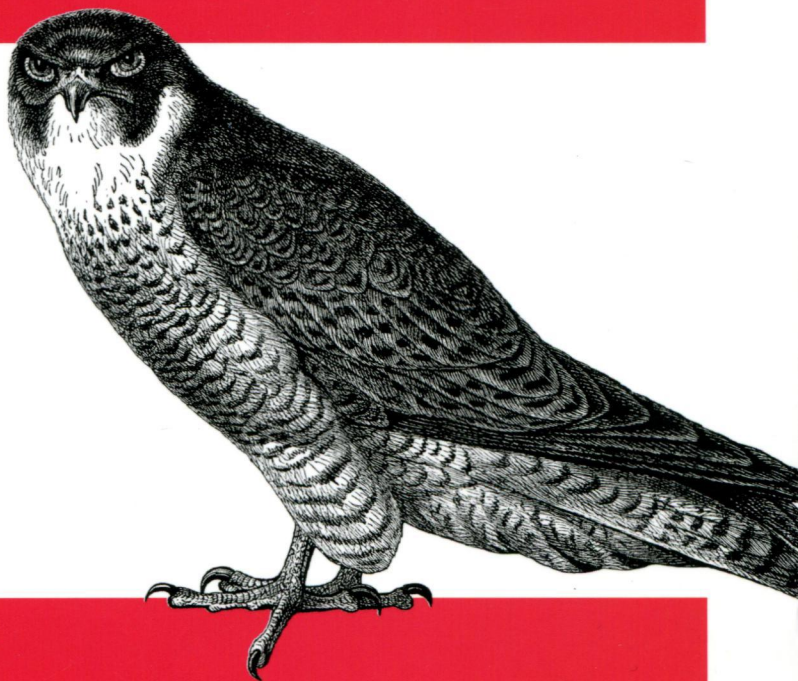


O'REILLY®



# SPARK

ДЛЯ ПРОФЕССИОНАЛОВ

Современные паттерны  
обработки больших данных

Сэнди Риза, Ури Лезерсон,  
Шон Оуэн, Джош Уиллс

 ПИТЕР®

Сэнди Риза, Ури Лезерсон,  
Шон Оуэн, Джош Уиллс

# SPARK

ДЛЯ ПРОФЕССИОНАЛОВ

---

Современные паттерны  
обработки больших данных



Санкт-Петербург · Москва · Екатеринбург · Воронеж  
Нижний Новгород · Ростов-на-Дону · Самара · Минск

2017

ББК 32.972.233.02  
УДК 004.62  
С71

**Сэнди Риза, Ури Лезерсон, Шон Оуэн, Джош Уиллс**

**С71** Spark для профессионалов: современные паттерны обработки больших данных. — СПб.: Питер, 2017. — 272 с.: ил. — (Серия «Бестселлеры O'Reilly»).  
ISBN 978-5-496-02401-3

В этой практической книге четверо специалистов Cloudera по анализу данных описывают самодостаточные паттерны для выполнения крупномасштабного анализа данных при помощи Spark. Авторы комплексно рассматривают Spark, статистические методы и множества данных, собранные в реальных условиях, и на этих примерах демонстрируют решения распространенных аналитических проблем.

**12+** (В соответствии с Федеральным законом от 29 декабря 2010 г. № 436-ФЗ.)

ББК 32.972.233.02  
УДК 004.62

Права на издание получены по соглашению с O'Reilly. Все права защищены. Никакая часть данной книги не может быть воспроизведена в какой бы то ни было форме без письменного разрешения владельцев авторских прав.

Информация, содержащаяся в данной книге, получена из источников, рассматриваемых издательством как надежные. Тем не менее, имея в виду возможные человеческие или технические ошибки, издательство не может гарантировать абсолютную точность и полноту приводимых сведений и не несет ответственности за возможные ошибки, связанные с использованием книги.

ISBN 978-1491912768 англ.

© 2016 Piter Press Ltd.

Authorized Russian translation of the English edition of Advanced Analytics with Spark, ISBN 9781491912768 © 2015 Sandy Ryza, Uri Laserson, Sean Owen and Josh Wills

This translation is published and sold by permission of O'Reilly Media, Inc., which owns or controls all rights to publish and sell the same.

ISBN 978-5-496-02401-3

© Перевод на русский язык ООО Издательство «Питер», 2017

© Издание на русском языке, оформление ООО Издательство «Питер», 2017

© Серия «Бестселлеры O'Reilly», 2017

# Краткое содержание

<b>Предисловие</b> .....	13
<b>Введение</b> .....	14
<b>Глава 1. Анализ больших данных</b> .....	17
<b>Глава 2. Введение в анализ данных с помощью Scala и Spark</b> . . . . .	25
<b>Глава 3. Рекомендация музыки и набор данных сервиса AudioScrobbler</b> .....	54
<b>Глава 4. Прогнозирование лесного покрова с использованием деревьев принятия решений</b> .....	74
<b>Глава 5. Обнаружение аномалий сетевого трафика с помощью кластеризации методом k-средних</b> .....	97
<b>Глава 6. Описание «Википедии» с помощью латентно-семантического анализа</b> .....	116
<b>Глава 7. Анализ сетей совместной встречаемости с помощью GraphX</b> .....	138
<b>Глава 8. Анализ геопространственных и временных данных на примере поездок нью-йоркских такси</b> .....	168

<b>Глава 9.</b> Оценка финансовых рисков с помощью моделирования по методу Монте-Карло . . . . .	191
<b>Глава 10.</b> Анализ геномных данных и проект BDG . . . . .	213
<b>Глава 11.</b> Анализ нейровизуальных данных с помощью PySpark и Thunder . . . . .	234
<b>Приложение А.</b> Spark: копнем поглубже . . . . .	253
<b>Приложение Б.</b> Новый API конвейеров библиотеки MLlib . . . . .	263