



مؤسسه فرهنگی هنری
دیبگران تهران

به نام خدا



مؤسسه فرهنگی هنری
دیبگران تهران

کلان داده ها

استخراج، ذخیره و پردازش

مؤلفان:

دکتر امین نظارات

(عضو هیئت علمی دانشگاه پیام نور)

مهندس محسن رنجبر

مهندس فرزاد ابراهیمی



هرگونه چاپ و تکثیر از محتویات این کتاب بدون اجازه کتبی ناشر ممنوع است. متخلفان به موجب قانون حمایت حقوق مؤلفان، مصنفان و هنرمندان تحت پیگرد قانونی قرار می‌گیرند.

◀ عنوان کتاب: کلان داده ها - استخراج، ذخیره و پردازش

◀ مولفان: دکتر امین نظارات

مهندس محسن رنجبر

مهندس فرزاد ابراهیمی

◀ ناشر: موسسه فرهنگی هنری دیباگران تهران

◀ صفحه آرای: نازنین نصیری

◀ طراح جلد: داریوش فرسای

◀ نوبت چاپ: اول

◀ تاریخ نشر: ۱۳۹۸

◀ چاپ و صحافی: درج عقیق

◀ تیراژ: ۱۰۰ جلد

◀ قیمت: ۷۲۰۰۰۰ ریال

◀ شابک: ۹۷۸-۶۲۲-۲۱۸-۱۴۰-۶

نشانی واحد فروش: تهران، میدان انقلاب،

خ کارگر جنوبی، روبروی پاساژ مهستان،

پلاک ۱۲۵۱

تلفن: ۲۲۰۸۵۱۱۱-۶۶۴۱۰۰۴۶

فروشگاههای اینترنتی دیباگران تهران :

WWW.MFTBOOK.IR

www.dibagarantehran.com

www.mftdibagaran.ir

نشانی تلگرام: @mftbook

اپلیکیشن دیباگران تهران را از سایت های اینترنتی دیباگران دریافت نمایید.

سرشناسه: نظارات، امین، ۱۳۵۹-

عنوان و نام پدیدآور: کلان داده ها استخراج، ذخیره و پردازش / مولفان: امین نظارات، محسن رنجبر، فرزاد ابراهیمی.

مشخصات نشر: تهران: دیباگران تهران: ۱۳۹۸

مشخصات ظاهری: ۲۲۴ص: مصور،

شابک: ۹۷۸-۶۲۲-۲۱۸-۱۴۰-۶

وضعیت فهرست نویسی: فیبا

موضوع: داده های کلان Big Data

موضوع: پایگاه های اطلاعاتی-مدیریت

موضوع: Database management

شناسه افزوده: ابراهیمی، فرزاد، ۱۳۷۰-

شناسه افزوده: رنجبر، محسن، ۱۳۶۳-

رده بندی کنگره: ۷۶/۹ QA

رده بندی دیویی: ۰۰۵/۷۴

شماره کتابشناسی ملی: ۵۷۰۴۱۶۳

فهرست مطالب

7 مقدمه ناشر

9 مقدمه مؤلف

فصل 1

10 تحلیل داده های بزرگ

13 چالش های علم داده

14 معرفی Apache Spark

فصل 2

19 مقدمه ای بر تجزیه و تحلیل داده ها با Spark و Scala

20 Scala برای متخصص داده

22 مدل برنامه نویسی Spark

23 ارتباط رکورد

24 شروع کردن: پسته Spark و SparkContext

30 داده های توزیع شده برگشت پذیر

31 آوردن داده از خوشه به کلاینت

37 انتقال کد از کلاینت به خوشه

38 از RDD به فریم های داده

43 تحلیل داده با DataFrame API

50 آمار خلاصه سریع برای دیتافریم

52reshaping و pivoting دیتافریم‌های
58 Join کردن دیتافریم‌ها و انتخاب ویژگی‌ها
60 آماده‌سازی مدل‌ها برای محیط‌های تولید
62 ارزیابی مدل

فصل 3

66 Audioscrobbler پیشنهاد موسیقی و مجموعه داده
68 مجموعه داده
69 الگوریتم حداقل مربعات متناوب پیشنهاد دهنده
72 آماده‌سازی داده‌ها
77 ایجاد اولین مدل
79 Broadcast متغیرهای
82 بررسی نقطه نظرات
85 ارزیابی کیفیت پیشنهاد
86 محاسبه AUC
89 hyperparameter انتخاب
92 ایجاد پیشنهادها

فصل 4

96 پیش‌بینی پوشش جنگل با درخت تصمیم
97 پیش‌بینی به سوی رگرسیون
98 بردارها و ویژگی‌ها
99 درختان تصمیم و جنگل‌ها
102 Covtype مجموعه داده

103.....	پیش پردازش داده‌ها
105.....	اولین درخت تصمیم
116.....	Hyperparameter های درخت تصمیم
117.....	تنظیم درختان تصمیم
124.....	جنگل تصمیم‌گیری تصادفی

فصل 5

129.....	تشخیص ناهنجاری در ترافیک شبکه با خوشه بندی K-Means
130.....	تشخیص ناهنجاری
131.....	خوشه‌بندی K-means
132.....	نفوذ شبکه
132.....	مجموعه داده KDD Cup 1999
134.....	برداشت اول در خوشه‌بندی
138.....	انتخاب K
141.....	مجازی‌سازی با SparkR
147.....	نرمال‌سازی ویژگی
150.....	مقادیر غیر عددی
151.....	استفاده از برچسب‌ها با آنتروپی
153.....	خوشه‌بندی در عمل

فصل 6

158.....	کتابخانه MLib
159.....	Decision Trees
164.....	Gradient-Boosted Trees (GBTs)

168.....	Linear Support Vector Machines (SVMs)
171.....	Logistic regression
174.....	Naive Bayes
176.....	Random Forests
180.....	Decision Trees
186.....	Gradient-Boosted Trees (GBTs)
189.....	Isotonic regression
192.....	Linear least squares, Lasso, and ridge regression
194.....	Random Forest Regression
198.....	K-means
200.....	Bisecting k-means
202.....	Gaussian mixture
203.....	Latent Dirichlet allocation (LDA)
206.....	Power iteration clustering (PIC)
208.....	Streaming k-means
210.....	Collaborative filtering
216.....	Singular value decomposition (SVD)
218.....	Principal component analysis (PCA)
220.....	FP-growth
222.....	Association Rules
223.....	PrefixSpan

خط مشی کیفیت انتشارات مؤسسه فرهنگی هنری دیباگران تهران در عرصه کتاب‌های است که بتواند خواسته‌های به روز جامعه فرهنگی و علمی کشور را تا حد امکان پوشش دهد.

حمد و سپاس ایزد منان را که با الطاف بی‌کران خود این توفیق را به ما ارزانی داشت تا بتوانیم در راه ارتقای دانش عمومی و فرهنگی این مرز و بوم در زمینه چاپ و نشر کتب علمی دانشگاهی، علوم پایه و به ویژه علوم کامپیوتر و انفورماتیک گام‌هایی هرچند کوچک برداشته و در انجام رسالتی که بر عهده داریم، مؤثر واقع شویم.

گسترده‌گی علوم و توسعه روزافزون آن، شرایطی را به وجود آورده که هر روز شاهد تحولات اساسی چشمگیری در سطح جهان هستیم. این گسترش و توسعه نیاز به منابع مختلف از جمله کتاب را به عنوان قدیمی‌ترین و راحت‌ترین راه دستیابی به اطلاعات و اطلاع‌رسانی، بیش از پیش روشن می‌نماید.

در این راستا، واحد انتشارات مؤسسه فرهنگی هنری دیباگران تهران با همکاری جمعی از اساتید، مؤلفان، مترجمان، متخصصان، پژوهشگران، محققان و نیز پرسنل ورزیده و ماهر در زمینه امور نشر درصدد هستند تا با تلاش‌های مستمر خود برای رفع کمبودها و نیازهای موجود، منابعی پُر بار، معتبر و با کیفیت مناسب در اختیار علاقمندان قرار دهند.

کتابی که در دست دارید با همت "آقایان دکتر امین نظارات - مهندس محسن رنجبر - مهندس فرزاد ابراهیمی" و تلاش جمعی از همکاران انتشارات میسر گشته که شایسته است از یکایک این گرامیان تشکر و قدردانی کنیم.

کارشناسی و نظارت بر محتوا: زهره قزلباش

در خاتمه ضمن سپاسگزاری از شما دانش‌پژوه گرامی درخواست می‌نماید با مراجعه به آدرس dibagaran.mft.info (ارتباط با مشتری) فرم نظرسنجی را برای کتابی که در دست دارید تکمیل و ارسال نموده، انتشارات دیباگران تهران را که جلب رضایت و وفاداری مشتریان را هدف خود می‌داند، یاری فرمایید.

امیدواریم همواره بهتر از گذشته خدمات و محصولات خود را تقدیم حضورتان نماییم.

مدیر انتشارات

مؤسسه فرهنگی هنری دیباگران تهران
bookmarket@mft.info

تقدیم به :

آنان که دوستشان دارم
پدر و مادرم عزیزم، همسر فداکار و مهربانم و فرزندانم نیلوفر و زهرا

دکتر امین نظارات

مقدمه مولف:

سالهاست که از نفت به عنوان ارزشمندترین ماده و ثروت بشر یاد می شود. کشورهای دارنده مخازن عظیم نفتی که در منطقه غرب آسیا قرار گرفته اند از جمله ثروتمندترین کشورهای دنیا هستند. اما آیا به راستی این تعبیر کماکان صحیح است؟ بر اساس تحلیلهای موسسات بزرگ اقتصادی دنیای همچون گارتنر، امروزه "داده از نفت ارزشمندتر است." در سال ۲۰۱۸ که به نام سال حکمرانی داده نام گذاری شده بود، کشورهایی ثروتمندتر بودند که داده بیشتری را در اختیار داشته و از پردازش این حجم عظیم از داده ها تولید ارزش افزوده نموده اند. صاحبان داده صاحبان قدرتند. علم داده که مجموعه ای از علوم آمار، ریاضی، کامپیوتر می باشد از جمله پر طرفدارترین علوم شده است و متخصصین این حوزه بیشترین درآمد سالانه را دارند. در این کتاب سعی کرده ام از تعاریف و مقدمات رایج در کتابهای کلان داده عبور کرده و به مفاهیم عملیاتی تر در پردازش، استخراج و ذخیره آنها بپردازم. شما در این کتاب به صورت عملی با روشها و ابزارهای پردازش داده های بزرگ آشنا خواهید شد. هر چند تمام تلاش نویسندگان در ارائه اطلاعات و مباحث به روز و به بلوغ رسیده بوده است اما به دلیل رشد بسیار سریع مباحث و ابزارهای این حوزه ممکن است زمانی برسد که تکنولوژی های مطرح شده در این کتاب قدیمی شده باشند. در هر حال هر گونه نظر شما می تواند باعث بالندگی بیشتر و ارائه کتابها و مطالب مفیدتر در آینده گردد.

راههای ارتباطی با بنده:

aminnezarat@gmail.com و وب سایت www.hpclab.ir می باشد.