

# Creating and executing a data preparation

## Overview

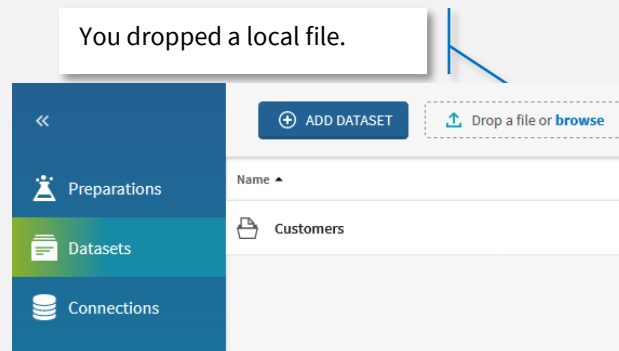
Using the Data Preparation application in Talend Cloud, you learned how to create a data preparation and dataset, discover data, and build a recipe. You also learned how to configure Talend Job components to execute data preparation.

## 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:

The screenshot shows a data sample table with the following columns: id, Name, last\_name, and CreditCardNum... The table contains 14 rows of data. The first row is highlighted in green.

	id	Name	last_name	CreditCardNum...
	integer	First Name	Last Name	text
1	771396	Kathryn	Garcia	347824415-58-832
2	718143	Jason	Alexander	4023539616-94-26
3	770396	Lillian	Simpson	30028301-05-8710
4	524952	WALTER	Ruiz	6011984475-45-53
5	744980	Joshua	Hunt	201496550-12-148
6	404656	Mildred	Flores	4134256895-08-58
7	9580018	Victor	Gonzalez	341463761-22-108
8	595042	Joshua	Simmons	345952598-26-126
9	149072	Beverly	Wright	6011204780-36-81
10	609026	Fred	Rodriguez	4974721038-86-93
11	761545	Joseph	Peterson	5610257277-85-27
12	31599	Denise	Martin	6011542819-46-91
13	955467	Jennifer	Sullivan	36249670-34-0414
14	A3873	Ronald	Gonzales	6378652589-19-83

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.

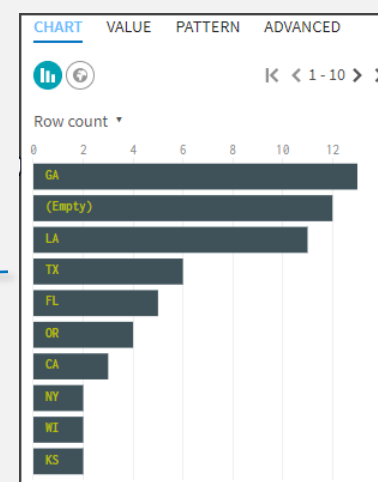
Filters 80/80

Add a filter...

		COMPANY text	CITY text	STATE US State Code	DATE date
2	r	Abata	Pearl City	HI	11/22/2015
3	T...	Camimbo	Wichita	KS	02/28/2015
4	e...	Yakitori	Fairbanks	AK	07/15/2015
5	r	Oyope	Wilmington	DE	03/16/2015
6		Edgeblab	Miami	FL	10/15/2015
7		Ntag	Atlanta	GA	12/17/2014

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.

Filters 80/80

Add a filter... EMAIL : [word][number]@[word].[word] x

	ID integer	NAME First Name	LAST_NAME Last Name	EMAIL Email	JOB_TITLE text
2	718143	Jason	ALEXANDER	jalexander44@gmail...	Chemical Engineer
27	952310	Bruce	PATTERSON	bpatterson11@upen...	Librarian
28	151131	Wayne	LAWRENCE	wlawrence19@bing...	Business Systems ...
44	495910	Patricia	FIELDS	pfields63@nih.gov	Nuclear Power Eng...
63	180671	Kathy	SIMS	ksims37@bing.com	Environmental Tech
69	520055	Bonnie	JACOBS	bjacobs13@google...	Electrical Engine...

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.

**JOB\_TITLE**

COLUMN ROW TABLE

Filter

Change to lower case ...

☐ Create new column

SUBMIT

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.

DATA PREPARATION

Customers Preparation

Dataset: Customers

- 1 Change to upper case on column LAST\_NAME
- 2 Remove negative values on column NAME
- 3 Remove trailing and leading characters on column LAST\_NAME
- 4 Search and replace on column STATE
- 5 Change date format on column DATE

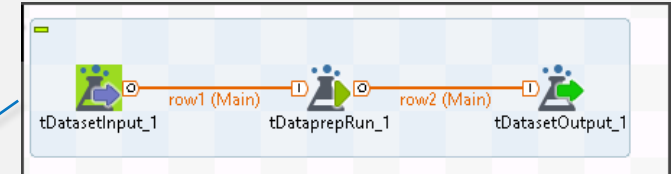
## 4

## 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.

