



Data Camp Live Training: Cleaning Data with Pyspark



Instructor Picture
Here

Mike Metzger
Data Engineer / Consultant



Introduction

Data Cleaning - the process of preparing & normalizing data for delivery / further analysis.

Data cleaning can be done with anything from a text editor, to Excel, to Python / Pandas.

But what happens when you have more data than you can process on a single system?

Expected contents:

- show_id: A unique integer identifier for the show
- type: The type of content, Movie or TV Show
- title: The title of the content
- director: The director (or directors)
- cast: The cast
- country: Country (or countries) where the content is available
- date_added: Date added to Netflix
- release_year: Year of content release
- rating: Content rating
- duration: The duration
- listed_in: The genres the content is listed in
- description: A description of the content

Dataset Overview

We'll be working with CSV data containing the movies and TV shows available on Netflix.

Except the data is corrupted in various ways (bad rows, missing data, etc).



Spark & Python

We'll be using the a combo of Spark and Python, mixed in with a bit of Bash scripting to analyze and clean our dataset.

jupyter

applayout_example

Last Checkpoint: 20 hours ago (autosaved)

Logout

File

Edit

View

Insert

Cell

Kernel

Widgets

Help


Trusted

Python 3

```
11         icon='backward',
12         layout=Layout(width='80%',
13                        height='30%'))
14 next_button = Button(description="Next",
15                       icon='forward',
16                       layout=Layout(width='80%',
17                                    height='30%'))
18 footer = HTML("Filename: {}".format(image_file))
19
20 AppLayout(header=header,
21           left_sidebar=prev_button,
22           center=image,
23           right_sidebar=next_button,
24           footer=footer,
25           grid_gap='20px',
26           justify_items='center',
27           align_items='center')
```

Simple Image Viewer

◀ Prev



▶ Next

Filename: images/cat.jpg



!! Requires a gmail account to edit !!

Session Agenda

- *Introduction*
- Initial loading of dataset
- Q & A
- Data filtering & cleanup
- Q & A
- More cleaning & formatting
- Q & A
- Recap/Closing Notes
- Next Steps/Take home assignment

The background features several abstract geometric shapes in teal and green. In the top-left corner, there is a large, rounded green shape. In the top-right, a teal triangle is partially visible, with a thin teal line extending from its base towards the center. The bottom-left contains a teal triangle with a thin teal line extending from its base towards the center. The bottom-right features a large, rounded green shape. The word "Notebook" is centered in the middle of the page.

Notebook

Recap and Closing Notes

What Did We Learn Today?

Spark can handle most any data format

Even ones that don't load correctly automatically!

Spark transformations are *lazy*

We primarily define the recipe of what we want to happen and let the framework do the rest.

Spark is customizable based on need

We can create any data layout we'd like using the built-in functions and our own user defined functions!

Coming Soon!

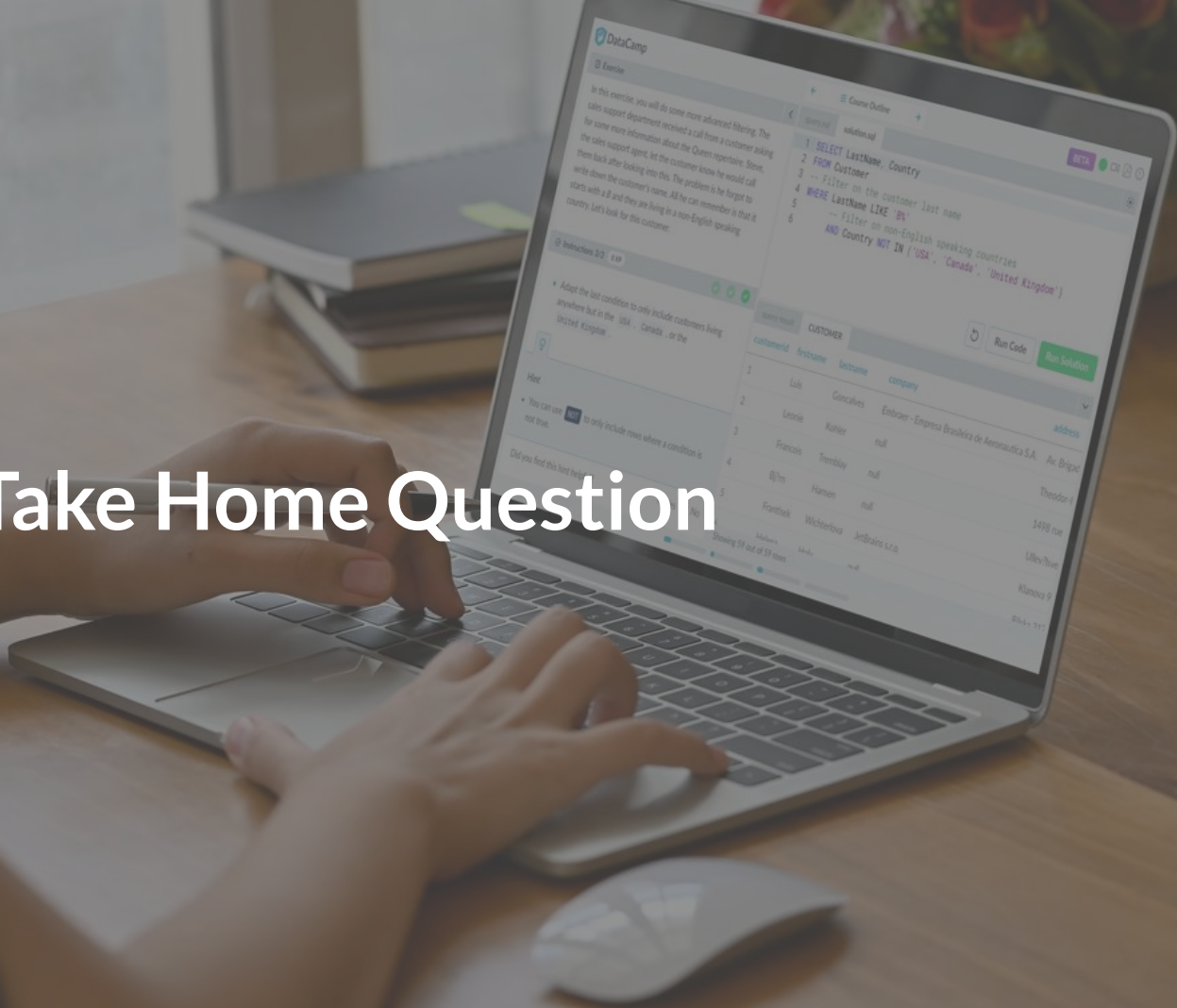


Don't miss these upcoming webinars and live training sessions!

- Session 1
- Session 2
- Session 3

(Note: Consult with Marketing for upcoming webinars, Kelsey for upcoming live trainings)

Take Home Question



Take Home Question

Use Spark to perform some further cleanup on the dataset:

- 1) *Split names* - You may have noticed that the names are combined for the cast and directors into a list. Consider how you would turn that data into a list / array column to easily access more detailed information (which shows have the largest cast, etc?)
- 2) *Splitting names further* - Consider taking any of the name fields and splitting it into first name, last name, etc. Take special consideration about how you would handle initials, names with more than 3 components, etc.
- 3) *Parsing dates* - Look at the `date_added` field and determine if and how you could reliably convert this to an actual datetime field.

Submission details:

- Share with us a code snippet with your output on LinkedIn, Twitter or Facebook
- Tag us `@DataCamp` with the hashtag `#datacamplive`
- [Optional] [Your social media handles or LinkedIn url]

Thank you

Mike Metzger
Data Engineer / Consultant
mike@flexiblecreations.com
[@mmetzger](https://twitter.com/mmetzger) - twitter

