Spis treści

Słowo wstępne XI

Wprowadzenie XIII

1. Wprowadzenie do analizy danych w Sparku 1
Czym jest Apache Spark? 1
Ujednolicony stos 2
Jądro Sparka 3
Spark SQL 3
Spark Streaming 3
MLlib 4
GraphX 4
Zarządzanie klastrami 4
Kto i po co korzysta ze Sparka? 4
Zadania z zakresu nauki o danych 5
Aplikacje przetwarzania danych 6
Krótka historia Sparka 6
Wersje i wydania Sparka 7
Warstwy pamięci w Sparku 7

2. Pobieranie Sparka i rozpoczęcie pracy 9
Pobieranie Sparka 9
Wprowadzenie do powłok Sparka dla Pythona i Scali 11
Wprowadzenie do podstawowych pojęć Sparka 14
Niezależne aplikacje 17
Inicjalizowanie kontekstu SparkContext 17
Tworzenie niezależnych aplikacji 19
Podsumowanie 21

3. Programowanie z rozproszonymi zbiorami danych RDD 23
Podstawy RDD 23
Tworzenie RDD 25
Działania na RDD 26
Transformacje 26
Akcje 28
Leniwa ewaluacja 29
Przekazywanie funkcji do Sparka 30
Python 30
Scala 31
Java 32
Popularne transformacje i akcje 33
Podstawowe RDD 34
Przekształcenia między typami RDD 42
Utrzymywanie (buforowanie) 44
Podsumowanie 46

4. Praca z parami klucz-wartość 47
Motywacja 47
Tworzenie RDD par 48
Transformacje na RDD par 49
Agregacje 51
Grupowanie danych 57
Złączenia 57
Sortowanie danych 59
Działania dostępne na RDD par 60
Partycjonowanie danych (zaawansowane) 60
Określanie partycjonera RDD 64
Działania, które zyskują dzięki partycjonowaniu 65
Działania, które mają wpływ na partycjonowanie 65
Przykład: PageRank 66
Niestandardowe partycjonery 68
Podsumowanie 70

5. Ładowanie i zapisywanie danych 71
Motywacja 71
Formaty plików 72
Pliki tekstowe 73

JSON 74
Wartości oddzielane przecinkami i tabulatorami 77
Pliki sekwencyjne 80
Pliki obiektowe 83
Formaty wejścia i wyjścia w Hadoop 84
Kompresja plików 88
Systemy plików 89
Lokalny lub „zwykły” 89
Amazon S3 90
HDFS 90
Dane strukturalne w Spark SQL 91
Apache Hive 91
JSON 92
Bazy danych 93
Łączniki z bazą danych Java 93
Cassandra 94
HBase 96
Elasticsearch 97
Podsumowanie 98

6. Zaawansowane programowanie w Sparku 99
Wprowadzenie 99
Akumulatory 100
Akumulatory i odporność na błędy 103
Akumulatory niestandardowe 104
Zmienne rozgłoszeniowe 104
Optymalizacja rozgłoszeń 107
Praca na poszczególnych partycjach 107
Potokowanie do programów zewnętrznych 110
Działania liczbowe na RDD 113
Podsumowanie 115

7. Uruchamianie Sparka w klastrze 117
Wprowadzenie 117
Spark Runtime Architecture 117
Sterownik 118
Wykonawcy 119
Menedżer klastrów 119

Uruchamianie programu 120
Podsumowanie 120
Wdrażanie aplikacji za pomocą spark-submit 121
Pakowanie kodu i elementów zależnych 123
Aplikacja Java Spark budowana za pomocą Mavena 124
Aplikacja Scala Spark budowana za pomocą sbt 126
Konflikty zależności 128
Harmonogramowanie w aplikacjach Sparka i między nimi 128
Menedżery klastrów 129
Menedżer klastrów Standalone 129
Hadoop YARN 133
Apache Mesos 134
Amazon EC2 135
Którego menedżera klastrów używać? 138
Podsumowanie 139

8. Dostrajanie i debugowanie Sparka 141
Konfigurowanie Sparka z SparkConf 141
Składniki wykonania: prace, zadania i etapy 145
Znajdowanie informacji 150
Sieciowy interfejs użytkownika w Sparku (Web UI) 150
Dzienniki sterownika i wykonawców 153
Kluczowe zagadnienia dotyczące wydajności 154
Poziom równoległości 155
Format serializacji 156
Zarządzanie pamięcią 157
Dostarczanie sprzętu 158
Podsumowanie 160

9. Spark SQL 161
Łączenie ze Spark SQL 162
Używanie Spark SQL w aplikacjach 164
Inicjalizacja Spark SQL 164
Przykład podstawowych zapytań 165
Elementy DataFrames 166
Buforowanie 169
Ładowanie i zapisywanie danych 170
Apache Hive 171

Źródła danych lub Parquet . . . 172
JSON . . . 173
Ze zbiorów RDD . . . 175
Serwer JDBC/ODBC . . . 176
Praca z Beeline . . . 178
Długotrwałe tablice i zapytania . . . 179
Funkcje definiowane przez użytkowników . . . 179
UDF w Spark SQL . . . 179
Funkcje UDF w Hive . . . 181
Wydajność Spark SQL . . . 181
Opcje dostrajania wydajności . . . 182
Podsumowanie . . . 183

10. Spark Streaming . . . 185
Prosty przykład . . . 186
Architektura i abstrakcja . . . 188
Transformacje . . . 192
Transformacje bezstanowe . . . 192
Transformacje stanowe . . . 195
Operacje wyjścia . . . 200
Źródła wejściowe . . . 201
Główne źródła . . . 202
Dodatkowe źródła . . . 203
Wiele źródeł i ustawianie rozmiaru klastra . . . 208
Działanie 24/7 . . . 208
Punkty kontrolne . . . 209
Odporność sterownika na błędy . . . 209
Odporność węzła roboczego na błędy . . . 211
Odporność odbiornika na błędy . . . 211
Gwarancje przetwarzania . . . 212
Interfejs użytkownika w strumieniowaniu . . . 213
Kwestie wydajności . . . 213
Rozmiary wsadu i okien . . . 213
Poziom równoległości . . . 214
Czyszczenie pamięci i jej wykorzystywanie . . . 214
Podsumowanie . . . 215

11. Systemy uczące się w MLlib 217
Przegląd 217
Wymagania dotyczące systemu 218
Podstawy systemów uczących się 219
Przykład: klasyfikacja spamu 220
Typy danych 223
Praca z wektorami 224
Algorytmy 225
Ekstrakcja cech 225
Statystyki 228
Klasyfikacja i regresja 229
Klastrowanie 234
Wspólne filtrowanie i rekomendacje 235
Zmniejszenie wymiarowości 237
Ewaluacja modelu 239
Wskazówki i kwestie wydajności 239
Przygotowanie cech 239
Konfigurowanie algorytmów 240
Bufrowanie zbiorów RDD do ponownego wykorzystania 240
Rozpoznawanie rzadkości 240
Poziom równoległości 241
API potoku 241
Podsumowanie 242

Skorowidz 243

O autorach 259