COMP 333 — Week 6 Data Wrangling Process

Data Wrangling Process

In Week 6 (this lecture) and Week 7 we will cover Data Wrangling which is the most time-consuming phase of Data Analytics.

Data Wrangling is the ETL process of data warehouses applied more generally as part of Data Analytics.

It is very important to clean and organize your data.

Remember GIGO (Garbage-In, Garbage-Out)

Definition *Data wrangling*, sometimes referred to as *data munging*, is the process of transforming and mapping data from one "raw" data form into another format with the intent of making it more appropriate and valuable for a variety of downstream purposes such as analytics. [wikipedia]

Process

There are several different perspectives of Data Wrangling
and how Data Wrangling fits into the broader Data Analytics.

In Chapter 2 of the `pandas` book
the Data Analytics process is defined as

- Interacting with the outside world
  Reading and writing with a variety of file formats and databases.
- Preparation
  Cleaning, munging, combining, normalizing, reshaping,
  slicing and dicing, and transforming data for analysis.
- Transformation
  Applying mathematical and statistical operations
  to groups of data sets to derive new data sets.
- Modeling and computation
  Connecting your data to statistical models, machine learning algorithms, or other computational tools
- Presentation
  Creating interactive or static graphical visualizations or textual summaries.

In Chapter 7 the Data Wrangling process is defined as

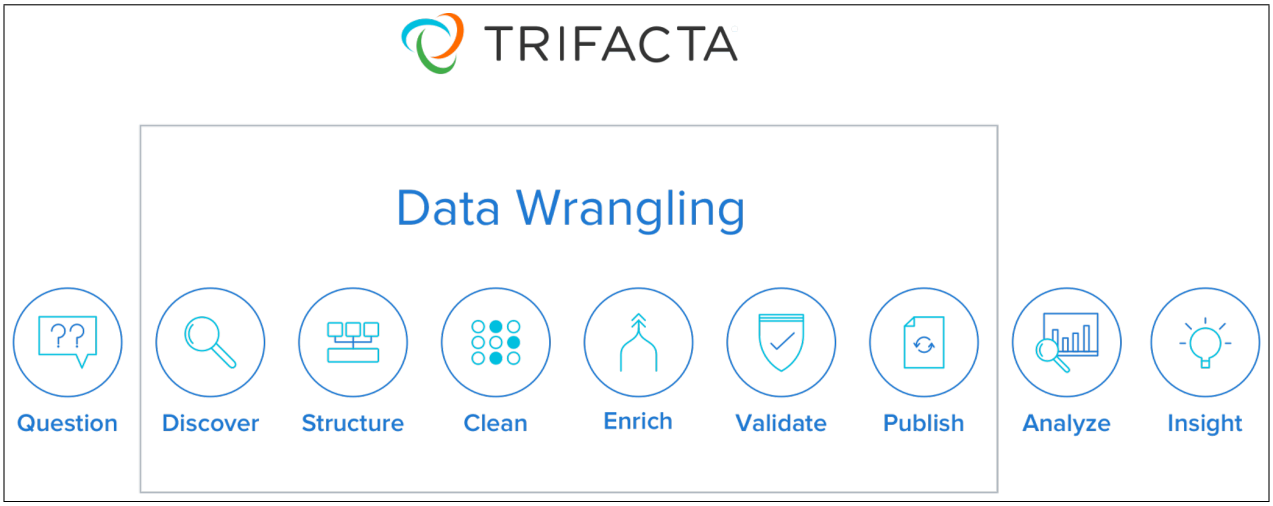
- clean
- transform
- merge
- reshape

In the video example, Isaac Vidas provides a workflow (process) for Data Wrangling

- content acquisition
- enrichment, which is adding new features from related data
- entity resolution
- combine, or integrate data from different sources

To see Data Wrangling inside Data Analytics, see the figure from Trifacta

Trifacta makes software for Data Wrangling



In our Nutshell overview, we follow the Trifacta website

`https://www.trifacta.com/data-wrangling/`

There are typically six iterative steps that make up the data wrangling process.

1. Discovering:

   Before you can dive deeply,
   you must better understand what is in your data,
   which will inform how you want to analyze it.
   How you wrangle customer data, for example, may be informed by
   where they are located,
   what they bought, or
   what promotions they received.

2. Structuring:

   This data wrangling step means organizing the data,
   which is necessary because raw data comes in many different shapes and sizes.
   A single column may turn into several rows for easier analysis.
   One column may become two.
   Movement of data is made for easier computation and analysis.

3. Cleaning:

   What happens when errors and outliers skew your data?
   You clean the data.
   What happens when state data is entered as CA or California or Calif.?
   You clean the data.
   Null values are changed and standard formatting implemented,
   ultimately increasing data quality,
   which is the goal of data wrangling.

4. Enriching:

   Here you take stock in your data and strategize about
   how other additional data might augment it.
   Questions asked during this data wrangling step might be:
   what new types of data can I derive from what I already have
   or what other information would better inform my decision making?

5. Validating:

   Validation rules are repetitive programming sequences
   that verify data consistency, quality, and security.
   Examples of validation include
   ensuring uniform distribution of attributes
   that should be distributed normally (e.g. birth dates)
   or confirming accuracy of fields through a check across data.

6. Publishing:

   Analysts prepare the wrangled data for use downstream
   — whether by a particular user or software —
   and document any particular steps taken or logic used to wrangle said data.
   Data wrangling gurus understand that implementation of insights
   relies upon the ease with which it can be accessed
   and utilized by others.

More on Validate

Check data is consistent and complete

Consistency Is your data consistent?

What does that mean:

Does your data fit into expected values for it?

Do field values match the data type for the column?

Are values within acceptable ranges?

Are rows unique? Duplicated?

Completeness Is your data complete?

What does that mean:

Are all expected values included in your data?

Are some fields missing values?

Are there expected values that are not present in the dataset?

Test your code Test your routines for your data wrangling process

Tidy Data

Your final data should be structured as *"Tidy Data"*

Recommended Book

Read the whole of Chapter 7 of the recommended book

Python for Data Analysis: Data Wrangling with Pandas, NumPy, and IPython, 2nd Edition, by Wes McKinney, O'Reilly Media, 2017.

`https://bedford-computing.co.uk/learning/wp-content/uploads/2015/10/Python-for-Data-Anal`
`pdf`