

Inhalt

Vorwort	XI
1 Einleitung	1
2 Big-Data	7
2.1 Historische Entstehung	8
2.2 Big-Data – ein passender Begriff?	9
2.2.1 Die drei V	10
2.2.2 Das vierte V – Veracity	13
2.2.3 Der Verarbeitungsaufwand ist big	14
2.2.4 Sicht der Industrien auf Big-Data	14
2.3 Eingliederung in BI und Data-Mining	15
3 Hadoop	19
3.1 Hadoop kurz vorgestellt	20
3.2 HDFS – das Hadoop Distributed File System	21
3.3 Hadoop 2.x und YARN	25
3.4 Hadoop als Single-Node-Cluster aufsetzen	26
3.4.1 Falls etwas nicht funktioniert	39
3.5 Map-Reduce	42
3.6 Aufsetzen einer Entwicklungsumgebung	44
3.7 Implementierung eines Map-Reduce-Jobs	51
3.8 Ausführen eines Jobs über Kommandozeile	63
3.9 Verarbeitung im Cluster	67
3.10 Aufsetzen eines Hadoop-Clusters	69
3.11 Starten eines Jobs via Hadoop-API	81
3.12 Verkettung von Map-Reduce-Jobs	94
3.13 Verarbeitung anderer Dateitypen	109
3.14 YARN-Anwendungen	125
3.14.1 Logging und Log-Aggregation in YARN	125
3.14.2 Eine einfache YARN-Anwendung	129

3.15	Vor- und Nachteile der verteilten Verarbeitung	153
3.16	Die Hadoop Java-API	154
3.16.1	Ein einfacher HDFS-Explorer	155
3.16.2	Cluster-Monitor	167
3.16.3	Überwachen der Anwendungen im Cluster	169
3.17	Gegenüberstellung zur traditionellen Verarbeitung	171
3.18	Big-Data aufbereiten	172
3.18.1	Optimieren der Algorithmen zur Datenauswertung	172
3.18.2	Ausdünnung und Gruppierung	174
3.19	Ausblick auf Apache Spark	176
3.20	Markt der Big-Data-Lösungen	178
4	Das Hadoop-Ecosystem	181
4.1	Ambari	182
4.2	Sqoop	183
4.3	Flume	183
4.4	HBase	184
4.5	Hive	184
4.6	Pig	185
4.7	Zookeeper	185
4.8	Mahout	186
4.9	Spark	187
4.10	Data Analytics und das Reporting	187
5	NoSQL und HBase	189
5.1	Historische Entstehung	189
5.2	Das CAP-Theorem	190
5.3	Typen von Datenbanken	191
5.4	Umstieg von SQL und Dateisystemen auf NoSQL oder HDFS	194
5.4.1	Methoden der Datenmigration	194
5.5	HBase	196
5.5.1	Das Datenmodell von HBase	196
5.5.2	Aufbau von HBase	198
5.5.3	Installation als Stand-alone	199
5.5.4	Arbeiten mit der HBase Shell	201
5.5.5	Verteilte Installation auf dem HDFS	203
5.5.6	Laden von Daten	206
5.5.6.1	HBase Bulk Loading über die Shell	207
5.5.6.2	Datenextrakt aus einer Datenbank über Sqoop	209
5.5.7	HBase Java-API	218
5.5.8	Der Umstieg von einem RDBMS auf HBase	242

6	Data-Warehousing mit Hive	245
6.1	Installation von Hive	246
6.2	Architektur von Hive	248
6.3	Das Command Line Interface (CLI)	249
6.4	HiveQL als Abfragesprache	251
6.4.1	Anlegen von Datenbanken	251
6.4.2	Primitive Datentypen	252
6.4.3	Komplexe Datentypen	252
6.4.4	Anlegen von Tabellen	253
6.4.5	Partitionierung von Tabellen	254
6.4.6	Externe und interne Tabellen	254
6.4.7	Löschen und leeren von Tabellen	255
6.4.8	Importieren von Daten	256
6.4.9	Zählen von Zeilen via count	257
6.4.10	Das SELECT-Statement	257
6.4.11	Beschränken von SELECT über DISTINCT	260
6.4.12	SELECT auf partitionierte Tabellen	261
6.4.13	SELECT sortieren mit SORT BY und ORDER BY	261
6.4.14	Partitionieren von Daten durch Bucketing	263
6.4.15	Gruppieren von Daten mittels GROUP BY	264
6.4.16	Subqueries – verschachtelte Abfragen	265
6.4.17	Ergebnismengen vereinigen mit UNION ALL	265
6.4.18	Mathematische Funktionen	266
6.4.19	String-Funktionen	267
6.4.20	Aggregatfunktionen	268
6.4.21	User-Defined Functions	269
6.4.22	HAVING	277
6.4.23	Datenstruktur im HDFS	277
6.4.24	Verändern von Tabellen	278
6.4.25	Erstellen von Views	281
6.4.26	Löschen einer View	281
6.4.27	Verändern einer View	281
6.4.28	Tabellen zusammenführen mit JOINS	282
6.5	Hive Security	284
6.5.1	Implementieren eines Authentication-Providers	290
6.5.2	Authentication-Provider für HiveServer2	294
6.5.3	Verwenden von PAM zur Benutzerauthentifizierung	295
6.6	Hive und JDBC	296
6.7	Datenimport mit Sqoop	314
6.8	Datenexport mit Sqoop	316
6.9	Hive und Impala	317
6.10	Unterschied zu Pig	318
6.11	Zusammenfassung	319

7	Big-Data-Visualisierung	321
7.1	Theorie der Datenvisualisierung	321
7.2	Diagrammauswahl gemäß Datenstruktur	327
7.3	Visualisieren von Big-Data erfordert ein Umdenken	328
7.3.1	Aufmerksamkeit lenken	329
7.3.2	Kontextsensitive Diagramme	331
7.3.3	3D-Diagramme	333
7.3.4	Ansätze, um Big-Data zu visualisieren	334
7.4	Neue Diagrammarten	336
7.5	Werkzeuge zur Datenvisualisierung	340
7.6	Entwicklung einer einfachen Visualisierungskomponente	344
8	Auf dem Weg zu neuem Wissen – aufbereiten, anreichern und empfehlen	357
8.1	Eine Big-Data-Table als zentrale Datenstruktur	360
8.2	Anreichern von Daten	362
8.2.1	Anlegen einer Wissensdatenbank	364
8.2.2	Passende Zuordnung von Daten	364
8.3	Diagrammempfehlungen über Datentypanalyse	368
8.3.1	Diagrammempfehlungen in der BDDTable	370
8.4	Textanalyse – Verarbeitung unstrukturierter Daten	376
8.4.1	Erkennung von Sprachen	377
8.4.2	Natural Language Processing	378
8.4.2.1	Klassifizierung	379
8.4.2.2	Sentiment-Analysis	384
8.4.3	Mustererkennung mit Apache UIMA	386
9	Zusammenfassung und Ausblick	405
10	Häufige Fehler	409
11	Anhang	415
11.1	Installation und Verwendung von Sqoop2	415
11.2	Hadoop für Windows 7 kompilieren	421
	Literaturverzeichnis	425
	Index	429