

Handbuch Data Engineering

Robuste Datensysteme planen und erstellen

DAS INHALTS- VERZEICHNIS

» Hier geht's
direkt
zum Buch

Vorwort	19
----------------------	-----------

Teil I Grundlagen und Bausteine

1 Data Engineering – eine Beschreibung	29
Was ist Data Engineering?	29
Data Engineering – eine Definition	30
Der Data Engineering Lifecycle	31
Die Entwicklung des Data Engineers	32
Data Engineering und Data Science	37
Data Engineering – Fähigkeiten und Tätigkeiten	39
Datenreife und der Data Engineer	40
Der berufliche Werdegang und die Kompetenzen eines Data Engineers	44
Geschäftliche Verantwortlichkeiten	44
Technische Verantwortlichkeiten	46
Das Kontinuum der Rollen im Data Engineering – von A nach B ..	49
Data Engineers innerhalb eines Unternehmens	50
Nach innen gerichtete versus nach außen gerichtete Data Engineers	50
Data Engineers und andere technische Rollen	51
Data Engineers und die Unternehmensführung	56
Fazit	59
Weitere Quellen	60
2 Der Data Engineering Lifecycle	63
Was ist der Data Engineering Lifecycle?	63
Datenlebenszyklus versus Data Engineering Lifecycle	64
Generierung: Quellsysteme	65

Speicherung	68
Ingestion.	70
Transformation	73
Bereitstellung	75
Die wesentlichen Unterströmungen innerhalb des Data Engineering Lifecycle	80
Sicherheit	81
Datenmanagement	82
DataOps.	92
Datenarchitektur	97
Orchestrierung	98
Softwareentwicklung	99
Fazit	102
Weitere Quellen	103
3 Konzeption einer guten Datenarchitektur.	105
Was ist Datenarchitektur?	105
Definition der Unternehmensarchitektur.	106
Definition der Datenarchitektur.	109
»Gute« Datenarchitektur	111
Die Grundsätze guter Datenarchitektur	112
Grundsatz 1: Wählen Sie gängige Komponenten mit Bedacht aus.	113
Grundsatz 2: Planen Sie Ausfälle ein.	113
Grundsatz 3: Planen Sie für Skalierbarkeit.	114
Grundsatz 4: Architektur heißt Führung.	115
Grundsatz 5: Seien Sie immer Architekt.	116
Grundsatz 6: Entwickeln Sie lose gekoppelte Systeme.	116
Grundsatz 7: Treffen Sie reversible Entscheidungen.	118
Grundsatz 8: Priorisieren Sie das Thema Sicherheit.	119
Grundsatz 9: Nutzen Sie FinOps.	120
Wichtige Konzepte der Architektur	122
Domänen und Dienste	122
Verteilte Systeme, Skalierbarkeit und Ausfallsicherheit.	124
Enge versus lockere Kopplung: Schichten, Monolithen und Microservices	126
Benutzerzugriff: Einzelmandant versus Mehrmandanten	131
Ereignisgesteuerte Architektur.	131
Brownfield- versus Greenfield-Projekte	132
Beispiele und Arten der Datenarchitektur	134
Data Warehouse	135
Data Lake	138

Konvergenz, Data Lakes der nächsten Generation und die Datenplattform	140
Modern Data Stack	140
Lambda-Architektur	141
Kappa-Architektur	142
Das Dataflow-Modell und die Vereinheitlichung von Batch und Streaming.	143
Architektur für IoT	144
Data Mesh.	147
Weitere Beispiele von Datenarchitekturen	148
Wer ist an der Gestaltung einer Datenarchitektur beteiligt?	149
Fazit	149
Weitere Quellen	150
4 Wahl der Technologien für den kompletten Data Engineering Lifecycle	155
Größe und Fähigkeiten des Teams	156
Schnelle Markteinführung	157
Interoperabilität.	157
Kostenoptimierung und Geschäftswert	158
Gesamtbetriebskosten.	158
Total Opportunity Cost of Ownership	160
FinOps	160
Gegenwart versus Zukunft: unveränderliche versus vergängliche Technologien	161
Unser Rat	163
Standort	164
Vor Ort	164
Cloud	165
Hybride Cloud	169
Multicloud	170
Dezentralisiert: Blockchain und Edge.	171
Unser Rat	171
Argumente für die Cloud-Rückführung	172
Eigenentwicklung oder Kauf	174
Open Source	175
Proprietäre geschlossene Systeme.	179
Unser Rat	181
Monolithisch versus modular	181
Monolith.	182
Modularität.	183
Der verteilte Monolith	184
Unser Rat	185

Serverlos versus Server	186
Serverless	186
Container	187
Wie bewertet man Server versus Serverless?	188
Unser Rat	189
Optimierung, Leistung und Benchmarking	190
Big Data ... für die 1990er	191
Absurde Kostenvergleiche	191
Asymmetrische Optimierung	192
Ausschluss der Gewährleistung	192
Die Unterströmungen und ihre Auswirkungen auf die Wahl der Technologien	192
Datenmanagement	192
DataOps	193
Datenarchitektur	193
Beispiel für Orchestrierung: Airflow	194
Softwareentwicklung	195
Fazit	195
Weitere Quellen	195

Teil II Der Data Engineering Lifecycle im Detail

5 Datengenerierung in den Quellsystemen	199
Datenquellen: Wie entstehen Daten?	200
Quellsysteme: zentrale Aspekte	201
Dateien und unstrukturierte Daten	201
APIs	201
Anwendungsdatenbanken (OLTP-Systeme)	201
Das OLAP-System	203
CDC	204
Protokolle	204
Datenbankprotokolle	206
CRUD	207
Insert-only	207
Nachrichten und Streams	208
Zeitypen	209
Quellsysteme – praktische Details	210
Datenbanken	211
APIs	220
Datenfreigabe	222
Datenquellen von Drittanbietern	223
Plattformen für das Streaming von Nachrichten und Ereignissen ...	224
Mit wem arbeiten Sie zusammen?	228

Die Bedeutung der Unterströmungen für Quellsysteme	230
Sicherheit	230
Datenmanagement	230
DataOps	231
Datenarchitektur	232
Orchestrierung	233
Softwareentwicklung	234
Fazit	234
Weitere Quellen	235
6 Speicherung	237
Komponenten der Datenspeicherung	239
Magnetische Festplatten	239
Solid State Drive	241
Direktzugriffsspeicher	242
Netzwerke und CPU	243
Serialisierung	244
Kompression	245
Caching	245
Datenspeichersysteme	246
Einzelner Rechner versus verteilte Speicherung	247
Eventuelle versus starke Konsistenz	247
Dateispeicher	249
Blockspeicher	251
Objektspeicher	255
Cache- und RAM-basierte Speichersysteme	261
Hadoop	262
Streaming-Storage	263
Indizes, Partitionen und Cluster	263
Speicherkonzepte im Data Engineering	266
Data Warehouse	266
Data Lake	267
Data Lakehouse	267
Datenplattformen	268
Stream-to-Batch-Speicherarchitektur	269
Große Ideen und Trends in der Speicherung	269
Datenkatalog	269
Datenfreigabe	270
Schema	271
Trennung von Verarbeitung und Speicherung	271
Lebenszyklus der Datenspeicherung und die Datenaufbewahrung	275
Mandantenfähiger versus mehrmandantenfähiger Speicher	279

Mit wem arbeiten Sie zusammen?	280
Unterströmungen	281
Sicherheit	281
Datenmanagement	281
DataOps	282
Datenarchitektur	283
Orchestrierung	283
Softwareentwicklung	283
Fazit	284
Weitere Quellen	284
7 Ingestion	285
Was versteht man unter Ingestion?	286
Wichtige technische Überlegungen für die Ingestionsphase	287
Begrenzte und nicht begrenzte Daten	288
Häufigkeit	289
Synchrone und asynchrone Ingestion	290
Serialisierung und Deserialisierung	291
Durchsatz und Skalierbarkeit	292
Zuverlässigkeit und Beständigkeit	292
Nutzdaten	293
Push, Pull und Polling	296
Überlegungen zur Batch-Ingestion	297
Snapshot oder differenzielle Extraktion	298
Dateibasierter Export und Ingestion	298
ETL und ELT	299
Inserts, Updates und Batch-Größe	299
Datenmigration	300
Überlegungen zur Ingestion von Nachrichten und Streams	301
Weiterentwicklung des Schemas	301
Verspätet eingegangene Daten	301
Reihenfolge und mehrfache Zustellung	302
Replay	302
Time to Live	302
Nachrichtengröße	303
Fehlerbehandlung und Dead-Letter-Queues	303
Pull und Push für Verbraucher	303
Standort	304
Möglichkeiten der Dateningestion	304
Direkte Datenbankverbindung	304
Change Data Capture	306
APIs	308

Nachrichtwarteschlangen und Event-Streaming-Plattformen . . .	309
Verwaltete Datenkonnektoren	310
Verschieben von Daten mithilfe des Objektspeichers	311
EDI	311
Datenbanken und Datelexport	312
Probleme mit gängigen Dateiformaten	312
Shell	313
SSH	313
SFTP und SCP	314
Webhooks.	314
Webinterface.	315
Web Scraping	315
Transfer Appliances für die Datenmigration	316
Datenfreigabe	317
Mit wem arbeiten Sie zusammen?	317
Vorgelagerte Stakeholder	317
Nachgelagerte Stakeholder	318
Unterströmungen	319
Sicherheit	319
Datenmanagement	319
DataOps	321
Orchestrierung	324
Softwareentwicklung	324
Fazit.	325
Weitere Quellen.	325
8 Queries, Modellierung und Transformation	327
Queries	328
Was ist eine Query?	329
Ablauf einer Abfrage	330
Der Abfrageoptimierer	331
Die Abfrageleistung verbessern.	331
Abfragen von Streaming-Daten.	338
Datenmodellierung	344
Was ist ein Datenmodell?	345
Konzeptuelle, logische und physische Datenmodelle.	346
Normalisierung.	348
Methoden der Datenmodellierung für die Batch-Analyse	352
Modellierung von Streaming-Daten	366
Transformationen	367
Batch-Transformationen.	368
Materialized Views, Federation und Query Virtualization.	384
Transformationen und Verarbeitung von Datenströmen	387

Mit wem arbeiten Sie zusammen?	390
Vorgelagerte Stakeholder	390
Nachgelagerte Stakeholder	391
Unterströmungen	391
Sicherheit	391
Datenmanagement	392
DataOps	393
Datenarchitektur	394
Orchestrierung	394
Softwareentwicklung	394
Fazit	395
Weitere Quellen	396

9 Bereitstellung von Daten für Analysen, Machine Learning und

Reverse ETL	399
Allgemeine Überlegungen zur Bereitstellung von Daten	400
Vertrauen	400
Was ist der Anwendungsfall, und wer ist der Anwender?	402
Datenprodukte	403
Self-Service oder nicht?	404
Datendefinitionen und -logik	405
Data Mesh	406
Analytik	406
Business Analytics	407
Operational Analytics	409
Embedded Analytics	411
Machine Learning	412
Was ein Data Engineer über ML wissen sollte	413
Wege der Datenbereitstellung für Analyse und ML	415
Austausch von Dateien	415
Datenbanken	416
Streaming-Systeme	418
Abfrageverbund	418
Datenfreigabe	419
Semantische und metrische Schichten	419
Datenbereitstellung in Notebooks	421
Reverse ETL	423
Mit wem arbeiten Sie zusammen?	425
Unterströmungen	425
Sicherheit	426
Datenmanagement	427
DataOps	428
Datenarchitektur	428

Orchestrierung	429
Softwareentwicklung	429
Fazit	430
Weitere Quellen	431

Teil III Sicherheit, Datenschutz und die Zukunft des Data Engineering

10 Sicherheit und Datenschutz	435
Menschen	436
Die Kraft des negativen Denkens	436
Seien Sie stets paranoid	437
Prozesse	437
Sicherheitstheater versus Sicherheitsgewohnheit	437
Aktive Sicherheit	438
Das Prinzip der geringsten Privilegien	438
Gemeinsame Verantwortung in der Cloud	439
Sichern Sie stets Ihre Daten	439
Ein Beispiel für eine Sicherheitsrichtlinie	439
Technologie	441
Systeme für Patches und Updates	441
Verschlüsselung	441
Protokollieren, überwachen und warnen	442
Netzwerkzugriff	444
Sicherheit für einfaches Data Engineering	444
Fazit	445
Weitere Quellen	445
11 Die Zukunft des Data Engineering	447
Der Data Engineering Lifecycle bleibt	448
Geringere Komplexität und benutzerfreundliche Datenwerkzeuge	448
Daten-OS in der Cloud und verbesserte Interoperabilität	449
»Unternehmerisches« Data Engineering	451
Titel und Zuständigkeiten verändern sich	452
Vom Modern Data Stack zum Live Data Stack	453
Live Data Stack	454
Streaming-Pipelines und analytische Echtzeit-Datenbanken	455
Die Verschmelzung von Daten und Anwendungen	456
Enge Rückkopplung zwischen Anwendungen und ML	456
Dark Matter Data und der Aufstieg der ... Spreadsheets?	457
Fazit	458

Anhang A	Serialisierung und Kompression – technische Details	461
Anhang B	Cloud-Vernetzung	469
Index	475