DATA WRANGLING

Concepts, Applications, and Tools

Edited by, M. Niranjanamurthy, Kavita Sheoran,
Geetika Dhand, *and* Prabhjot Kaur

Scrivener Publishing

WILEY

Data Wrangling

Scrivener Publishing
100 Cummings Center, Suite 541J
Beverly, MA 01915-6106

Publishers at Scrivener
Martin Scrivener (martin@scrivenerpublishing.com)
Phillip Carmical (pcarmical@scrivenerpublishing.com)

Data Wrangling

Concepts, Applications and Tools

Edited by
M. Niranjanamurthy
Kavita Sheoran
Geetika Dhand
and
Prabhjot Kaur

WILEY

This edition first published 2023 by John Wiley & Sons, Inc., 111 River Street, Hoboken, NJ 07030, USA and Scrivener Publishing LLC, 100 Cummings Center, Suite 541J, Beverly, MA 01915, USA

For more information about Scrivener publications please visit www.scrivenerpublishing.com.

Wiley Global Headquarters
111 River Street, Hoboken, NJ 07030, USA

For details of our global editorial offices, customer services, and more information about Wiley products visit us at www.wiley.com.

Library of Congress Cataloging-in-Publication Data

ISBN 978-1-119-87968-8

Cover images: Color Grid Background | Anatoly Stojko | Dreamstime.com
Data Center Platform | Siarhei Yurchanka | Dreamstime.com
Cover design: Kris Hackerott

Set in size of 11pt and Minion Pro by Manila Typesetting Company, Makati, Philippines

Printed in the USA

10 9 8 7 6 5 4 3 2 1

Contents

1

Basic Principles of Data Wrangling

Akshay Singh*, Surender Singh and Jyotsna Rathee

Department of Information Technology, Maharaja Surajmal Institute of Technology, Janakpuri, New Delhi, India

Abstract

Data wrangling is considered to be a crucial step of data science lifecycle. The quality of data analysis directly depends on the quality of data itself. As the data sources are increasing with a fast pace, it is more than essential to organize the data for analysis. The process of cleaning, structuring, and enriching raw data into the required data format in order to make better judgments in less time is known as data wrangling. It entails the manual conversion and mapping of data from one raw form to another in order to facilitate data consumption and organization. It is also known as data munging, meaning "digestible" data. The iterative process of gathering, filtering, converting, exploring, and integrating data come under the data wrangling pipeline. The foundation of data wrangling is data gathering. The data is extracted, parsed, and scraped before the process of removing unnecessary information from raw data. Data filtering or scrubbing includes removing corrupt and invalid data, thus keeping only the needful data. The data is transformed from unstructured to a bit structured form. Then, the data is converted from one format to another format. To name a few, some common formats are CSV, JSON, XML, SQL, etc. The preanalysis of data is to be done in data exploration step. Some preliminary queries are applied on the data to get the sense of the available data. The hypothesis and statistical analysis can be formed after basic exploration. After exploring the data, the process of integrating data begins in which the smaller pieces of data are added up to form big data. After that, validation rules are applied on data to verify its quality, consistency, and security. In the end, analysts prepare and publish the wrangled data for further analysis. Various platforms available for publishing the wrangled data are GitHub, Kaggle, Data Studio, personal blogs, websites, etc.

***Keywords*:** Data wrangling, big data, data analysis, cleaning, structuring, validating, optimization

**Corresponding author*: akshaysingh@msit.in

M. Niranjanamurthy, Kavita Sheoran, Geetika Dhand, and Prabhjot Kaur (eds.) Data Wrangling: Concepts, Applications and Tools, (1–18) © 2023 Scrivener Publishing LLC

1.1 Introduction

Meaningless raw facts and figures are termed as data which are of no use. Data are analyzed so that it provides certain meaning to raw facts, which is known as information. In current scenario, we have ample amount of data that is increasing many folds day by day which is to be managed and examined for better performance for meaningful analysis of data. To answer such inquiries, we must first wrangle our data into the appropriate format. The most time-consuming part and essential part is wrangling of data [1].

Definition 1—"Data wrangling is the process by which the data required by an application is identified, extracted, cleaned and integrated, to yield a data set that is suitable for exploration and analysis." [2]

Definition 2—"Data wrangling/data munging/data cleaning can be defined as the process of cleaning, organizing, and transforming raw data into the desired format for analysts to use for prompt decision making."

Definition 3—"Data wrangling is defined as an art of data transformation or data preparation." [3]

Definition 4—"Data wrangling term is derived and defined as a process to prepare the data for analysis with data visualization aids that accelerates the faster process." [4]

Definition 5—"Data wrangling is defined as a process of iterative data exploration and transformation that enables analysis." [1]

Although data wrangling is sometimes misunderstood as ETL techniques, these two are totally different with each other. Extract, transform, and load ETL techniques require handiwork from professionals and professionals at different levels of the process. Volume, velocity, variety, and veracity, i.e., 4 V's of big data becomes exorbitant in ETL technology [2].

We can categorize values into two sorts along a temporal dimension in any phase of life where we have to deal with data: near-term value and long-term value. We probably have a long list of questions we want to address with our data in the near future. Some of these inquiries may be ambiguous, such as "Are consumers actually changing toward communicating with us via their mobile devices?" Other, more precise inquiries can include: "When will our clients' interactions largely originate from mobile devices rather than desktops or laptops?" Various research work, different projects, product sale, company's new product to be launched, different businesses etc. can be tackled in less time with more efficiency using data wrangling.

- Aim of Data Wrangling: Data wrangling aims are as follows:
  a) Improves data usage.
  b) Makes data compatible for end users.
  c) Makes analysis of data easy.
  d) Integrates data from different sources, different file formats.
  e) Better audience/customer coverage.
  f) Takes less time to organize raw data.
  g) Clear visualization of data.

In the first section, we demonstrate the workflow framework of all the activities that fit into the process of data wrangling by providing a workflow structure that integrates actions focused on both sorts of values. The key building pieces for the same are introduced: data flow, data wrangling activities, roles, and responsibilities [10]. When commencing on a project that involves data wrangling, we will consider all of these factors at a high level.

The main aim is to ensure that our efforts are constructive rather than redundant or conflicting, as well as within a single project by leveraging formal language as well as processes to boost efficiency and continuity. Effective data wrangling necessitates more than just well-defined workflows and processes.

Another aspect of value to think about is how it will be provided within an organization. Will organizations use the exact values provided to them and analyze the data using some automated tools? Will organizations use the values provided to them in an indirect manner, such as by allowing employees in your company to pursue a different path than the usual?

- Indirect Value: By influencing the decisions of others and motivating process adjustments. In the insurance industry, for example, risk modeling is used.
- Direct Value: By feeding automated processes, data adds value to a company. Consider Netflix's recommendation engine [6].

Data has a long history of providing indirect value. Accounting, insurance risk modeling, medical research experimental design, and intelligence analytics are all based on it. The data used to generate reports and visualizations come under the category of indirect value. This can be accomplished when people read our report or visualization, assimilate the information into their existing world knowledge, and then apply that knowledge to improve their behaviors. The data here has an indirect influence on other people's judgments. The majority of our data's known potential value will be given indirectly in the near future.

Giving data-driven systems decisions for speed, accuracy, or customization provides direct value from data. The most common example is resource distribution and routing that is automated. This resource is primarily money in the field of high-frequency trading and modern finance. Physical goods are routed automatically in some industries, such as Amazon or Flipkart. Hotstar and Netflix, for example, employ automated processes to optimize the distribution of digital content to their customers. For example, antilock brakes in automobiles employ sensor data to channel energy to individual wheels on a smaller scale. Modern testing systems, such as the GRE graduate school admission exam, dynamically order questions based on the tester's progress. A considerable percentage of operational choices is directly handled by data-driven systems in all of these situations, with no human input.

1.2 Data Workflow Structure

In order to derive direct, automated value from our data, we must first derive indirect, human-mediated value. To begin, human monitoring is essential to determine what is "in" our data and whether the data's quality is high enough to be used in direct and automated methods. We cannot anticipate valuable outcomes from sending data into an automated system blindly. To fully comprehend the possibilities of the data, reports must be written and studied. As the potential of the data becomes clearer, automated methods can be built to utilize it directly. This is the logical evolution of information sets: from immediate solutions to identified problems to longer-term analyses of a dataset's fundamental quality and potential applications, and finally to automated data creation systems. The passage of data through three primary data stages:

a) raw,
b) refined,
c) produced,

is at the heart of this progression.

1.3 Raw Data Stage

In the raw data stage, there are three main actions: data input, generic metadata creation, and proprietary metadata creation. As illustrated in

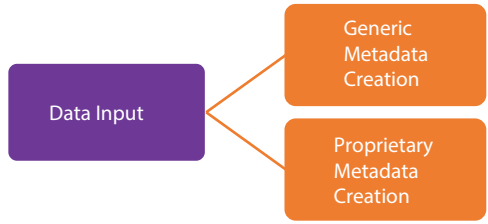

Figure 1.1 Actions in the raw data stage.

Figure 1.1, based on their production, we can classify these actions into two groups. The two ingestion actions are split into two categories, one of which is dedicated to data output. The second group of tasks is metadata production, which is responsible for extracting information and insights from the dataset.

The major purpose of the raw stage is to uncover the data. We ask questions to understand what our data looks like when we examine raw data. Consider the following scenario:

- What are the different types of records in the data?
- How are the fields in the records encoded?
- What is the relationship between the data and our organization, the kind of processes we have, and the other data we already have?

1.3.1 Data Input

The ingestion procedure in traditional enterprise data warehouses includes certain early data transformation processes. The primary goal of these transformations is to transfer inbound components to their standard representations in the data warehouse.

Consider the case when you are ingesting a comma separated file. The data in the CSV file is saved in predetermined locations after it has been modified to fit the warehouse's syntactic criteria. This frequently entails adding additional data to already collected data. In certain cases, appends might be as simple as putting new records to the "end" of a dataset. The add procedure gets more complicated when the incoming data contains both changes to old data and new data. In many of these instances, you will need to ingest fresh data into a separate place, where you can apply more intricate merging criteria during the refined data stage. It is important to highlight, however, that a separate refined data stage will be required

throughout the entire spectrum of ingestion infrastructures. This is due to the fact that refined data has been wrangled even further to coincide with anticipated analysis.

Data from multiple partners is frequently ingested into separate datasets, in addition to being stored in time-versioned partitions. The ingestion logic is substantially simplified as a result of this. As the data progresses through the refinement stage, the individual partner data is harmonized to a uniform data format, enabling for quick cross-partner analytics.

1.3.2 Output Actions at Raw Data Stage

In most circumstances, the data you are consuming in first stage is predefined, i.e., what you will obtain and how to use it are known to you. What will when some new data is added to the database by the company? To put it another way, what can be done when the data is unknown in part or in whole? When unknown data is consumed, two additional events are triggered, both of which are linked to metadata production. This process is referred to as "generic metadata creation." A second activity focuses on determining the value of your data based on the qualities of your data. This process is referred to as "custom metadata creation."

Let us go over some fundamentals before we get into the two metadata-generating activities. Records are the building blocks of datasets. Fields are what make up records. People, items, relationships, and events are frequently represented or corresponded to in records. The fields of a record describe the measurable characteristics of an individual, item, connection, or incident. In a dataset of retail transactions, for example, every entry could represent a particular transaction, with fields denoting the purchase's monetary amount, the purchase time, the specific commodities purchased, etc.

In relational database, you are probably familiar with the terms "rows" and "columns." Rows contain records and columns contain fields. Representational consistency is defined by structure, granularity, accuracy, temporality, and scope. As a result, there are also features of a dataset that your wrangling efforts must tune or improve. The data discovery process frequently necessitates inferring and developing specific information linked to the potential value of your data, in addition to basic metadata descriptions.

1.3.3 Structure

The format and encoding of a dataset's records and fields are referred to as the dataset's structure. We can place datasets on a scale based on how

homogeneous their records and fields are. The dataset is "rectangular" at one end of the spectrum and can be represented as a table. The table's rows contain records and columns contain fields in this format. You may be dealing with a "jagged" table when the data is inconsistent. A table like this is not completely rectangular any longer. Data formats like XML and JSON can handle data like this with inconsistent values.

Datasets containing a diverse set of records are further along the range. A heterogeneous dataset from a retail firm, for example, can include both customer information and customer transactions. When considering the tabs in a complex Excel spreadsheet, this is a regular occurrence. The majority of analysis and visualization software will need that these various types of records be separated and separate files are formed.

1.3.4 Granularity

A dataset's granularity relates to the different types of things that represents the data. Data entries represent information about a large number of different instances of the same type of item. The roughness and refinement of granularity are often used phrases. This refers to the depth of your dataset's records, or the number of unique entities associated with a single entry, in the context of data. A data with fine granularity might contain an entry indicating one transaction by only one consumer.

You might have a dataset with even finer granularity, with each record representing weekly combined revenue by location. The granularity of the dataset may be coarse or fine, depending on your intended purpose. Assessing the granularity of a dataset is a delicate process that necessitates the use of organizational expertise. These are some examples of granularity-related custom metadata.

1.3.5 Accuracy

The quality of a data is measured by the accuracy. The records used to populate the dataset's fields should be consistent and correct. Consider the case of a customer activities dataset. This collection of records includes information on when clients purchased goods. The record's identification may be erroneous in some cases; for example, a UPC number can have missing digits or it can be expired. Any analysis of the dataset would be limited by inaccuracies, of course. Spelling mistakes, unavailability of the variables, numerical floating value mistakes, are all examples of common inaccuracies.

Some values can appear more frequently and some can appear less frequently in a database. This condition is called frequency outliers which

can also be assessed with accuracy. Because such assessments are based on the knowledge of an individual organization and making frequency assessments is essentially a custom metadata matter.

1.3.6 Temporality

A record present in the table is a snapshot of a commodity at a specific point of time. As a result, even if a dataset had a consistent representation at the development phase and later some changes may cause it to become inaccurate or inconsistent. You could, for example, utilize a dataset of consumer actions to figure out how many goods people own. However, some of these things may be returned weeks or months after the initial transaction. The initial dataset is not the accurate depiction of objects purchased by a customer, despite being an exact record of the original sales transaction.

The time-sensitive character of representations, and thus datasets, is a crucial consideration that should be mentioned explicitly. Even if time is not clearly recorded, then also it is very crucial to know the influence of time on the data.

1.3.7 Scope

A dataset's scope has two major aspects. The number of distinct properties represented in a dataset is the first dimension. For example, we might know when a customer action occurred and some details about it. The second dimension is population coverage by attribute. Let us start with the number of distinct attributes in a dataset before moving on to the importance of scope. In most datasets, each individual attribute is represented by a separate field. There exists a variety of fields in a dataset with broad scope and in case of datasets with narrow scope, there exists a few fields.

The scope of a dataset can be expanded by including extra field attributes. Depending on your analytics methodology, the level of detail necessary may vary. Some procedures, such as deep learning, demand for keeping a large number of redundant attributes and using statistical methods to reduce them to a smaller number. Other approaches work effectively with a small number of qualities. It is critical to recognize the systematic biasness in a dataset since any analytical inferences generated from the biased dataset would be incorrect. Drug trial datasets are usually detailed to the patient level. If, on the other hand, the scope of the dataset has been deliberately changed by tampering the records of patients due to their death during trial or due to abnormalities shown by the machine, the analysis of the used medical dataset is shown misrepresented.

Also of Interest

By the same editors

ADVANCES IN DATA SCIENCE AND ANALYTICS, Edited by M. Niranjanamurthy, Hemant Kumar Gianey, and Amir H. Gandomi, ISBN: 9781119791881. Presenting the concepts and advances of data science and analytics, this volume, written and edited by a global team of experts, also goes into the practical applications that can be utilized across multiple disciplines and industries, for both the engineer and the student, focusing on machining learning, big data, business intelligence, and analytics.

WIRELESS COMMUNICATION SECURITY: Mobile and Network Security Protocols, Edited by Manju Khari, Manisha Bharti, and M. Niranjanamurthy, ISBN: 9781119777144. Presenting the concepts and advances of wireless communication security, this volume, written and edited by a global team of experts, also goes into the practical applications for the engineer, student, and other industry professionals.

MEDICAL IMAGING, Edited by H. S. Sanjay, and M. Niranjanamurthy ISBN: 9781119785392. Written and edited by a team of experts in the field, this is the most comprehensive and up-to-date study of and reference for the practical applications of medical imaging for engineers, scientists, students, and medical professionals.

SECURITY ISSUES AND PRIVACY CONCERNS IN INDUSTRY 4.0 APPLICATIONS, Edited by Shibin David, R. S. Anand, V. Jeyakrishnan, and M. Niranjanamurthy, ISBN: 9781119775621. Written and edited by a team of international experts, this is the most comprehensive and up-to-date coverage of the security and privacy issues surrounding Industry 4.0 applications, a must-have for any library.

Check out these other related titles from Scrivener Publishing

CONVERGENCE OF DEEP LEARNING IN CYBER-IOT SYSTEMS AND SECURITY, Edited by Rajdeep Chakraborty, Anupam Ghosh, Jyotsna Kumar Mandal and S. Balamurugan, ISBN: 9781119857211. In-depth analysis of Deep Learning-based cyber-IoT systems and security which will be the industry leader for the next ten years.

MACHINE INTELLIGENCE, BIG DATA ANALYTICS, AND IOT IN IMAGE PROCESSING: Practical Applications, Edited by Ashok Kumar, Megha Bhushan, José A. Galindo, Lalit Garg and Yu-Chen Hu, ISBN: 9781119865049. Discusses both theoretical and practical aspects of how to harness advanced technologies to develop practical applications such as drone-based surveillance, smart transportation, healthcare, farming solutions, and robotics used in automation.

MACHINE LEARNING TECHNIQUES AND ANALYTICS FOR CLOUD SECURITY, Edited by Rajdeep Chakraborty, Anupam Ghosh and Jyotsna Kumar Mandal, ISBN: 9781119762256. This book covers new methods, surveys, case studies, and policy with almost all machine learning techniques and analytics for cloud security solutions.