Addison Wesley Data & Analytics Series

Practical **Data Science** with Hadoop® and Spark

Designing
and Building
Effective Analytics
at Scale

Ofer Mendelevitch
Casey Stella
Douglas Eadline

Practical Data Science with Hadoop® and Spark

Practical Data Science with Hadoop and Spark: Designing and Building Effective Analytics at Scale

Table of Contents

Cover
Title Page
Copyright Page
Contents
Foreword
Preface
Acknowledgments
About the Authors
I: Data Science with HadoopAn Overview
1 Introduction to Data Science
What Is Data Science?
Example: Search Advertising
A Bit of Data Science History
Statistics and Machine Learning
Innovation from Internet Giants
Data Science in the Modern Enterprise
Becoming a Data Scientist
The Data Engineer
The Applied Scientist
Transitioning to a Data Scientist Role
Soft Skills of a Data Scientist

Pearson

Table of Contents

Building a Data Science Team

The Data Science Project Life Cycle

Ask the Right Question

Data Acquisition

Data Cleaning: Taking Care of Data Quality

Explore the Data and Design Model Features

Building and Tuning the Model

Deploy to Production

Managing a Data Science Project

Summary

2 Use Cases for Data Science

Big DataA Driver of Change

Volume: More Data Is Now Available

Variety: More Data Types

Velocity: Fast Data Ingest

Business Use Cases

Product Recommendation

Customer Churn Analysis

Customer Segmentation

Sales Leads Prioritization

Sentiment Analysis

Fraud Detection

Predictive Maintenance

Market Basket Analysis

Predictive Medical Diagnosis

Predicting Patient Re-admission

Detecting Anomalous Record Access

Insurance Risk Analysis

Predicting Oil and Gas Well Production Levels

Summary

3 Hadoop and Data Science

Table of Contents

What Is Hadoop?

Distributed File System

Resource Manager and Scheduler

Distributed Data Processing Frameworks

Hadoops Evolution

Hadoop Tools for Data Science

Apache Sqoop

Apache Flume

Apache Hive

Apache Pig

Apache Spark

R

Python

Java Machine Learning Packages

Why Hadoop Is Useful to Data Scientists

Cost Effective Storage

Schema on Read

Unstructured and Semi-Structured Data

Multi-Language Tooling

Robust Scheduling and Resource Management

Levels of Distributed Systems Abstractions

Scalable Creation of Models

Scalable Application of Models

Summary

II: Preparing and Visualizing Data with Hadoop

4 Getting Data into Hadoop

Hadoop as a Data Lake

The Hadoop Distributed File System (HDFS)

Direct File Transfer to Hadoop HDFS

Importing Data from Files into Hive Tables

Import CSV Files into Hive Tables

Table of Contents

Importing Data into Hive Tables Using Spark

Import CSV Files into HIVE Using Spark

Import a JSON File into HIVE Using Spark

Using Apache Sqoop to Acquire Relational Data

Data Import and Export with Sqoop

Apache Sqoop Version Changes

Using Sqoop V2: A Basic Example

Using Apache Flume to Acquire Data Streams

Using Flume: A Web Log Example Overview

Manage Hadoop Work and Data Flows with Apache Oozie

Apache Falcon

Whats Next in Data Ingestion?

Summary

5 Data Munging with Hadoop

Why Hadoop for Data Munging?

Data Quality

What Is Data Quality?

Dealing with Data Quality Issues

Using Hadoop for Data Quality

The Feature Matrix

Choosing the Right Features

Sampling: Choosing Instances

Generating Features

Text Features

Time-Series Features

Features from Complex Data Types

Feature Manipulation

Dimensionality Reduction

Summary

6 Exploring and Visualizing Data

Why Visualize Data?

Table of Contents

Motivating Example: Visualizing Network Throughput

Visualizing the Breakthrough That Never Happened

Creating Visualizations

Comparison Charts

Composition Charts

Distribution Charts

Relationship Charts

Using Visualization for Data Science

Popular Visualization Tools

R

Python: Matplotlib, Seaborn, and Others

SAS

Matlab

Julia

Other Visualization Tools

Visualizing Big Data with Hadoop

Summary

III: Applying Data Modeling with Hadoop

7 Machine Learning with Hadoop

Overview of Machine Learning

Terminology

Task Types in Machine Learning

Big Data and Machine Learning

Tools for Machine Learning

The Future of Machine Learning and Artificial Intelligence

Summary

8 Predictive Modeling

Overview of Predictive Modeling

Classification Versus Regression

Evaluating Predictive Models

Table of Contents

        - Evaluating Classifiers
        - Evaluating Regression Models
        - Cross Validation
    - Supervised Learning Algorithms
    - Building Big Data Predictive Model Solutions
        - Model Training
        - Batch Prediction
        - Real-Time Prediction
    - Example: Sentiment Analysis
        - Tweets Dataset
        - Data Preparation
        - Feature Generation
        - Building a Classifier
    - Summary
- 9 Clustering
    - Overview of Clustering
    - Uses of Clustering
    - Designing a Similarity Measure
        - Distance Functions
        - Similarity Functions
    - Clustering Algorithms
    - Example: Clustering Algorithms
        - k-means Clustering
        - Latent Dirichlet Allocation
    - Evaluating the Clusters and Choosing the Number of Clusters
    - Building Big Data Clustering Solutions
    - Example: Topic Modeling with Latent Dirichlet Allocation
        - Feature Generation
        - Running Latent Dirichlet Allocation
    - Summary
- 10 Anomaly Detection with Hadoop

Table of Contents

- Overview
- Uses of Anomaly Detection
- Types of Anomalies in Data
- Approaches to Anomaly Detection
  - Rules-based Methods
  - Supervised Learning Methods
  - Unsupervised Learning Methods
  - Semi-Supervised Learning Methods
- Tuning Anomaly Detection Systems
- Building a Big Data Anomaly Detection Solution with Hadoop
- Example: Detecting Network Intrusions
  - Data Ingestion
  - Building a Classifier
  - Evaluating Performance
- Summary

11 Natural Language Processing

- Natural Language Processing
  - Historical Approaches
  - NLP Use Cases
  - Text Segmentation
  - Part-of-Speech Tagging
  - Named Entity Recognition
  - Sentiment Analysis
  - Topic Modeling
- Tooling for NLP in Hadoop
  - Small-Model NLP
  - Big-Model NLP
- Textual Representations
  - Bag-of-Words
  - Word2vec
- Sentiment Analysis Example

Table of Contents

Stanford CoreNLP

Using Spark for Sentiment Analysis

Summary

12 Data Science with HadoopThe Next Frontier

Automated Data Discovery

Deep Learning

Summary

Appendix A: Book Web Page and Code Download

Appendix B: HDFS Quick Start

Quick Command Dereference

General User HDFS Commands

List Files in HDFS

Make a Directory in HDFS

Copy Files to HDFS

Copy Files from HDFS

Copy Files within HDFS

Delete a File within HDFS

Delete a Directory in HDFS

Get an HDFS Status Report (Administrators)

Perform an FSCK on HDFS (Administrators)

Appendix C: Additional Background on Data Science and Apache Hadoop and Spark

General Hadoop/Spark Information

Hadoop/Spark Installation Recipes

HDFS

MapReduce

Spark

Essential Tools

Table of Contents

- Machine Learning

Index

Pearson