WIKIPEDIA
The Free Encyclopedia

Main page
Contents
Featured content
Current events
Random article
Donate to Wikipedia
Wikipedia store

Interaction

Help
About Wikipedia
Community portal
Recent changes
Contact page

Tools

What links here
Related changes
Upload file
Special pages
Permanent link
Page information
Wikidata item
Cite this page

Print/export

Create a book
Download as PDF
Printable version

Languages

한국어
Íslenska
Norsk
Edit links

Data wrangling

From Wikipedia, the free encyclopedia

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. A **data wrangler** is a person who performs these transformation operations.

This may include further munging, data visualization, data aggregation, training a statistical model, as well as many other potential uses. Data munging as a process typically follows a set of general steps which begin with extracting the data in a raw form from the data source, "munging" the raw data using algorithms (e.g. sorting) or parsing the data into predefined data structures, and finally depositing the resulting content into a data sink for storage and future use.[1]

Contents

Background [edit]

The "wrangler" non-technical term is often said to derive from work done by the United States Library of Congress's National Digital Information Infrastructure and Preservation Program (NDIIPP) and their program partner the Emory University Libraries based MetaArchive Partnership. The term "mung" has roots in munging as described in the Jargon File.[2] The term "Data Wrangler" was also suggested as the best analogy to coder for someone working with data.[3]

The terms data wrangling and data wrangler had sporadic use in the 1990s and early 2000s. One of the earliest business mentions of data wrangling was in an article in Byte Magazine in 1997 (Volume 22 issue 4) referencing “Perl’s data wrangling services”. In 2001 it was reported that CNN hired[4] “a dozen data wranglers” to help track down information for news stories.

One of the first mentions of data wrangling in a scientific context was by Donald Cline during the NASA/NOAA Cold Lands Processes Experiment.[5] Cline stated the data wranglers “coordinate the acquisition of the entire collection of the experiment data.” Cline also specifies duties typically handled by a **storage administrator** for working with large amounts of data. This can occur in areas like major research projects and the making of films with a large amount of complex computer-generated imagery. In research, this involves both data transfer from research instrument to storage grid or storage facility as well as data manipulation for re-analysis via high performance computing instruments or access via cyberinfrastructure-based digital libraries.

Typical use [edit]

The data transformations are typically applied to distinct entities (e.g. fields, rows, columns, data values etc.) within a data set, and could include such actions as extractions, parsing, joining, standardizing, augmenting, cleansing, consolidating and filtering to create desired wrangling outputs that can be leveraged downstream.

The recipients could be individuals, such as data architects or data scientists who will investigate the data further, business users who will consume the data directly in reports, or systems that will further process the data and write it into targets such as data warehouses, data lakes or downstream applications.

Modus operandi [edit]

Depending on the amount and format of the incoming data, data wrangling has traditionally been performed manually (e.g. via spreadsheets such as Excel) or via hand-written scripts in languages such as

Python or SQL. R, a language often used in data mining and statistical data analysis, is now also often[6] used for data wrangling. Other terms for these processes have included data franchising[7], data preparation and data munging.

Wrangler project [edit]

In 2011, researchers from Stanford University and UC Berkeley published a paper entitled *Wrangler: Interactive Visual Specification of Data Transformation Scripts*.[8] In it, the authors described a research project called Wrangler[9], which was "an interactive system for creating data transformations." Wrangler introduced a new way to perform data wrangling through direct interaction with data presented in a visual interface. Analysts could interactively explore, change and manipulate the data and immediately see results. Wrangler tracked the user's data transformations and could then automatically generate code or scripts that could be applied repeatedly on other datasets.

In 2012, several of the authors (Kandel, Hellerstein, Heer) went on to found Trifacta, which is a commercialization of the software in the Wrangler project. Since then, a number of other companies have developed products. to perform data wrangling. These include both commercial and freely available offerings.

See also [edit]

- Data cleaning, removing erroneous data in a corpus of data.
- Data editing, correcting errors in a corpus of data.
- Data scraping, extracting parts of a corpus of data with automated tools.
- Data curation, a more general and abstract activity
- Data pre-processing, a step of cleaning data in data mining for analysis purposes
- Data fusion and data integration
- Data preparation
- OpenRefine
- Semantic mapping (data integration)
- Simultaneous editing, efficient repeated edition of text in a multiple selection through direct manipulation.
- Extract, transform, load
- Big Data

References [edit]

1. ^ What Is Data Munging?
2. ^ Jargon File entry for Mung
3. ^ Open Knowledge Foundation Blog Post
4. ^ Behind the Headlines at Revamped News
5. ^ Parsons, MA, MJ Brodzik, and NJ Rutter. 2004. Data management for the cold land processes experiment: improving hydrological science. HYDROL PROCESS. 18:3637-653. http://onlinelibrary.wiley.com/doi/10.1002/hyp.5801/abstract
6. ^ O'Reilly 2016 Data Science Survey
7. ^ What is Data Franchising? (2003 and 2017 IRI)
8. ^ Kandel, Paepcke, Hellerstein Heer, 2011 Wrangler: Interactive Visual Specification of Data Transformation Scripts
9. ^ Stanford Wrangler Research Project

v · t · e **Data**

Analysis · Archaeology · Cleansing · Collection · Compression · Corruption · Curation · Degradation · Editing · Farming · Format management · Fusion · Integration · Integrity · Library · Loss · Management · Migration · Mining · Pre-processing · Preservation · Protection (privacy) · Recovery · Reduction · Retention · Quality · Science · Scraping · Scrubbing · Security · Stewardship · Storage · Validation · Warehouse · **Wrangling/munging**

Categories: Computer occupations | Data mapping