

# 데이터 과학과 하둡&스파크로 안내하는 완벽한 가이드

데이터 과학과 현대 하둡 생태계의 정수를 소개한다.

하둡과 스파크가 대규모 데이터 과학 문제를 해결할 수 있는 효과적인 플랫폼으로 진화해 온 과정과 그 이유를 설명한다. 더불어 데이터 과학의 응용 사례를 포괄적으로 다루며, 데이터 입수, 데이터 개조 그리고 시각화의 여러 단계에 대한 실용적인 가이드라인을 제시한다.

데이터에 대한 기초 지식을 다진 다음, 구체적인 응용 사례에 집중한다.

머신러닝, 감성 분석을 위한 예측 모델링, 문서 분석을 위한 군집화, 이상 탐지 그리고 자연어 처리에 이르는 다양한 응용 사례를 접할 수 있다.

마지막으로 데이터 과학의 실무자가 되고 싶은 당신에게!

하둡과 스파크로 어떻게 데이터 과학 과제의 ROI를 끌어올릴 수 있는지에 대한 비즈니스 지향의 가이드라인을 제시한다.

## 이 책에서 배우는 것들

- 데이터 과학이란 무엇이고 어떻게 진화를 거듭했는지 그리고 데이터 과학자의 커리어를 계획하는 방법
- 빅데이터의 특징(규모, 다양성, 속도)이 데이터 과학 활용에 어떤 영향을 주었는지
- HDFS, 맵리듀스, YARN, 스파크를 아우르는 하둡과 하둡 생태계
- 하이브와 스파크를 활용한 데이터 입수 방법
- 데이터 품질과 데이터 전처리, 데이터 준비 그리고 모델링 방법
- 대규모 데이터셋에서 통찰을 도출하는 데이터 시각화 방법
- 머신러닝: 분류, 회귀, 군집화 그리고 이상 탐지
- 예측 모델링을 위한 알고리즘과 하둡 도구
- 군집 분석과 유사도 함수들
- 대규모 이상 탐지 시스템
- 데이터 과학을 인간의 언어에 적용하는 자연어 처리 방법



하둡과 스파크를 활용한 실용 데이터 과학  
PRACTICAL DATA SCIENCE WITH HADOOP® AND SPARK

예제 소스 <https://github.com/gilbutITbook/006811>

정가 28,000원



9 791160 502213  
ISBN 979-11-6050-221-3