JOB AID

Creating and executing a data preparation

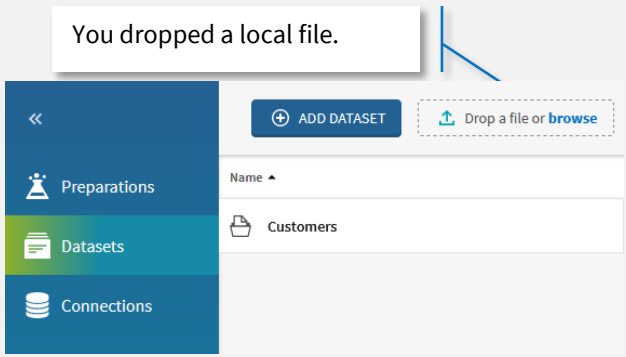
Overview

Using the Data Preparation application in Talend Cloud, you learned how to create a data preparation and dataset, discover data, and build a recipe. You also learned how to configure Talend Job components to execute data preparation.

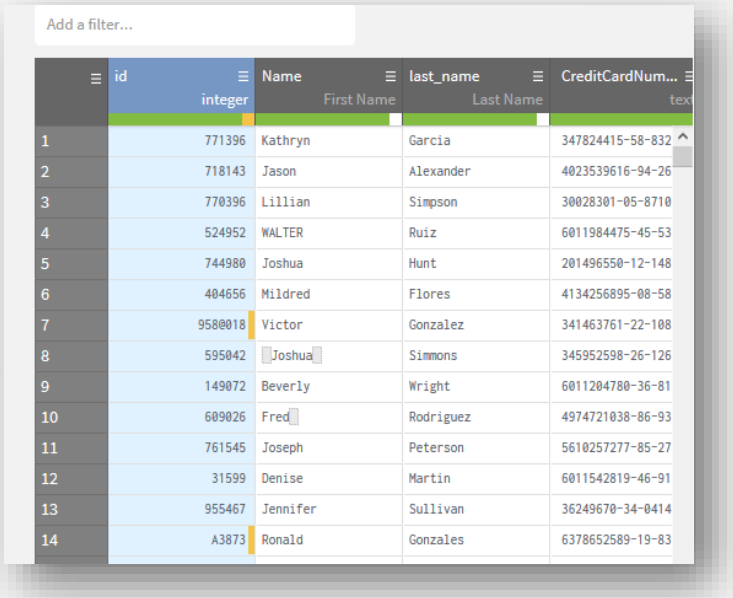
Key steps

1 To access your first dataset, you can:

1. Drop a file.
2. Create a new dataset using a connection.



The data sample is displayed in Data Preparation:



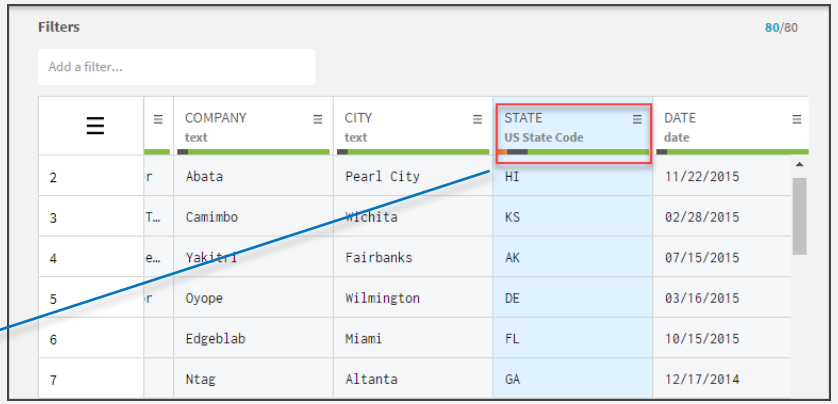
JOB AID

2 Using Data Preparation to discover data

There are several tools for discovering data available in Data Preparation.

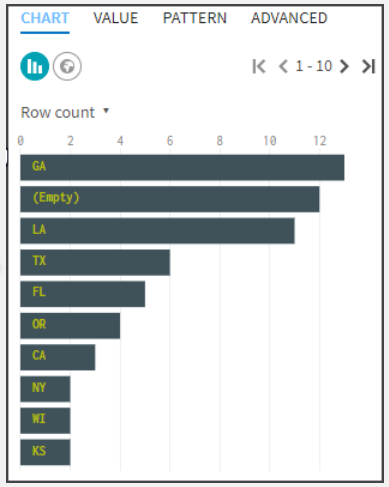
Each dataset column in Data Preparation is associated with a semantic type.

The **data quality bar** displays (by column) the number of fields with correct data in green, empty fields in white, and incorrect data in orange.
Here the **State** column is recognized as a US State Code.



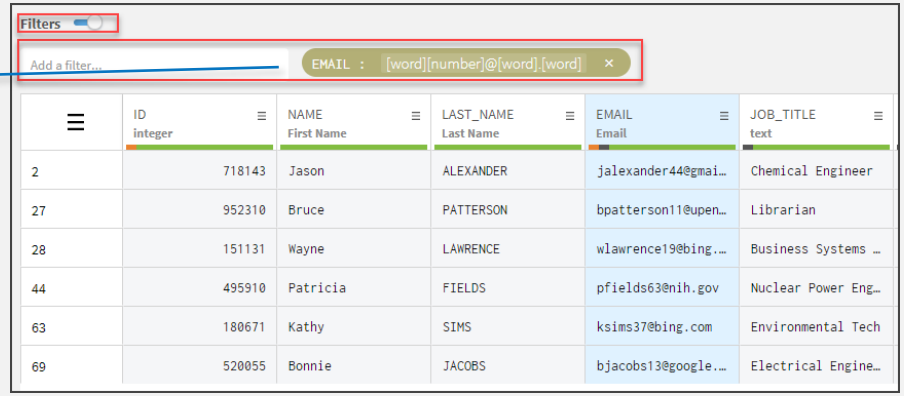
You can also explore **column statistics**.

For example, you can view the number of occurrences per US State Code.



You can **apply filters** to have specific data contained in your dataset.

Here, the filter is applied on the **EMAIL** column.

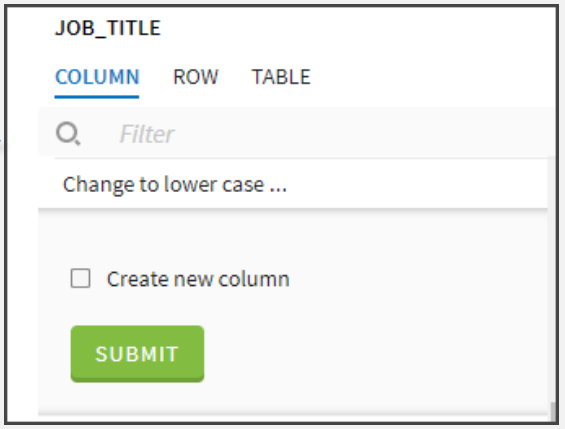


JOB AID

3 Building a recipe

Using Data Preparation, you can apply cleansing and conversion functions to the dataset columns.

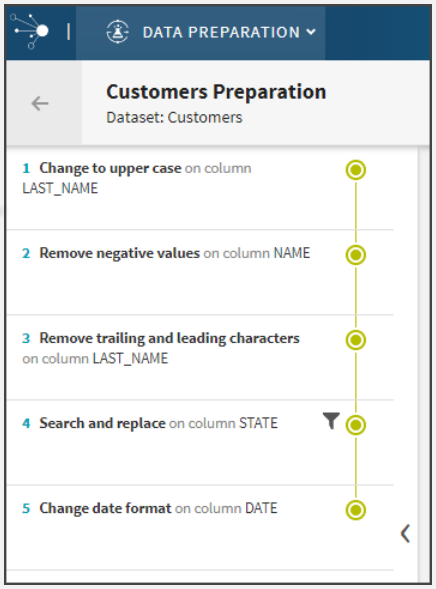
The function **Change to lower case** will be applied to the **JOB_TITLE** column.



The transformation steps (functions) are logged in a recipe saved in the data preparation.

When executing a data preparation against a dataset, the data preparation recipe is applied to the dataset.

This data preparation recipe has five steps.
Here, the first step consists of applying the function **Change to upper case** on the **LAST_NAME** column.

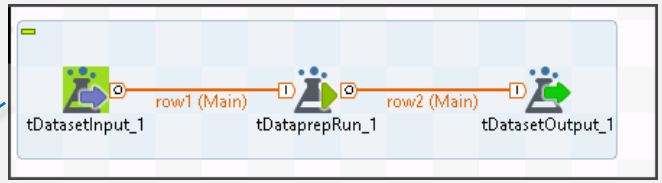


JOB AID

4 Executing a data preparation in a Talend Job

In a Talend Job, you can use the following components to access a data preparation and a dataset defined in Data Preparation:

- **tDatasetInput**: reads data from a dataset
- **tDataprepRun**: processes data from flow through preparation steps (recipe)
- **tDatasetOutput**: creates a dataset in Data Preparation



For example, this Job reads the input data from Talend Data Preparation, executes the preparation, and publishes the results file to another dataset in Talend Data Preparation.

Talend Cloud credentials are required to access datasets and data preparation.

When using a **tDataprepRun** component, the URL, Username, and Password properties are required to run a data preparation to connect Data Preparation in Talend Cloud.