

Przedmowa	17
-----------------	----

Część I. Podstawy i bloki budulcowe	23
--	-----------

1. Czym jest inżynieria danych?	25
--	-----------

Czym jest inżynieria danych?	25
Definicja inżynierii danych	26
Cykl życia inżynierii danych	27
Ewolucja inżyniera danych	28
Inżynieria danych a nauka o danych	33
Umiejętności w zakresie inżynierii danych i wykonywane działania	34
Znaczenie dojrzałości danych dla inżyniera danych	35
Umiejętności inżyniera danych	39
Obowiązki biznesowe	40
Obowiązki techniczne	41
Kontinuum ról inżynierii danych od A do B	43
Inżynierowie danych wewnątrz organizacji	44
Inżynierowie danych wewnętrznych systemów firmy	
a inżynierowie danych systemów zewnętrznych	45
Inżynierowie danych a inne role techniczne	46
Inżynierowie danych a kierownictwo biznesowe	50
Podsumowanie	53
Zasoby dodatkowe	53

2. Cykl życia inżynierii danych?	55
---	-----------

Czym jest cykl życia inżynierii danych?	55
Cykl życia danych a cykl życia inżynierii danych	56
Generowanie — systemy źródłowe	57
Przechowywanie	59
Pozyskiwanie	62

Przekształcanie	65
Serwowanie danych	66
Główne nurty w cyklu życia inżynierii danych	71
Bezpieczeństwo	71
Zarządzanie danymi	72
DataOps	82
Architektura danych	86
Orkiestracja	87
Inżynieria oprogramowania	88
Podsumowanie	91
Zasoby dodatkowe	92
3. Projektowanie dobrej architektury danych	93
Czym jest architektura danych?	93
Definicja architektury korporacyjnej	93
Definicja architektury danych	97
„Dobra” architektura danych	98
Zasady dobrej architektury danych	99
Zasada 1. Mądrze dobieraj wspólne komponenty	100
Zasada 2. Przygotuj się na awarie	101
Zasada 3. Tworzenie architektury z myślą o skalowalności	101
Zasada 4. Architektura to przywództwo	102
Zasada 5. Pracuj nad architekturą ciągle	103
Zasada 6. Buduj luźno powiązane systemy	103
Zasada 7. Podejmuj odwracalne decyzje	105
Zasada 8. Traktuj bezpieczeństwo priorytetowo	105
Zasada 9. Korzystaj z FinOps	107
Główne pojęcia dotyczące architektury danych	108
Dziedziny i usługi	109
Systemy rozproszone, skalowalność i projektowanie z uwzględnieniem awarii	110
Sprzężenia ścisłe a sprzężenia luźne: warstwy, monolity i mikrousługi	111
Dostęp użytkowników — pojedynczy użytkownik a wielodostęp	116
Architektura sterowana zdarzeniami	116
Projekty typu brownfield kontra projekty typu greenfield	117
Przykłady i typy architektury danych	119
Hurtownia danych	119
Jeziora danych	123
Konwergencja, jeziora danych nowej generacji i platforma danych	124
Nowoczesny stos danych	125
Architektura Lambda	126
Architektura Kappa	127
Model przepływu danych oraz ujednolicone przetwarzanie wsadowe i strumieniowe	127
Architektura dla IoT	128

Siatka danych	131
Przykłady innych architektur danych	132
Kto jest zaangażowany w projektowanie architektury danych?	133
Podsumowanie	133
Zasoby dodatkowe	133
4. Wybór technologii w całym cyklu życia inżynierii danych	137
Wielkość i możliwości zespołu	138
Szybkość wprowadzania produktów na rynek	139
Interoperacyjność	139
Optymalizacja kosztów i wartości biznesowej	140
Całkowity koszt posiadania	140
Całkowity koszt alternatywny posiadania	141
FinOps	142
Teraźniejszość kontra przyszłość — technologie niezmiennie kontra przejściowe	142
Nasza rada	144
Lokalizacja	145
Lokalnie	145
Chmura	146
Chmura hybrydowa	149
Rozwiązania wielochmurowe	150
Decentralizacja. Blockchain i przetwarzanie brzegowe	151
Nasza rada	151
Argumenty za „repatriacją” z chmury	152
Budowanie zamiast kupowania	154
Oprogramowanie open source	155
Własne ogrody otoczone murem	159
Nasza rada	160
Monolit czy rozwiązanie modułowe	161
Monolit	161
Architektura modułowa	162
Wzorzec rozproszonego monolitu	164
Nasza rada	164
Rozwiązania bezserwerowe kontra rozwiązania oparte na serwerach	165
Rozwiązania bezserwerowe	165
Kontenery	166
Jak ocenić rozwiązanie serwerowe w porównaniu z bezserwerowym?	167
Nasza rada	168
Optymalizacja, wydajność i wojny testów porównawczych	169
Big data... na lata dziewięćdziesiąte	170
Bezsensowne porównania kosztów	170

Asymetryczna optymalizacja	171
Niech kupujący się strzeże	171
Nurty cyklu życia inżynierii danych i ich wpływ na wybór technologii	171
Zarządzanie danymi	171
DataOps	172
Architektura danych	172
Przykład orkiestracji — Airflow	172
Inżynieria oprogramowania	173
Podsumowanie	173
Zasoby dodatkowe	174

Część II. Cykl życia inżynierii danych w szczegółach 175

5. Generowanie danych w systemach źródłowych	177
Źródła danych — jak tworzone są dane?	178
Systemy źródłowe. Najważniejsze pojęcia	178
Pliki i dane bez struktury	179
Interfejsy API	179
Bazy danych aplikacji (systemy OLTP)	179
Systemy przetwarzania analitycznego online (OLAP)	181
Przechwytywanie zdarzeń zmiany danych	181
Logi	182
Logi bazy danych	183
CRUD	184
Tylko wstawianie	184
Komunikaty i strumienie	185
Rodzaje czasu	186
Praktyczne szczegóły dotyczące systemów źródłowych	187
Bazy danych	188
Interfejsy API	196
Współdzielenie danych	198
Zewnętrzne źródła danych	199
Kolejki komunikatów i platformy strumieniowego przesyłania zdarzeń	200
Z kim będziesz pracować?	203
Nurty inżynierii danych i ich wpływ na systemy źródłowe	205
Bezpieczeństwo	205
Zarządzanie danymi	205
DataOps	206
Architektura danych	207
Orkiestracja	208
Inżynieria oprogramowania	209
Podsumowanie	209
Zasoby dodatkowe	210

6. Składowanie	211
Podstawowe elementy systemów składowania danych	213
Dyski magnetyczne	213
Dyski SSD	215
Pamięć operacyjna	216
Infrastruktura sieci i procesor	217
Serializacja	217
Kompresja	218
Buforowanie	218
Systemy składowania danych	219
Składowanie na pojedynczym serwerze a składowanie rozproszone	220
Spójność ostateczna kontra spójność silna	220
Składowanie w plikach	221
Blokowe systemy składowania	223
Magazyn obiektów	227
Systemy składowania oparte na pamięci podręcznej i pamięci operacyjnej	233
Rozproszony system plików Hadoop	233
Składowanie strumieniowe	234
Indeksy, partycjonowanie i klastrowanie	235
Abstrakcje składowania w inżynierii danych	236
Hurtownia danych	237
Jeziora danych	238
Data lakehouse	238
Platformy danych	239
Architektura pamięci masowej stream-to-batch	239
Wielkie pomysły i trendy dotyczące składowania	240
Katalog danych	240
Współdzielenie danych	241
Schemat	241
Oddzielenie przetwarzania od składowania	242
Cykl życia systemów składowania i utrzymywanie danych	245
Magazyny dla jednego i wielu dzierżawców	249
Z kim będziesz pracować?	250
Główne nurty	250
Bezpieczeństwo	250
Zarządzanie danymi	251
DataOps	252
Architektura danych	252
Orkiestracja	253
Inżynieria oprogramowania	253
Podsumowanie	253
Zasoby dodatkowe	253

7. Pozyskiwanie danych	255
Czym jest pozyskiwanie danych?	255
Kluczowe zagadnienia inżynieryjne dotyczące fazy pozyskiwania danych	257
Dane związane kontra dane niezwiązane	257
Częstość	258
Pozyskiwanie synchroniczne a asynchroniczne	260
Serializacja i deserializacja	261
Przepustowość i skalowalność	261
Niezawodność i trwałość	262
Ładunek danych	262
Wzorce pozyskiwania pull, push czy odpytywanie?	265
Zagadnienia dotyczące pozyskiwania danych partiami	266
Ekstrakcja migawkowa lub różnicowa	267
Eksportowanie i pozyskiwanie oparte na plikach	267
Systemy ETL kontra ELT	268
Wstawianie, aktualizacje i rozmiar partii	268
Migracje danych	268
Zagadnienia dotyczące pozyskiwania komunikatów i pozyskiwania strumieniowego	269
Ewolucje schematu	269
Spóźnione dane	270
Kolejność zdarzeń i wielokrotne dostarczanie	270
Ponowne odtwarzanie	270
Czas życia	270
Rozmiar wiadomości	271
Obsługa błędów i kolejki utraconych wiadomości	271
Konsumenci typu pull kontra konsumenci typu push	272
Lokalizacja	272
Sposoby pozyskiwania danych	272
Bezpośrednie połączenie z bazą danych	272
Przechwytywanie zdarzeń zmian danych	274
Interfejsy API	276
Kolejki komunikatów i platformy strumieniowego przesyłania zdarzeń	277
Zarządzane łączniki danych	278
Przenoszenie danych za pomocą obiektowego magazynu danych	279
EDI	279
Bazy danych i eksportowanie plików	279
Problemy z popularnymi formatami plików	280
Powłoka	280
SSH	281
SFTP i SCP	281
Webhooki	281
Interfejs webowy	282

Web scraping	282
Urządzenia do przesyłania danych wykorzystywane do migracji	283
Współdzielenie danych	284
Z kim będziesz pracować?	284
Interesariusze w górnej części strumienia przetwarzania	284
Interesariusze z dolnej części strumienia przetwarzania	285
Główne nurty	286
Bezpieczeństwo	286
Zarządzanie danymi	286
DataOps	288
Orkiestracja	290
Inżynieria oprogramowania	290
Podsumowanie	291
Zasoby dodatkowe	291
8. Zapytania, modelowanie i przekształcenia	292
Zapytania	293
Czym jest zapytanie?	293
Cykl życia zapytania	295
Optymalizator zapytań	295
Poprawa wydajności zapytań	296
Zapytania do danych przekazywanych strumieniowo	301
Modelowanie danych	308
Co to jest model danych?	308
Pojęciowe, logiczne i fizyczne modele danych	309
Normalizacja	310
Techniki modelowania danych analitycznych pozyskiwanych partiami	314
Modelowanie danych pozyskiwanych strumieniowo	328
Przekształcenia	329
Przekształcenia wsadowe	330
Widoki zmateriaлизованe, federacja i wirtualizacja zapytań	343
Przekształcanie i przetwarzanie danych przekazywanych strumieniowo	346
Z kim będziesz pracować?	349
Interesariusze w górnej części strumienia przetwarzania	349
Interesariusze z dolnej części strumienia przetwarzania	350
Główne nurty	350
Bezpieczeństwo	350
Zarządzanie danymi	351
DataOps	352
Architektura danych	352

Orkiestracja	353
Inżynieria oprogramowania	353
Podsumowanie	354
Zasoby dodatkowe	355

9. Serwowanie danych na potrzeby analizy, uczenia maszynowego i odwróconych procesów ETL	357
Ogólne uwagi dotyczące serwowania danych	358
Zaufanie	358
Jaki jest przypadek użycia i kto jest użytkownikiem?	360
Produkty danych	360
Produkt samoobsługowy czy nie?	361
Definicje danych i logika	363
Siatki danych	363
Analityka	364
Analityka biznesowa	364
Analityka operacyjna	366
Analityka wbudowana	368
Uczenie maszynowe	369
Co inżynier danych powinien wiedzieć o ML?	370
Sposoby serwowania danych na potrzeby analityki i uczenia maszynowego	371
Wymiana za pomocą plików	372
Bazy danych	373
Systemy strumieniowe	374
Zapytania federacyjne	374
Współdzielenie danych	375
Warstwy semantyki i metryk	376
Serwowanie danych w notatnikach	377
Odwrócony ETL	379
Z kim będziesz pracować?	380
Główne nurty	381
Bezpieczeństwo	381
Zarządzanie danymi	382
DataOps	383
Architektura danych	384
Orkiestracja	384
Inżynieria oprogramowania	385
Podsumowanie	386
Zasoby dodatkowe	386

10. Bezpieczeństwo i prywatność	391
Ludzie	392
Moc negatywnego myślenia	392
Zawsze bądź paranoikiem	393
Procesy	393
Teatr bezpieczeństwa kontra nawyki bezpieczeństwa	393
Aktywne zabezpieczenia	394
Zasada najmniejszych uprawnień	394
Wspólna odpowiedzialność w chmurze	394
Zawsze twórz kopie zapasowe danych	395
Przykładowa polityka bezpieczeństwa	395
Technologia	396
Wdrażanie poprawek i aktualizacji	397
Szyfrowanie	397
Logowanie, monitorowanie i ostrzeganie	398
Dostęp do sieci	399
Bezpieczeństwo niskopoziomowej inżynierii danych	399
Podsumowanie	400
Zasoby dodatkowe	400
11. Przyszłość inżynierii danych	401
Cykl życia inżynierii danych nie zniknie	401
Zmniejszenie złożoności i rozwój łatwych w użyciu narzędzi danych	402
System operacyjny danych w skali chmury i lepsza interoperacyjność	403
Korporacyjna inżynieria danych	405
Tytuły zawodowe i zakresy obowiązków będą się zmieniać... ..	406
Ewolucja nowoczesnego stosu danych w kierunku stosu danych na żywo	407
Stos danych na żywo	407
Potoki strumieniowe i analityczne bazy danych czasu rzeczywistego	408
Fuzja danych z aplikacjami	409
Ścisłe sprzężenie zwrotne między aplikacjami a uczeniem maszynowym	410
Dane ciemnej materii i rozwój... arkuszy kalkulacyjnych?!	410
Podsumowanie	411
A Serializacja i kompresja. Szczegóły techniczne	413
B Sieć w chmurze	420
Skorowidz	424