

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
© 2023 Scrivener Publishing LLC
For more information about Scrivener publications please visit www.scrivenerpublishing.com.

All rights reserved. No part of this publication may be reproduced, stored in a retrieval system, or transmitted, in any form or by any means, electronic, mechanical, photocopying, recording, or otherwise, except as permitted by law. Advice on how to obtain permission to reuse material from this title is available at <http://www.wiley.com/go/permissions>.

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.

Limit of Liability/Disclaimer of Warranty

While the publisher and authors have used their best efforts in preparing this work, they make no representations or warranties with respect to the accuracy or completeness of the contents of this work and specifically disclaim all warranties, including without limitation any implied warranties of merchantability or fitness for a particular purpose. No warranty may be created or extended by sales representatives, written sales materials, or promotional statements for this work. The fact that an organization, website, or product is referred to in this work as a citation and/or potential source of further information does not mean that the publisher and authors endorse the information or services the organization, website, or product may provide or recommendations it may make. This work is sold with the understanding that the publisher is not engaged in rendering professional services. The advice and strategies contained herein may not be suitable for your situation. You should consult with a specialist where appropriate. Neither the publisher nor authors shall be liable for any loss of profit or any other commercial damages, including but not limited to special, incidental, consequential, or other damages. Further, readers should be aware that websites listed in this work may have changed or disappeared between when this work was written and when it is read.

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	1
	<i>Akshay Singh, Surender Singh and Jyotsna Rathee</i>	
1.1	Introduction	2
1.2	Data Workflow Structure	4
1.3	Raw Data Stage	4
1.3.1	Data Input	5
1.3.2	Output Actions at Raw Data Stage	6
1.3.3	Structure	6
1.3.4	Granularity	7
1.3.5	Accuracy	7
1.3.6	Temporality	8
1.3.7	Scope	8
1.4	Refined Stage	9
1.4.1	Data Design and Preparation	9
1.4.2	Structure Issues	9
1.4.3	Granularity Issues	10
1.4.4	Accuracy Issues	10
1.4.5	Scope Issues	11
1.4.6	Output Actions at Refined Stage	11
1.5	Produced Stage	12
1.5.1	Data Optimization	13
1.5.2	Output Actions at Produced Stage	13
1.6	Steps of Data Wrangling	14
1.7	Do's for Data Wrangling	16
1.8	Tools for Data Wrangling	16
	References	17

2	Skills and Responsibilities of Data Wrangler	19
	<i>Prabhjot Kaur, Anupama Kaushik and Aditya Kapoor</i>	
2.1	Introduction	20
2.2	Role as an Administrator (Data and Database)	21
2.3	Skills Required	22
2.3.1	Technical Skills	22
2.3.1.1	Python	22
2.3.1.2	R Programming Language	25
2.3.1.3	SQL	26
2.3.1.4	MATLAB	27
2.3.1.5	Scala	27
2.3.1.6	EXCEL	28
2.3.1.7	Tableau	28
2.3.1.8	Power BI	29
2.3.2	Soft Skills	31
2.3.2.1	Presentation Skills	31
2.3.2.2	Storytelling	32
2.3.2.3	Business Insights	32
2.3.2.4	Writing/Publishing Skills	32
2.3.2.5	Listening	33
2.3.2.6	Stop and Think	33
2.3.2.7	Soft Issues	33
2.4	Responsibilities as Database Administrator	34
2.4.1	Software Installation and Maintenance	34
2.4.2	Data Extraction, Transformation, and Loading	34
2.4.3	Data Handling	35
2.4.4	Data Security	35
2.4.5	Data Authentication	35
2.4.6	Data Backup and Recovery	35
2.4.7	Security and Performance Monitoring	36
2.4.8	Effective Use of Human Resource	36
2.4.9	Capacity Planning	36
2.4.10	Troubleshooting	36
2.4.11	Database Tuning	36
2.5	Concerns for a DBA	37
2.6	Data Mishandling and Its Consequences	39
2.6.1	Phases of Data Breaching	40
2.6.2	Data Breach Laws	41
2.6.3	Best Practices For Enterprises	41

2.7	The Long-Term Consequences: Loss of Trust and Diminished Reputation	42
2.8	Solution to the Problem	42
2.9	Case Studies	42
2.9.1	UBER Case Study	42
2.9.1.1	Role of Analytics and Business Intelligence in Optimization	44
2.9.1.2	Mapping Applications for City Ops Teams	46
2.9.1.3	Marketplace Forecasting	47
2.9.1.4	Learnings from Data	48
2.9.2	PepsiCo Case Study	48
2.9.2.1	Searching for a Single Source of Truth	49
2.9.2.2	Finding the Right Solution for Better Data	49
2.9.2.3	Enabling Powerful Results with Self-Service Analytics	50
2.10	Conclusion	50
	References	50
3	Data Wrangling Dynamics	53
	<i>Simarjit Kaur, Anju Bala and Anupam Garg</i>	
3.1	Introduction	53
3.2	Related Work	54
3.3	Challenges: Data Wrangling	55
3.4	Data Wrangling Architecture	56
3.4.1	Data Sources	57
3.4.2	Auxiliary Data	57
3.4.3	Data Extraction	58
3.4.4	Data Wrangling	58
3.4.4.1	Data Accessing	58
3.4.4.2	Data Structuring	58
3.4.4.3	Data Cleaning	58
3.4.4.4	Data Enriching	59
3.4.4.5	Data Validation	59
3.4.4.6	Data Publication	59
3.5	Data Wrangling Tools	59
3.5.1	Excel	59
3.5.2	Altair Monarch	60
3.5.3	Anzo	60
3.5.4	Tabula	61

3.5.5	Trifacta	61
3.5.6	Datameer	63
3.5.7	Paxata	63
3.5.8	Talend	65
3.6	Data Wrangling Application Areas	65
3.7	Future Directions and Conclusion	67
	References	68
4	Essentials of Data Wrangling	71
	<i>Menal Dahiya, Nikita Malik and Sakshi Rana</i>	
4.1	Introduction	71
4.2	Holistic Workflow Framework for Data Projects	72
4.2.1	Raw Stage	73
4.2.2	Refined Stage	74
4.2.3	Production Stage	74
4.3	The Actions in Holistic Workflow Framework	74
4.3.1	Raw Data Stage Actions	74
4.3.1.1	Data Ingestion	75
4.3.1.2	Creating Metadata	75
4.3.2	Refined Data Stage Actions	76
4.3.3	Production Data Stage Actions	77
4.4	Transformation Tasks Involved in Data Wrangling	78
4.4.1	Structuring	78
4.4.2	Enriching	78
4.4.3	Cleansing	79
4.5	Description of Two Types of Core Profiling	79
4.5.1	Individual Values Profiling	80
4.5.1.1	Syntactic	80
4.5.1.2	Semantic	80
4.5.2	Set-Based Profiling	80
4.6	Case Study	80
4.6.1	Importing Required Libraries	81
4.6.2	Changing the Order of the Columns in the Dataset	82
4.6.3	To Display the DataFrame (Top 10 Rows) and Verify that the Columns are in Order	82
4.6.4	To Display the DataFrame (Bottom 10 rows) and Verify that the Columns Are in Order	83
4.6.5	Generate the Statistical Summary of the DataFrame for All the Columns	83
4.7	Quantitative Analysis	84
4.7.1	Maximum Number of Fires on Any Given Day	84

4.7.2	Total Number of Fires for the Entire Duration for Every State	85
4.7.3	Summary Statistics	86
4.8	Graphical Representation	86
4.8.1	Line Graph	86
4.8.2	Pie Chart	86
4.8.3	Bar Graph	87
4.9	Conclusion	89
	References	90
5	Data Leakage and Data Wrangling in Machine Learning for Medical Treatment	91
	<i>P.T. Jamuna Devi and B.R. Kavitha</i>	
5.1	Introduction	91
5.2	Data Wrangling and Data Leakage	93
5.3	Data Wrangling Stages	94
5.3.1	Discovery	94
5.3.2	Structuring	95
5.3.3	Cleaning	95
5.3.4	Improving	95
5.3.5	Validating	95
5.3.6	Publishing	95
5.4	Significance of Data Wrangling	96
5.5	Data Wrangling Examples	96
5.6	Data Wrangling Tools for Python	96
5.7	Data Wrangling Tools and Methods	99
5.8	Use of Data Preprocessing	100
5.9	Use of Data Wrangling	101
5.10	Data Wrangling in Machine Learning	104
5.11	Enhancement of Express Analytics Using Data Wrangling Process	106
5.12	Conclusion	106
	References	106
6	Importance of Data Wrangling in Industry 4.0	109
	<i>Rachna Jain, Geetika Dhand, Kavita Sheoran and Nisha Aggarwal</i>	
6.1	Introduction	110
6.1.1	Data Wrangling Entails	110
6.2	Steps in Data Wrangling	111
6.2.1	Obstacles Surrounding Data Wrangling	113

6.3	Data Wrangling Goals	114
6.4	Tools and Techniques of Data Wrangling	115
6.4.1	Basic Data Munging Tools	115
6.4.2	Data Wrangling in Python	115
6.4.3	Data Wrangling in R	116
6.5	Ways for Effective Data Wrangling	116
6.5.1	Ways to Enhance Data Wrangling Pace	117
6.6	Future Directions	119
	References	120
7	Managing Data Structure in R	123
	<i>Mittal Desai and Chetan Dudhagara</i>	
7.1	Introduction to Data Structure	123
7.2	Homogeneous Data Structures	125
7.2.1	Vector	125
7.2.2	Factor	131
7.2.3	Matrix	132
7.2.4	Array	136
7.3	Heterogeneous Data Structures	138
7.3.1	List	139
7.3.2	Dataframe	144
	References	146
8	Dimension Reduction Techniques in Distributional Semantics: An Application Specific Review	147
	<i>Pooja Kherwa, Jyoti Khurana, Rahul Budhraj, Sakshi Gill, Shreyansh Sharma and Sonia Rathee</i>	
8.1	Introduction	148
8.2	Application Based Literature Review	150
8.3	Dimensionality Reduction Techniques	158
8.3.1	Principal Component Analysis	158
8.3.2	Linear Discriminant Analysis	161
8.3.2.1	Two-Class LDA	162
8.3.2.2	Three-Class LDA	162
8.3.3	Kernel Principal Component Analysis	165
8.3.4	Locally Linear Embedding	169
8.3.5	Independent Component Analysis	171
8.3.6	Isometric Mapping (Isomap)	172
8.3.7	Self-Organising Maps	173
8.3.8	Singular Value Decomposition	174
8.3.9	Factor Analysis	175
8.3.10	Auto-Encoders	176

8.4	Experimental Analysis	178
8.4.1	Datasets Used	178
8.4.2	Techniques Used	178
8.4.3	Classifiers Used	179
8.4.4	Observations	179
8.4.5	Results Analysis Red-Wine Quality Dataset	179
8.5	Conclusion	182
	References	182
9	Big Data Analytics in Real Time for Enterprise Applications to Produce Useful Intelligence	187
	<i>Prashant Vats and Siddhartha Sankar Biswas</i>	
9.1	Introduction	188
9.2	The Internet of Things and Big Data Correlation	190
9.3	Design, Structure, and Techniques for Big Data Technology	191
9.4	Aspiration for Meaningful Analyses and Big Data Visualization Tools	193
9.4.1	From Information to Guidance	194
9.4.2	The Transition from Information Management to Valuation Offerings	195
9.5	Big Data Applications in the Commercial Surroundings	196
9.5.1	IoT and Data Science Applications in the Production Industry	197
9.5.1.1	Devices that are Inter Linked	199
9.5.1.2	Data Transformation	199
9.5.2	Predictive Analysis for Corporate Enterprise Applications in the Industrial Sector	204
9.6	Big Data Insights' Constraints	207
9.6.1	Technological Developments	207
9.6.2	Representation of Data	207
9.6.3	Data That Is Fragmented and Imprecise	208
9.6.4	Extensibility	208
9.6.5	Implementation in Real Time Scenarios	208
9.7	Conclusion	209
	References	210
10	Generative Adversarial Networks: A Comprehensive Review	213
	<i>Jyoti Arora, Meena Tushir, Pooja Kherwa and Sonia Rathee</i>	
	List of Abbreviations	213
10.1	Introduction	214
10.2	Background	215

10.2.1	Supervised vs Unsupervised Learning	215
10.2.2	Generative Modeling vs Discriminative Modeling	216
10.3	Anatomy of a GAN	217
10.4	Types of GANs	218
10.4.1	Conditional GAN (CGAN)	218
10.4.2	Deep Convolutional GAN (DCGAN)	220
10.4.3	Wasserstein GAN (WGAN)	221
10.4.4	Stack GAN	222
10.4.5	Least Square GAN (LSGANs)	222
10.4.6	Information Maximizing GAN (INFOGAN)	223
10.5	Shortcomings of GANs	224
10.6	Areas of Application	226
10.6.1	Image	226
10.6.2	Video	226
10.6.3	Artwork	227
10.6.4	Music	227
10.6.5	Medicine	227
10.6.6	Security	227
10.7	Conclusion	228
	References	228
11	Analysis of Machine Learning Frameworks Used in Image Processing: A Review	235
	<i>Gurpreet Kaur and Kamaljit Singh Saini</i>	
11.1	Introduction	235
11.2	Types of ML Algorithms	236
11.2.1	Supervised Learning	236
11.2.2	Unsupervised Learning	237
11.2.3	Reinforcement Learning	238
11.3	Applications of Machine Learning Techniques	238
11.3.1	Personal Assistants	238
11.3.2	Predictions	238
11.3.3	Social Media	240
11.3.4	Fraud Detection	240
11.3.5	Google Translator	242
11.3.6	Product Recommendations	242
11.3.7	Videos Surveillance	243
11.4	Solution to a Problem Using ML	243
11.4.1	Classification Algorithms	243
11.4.2	Anomaly Detection Algorithm	244
11.4.3	Regression Algorithm	244

11.4.4	Clustering Algorithms	245
11.4.5	Reinforcement Algorithms	245
11.5	ML in Image Processing	246
11.5.1	Frameworks and Libraries Used for ML Image Processing	246
11.6	Conclusion	248
	References	248
12	Use and Application of Artificial Intelligence in Accounting and Finance: Benefits and Challenges	251
	<i>Ram Singh, Rohit Bansal and Niranjanamurthy M.</i>	
12.1	Introduction	252
12.1.1	Artificial Intelligence in Accounting and Finance Sector	252
12.2	Uses of AI in Accounting & Finance Sector	254
12.2.1	Pay and Receive Processing	254
12.2.2	Supplier on Boarding and Procurement	255
12.2.3	Audits	255
12.2.4	Monthly, Quarterly Cash Flows, and Expense Management	255
12.2.5	AI Chatbots	255
12.3	Applications of AI in Accounting and Finance Sector	256
12.3.1	AI in Personal Finance	257
12.3.2	AI in Consumer Finance	257
12.3.3	AI in Corporate Finance	257
12.4	Benefits and Advantages of AI in Accounting and Finance	258
12.4.1	Changing the Human Mindset	259
12.4.2	Machines Imitate the Human Brain	260
12.4.3	Fighting Misrepresentation	260
12.4.4	AI Machines Make Accounting Tasks Easier	260
12.4.5	Invisible Accounting	261
12.4.6	Build Trust through Better Financial Protection and Control	261
12.4.7	Active Insights Help Drive Better Decisions	261
12.4.8	Fraud Protection, Auditing, and Compliance	262
12.4.9	Machines as Financial Guardians	263
12.4.10	Intelligent Investments	264
12.4.11	Consider the “Runaway Effect”	264
12.4.12	Artificial Control and Effective Fiduciaries	264
12.4.13	Accounting Automation Avenues and Investment Management	265

12.5	Challenges of AI Application in Accounting and Finance	265
12.5.1	Data Quality and Management	267
12.5.2	Cyber and Data Privacy	267
12.5.3	Legal Risks, Liability, and Culture Transformation	267
12.5.4	Practical Challenges	268
12.5.5	Limits of Machine Learning and AI	269
12.5.6	Roles and Skills	269
12.5.7	Institutional Issues	270
12.6	Suggestions and Recommendation	271
12.7	Conclusion and Future Scope of the Study	272
	References	272
13	Obstacle Avoidance Simulation and Real-Time Lane Detection for AI-Based Self-Driving Car	275
	<i>B. Eshwar, Harshaditya Sheoran, Shivansh Pathak and Meena Rao</i>	
13.1	Introduction	275
13.1.1	Environment Overview	277
13.1.1.1	Simulation Overview	277
13.1.1.2	Agent Overview	278
13.1.1.3	Brain Overview	279
13.1.2	Algorithm Used	279
13.1.2.1	Markovs Decision Process (MDP)	279
13.1.2.2	Adding a Living Penalty	280
13.1.2.3	Implementing a Neural Network	280
13.2	Simulations and Results	281
13.2.1	Self-Driving Car Simulation	281
13.2.2	Real-Time Lane Detection and Obstacle Avoidance	283
13.2.3	About the Model	283
13.2.4	Preprocessing the Image/Frame	285
13.3	Conclusion	286
	References	287
14	Impact of Suppliers Network on SCM of Indian Auto Industry: A Case of Maruti Suzuki India Limited	289
	<i>Ruchika Pharswan, Ashish Negi and Tridib Basak</i>	
14.1	Introduction	290
14.2	Literature Review	292
14.2.1	Prior Pandemic Automobile Industry/COVID-19 Thump on the Automobile Sector	294

14.2.2	Maruti Suzuki India Limited (MSIL) During COVID-19 and Other Players in the Automobile Industry and How MSIL Prevailed	296
14.3	Methodology	297
14.4	Findings	298
14.4.1	Worldwide Economic Impact of the Epidemic	298
14.4.2	Effect on Global Automobile Industry	298
14.4.3	Effect on Indian Automobile Industry	301
14.4.4	Automobile Industry Scenario That Can Be Expected Post COVID-19 Recovery	306
14.5	Discussion	306
14.5.1	Competitive Dimensions	306
14.5.2	MSIL Strategies	307
14.5.3	MSIL Operations and Supply Chain Management	308
14.5.4	MSIL Suppliers Network	309
14.5.5	MSIL Manufacturing	310
14.5.5	MSIL Distributors Network	311
14.5.6	MSIL Logistics Management	312
14.6	Conclusion	312
	References	312
	About the Editors	315
	Index	317

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. Niranjnamurthy, 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

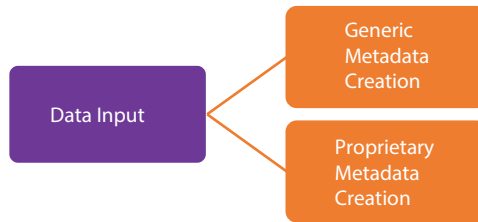


Figure 1.1 Actions in the raw data stage.

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 pre-defined, 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.

- Tableau, 28–29, 49, 50, 100
- Tabula, 61, 62f, 115
- .tail()* function, 83, 84f
- Talend, 65, 75
- Tang, W., 224
- TanH activation function, 221
- Tata motors, 290–291, 296, 302, 304t, 305t, 306
- Tata –Nano, 308
- Technical skills, of data wrangler, 22–30
 - Excel, 28
 - MATLAB, 27
 - Power BI, 29–30
 - python, 22–25
 - R programming language, 25–26
 - Scala, 27–28
 - SQL, 26–27
 - Tableau, 28–29
- Temporal difference (TD), 280
- Temporality, 8
- Tenenbaum, J.B., 149
- Tensorflow, 247
- TensorFlow K-NN classification technique, 194
- Tesla, 292
- Test dataset, 237
- Text mining, 192
- t()* function, 136
- Theano, 116
- Theft, data, 40
- Thermal imaging sensor, 199
- Tokuda, K., 168–169
- Tomer, S., 294
- Tools, data wrangling, 59–65
 - Altair Monarch, 60, 61f
 - Anzo, 60, 61, 62f
 - basic data munging tools, 115
 - cleaning and consolidating data, 100
 - Datameer, 63, 64f
 - Excel, 59–60
 - extracting insights from data, 100
 - Paxata, 63, 64f
 - processing and organizing data, 99–100
 - for python, 96–99, 115–116
 - R tool, 116
 - Tabula, 61, 62f, 115
 - Talend, 65
 - Trifacta, 61, 63
- Toyota, 290, 291, 294, 301, 302, 304t, 305t
- #ToyotaWithIndia, 294
- Traffic data, 66–67
- Training dataset, 237
- Transformation, data, 2, 21, 26–27, 34, 54, 63, 66, 71, 117
- Transformation tasks, in data wrangling, 78–79
 - cleansing, 79
 - enriching, 78–79
 - structuring, 78
- Transpose of matrix, 136
- Trifacta, 49, 50, 55, 61, 63
- Trifacta wrangler, 55, 61, 66
- Troubleshooting, 36
- Trust, loss of, 42
- Tuytelaars, T., 226
- Twitter, 119, 194

- Uber (case study), 42–48
- UberPOOL, 46
- UL (unsupervised learning), 236, 237, 239t, 245
- Unions, 79
- United States, COVID-19 on
 - automotive sector, 300–301
- Unsupervised learning (UL), 236, 237, 239t, 245
 - supervised vs, 215–216
- Unsupervised machine learning algorithms, 99, 105

- VAEs (variational autoencoders), 67, 215, 224
- Validation,
 - data, 15, 59, 95, 111
 - dataset, 104
- Valkov, L., 224
- Valuation offerings, information management to, 195–196
- Value-added data system (VADA), 66
- van der Maaten, L.J.P., 148
- van Ham, F., 54, 81
- Varghese, S., 293
- Variances, defined, 159
- Variational autoencoders (VAEs), 67, 215, 224
- Vectors, data structure in R, 124, 125–131
 - arithmetic operations, 129–130
 - atomic vectors, types, 125–126
 - element recycling, 130
 - elements, accessing, 128–129
 - nesting of, 129
 - sorting of, 130–131
 - using `c()` function, 127–128
 - using colon operator, 126
 - using sequence (`seq`) operator, 127
- VEEGAN, 224
- Verizon, 42
- Videos, 226
 - surveillance, 243
- Vidya, R., 292
- Visa exchange, 257
- Visualization,
 - data, 24, 45, 48–49
 - map, 46–47
- VLOOKUP function, 28
- Volkswagen, 293, 306

- Waldstein, S.M., 227
- Wang, L., 226
- Wang, Z., 222
- WannaCry, 38
- Warde-Farley, D., 214
- Warehouse administrator, 21
- Wasserstein distance, 221
- Wasserstein GANs (WGANs), 218, 221–222
- WebGazer, 247–248
- Websites, online shopping, 242
- Wei, X., 226
- #WePledgeToBeSafe, 294
- WGANs (Wasserstein GANs), 218, 221–222
- Wikiart dataset, 227
- Wisconsin breast cancer dataset, 178, 179, 181t
- Within-class scatter matrix, 163, 164
- Wood inspection, 173
- Workflow framework, holistic,
 - actions in, 74–78
 - production data stage, 77–78
 - raw data stage, 74–76
 - refined data stage, 76–77
 - for data projects, 72–74
- World Health Organization (WHO), 294
- Wrangler edge, 61
- Wrangler enterprise, 61
- Writing skills, 32–33

- Xero, 261
- Xie, H., 222
- XML, data format, 7
- Xu, B., 214
- Xu, T., 222
- Xu, Z., 293

- Yan, X., 222
- Yates, A., 66
- Yazdanbakhsh, A., 224
- YFCC100M dataset, 219
- Yoo, H., 276

- Zaremba, W., 225
- Zen, H., 168–169
- Zeng, C., 224
- Zhang, H., 222, 225
- Zhou, F., 224

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 Shubin 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.