

1	Einleitung	1
2	Big Data	7
2.1	Historische Entstehung	9
2.2	Big Data – ein passender Begriff?	10
2.2.1	Die drei V	11
2.2.2	Weitere Vs	14
2.2.3	Der Verarbeitungsaufwand ist big	14
2.2.4	Sicht der Industrie auf Big Data	15
2.3	Eingliederung in BI und Data Mining	16
3	Hadoop	21
3.1	Hadoop kurz vorgestellt	21
3.2	HDFS – das Hadoop Distributed File System	23
3.3	Hadoop 2.x und YARN	28
3.4	Hadoop als Single-Node-Cluster aufsetzen	30
3.4.1	Falls etwas nicht funktioniert	44
3.5	Map Reduce	46
3.6	Aufsetzen einer Entwicklungsumgebung	49
3.7	Implementierung eines Map-Reduce-Jobs	56
3.8	Ausführen eines Jobs über Kommandozeile	68
3.9	Verarbeitung im Cluster	72
3.10	Aufsetzen eines Hadoop-Clusters	74
3.11	Starten eines Jobs via Hadoop-API	86
3.12	Verketteten von Map-Reduce-Jobs	99
3.13	Verarbeitung anderer Dateitypen	115
3.14	YARN-Anwendungen	130
3.14.1	Logging und Log-Aggregation in YARN	131
3.14.2	Eine einfache YARN-Anwendung	134
3.15	Vor- und Nachteile der verteilten Verarbeitung	159

3.16 Die Hadoop Java-API	160
3.16.1 Ein einfacher HDFS-Explorer	161
3.16.2 Cluster-Monitor	173
3.16.3 Überwachen der Anwendungen im Cluster	175
3.17 Gegenüberstellung zur traditionellen Verarbeitung	177
3.18 Big Data aufbereiten	178
3.18.1 Optimieren der Algorithmen zur Datenauswertung	178
3.18.2 Ausdünnung und Gruppierung	180
3.19 Ausblick auf Apache Spark	182
3.20 Markt der Big-Data-Lösungen	184
4 Das Hadoop-Ecosystem	187
4.1 Ambari	188
4.2 Sqoop	189
4.3 Flume	189
4.4 HBase	190
4.5 Hive	191
4.6 Pig	191
4.7 ZooKeeper	191
4.8 Oozie	192
4.9 Mahout	193
4.10 Data Analytics und das Reporting	193
5 NoSQL und HBase	195
5.1 Historische Entstehung	195
5.2 Das CAP-Theorem	196
5.3 ACID und BASE	197
5.4 Typen von Datenbanken	198
5.5 Umstieg von SQL und Dateisystemen auf NoSQL oder HDFS	201
5.5.1 Methoden der Datenmigration	201
5.6 HBase	203
5.6.1 Das Datenmodell von HBase	203
5.6.2 Aufbau von HBase	206
5.6.3 Installation als Stand-alone	207
5.6.4 Arbeiten mit der HBase Shell	209
5.6.5 Verteilte Installation auf dem HDFS	211
5.6.6 Laden von Daten	214
5.6.7 HBase Java-API	226
5.6.8 Der Umstieg von einem RDBMS auf HBase	249
6 Data Warehousing mit Hive	253
6.1 Installation von Hive	254

6.2	Architektur von Hive	256
6.3	Das Command Line Interface (CLI)	257
6.4	HiveQL als Abfragesprache	259
6.4.1	Anlegen von Datenbanken	259
6.4.2	Primitive Datentypen	260
6.4.3	Komplexe Datentypen	260
6.4.4	Anlegen von Tabellen	261
6.4.5	Partitionierung von Tabellen	262
6.4.6	Externe und interne Tabellen	262
6.4.7	Löschen und Leeren von Tabellen	263
6.4.8	Importieren von Daten	264
6.4.9	Zählen von Zeilen via count	265
6.4.10	Das SELECT-Statement	265
6.4.11	Beschränken von SELECT über DISTINCT	269
6.4.12	SELECT auf partitionierte Tabellen	269
6.4.13	SELECT sortieren mit SORT BY und ORDER BY	270
6.4.14	Partitionieren von Daten durch Bucketing	271
6.4.15	Gruppieren von Daten mittels GROUP BY	272
6.4.16	Subqueries - verschachtelte Abfragen	273
6.4.17	Ergebnismengen vereinigen mit UNION ALL	273
6.4.18	Mathematische Funktionen	274
6.4.19	String-Funktionen	276
6.4.20	Aggregatfunktionen	276
6.4.21	User-Defined Functions	277
6.4.22	HAVING	285
6.4.23	Datenstruktur im HDFS	286
6.4.24	Verändern von Tabellen	286
6.4.25	Erstellen von Views	289
6.4.26	Löschen einer View	289
6.4.27	Verändern einer View	289
6.4.28	Tabellen zusammenführen mit JOINS	290
6.5	Hive Security	292
6.5.1	Implementieren eines Authentication-Providers	298
6.5.2	Authentication-Provider für HiveServer2	303
6.5.3	Verwenden von PAM zur Benutzerauthentifizierung	303
6.6	Hive und JDBC	304
6.7	Datenimport mit Sqoop	322
6.8	Datenexport mit Sqoop	324
6.9	Hive und Impala	325
6.10	Unterschied zu Pig	326
6.11	Zusammenfassung	327

7	Big-Data-Visualisierung	329
7.1	Theorie der Datenvisualisierung	329
7.2	Diagrammauswahl gemäß Datenstruktur	335
7.3	Visualisieren von Big Data erfordert ein Umdenken	336
7.3.1	Aufmerksamkeit lenken	337
7.3.2	Kontextsensitive Diagramme	339
7.3.3	3D-Diagramme	341
7.3.4	Ansätze, um Big-Data zu visualisieren	342
7.4	Neue Diagrammart	344
7.5	Werkzeuge zur Datenvisualisierung	348
7.6	Entwicklung einer einfachen Visualisierungskomponente	352
8	Auf dem Weg zu neuem Wissen – Aufbereiten, Anreichern und Empfehlen	365
8.1	Eine Big-Data-Table als zentrale Datenstruktur	368
8.2	Anreichern von Daten	370
8.2.1	Anlegen einer Wissensdatenbank	371
8.2.2	Passende Zuordnung von Daten	372
8.3	Diagrammempfehlungen über Datentypanalyse	376
8.3.1	Diagrammempfehlungen in der BDataTable	378
8.4	Textanalyse – Verarbeitung unstrukturierter Daten	384
8.4.1	Erkennung von Sprachen	385
8.4.2	Natural Language Processing	386
8.4.3	Mustererkennung mit Apache UIMA	394
9	Infrastruktur	415
9.1	Hardware	416
9.2	Betriebssystem	417
9.2.1	Paketmanager	417
9.2.2	Git	418
9.2.3	VIM	419
9.2.4	Terminalumgebung	419
9.3	Virtualisierung	420
9.4	Container	420
9.4.1	Docker-Crashkurs	421
9.4.2	Infrastructure as Code	424
9.5	Distributionen	424
9.6	Reproduzierbarkeit	425
9.7	Zusammenfassung	425
10	Programmiersprachen	427
10.1	Merkmale	428
10.1.1	Funktionale Paradigmen	428

10.2	Big-Data-Programmiersprachen	429
10.2.1	Java	429
10.2.2	Scala	430
10.2.3	Python	433
10.2.4	R	436
10.2.5	Weitere Programmiersprachen	437
10.3	Zusammenfassung	438
11	Polyglot Persistence	439
11.1	Praxis	440
11.1.1	Redis	440
11.1.2	MongoDB	443
11.1.3	Neo4j	443
11.1.4	S3	444
11.1.5	Apache Kudu	447
11.2	Zusammenfassung	447
12	Apache Kafka	449
12.1	Der Kern	450
12.2	Erste Schritte	450
12.3	Dockerfile	454
12.4	Clients	454
12.5	Python Chat Client	454
12.6	Zusammenfassung	456
13	Data Processing Engines	457
13.1	Von Map Reduce zu GPPEs	457
13.1.1	Herausforderungen	458
13.1.2	Verfahren zur Verbesserung	459
13.1.3	Von Batch und Streaming zu Lambda	461
13.1.4	Frameworks in a Nutshell	462
13.2	Apache Spark	462
13.2.1	Datasets	462
13.2.2	Von RDDs zu Data Frames	463
13.2.3	Hands On Apache Spark	463
13.2.4	Client-Programme schreiben	465
13.2.5	Das Spark-Ecosystem	470
13.3	Zusammenfassung	474
14	Streaming	475
14.1	Kernparadigmen	475
14.2	Spark Streaming	478
14.2.1	Beispiel	479

14.3	Apache Flink	480
14.4	Zusammenfassung	483
15	Data Governance	485
15.1	Begriffdschungel	486
15.2	Governance-Pfeiler	487
15.2.1	Transparenz	487
15.2.2	Verantwortung	488
15.2.3	Standardisierung	489
15.3	Fokusthemen von Data Governance	489
15.3.1	Policies	489
15.3.2	Quality	490
15.3.3	Compliance	490
15.3.4	Business Intelligence	490
15.4	Datenschutz	491
15.4.1	Werkzeuge	492
15.5	Sicherheit im Hadoop-Ecosystem	497
15.6	Metadatenmanagement	498
15.6.1	Open-Source-Werkzeuge	499
15.6.2	Kommerzielle Datenkataloge	500
15.7	Organisatorische Themen	500
15.7.1	Privacy by Design	501
15.7.2	k-Anonymity	501
15.7.3	Standards	503
15.8	Zusammenfassung	503
16	Zusammenfassung und Ausblick	505
16.1	Zur zweiten Auflage 2018	505
16.2	Zur ersten Auflage 2014	507
17	Häufige Fehler	511
18	Anleitungen	517
18.1	Installation und Verwendung von Sqoop2	517
18.2	Hadoop für Windows 7 kompilieren	523
19	Literaturverzeichnis	527
Index		531