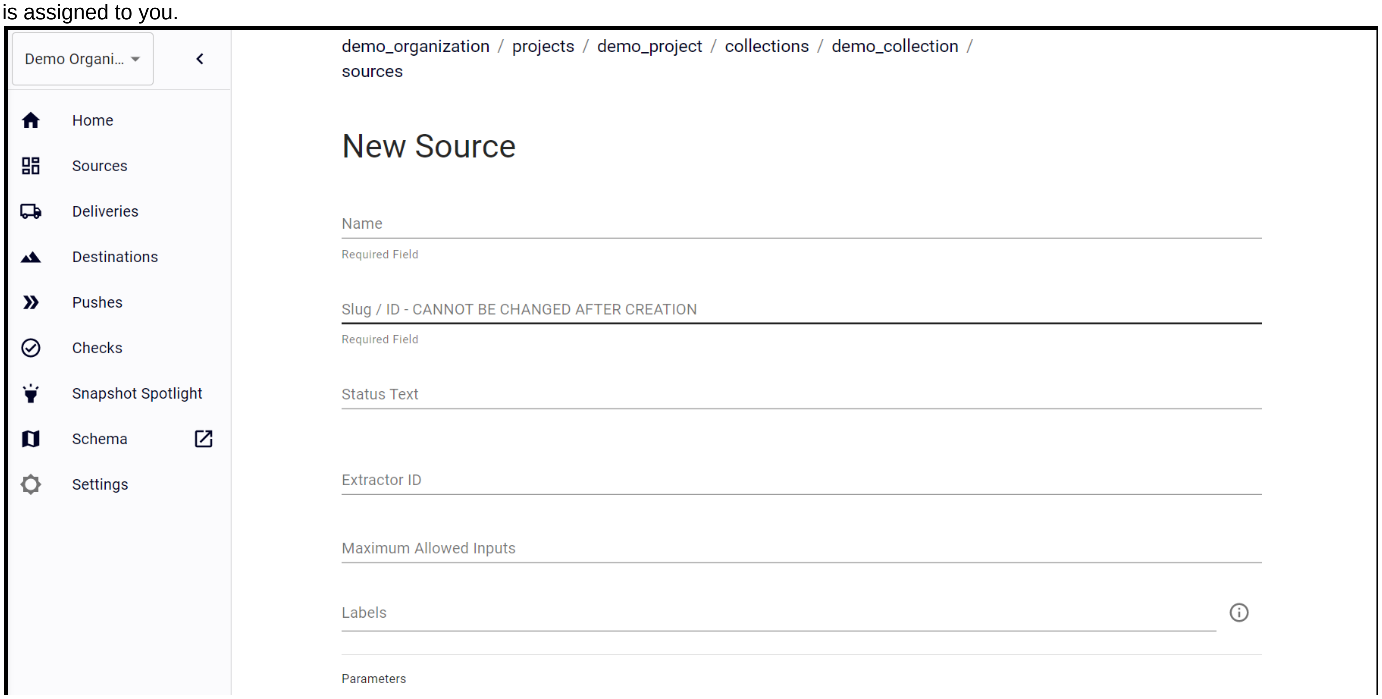
   Parameters help you further group or distinguish Sources and dictate behavior. **Locale** and **Domain** are the default Parameters. If you have an Extractor that performs http://ebay.com searches, for example, *ebay.com-uk* and *ebay.com-fr*, might represent locales. Domain is the website; in this case, http://ebay.com . Parameters are extremely powerful and have numerous downstream uses. Using Parameters, you can tag Sources and establish key/value pairs. You also can use Parameters as filters. You can add the *stage* Parameter and associate it with the *dev* environment. You can add Parameters at this Collection level. For each Parameter you add here, you can provide a value for the Parameter on the related **Sources** page. For example, **Locale=en_us**, **Domain=ebay.com**, **Stage=dev**.
6. To share helpful Collection information with team members, enter text in the **README** section.

   This section, which supports the *markdown* syntax, allows you to provide additional Collection context and insight.
7. To store content, click **Save**. To disregard, click **Cancel**.

Sources

A Source is an Extractor or web crawling tool. A Source maps to an Extractor ID.
A group of Sources is a Collection.

Projects can be locked, restricting write and edit operations under each subsequent DOC layer (such as Collections, Sources, and Snapshots.) You will not have Project access unless you are an ORG_ADMIN, OPS, or assigned to work on a specific layer of this Project. If your role is ORG MEMBER and you are attempting to edit a Source, for example, you cannot make changes unless this Source is assigned to you.



To add a Source:

1. From the left navigation pane, click **Sources**.
2. From the top right of the **Collection Sources** page, click the **Add Source** icon or plus (**+**) symbol.
3. From the **New Source** page, enter text in the **Name** field.

   The name you enter will autofill the **Slug/ID** field. You can associate Slug/IDs with many DOC platform objects, which serve as self-defined identifiers. Slug/IDs can be useful as you reference APIs or create variable names. You cannot change Slug/IDs. That noted, ensure they are meaningful.
4. Enter content in the **Status Text** field.

   A Source may transition from one state to another. QUEUED, IN_PROGRESS, ISSUE, READY, and ACTIVE (among others) represent states. The **Status Text** field allows you to associate a description with the state. If the state is ISSUE, for example, the following text might appear in this field: *Returns 500 responses instead of 404. This is a block and is no longer used. Lambdas now perform this job.* The content in the **Status Text** field also might indicate that the website changed (and is no longer accessible), or the Extractor must be retrained. The text, in these cases, describes the ISSUE. When you initially establish a Source, you may leave this field blank. When you save this entry, the text in the **State** field defaults to QUEUED; however, as the import process progresses (and the state changes), you may choose to enter clarifying text in this field.
5. Enter a value in the **Extractor ID** field.

   This value represents the alphanumeric assignment of the Extractor you designate. You can retrieve this value from the browser of the Extractor. For example:

   https://app.import.io/dash/extractors/**32332031-6ba6-444f-885a-b9f99abcf70e**/history

6. Enter a value in the **Maximum Allowed Inputs** field or use the arrow keys at the end of this line to make a selection.

   Excessive inputs can hog resources required by other projects. When running a Flow, if the number of inputs for a Source breaches the Maximum Allowed Inputs, the inputs file will be trimmed to allow the Snapshot to continue to run (as opposed to failing to start the Snapshot entirely). This information is captured in the **Trimmed Input Count** field on the **Snapshots** page.
7. Enter **Labels**.

   Labels are user-defined tags. *High frequency*, *dev*, *development*, and *staging* are commonly used labels. Labels are similar, in use, to Parameters.
8. Enter **Parameters**.

   Parameters help you further group or distinguish Sources and dictate behavior. **Locale** and **Domain** are the default Parameters. If you have an Extractor that performs http://ebay.com searches, for example, *ebay.com-uk* and *ebay.com-fr*, might represent locales. Domain is the website; in this case, http://ebay.com . Parameters are extremely powerful and have numerous downstream uses. Using Parameters, you can tag Sources and establish key/value pairs. You also can use Parameters as filters. You can add the *stage* Parameter and associate it with the *dev* environment. You can add Parameters on the Collections page. For each Parameter, you can provide a value. For example, **Locale=en_us**, **Domain=ebay.com**, **Stage=dev**.
9. Enter content in the **Status Code Formula** field.

   Commonly, the *status code* refers to the HTTP response code. For example, a *200* response code is usually returned when the page fetch is successful. However, some sites return unusual status codes which may cause the system to register failures when, perhaps, there are none. A *200* response code may still be returned when the intended data is not provided/returned. The Status Code Formula allows the status code to be rewritten based on various clues or indicators.

   According to the screen help text or tooltip for this field:

   "Falsy value means no change, -1 means *blocked*, otherwise return a number. Supported functions, plus PARAM('name') to get a Source Parameter and INPUT('name') to get an input value and all the page row columns as variables, such as exceptionType."

   This help text, therefore, indicates that the Formula may yield a result based on the Source Parameter and values in the output columns; in addition, you simply could rewrite the status code directly. For example, the following Status Code Formula could be written for a site that returned a *101* status code instead of *200*.

```
IF(statusCode = 101, 200, statusCode)
```

The statement above changes the status code to *200* if it was originally *101*.

In another example, the statement also could reference columns of data to ensure they have meaningful values:

```
IF(productId = NULL, statusCode, 200)
```

10. To share helpful Source information with team members, enter text in the **README** section.

    This section, which supports the *markdown* syntax, allows you to provide additional context and insight.
11. To store content, click **Save**. To disregard, click **Cancel**.

Snapshots

A Snapshot, which maps to a crawl run, is the extracted data or output.

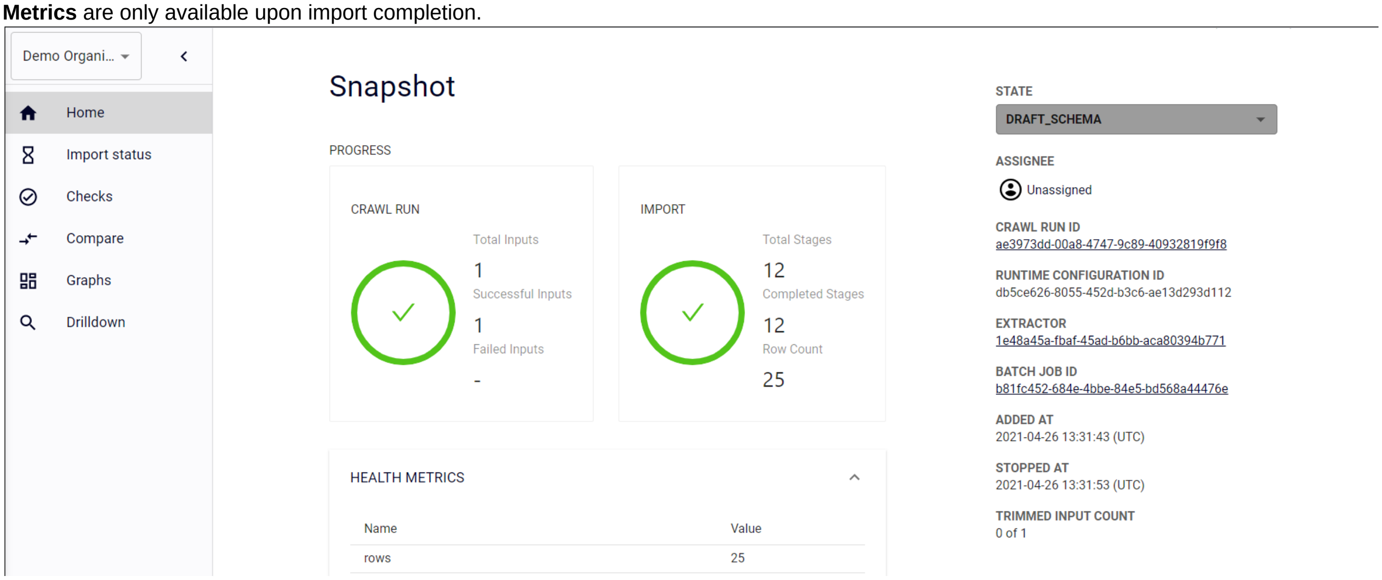
Projects can be locked, restricting write and edit operations under each subsequent DOC layer (such as Collections, Sources, and Snapshots.) You will not have Project access unless you are an ORG_ADMIN, OPS, or assigned to work on a specific layer of this Project. If your role is ORG MEMBER and you are attempting to edit a Snapshot, for example, you cannot make changes unless this Snapshot is assigned to you.

There are three ways to import extracted data into the DOC environment:

- Import by Crawl Run ID / Import Latest Crawl Run
- Modify **Source Settings**
- Access the **Run Source** option (Recommended)

With each option, you can track the progress of the **Crawl Run** and the **Import** – each identified by an in-progress chart. While the *crawl run* is the process whereby a designated website is accessed to retrieve data, the *import* occurs when this data is received and brought into the DOC environment. After a successful crawl run and import, **Health Metrics** are available. This information, located below the crawl run and import charts, reflects the data fitness. Below **Health Metrics** resides an **Activity** section which may provide additional data import detail.

After you initiate the import, you can track the status by reviewing the progress charts. **Health Metrics** are only available upon import completion.



okPct	100
pages	1
dataPct	100
dupePct	0
errorPct	0

Import by Crawl Run ID

1. From the **Snapshots** page, click the ellipses or **More Options** icon at the top right of the page.
2. From the list that appears, select **Import by Crawl Run ID**.
3. From the **Import Crawl Run by ID** modal, enter the Crawl Run ID.

   You can retrieve this ID from the Extractor application. Identify the Extractor whose data you want to import, access the **Run History** tab, click the **Preview Data** icon (right), and copy the now visible Crawl Run ID located near the top center of this tab.
4. Return to the DOC environment and paste this ID into the modal before clicking the **OK** button.

Import Latest Crawl Run

1. From the **Snapshots** page, click the ellipses or **More Options** icon at the top right of the page.
2. From the list that appears, select **Import Latest Crawl Run**

   The system will examine your existing list of crawl runs and import the most recent.

Source Settings

1. From the **Source** level, click **Settings**.
2. From the **Source Settings** page, select the **Automatically Import Data** checkbox.

   A **Source Saved** message appears near the top right of the page, confirming your selection.
3. Return to the Extractor and run.

   This action fires an event to the DOC environment which triggers the data extraction import.

Run Source (Recommended)

This method is preferred, as it is the most efficient way to import extracted data into the DOC environment.

1. From the **Snapshots** page, click the **Run Source** icon located at the top right of the page.

   Running the Source establishes background windows, permits, and other processes which render this extraction method the most efficient.

As data is imported into the DOC environment, the status of the Snapshot commonly transitions to a number of states from PENDING_QUEUE to DRAFT_SCHEMA. The DRAFT_SCHEMA state indicates that the Schema has not been published. You cannot push data to its Destination until a Schema is published. You must review the data carefully before publishing the Schema. You cannot make any breaking changes after the Schema is published. You can add columns without issue; however, changing a column type from Text to Currency, for example, is a breaking change.

There are several pages you can access to view the import status or check the data. The **Import Status** page enables you to see the import pipeline stages. For each stage, you can view Start and Finish times along with the Progress, Errors, and Completion. You also can access the **Checks** page to determine the data validity. If an alert indicates that a certain number of rows is expected for a particular run (but significantly less rows are output), you can view this page to troubleshoot and determine any data issues.

After import, a **Download** icon appears at the top right of the **Snapshots** page. Upon selection, you can choose from various file types. You can download this Snapshot file to further evaluate the extracted data or perform troubleshooting.

Retry Snapshot

If a Snapshot import fails, the JQ transform for custom files changes or fails, or the runtime configuration of a Snapshot changes, you might need to retry certain stages of the Snapshot lifecycle.

The **Snapshots** page includes ellipses or the **More Options** icon at the top right of the page.
Upon selection (after a data import), four options appear. The table below describes each.

Option	Purpose
Re-extract Snapshot	Perform this action if you have updated the Extractor. The original Snapshot will assume a SUPERCEDED state, and the new Snapshot will include both a new crawl run and import. In addition, the **Re-extract** stage will be added to the import pipeline of the new Snapshot.
Re-import Snapshot	Perform this action if a Snapshot fails to import prior to the **Generate Assets** stage of the import pipeline. You can re-import data if there are inconsistencies between the Extractor and DOC Schemas. The original import will assume a SUPERCEDED state, and a new Snapshot will include both a new crawl run and import.
Re-run Snapshot	Perform this action if an entirely new crawl run is necessary without a revision to an Extractor's runtime configuration. The original Snapshot will assume a SUPERCEDED state, and a new Snapshot will be created along with a new crawl run.
Regenerate Custom Assets	**Generate Assets** is the final stage of the import pipeline for a Snapshot. If the Snapshot's Collection has any linked Destinations with a custom file type selected (custom text or custom parquet), they will be created during this stage. If the JQ transform associated with a Snapshot fails for a custom file, the Snapshot will fail to import (for example, status will be FAILED) with a relevant error message. If a Snapshot fails because of the **Generate Assets** stage, this stage can be retried with the **Regenerate Custom Assets** option. This action will only retry the **Generate Assets** stage. If the regeneration is successful, a Snapshot can continue to QA to be pushed.

Flows

A Flow represents the data Delivery pipeline or process of rendering data to customers. Since customers may have varying Delivery specifications, you must configure the DOC environment such that data Delivery is aligned with customer requirements. In brief, when you create a Flow, you configure a data Delivery for a Collection (group of Sources/Extractors) within a designated timeframe. A Flow allows you to run numerous Sources simultaneously in a scheduled window.

Dependencies

- A Flow belongs to a Project and, in this case, is the child of a Project.
- A Flow requires a Collection, which extends itself to a necessary Schema.

Flow Types
There are three Flow types: **Legacy**, **Simple**, and **Chained**

Less automated than the Simple Flow type, **Legacy Flows** involve establishing **Collection Information** and configuring data Delivery for this Collection within a designated timeframe. A Legacy Flow requires that you start any crawl run manually. During this process, Snapshots (extracted data) are captured. While currently the default Flow type, Legacy is slated for deprecation.

Unlike Legacy Flows, **Simple Flows** automatically start each Source associated with the Collection. You then must start or schedule the Deliveries (which triggers the crawl runs) – making this Flow type more convenient than Legacy. While you can cancel Simple Flows, you cannot cancel a Legacy. For Simple Flows, this action cancels the Snapshot crawl runs in progress and does not import this data. Simple Flows provide more options for customization and, as such, prompt you to respond to additional questions during configuration. Simple and Chained are the most common Flow types.

The **Chained Flow** type is the most complex, as it involves two or more chained Extractors – so named because one is dependent on the other. Based on the amount of website data, two or more extractions might be required for complete data retrieval. For example, Extractor 1 might perform a search for a list of books and categories (performing an extraction for each page of search results), outputting book names and categories. To run, Extractor 2 would require the output from Extractor 1.

Extractor 2 would take the search results, access each page, and retrieve the product details such as name, product URL, rank, price, and availability. The Extractors are chained such that when Extractor 1 finishes, Extractor 2 starts by using the output from Extractor 1. Chained Flows are commonly used when performing a search and providing related product details. This action is more challenging using a single Extractor, hence Chained Flows. Chained Flows include Segments, and each correlates to the set of first Extractors. Each Segment is attached to a Collection, which may be comprised of several Sources.

For detailed insight on each Flow configuration, access **Flows** in the primary User Guide.

Deliveries

The **Deliveries** page, commonly referred to as the Delivery Dashboard, populates after the execution of **scheduled** and **unscheduled** Flows.

To understand scheduled and unscheduled Flows, you must recall the setup you designated when you configured the Flows.

On the **Flows** page, there is an **Active** checkbox. This selected checkbox ensures Deliveries run as **scheduled**, per the **Cron Expression**, visible after you adjust the toggle switch: **Do you want to add scheduling?** The **Active** checkbox is selected by default. If you clear this checkbox, no Deliveries will run even if specified in the **Cron Expression**. Clearing the **Active** checkbox provides a method by which to maintain the Flow while *not* running scheduled Deliveries. In this case, you can run this **unscheduled** Flow manually (as needed) by clicking the **Run Flow** button located at the top right of the page specific to that configured Flow.

Deliveries cannot exist in the DOC environment without Flows.

Each Delivery represents a time period set to collect the data on the parent Flow. If the data has been chunked, the Delivery is sent to the client as such. Each Snapshot that corresponds to a Delivery will have a window of time during which it will collect the data along with a window of time for this data to be delivered.

The **Deliveries** page displays all **Open** and **Closed** Deliveries. While an Open Delivery has not yet reached the **Hours to Finish** value as designated in the Flow configuration, a Closed Delivery has reached this **Hours to Finish** designation.

The **Share Deliveries** icon not only may be visible on this primary page but also might be visible at the top right of the page that opens after you click the **Drilldown** button in the right pane. Internally, selection of this icon generates tokens used to authenticate each user. Only SUPER_ADMIN and ORG_ADMINs can create public tokens. Externally, selection of this icon allows access to a URL which you must first copy then paste into a browser to display a public **Deliveries** page. These public pages are READ only.