

دادہ ہا می بزرگ

طہندی آن ہا

مؤلفان:

نجمہ ملایی

سعید طاہری

داده‌های بزرگ و طبقه‌بندی آن‌ها

مؤلفان: نجمه ملایی، سعید طاهری
ویراستار علمی: دکتر رامین مولاناپور
طراحی جلد، حروفچینی و صفحه‌آایی: همتا بیداریان
ناشر: انتشارات آتی‌نگر/ ناشر همکار: انتشارات وینا / پ اول، ۱۳۹۶ / شمارگان: ۵۰۰ نسخه
قیمت: ۶۰,۰۰۰ ریال
شابک: ۹۷۸-۶۰۰-۷۶۳۱-۵۳-۹

ISBN: 978-600-7631-53-9

نشانی دفتر فروش: خیابان جمالزاده جنوبی، روبه‌روی کوچه مستحق، ۱۴۴، واحد ۱ / تلفن: ۸-۶۶۵۶۵۳۳۶ / نمابر: ۶۶۵۶۵۳۳۷

www.ati-negar.com * info@ati-negar.com



موضوع: داده‌های کلان (Big Data) - پایگاه‌های اطلاعاتی، مدیریت (Database Management) - داده‌کاوی (Data Mining)

شناسه افزوده: طاهری، سعید، ۱۳۵۷ - مؤلف

رده‌بندی کنگره

رده‌بندی دیویی

شماره کتابشناسی ملی

QA ۱۳۹۶ ۵۳/۹/۵۱۲

۰۵۱۲

۷۱۸۳

فهرست

۵	مقدمه
۷	داده
۷	الزامات نهفتگی
۱۳	ساختار
۱۸	زیرساخت رایانش
۲۳	Hadoop1.0
۲۷	Hadoop2.0
۲۹	Berkeley Spark
۳۲	زیرساخت ذخیره‌گاه
۳۸	ارزیابی عملکرد و بهینه‌سازی‌های منتخب

۴۱	رویکردهای تحلیل
۴۱	الگوریتم‌های یادگیری ماشین
۴۱	انواع الگوریتم‌های یادگیری ماشین
۴۵	داده‌کاوی از طریق تنوع
۴۸	تکنیک‌های آماری
۵۱	معیارهای اندازه‌گیری برای ارزیابی دسته‌بندی کننده‌ها بیش‌بینی‌کننده‌ها
۵۴	مصورسازی
۵۶	امنیت و حریم خصوصی
۶۳	نتیجه‌گیری
۶۴	منابع

مقدمه

اصطلاح داده‌های بزرگ اشاره به مقدار عظیم اطلاعات دیجیتالی شرکت‌ها و دولت‌ها دارد که در مورد ما و پیرامون ما انباشته شده است. این داده‌ها تنها از طریق تبادل اطلاعات سنتی و با استفاده از نرم‌افزارهای کامپیوتری، تلفن‌های همراه و غیره تولید نمی‌شوند، بلکه از تعداد زیادی از حسگرهای متفاوت تعبیه‌شده در محیط‌های مختلف و در خیابان‌های شهر (دوربین‌ها و میکروفون‌ها) یا موتورهای جت (حسگرهای دما) و اینترنت اشیا به سرعت در حال افزایش؛ جایی که تقریباً هر وسیله برقی به صورت مجازی به اینترنت وصل شده و داده‌هایی تولید می‌کند، به وجود می‌آیند.

هر روزه ۲/۵ کوئینتیلین^۱ (۱۰^{۲۶}) بایت از داده‌ها تولید می‌شود. ما تخمین بیش از ۹۰٪ داده‌ها در جهان تنها در دو سال اخیر تولید شده‌اند (از سال ۲۰۱۱ تا ۲۰۱۸). مباحث ذخیره‌سازی، رایانش، امنیت و حریم خصوصی و دیگردهای تحلیلی همگی از طریق نرخ، حجم و تنوع داده‌های بزرگ معنادار می‌شوند، مانند زیرساخت‌های ابری در مقیاس وسیع، تنوع داده‌ها و سرعت داده‌ها، ماهیت جریانی اکتساب داده‌ها و انتقال میان ابری با حجم بالا.

یک دسته‌بندی شش بعدی در شکل ۱ نشان داده شده است. این شش بعد از شش جنبه به وجود می‌آیند که برای ایجاد زیرساخت‌های داده‌های بزرگ مورد نیازند. هر یک از ابعاد، در ادامه متن توضیح داده شده است.

1- Quintillion