Data Science mit Big Data

Techniken, Werkzeuge und Algorithmen zur Analyse großer Datenmengen

DAS INHALTS-VERZEICHNIS

Inhalt

Vorwort ..... **XV**

Über die Autoren ..... **XVII**

Teil 1 Grundlagen **1**

1 Gestatten, Data – Big Data **3**

1.1 Die Facetten von Big Data 4
1.2 Definition mit 3Vs 12
1.3 Berührungspunkte mit Data Science 15
1.4 Neue Rollen und Tätigkeitsfelder 26
1.5 Erwartungsmanagement 33
1.6 Voraussetzungen 34
1.7 Überblick 35
Literatur 38

2 Grundbegriffe verteilter Systeme **41**

2.1 Historie 42
2.2 Skalierbarkeit 43
2.2.1 Amdahl's Law 43
2.2.2 Algorithmenkomplexität 44
2.2.3 Flynn's Taxonomy 46
2.2.4 PRAM-Modell 47
2.3 Immer schön skalierbar bleiben 49
2.3.1 Geiz ist geil: Share nothing! 49
2.3.2 Datenlokalität und Lastverteilung 50
2.3.3 Lockfreies Schreiben 50
2.4 Resilienz durch Redundanz 51
Literatur 54

3 Daten ... 55
3.1 Daten sind nicht gleich Daten ... 55
3.2 Schlüssel-Wert-Paare und Metadaten ... 57
3.3 Datenquellen: Woher kommen die Daten? ... 60
3.3.1 Öffentliche und kuratierte Datenplattformen ... 60
3.3.2 Operative und generierte Unternehmensdaten ... 62
3.3.3 Web Mining und Scraping ... 63
3.3.4 Datenhandel: Externe Dienste und aktiver Erwerb ... 63
3.4 Der Grad der Strukturierung ... 65
3.4.1 Unstrukturierte Daten: Das „Chaos“ ... 65
3.4.2 Strukturierte Daten: Die „Tabelle“ ... 66
3.4.3 Semistrukturierte Daten: Der Mittelweg ... 68
3.4.4 Quasistrukturierte und polystrukturierte Daten ... 70
3.4.5 Datentypen und deren Wichtigkeit ... 71
3.5 Zeilen- vs. spaltenorientierte Speicherung ... 72
3.5.1 Zeilenorientierte Speicherung ... 73
3.5.2 Spaltenorientierte Speicherung ... 73
3.5.3 Zusammenfassender Vergleich ... 74
3.6 Zusammenfassung und Ausblick ... 75
Literatur ... 75

4 Vorgehensmodelle für Big Data Analytics ... 77
4.1 Motivation ... 77
4.2 Knowledge Discovery in Databases ... 81
4.3 Cross-Industry Standard Model for Data Mining ... 84
4.4 Analytics Solutions Unified Method for Data Mining / Predictive Analytics 89
4.5 Big Data Analysis Lifecycle ... 91
4.6 Team Data Science Process ... 95
4.7 Data Science Process Model ... 100
Literatur ... 115

Teil 2 Datenhaltung ... 117

5 Datenformate ... 119
5.1 Die wichtigsten Formate im Big-Data-Umfeld ... 120
5.1.1 Zeilenorientierte und textbasierte Formate ... 120
5.1.2 Spaltenorientierte und binäre Formate ... 132
5.1.3 Data Frames: Die In-Memory-Abstraktion ... 135
5.2 Datenkonvertierung ... 139
5.2.1 Praxisbeispiel: Von CSV zu Parquet mit PySpark ... 139
5.3 Praktischer Leitfaden: Welches Format für welchen Zweck? ... 141
Literatur ... 143

6 Big Daten-Management ... 145
6.1 Relationale Datenbanken ... 146
6.1.1 Möglichkeiten & Grenzen ... 148
6.2 NoSQL ... 150
6.2.1 Kurzüberblick ... 151
6.2.2 Technische Grundlagen von NoSQL-Datenbanken ... 152
6.3 CAP-Theorem und Eventual Consistency ... 159
6.3.1 Eventual Consistency ... 161
6.4 Völlig schemalos ... 162
6.5 In-Memory: Schneller geht's im Arbeitsspeicher ... 163
Literatur ... 164

7 Data Warehouse ... 167
7.1 Historischer Rückblick und Ausgangslage ... 167
7.2 Ursprünge und Vordenker des Data-Warehouse-Ansatzes ... 171
7.3 Definitions- und Modellierungsansatz nach Inmon ... 172
7.4 Architektur nach Inmon ... 175
7.5 Kimballs dimensionaler Ansatz ... 179
7.6 Dimensionale Modellierung ... 182
7.7 Multidimensionale Datenanalyse mit OLAP-Systemen ... 189
7.8 Inmon und Kimball im Vergleich ... 204
7.9 Business Intelligence ... 206
7.10 Data Mining ... 212
Literatur ... 214

8 Data Lake ... 217
8.1 Ausgangslage und Motivation ... 217
8.2 Von der Flasche zum See: Die Data-Lake-Metapher ... 218
8.3 ETL vs. ELT ... 222
8.4 Architektur von Data Lakes ... 224
8.5 Rolle und Bedeutung von Data Lakes ... 234
Literatur ... 238

9 Data Lakehouse ... 239
9.1 Ausgangslage und Motivation ... 239
9.2 Haus mit Seeblick – Konvergenz von Data Warehouse und Data Lake ... 242
9.3 Prinzipien einer Data-Lakehouse-Architektur ... 246
9.4 Data-Lakehouse-Architektur ... 249
Literatur ... 262

10 Data Mesh ... 263
10.1 Ausgangslage und Motivation ... 263
10.2 Probleme bisheriger Ansätze ... 265
10.3 Prinzipien und Architektur ... 267
Literatur ... 282

11 NoSQL ... 283
11.1 Key-Value-Stores ... 284
11.2 Object Stores ... 294
11.3 Wide-Column Stores ... 319
11.4 Document Stores ... 333
11.5 Suchmaschinen ... 345
11.6 Graph-Datenbanken ... 353
11.7 Zeitreihendatenbanken ... 376
11.8 Vektordatenbanken ... 382
11.9 Multi-Model-Datenbanken ... 393
11.10 Praktischer Leitfaden: Welche Datenbank für welchen Zweck? ... 403
Literatur ... 404

12 Daten-Broker ... 407
12.1 Konzepte und Funktionsweise von Message Queues ... 407
12.2 Log-basierte Message Broker ... 410
12.3 Beispiel: Kafka ... 413
12.3.1 Grundbegriffe und Komponenten ... 415
12.3.2 Architektur ... 417
12.3.3 Verarbeitungs- und Zustellgarantien: Von At-most-once bis Exactly-once ... 420
12.3.4 Kafka mit Docker – Erste Schritte ... 421
Literatur ... 427

Teil 3 Datenverarbeitung ... 429

13 Verarbeitungsparadigmen ... 431
13.1 Historie ... 431
13.2 Überblick ... 432
13.3 Batch-Verarbeitung ... 434
13.3.1 MapReduce und Hadoop ... 435
13.3.2 Einleitendes Beispiel ... 436
13.3.3 Einführendes Codebeispiel ... 437
13.4 (Micro-)Batch-Verarbeitung mit Spark ... 455
13.4.1 Architektur ... 457
13.4.2 Spark starten ... 458
13.4.3 Grundlegende Konzepte oder der Kern von Spark ... 460
13.4.4 DAGs oder die Bestandteile eines Spark Jobs ... 468
13.4.5 DataFrames und SQL ... 469
13.4.6 MLlib: Verteiltes Machine Learning ... 474
13.4.7 Structured Streaming: Mikro-Batch-Verarbeitung von Datenströmen ... 480
13.5 „Echtzeit“-Verarbeitung von Datenströmen ... 488
13.5.1 Grundkonzepte ... 489
13.5.2 Verarbeitungsmuster in Streaming-Systemen ... 491
13.5.3 Stream-Processing mit Apache Flink ... 498

13.6 Apache Ignite – In-Memory-Computing-Plattform ... 518
13.6.1 Einordnung in das Big-Data-Ökosystem ... 519
13.6.2 Partitionierung und der Affinity Key ... 519
13.6.3 Praxisbeispiel: Sensordaten ... 522
13.6.4 Die Grenzen von verteiltem SQL ... 528
13.6.5 Ignite Thin Clients ... 529
13.6.6 Positionierung im Markt ... 530
13.6.7 Fazit ... 530
Literatur ... 531

14 Skalierbare Algorithmen und Datenstrukturen ... 533
14.1 Morris-Counter als Hinführung ... 534
14.2 HyperLogLog für Mengen-Kardinalitäten ... 537
14.3 In Bloom… Mengenzuhörigkeit ... 540
14.4 Häufigkeitsabschätzungen ... 542
14.5 Median und andere statistische Lagemaße ... 544
14.6 Information Retrieval und Text Mining ... 547
14.6.1 Textvorverarbeitung ... 548
14.6.2 Token-Repräsentation ... 548
14.6.3 Retrieval-Modelle ... 550
14.6.4 Textanalysealgorithmen ... 556
14.7 Schnelle Ähnlichkeitssuche ... 560
14.8 Websuche mit PageRank ... 564
Literatur ... 568

15 Datenanalyse und Visualisierung ... 571
15.1 Klassische Visualisierung ... 572
15.1.1 Die zentralen Disziplinen und ihre Abgrenzung ... 574
15.2 Human-in-the-Loop: Visualisierung ... 576
15.3 Vorgehensmodelle für visuelle Analysen ... 578
15.4 Strategien der Big-Data-Visualisierung ... 579
15.4.1 Filtern und verdichten ... 580
15.4.2 Smarte Dashboards und verknüpfte Ansichten ... 580
15.4.3 Server-Side und Client-Side GPU Rendering ... 582

15.5 Praktische Werkzeuge und Plattformen 584
15.5.1 Data Frames und Visualisierungssprachen 585
15.5.2 Interaktive Analytik und Data Apps 589
15.5.3 Echtzeit-Monitoring 591
15.5.4 Self-Service Business Intelligence mit Anwendungsbeispiel 598
15.5.5 Visualisierung für Explainable AI (XAI) 607
15.6 Zusammenfassung: Das Spektrum der Big-Data-Visualisierung 608
15.6.1 Ziel der Analyse 608
15.6.2 Zielgruppe und Anwendungszweck 609
15.6.3 Der Datenfluss 609
15.6.4 Ethische Grundsätze 610
Literatur 611

Teil 4 Praktische Umsetzung 613

16 Architekturwissen für Big Data 615
16.1 Begriffsdefinitionen 616
16.2 Microservices als Architekturstil 618
16.3 Referenzarchitekturen 620
16.3.1 Batch-„Architektur“ 620
16.3.2 Lambda-Architektur 622
16.3.3 Streaming- und Kappa-Architekturen 624
16.3.4 Quo vadis, Datenarchitektur? 628
16.3.5 Kommt nun das Data Streamhouse? 633
16.3.6 Große Webanwendungen 636
16.4 Architekturmuster für große Datenmengen 639
16.4.1 Tiered Storage 639
16.4.2 Polyglot Persistence 641
16.4.3 Event Sourcing 642
16.4.4 Dead Letter Queue 643
16.4.5 Circuit Breaker 644
Literatur 645

17 Test und Betrieb von Big-Data-Systemen 647
17.1 DevOps für skalierbare Datenplattformen 648
17.1.1 Grundlegende Technologien: Container und Orchestrierung 649
17.1.2 Die drei Säulen von DevOps in der Praxis 658
17.1.3 Überblick: Wer macht was? 667
17.2 Betriebsmodelle für Big-Data-Systeme 669
17.2.1 Grundlagen: Cloud-Deployment- und Service-Modelle 669
17.2.2 Der Self-Managed-Ansatz 671
17.2.3 Der Cloud-Managed-Ansatz 672
17.2.4 Der Hybrid-Ansatz 672
17.2.5 Fazit 674
17.3 Teststrategien für Big Data 674
17.3.1 Funktionales Testen 675
17.3.2 Nicht funktionale Tests 677
17.3.3 Chaos Engineering: Testing in Production 681
17.3.4 Beschaffung von Testdaten 682
Literatur 684

18 KI-basierte Systeme mit Generativer KI 685
18.1 Große Sprachmodelle – Large Language Models (LLMs) 686
18.1.1 Der Durchbruch: Die Transformer-Architektur 687
18.1.2 Sprachmodelle für alle? 690
18.1.3 Große Datenmengen als Erfolgsgeheimnis 691
18.2 Training eines Sprachmodells 694
18.2.1 Training (Fine-Tuning) eines vortrainierten Modells 699
18.2.2 Zwischenfazit: Sprachmodelle im Kontext von Big Data 703
18.3 Retrieval-Augmented Generation (RAG) 703
18.4 Function Calling und Agenten-basierte LLMs 706
18.4.1 Function Calling 706
18.4.2 Agenten 709
18.5 Model Context Protocol (MCP) 712
18.5.1 Das Problem: Zusammenführung heterogener Datenquellen 713
18.5.2 MCP-Beispiele mit FastMCP und LangChain 715
18.5.3 Analyse des Superstore-Datensatzes mit MCP 719
Literatur 725

Teil 5 Epilog ... **727**

19 Gedanken zum Abschluss ... **729**

19.1 Praxisbeispiele ... 729

19.2 Key Takeaways ... 733

19.3 Soziale und ethische Aspekte ... 734

19.3.1 Arbeitsplatz der Zukunft? ... 735

19.3.2 Ethische Überlegungen zur Datennutzung ... 735

19.4 Farewell ... 739

Literatur ... 740

Abbildungsverzeichnis ... **741**

Tabellenverzeichnis ... **749**

Listingsverzeichnis ... **751**

Index ... **757**