

FOR PROFESSIONALS BY PROFESSIONALS®

Practical Data Science

Learn how to build a data science technology stack and perform good data science with repeatable methods. You will learn how to turn data lakes into business assets.

The data science technology stack demonstrated in *Practical Data Science* is built from components in general use in the industry. Data scientist Andreas Vermeulen demonstrates in detail how to build and provision a technology stack to yield repeatable results. He shows you how to apply practical methods to extract actionable business knowledge from data lakes consisting of data from a polyglot of data types and dimensions.

What You'll Learn:

- Become fluent in the essential concepts and terminology of data science and data engineering
- Build and use a technology stack that meets industry criteria
- Master the methods for retrieving actionable business knowledge
- Coordinate the handling of polyglot data types in a data lake for repeatable results

US \$39.99

Shelve in:
Big Data/Analytics

User level:
Beginning–Intermediate

SOURCE CODE ONLINE



About the Author	xv
About the Technical Reviewer	xvii
Acknowledgments	xix
Introduction	xxi
Chapter 1: Data Science Technology Stack	1
Rapid Information Factory Ecosystem	1
Data Science Storage Tools.....	2
Schema-on-Write and Schema-on-Read	2
Data Lake	4
Data Vault.....	4
Hubs	5
Links.....	5
Satellites.....	5
Data Warehouse Bus Matrix.....	6
Data Science Processing Tools	6
Spark.....	6
Spark Core.....	7
Spark SQL.....	7
Spark Streaming.....	7
MLlib Machine Learning Library	7
GraphX.....	8
Mesos.....	9
Akka	9
Cassandra	9

Kafka.....	10
Kafka Core	10
Kafka Streams	10
Kafka Connect	10
Elastic Search	11
R.....	11
Scala	12
Python	12
MQTT (MQ Telemetry Transport).....	13
What's Next?	13
Chapter 2: Vermeulen-Krennwallner-Hillman-Clark	15
Windows	15
Linux	15
It's Now Time to Meet Your Customer	16
Vermeulen PLC	16
Krennwallner AG	17
Hillman Ltd	18
Clark Ltd	18
Processing Ecosystem	19
Scala.....	20
Apache Spark	20
Apache Mesos	21
Akka	21
Apache Cassandra.....	21
Kafka	22
Message Queue Telemetry Transport	22
Example Ecosystem.....	22
Python	23
Is Python3 Ready?	23

R	26
Development Environment	27
R Packages	28
Sample Data	30
IP Addresses Data Sets	32
Customer Data Sets	35
Logistics Data Sets	35
Summary	38
Chapter 3: Layered Framework	39
Definition of Data Science Framework	40
Cross-Industry Standard Process for Data Mining (CRISP-DM)	40
Business Understanding	41
Data Understanding	42
Data Preparation	42
Modeling	42
Evaluation	42
Deployment	43
Homogeneous Ontology for Recursive Uniform Schema	43
The Top Layers of a Layered Framework	44
The Basics for Business Layer	45
The Basics for Utility Layer	46
The Basics for Operational Management Layer	47
The Basics for Audit, Balance, and Control Layer	48
The Basics for Functional Layer	49
Layered Framework for High-Level Data Science and Engineering	50
Windows	51
Linux	51
Summary	51

Chapter 4: Business Layer	53
Business Layer	53
The Functional Requirements	54
The Nonfunctional Requirements	63
Common Pitfalls with Requirements	78
Engineering a Practical Business Layer	81
Requirements	81
Requirements Registry	81
Traceability Matrix	82
Summary	83
Chapter 5: Utility Layer	85
Basic Utility Design	87
Data Processing Utilities	89
Maintenance Utilities	112
Processing Utilities	114
Engineering a Practical Utility Layer	115
Maintenance Utility	115
Data Utility	116
Processing Utility	116
Summary	117
Chapter 6: Three Management Layers	119
Operational Management Layer	119
Processing-Stream Definition and Management	119
Parameters	120
Scheduling	121
Monitoring	123
Communication	124
Alerting	124
Audit, Balance, and Control Layer	125
Audit	125
Process Tracking	130

Data Provenance	130
Data Lineage.....	130
Balance	131
Control.....	131
Yoke Solution	132
Producer	132
Consumer	133
Directed Acyclic Graph Scheduling.....	134
Yoke Example	134
Cause-and-Effect Analysis System	140
Functional Layer.....	141
Data Science Process	141
Start with a What-if Question	141
Take a Guess at a Potential Pattern.....	141
Gather Observations and Use Them to Produce a Hypothesis.....	142
Use Real-World Evidence to Verify the Hypothesis.....	142
Collaborate Promptly and Regularly with Customers and Subject Matter Experts As You Gain Insights	142
Summary.....	144
Chapter 7: Retrieve Superstep.....	147
Data Lakes	148
Data Swamps.....	149
1. Start with Concrete Business Questions	149
2. Data Governance	150
3. Data Quality	172
4. Audit and Version Management.....	172
Training the Trainer Model	172
Understanding the Business Dynamics of the Data Lake	173
R Retrieve Solution	173
Vermeulen PLC	173

Krennwallner AG	186
Hillman Ltd	188
Clark Ltd	194
Actionable Business Knowledge from Data Lakes.....	202
Engineering a Practical Retrieve Superstep.....	202
Vermeulen PLC	203
Krennwallner AG	209
Hillman Ltd	222
Clark Ltd	259
Connecting to Other Data Sources.....	261
SQLite	261
Date and Time	264
Other Databases	266
PostgreSQL.....	267
Microsoft SQL Server.....	267
MySQL	267
Oracle	268
Microsoft Excel.....	268
Apache Spark	269
Apache Cassandra.....	269
Apache Hive.....	270
Apache Hadoop	270
Amazon S3 Storage	271
Amazon Redshift	271
Amazon Web Services	272
Summary.....	273
Chapter 8: Assess Superstep.....	275
Assess Superstep	275
Errors	276
Accept the Error.....	276
Reject the Error	276

Correct the Error	277
Create a Default Value	277
Analysis of Data	277
Completeness	277
Uniqueness	278
Timeliness	278
Validity	278
Accuracy	278
Consistency	279
Practical Actions	279
Missing Values in Pandas	279
Engineering a Practical Assess Superstep	296
Vermeulen PLC	296
Krennwallner AG	339
Hillman Ltd	356
Clark Ltd	406
Summary	420
Chapter 9: Process Superstep	421
Data Vault	422
Hubs	422
Links	422
Satellites	422
Reference Satellites	423
Time-Person-Object-Location-Event Data Vault	423
Time Section	424
Person Section	427
Object Section	430
Location Section	433
Event Section	436
Engineering a Practical Process Superstep	439
Time	439

Person	463
Object	476
Location	499
Event.....	507
Data Science Process	510
Roots of Data Science	510
Monte Carlo Simulation	515
Causal Loop Diagrams.....	515
Pareto Chart.....	517
Correlation Analysis	519
Forecasting.....	519
Data Science	524
Summary.....	525
Chapter 10: Transform Superstep.....	527
Transform Superstep.....	527
Dimension Consolidation	528
Sun Model	534
Building a Data Warehouse	542
Transforming with Data Science	547
Steps of Data Exploration and Preparation.....	547
Missing Value Treatment.....	547
Common Feature Extraction Techniques	556
Hypothesis Testing	562
T-Test.....	562
Chi-Square Test	565
Overfitting and Underfitting.....	566
Polynomial Features	567
Common Data-Fitting Issue	568
Precision-Recall	570
Precision-Recall Curve	570
Sensitivity and Specificity	571

F1-Measure	571
Receiver Operating Characteristic (ROC) Analysis Curves.....	577
Cross-Validation Test.....	579
Univariate Analysis.....	580
Bivariate Analysis.....	581
Multivariate Analysis	581
Linear Regression	581
Simple Linear Regression.....	582
RANSAC Linear Regression	587
Hough Transform	589
Logistic Regression.....	590
Simple Logistic Regression	590
Multinomial Logistic Regression	592
Ordinal Logistic Regression.....	594
Clustering Techniques.....	601
Hierarchical Clustering	601
Partitional Clustering.....	608
ANOVA	613
Principal Component Analysis (PCA)	615
Factor Analysis	615
Conjoint Analysis	619
Decision Trees	624
Support Vector Machines, Networks, Clusters, and Grids	629
Support Vector Machines	629
Support Vector Networks.....	632
Support Vector Clustering.....	632
Data Mining.....	633
Association Patterns.....	633
Classification Patterns.....	635
Bayesian Classification.....	638

Sequence or Path Analysis	640
Forecasting.....	646
Pattern Recognition	646
Machine Learning	649
Supervised Learning	649
Unsupervised Learning.....	651
Reinforcement Learning	651
Bagging Data	652
Random Forests.....	655
Computer Vision (CV).....	660
Natural Language Processing (NLP).....	661
Text-Based.....	661
Speech-Based	662
Neural Networks	662
Gradient Descent.....	664
Regularization Strength.....	665
Simple Neural Network	665
TensorFlow.....	670
Basic TensorFlow.....	671
Real-World Uses of TensorFlow.....	676
The One-Arm Bandits (Contextual Version)	676
Processing the Digits on a Container	679
Summary.....	684
Chapter 11: Organize and Report Supersteps	685
Organize Superstep.....	685
Horizontal Style	686
Vertical Style.....	690
Island Style.....	693
Secure Vault Style.....	696
Association Rule Mining	699
Engineering a Practical Organize Superstep.	703

Report Superstep	718
Summary of the Results	719
Engineering a Practical Report Superstep.....	722
Graphics	741
Plot Options	742
More Complex Graphs	752
Summary	775
Pictures	776
Channels of Images	776
Cutting the Edge	778
One Size Does Not Fit All	781
Showing the Difference.....	783
Summary.....	786
Closing Words	786
Index.....	787