

УДК 004.65:004.43 Spark

ББК 32.972.34

K21

**Карау, Холден.**

K21 Изучаем Spark. Молниеносный анализ данных / Х. Карау, Э. Конвински, П. Венделл и др. ; пер. с англ. — 2-е изд., эл. — 1 файл pdf : 305 с. — Москва : ДМК Пресс, 2023. — Систем. требования: Adobe Reader XI либо Adobe Digital Editions 4.5 ; экран 10". — Текст : электронный.

ISBN 978-5-89818-320-2

Объем обрабатываемых данных во всех областях человеческой деятельности продолжает расти быстрыми темпами. Существуют ли эффективные приемы работы с ним? В этой книге рассказывается об Apache Spark, открытой системе кластерных вычислений, которая позволяет быстро создавать высокопроизводительные программы анализа данных. С помощью Spark вы сможете манипулировать огромными объемами данных посредством простого API на Python, Java и Scala.

Написанная разработчиками Spark, эта книга поможет исследователям данных и программистам быстро включиться в работу. Она рассказывает, как организовать параллельное выполнение заданий всего несколькими строчками кода, и охватывает примеры от простых пакетных приложений до программ, осуществляющих обработку потоковых данных и использующих алгоритмы машинного обучения.

УДК 004.65:004.43 Spark

ББК 32.972.34

**Электронное издание на основе печатного издания:** Изучаем Spark. Молниеносный анализ данных / Х. Карау, Э. Конвински, П. Венделл и др. ; пер. с англ. — Москва : ДМК Пресс, 2015. — 304 с. — ISBN 978-5-97060-323-9. — Текст : непосредственный.

Все права защищены. Любая часть этой книги не может быть воспроизведена в какой бы то ни было форме и какими бы то ни было средствами без письменного разрешения владельцев авторских прав.

Материал, изложенный в данной книге, многократно проверен. Но поскольку вероятность технических ошибок все равно существует, издательство не может гарантировать абсолютную точность и правильность приводимых сведений. В связи с этим издательство не несет ответственности за возможные ошибки, связанные с использованием книги

В соответствии со ст. 1299 и 1301 ГК РФ при устранении ограничений, установленных техническими средствами защиты авторских прав, правообладатель вправе требовать от нарушителя возмещения убытков или выплаты компенсации.

ISBN 978-5-89818-320-2

© 2015 Databricks

© Оформление, издание, ДМК Пресс, 2015

# Содержание

<b>Предисловие .....</b>	<b>10</b>
<b>Вступление .....</b>	<b>11</b>
<b>Глава 1. Введение в анализ данных с помощью Spark .....</b>	<b>18</b>
Что такое Apache Spark? .....	18
Унифицированный стек .....	19
Spark Core .....	20
Spark SQL .....	20
Spark Streaming .....	21
MLlib .....	21
GraphX .....	22
Диспетчеры кластеров .....	22
Кто и с какой целью использует Spark? .....	22
Исследование данных .....	23
Обработка данных .....	24
Краткая история развития Spark .....	24
Версии Spark .....	26
Механизмы хранения данных для Spark .....	26
<b>Глава 2. Загрузка и настройка Spark .....</b>	<b>27</b>
Загрузка Spark .....	27
Введение в командные оболочки Spark для Python и Scala .....	29
Введение в основные понятия Spark .....	33
Автономные приложения .....	35
Инициализация SparkContext .....	36
Сборка автономных приложений .....	38
В заключение .....	41
<b>Глава 3. Программирование операций с RDD .....</b>	<b>42</b>
Основы RDD .....	42
Создание RDD .....	45
Операции с RDD .....	46
Преобразования .....	46
Действия .....	47
Отложенные вычисления .....	49
Передача функций в Spark .....	50
Python .....	50
Scala .....	51
Java .....	52

Часто используемые преобразования и действия .....	54
Простые наборы RDD .....	54
Преобразование типов RDD .....	63
Сохранение (кэширование).....	65
В заключение.....	68
<b>Глава 4. Работа с парами ключ/значение .....</b>	<b>69</b>
Вступление .....	69
Создание наборов пар.....	70
Преобразования наборов пар.....	71
Агрегирование.....	73
Группировка данных .....	80
Соединения .....	81
Сортировка.....	82
Действия над наборами пар ключ/значение .....	83
Управление распределением данных.....	84
Определение объекта управления распределением RDD .....	88
Операции, получающие выгоды от наличия информации о распределении.....	89
Операции, на которые влияет порядок распределения.....	90
Пример: PageRank.....	91
Собственные объекты управления распределением .....	93
В заключение.....	96
<b>Глава 5. Загрузка и сохранение данных .....</b>	<b>97</b>
Вступление .....	97
Форматы файлов.....	98
Текстовые файлы.....	99
JSON .....	101
Значения, разделенные запятыми, и значения, разделенные табуляциями .....	104
SequenceFiles.....	108
Объектные файлы.....	111
Форматы Nadoop для ввода и вывода .....	112
Сжатие файлов.....	117
Файловые системы.....	118
Локальная/«обычная» файловая система.....	118
Amazon S3 .....	119
HDFS.....	119
Структурированные данные и Spark SQL.....	120
Apache Hive .....	121
JSON .....	122
Базы данных.....	123

Java Database Connectivity .....	123
Cassandra .....	124
HBase .....	127
Elasticsearch.....	127
В заключение .....	129
<b>Глава 6. Дополнительные возможности Spark.....</b>	<b>130</b>
Введение .....	130
Аккумуляторы.....	131
Аккумуляторы и отказоустойчивость .....	135
Собственные аккумуляторы .....	136
Широковещательные переменные.....	136
Оптимизация широковещательных рассылок.....	139
Работа с разделами по отдельности.....	140
Взаимодействие с внешними программами .....	143
Числовые операции над наборами RDD .....	147
В заключение .....	149
<b>Глава 7. Выполнение в кластере .....</b>	<b>150</b>
Введение .....	150
Архитектура среды Spark времени выполнения.....	151
Драйвер.....	151
Исполнители.....	153
Диспетчер кластера .....	153
Запуск программы .....	154
Итоги .....	154
Развертывание приложений с помощью spark-submit.....	155
Упаковка программного кода и зависимостей.....	158
Сборка приложения на Java с помощью Maven .....	159
Сборка приложения на Scala с помощью sbt.....	161
Конфликты зависимостей.....	163
Планирование приложений и в приложениях Spark .....	163
Диспетчеры кластеров.....	164
Диспетчер кластера Spark Standalone.....	165
Hadoop YARN .....	169
Apache Mesos.....	171
Amazon EC2.....	173
Выбор диспетчера кластера .....	176
В заключение .....	177
<b>Глава 8. Настройка и отладка Spark.....</b>	<b>178</b>
Настройка Spark с помощью SparkConf.....	178
Компоненты выполнения: задания, задачи и стадии .....	181

## 8 ❖ Содержание

Поиск информации .....	189
Веб-интерфейс Spark .....	189
Журналы драйверов и исполнителей.....	193
Ключевые факторы, влияющие на производительность .....	195
Степень параллелизма .....	195
Формат сериализации .....	196
Управление памятью.....	198
Аппаратное обеспечение.....	199
В заключение.....	201
<b>Глава 9. Spark SQL .....</b>	<b>202</b>
Включение Spark SQL в приложения.....	203
Использование Spark SQL в приложениях .....	205
Инициализация Spark SQL .....	205
Пример простого запроса.....	207
Наборы данных SchemaRDD.....	208
Кэширование .....	210
Загрузка и сохранение данных.....	211
Apache Hive .....	212
Parquet.....	213
JSON.....	214
Из RDD.....	216
Сервер JDBC/ODBC.....	217
Работа с программой Beeline.....	219
Долгоживущие таблицы и запросы .....	220
Функции, определяемые пользователем .....	221
Spark SQL UDF .....	221
Hive UDF .....	222
Производительность Spark SQL.....	223
Параметры настройки производительности .....	223
В заключение.....	225
<b>Глава 10. Spark Streaming.....</b>	<b>226</b>
Простой пример.....	227
Архитектура и абстракция.....	230
Преобразования.....	234
Преобразования без сохранения состояния .....	234
Преобразования с сохранением состояния.....	238
Операции вывода.....	244
Источники исходных данных .....	245
Основные источники .....	246
Дополнительные источники .....	247
Множество источников и размеры кластера.....	252

Круглосуточная работа .....	252
Копирование в контрольных точках .....	253
Повышение отказоустойчивости драйвера.....	254
Отказоустойчивость рабочих узлов .....	255
Отказоустойчивость приемников .....	256
Гарантированная обработка.....	257
Веб-интерфейс Spark Streaming.....	257
Проблемы производительности.....	258
Интервал пакетирования и протяженность окна .....	258
Степень параллелизма .....	259
Сборка мусора и использование памяти .....	259
В заключение .....	260
<b>Глава 11. Машинное обучение с MLlib .....</b>	<b>261</b>
Обзор.....	261
Системные требования .....	263
Основы машинного обучения.....	263
Пример: классификация спама .....	265
Типы данных.....	269
Векторы .....	269
Алгоритмы.....	271
Извлечение признаков .....	271
Статистики .....	275
Классификация и регрессия.....	276
Кластеризация .....	282
Коллаборативная фильтрация и рекомендации .....	283
Понижение размерности .....	285
Оценка модели .....	287
Советы и вопросы производительности .....	288
Выбор признаков.....	288
Настройка алгоритмов .....	289
Кэширование наборов RDD для повторного использования .....	289
Разреженные векторы .....	290
Степень параллелизма .....	290
Высокоуровневый API машинного обучения.....	290
В заключение .....	292
<b>Предметный указатель .....</b>	<b>293</b>