

کلان داده؛ آموزش عملی، چالش‌ها و چشم‌اندازها

تألیف

عباس میرزایی

(ارشدیار دانشگاه آزاد اسلامی واحد اردبیل)

فرامرز نیکجو



انتشارات جهاد دانشگاهی استان اردبیل

۱۳۹۸

سرشناسه	: میرزایی، عباس، ۱۳۶۱ -
عنوان و نام پدیدآور	: کلان‌داده: آموزش عملی، چالش‌ها و چشماندازها/ تألیف عباس میرزایی، فرامرز نیکجو.
مشخصات نشر	: اردبیل: جهاد دانشگاهی، واحد اردبیل، سازمان انتشارات، ۱۳۹۸.
مشخصات ظاهری	: ۱۷۲ ص.
شابک	: ۱ - ۴۱ - ۶۲۳۱ - ۶۲۲ - ۹۷۸ - قیمت: ۳۹۰۰۰۰ ریال
وضعیت فهرست نویسی	: فیا
موضوع	: داده‌های کلان
موضوع	: Big data
شناسه افزوده	: نیکجو، فرامرز، ۱۳۵۵ -
شناسه افزوده	: جهاد دانشگاهی. واحد استان اردبیل. سازمان انتشارات
شناسه افزوده	: <i>Jahad-e Daneshgahi. Vahed-e Ostan Ardabil. Press Organization</i>
رده‌بندی کنگره	: ۵۷۶/۹
رده‌بندی دیویی	: ۰۵/۷۴
شماره کتابشناسی ملی	: ۶۱۱۴۸۰۳

کلان داده؛ آموزش عملی، چالش‌ها و چشم‌اندازها

تألیف: عباس میرزایی و فرامرز نیکجو

صفحه آرایی و طرح روی جلد: محمود خروشی

ناشر: انتشارات جهاد دانشگاهی استان اردبیل

چاپ: اول - ۱۳۹۸

شمارگان: ۱۰۰۰ نسخه

قیمت: ۳۹۰۰۰ تومان

سازش: ۹۷۸-۶۲۲-۶۲۳۱-۴۱-۱

این اثر، مشمول قانون حمایت مؤلفان و مصنفان و هنرمندان مصوب ۱۳۴۸ است، هر کسی تمام یا قسمتی از این اثر را بدون اجازه مؤلف (ناشر) نشر یا پخش یا عرضه نماید مورد پیگرد قانونی قرار خواهد گرفت.



واحد استان اردبیل

نشانی انتشارات: اردبیل - شهرک کارشناسان - میدان شفا، تلفن: ۰۳۰۹۵۰۶ - وبسایت: ard.chap@chmail.ir
نشانی سازمان انتشارات: تهران - خیابان انقلاب اسلامی - خیابان فخررازی - خیابان شهدای اسلامبول - پلاک ۷۲، تلفن: ۶۶۹۵۲۹۴۸

مقدمه.....	۱۱
فصل ۱: مفاهیم اساسی داده های عظیم	۱۳
داده های عظیم چیست.....	۱۵
داده های عظیم شبیه چیست.....	۱۶
تعریف و ویژگی های داده های عظیم.....	۲۲
ارزش داده های عظیم.....	۲۵
توسعه داده های عظیم.....	۲۶
چالش های داده های عظیم.....	۲۸
شناسایی محرک های تغییر.....	۳۲
کاربردهای اساسی داده های عظیم.....	۳۸
سیر تکامل کاربرد.....	۳۸
حوزه های تحلیل داده های عظیم.....	۴۰
تحلیل داده ساخت یافته.....	۴۰
تحلیل داده متنی.....	۴۰
تحلیل داده وب.....	۴۲
تحلیل داده چند رسانه ای.....	۴۴
تحلیل داده شبکه.....	۴۶
تحلیل ترافیک سیار.....	۴۹
کاربردهای اصلی.....	۵۱
کاربرد داده های عظیم در شرکت ها.....	۵۱
کاربرد داده های عظیم مبتنی بر اینترنت اشیا.....	۵۲
کاربرد داده های عظیم مبتنی بر شبکه اجتماعی آنلاین.....	۵۴
کاربرد داده های عظیم در مراقبت های بهداشتی و پزشکی.....	۵۷
هوش جمعی.....	۵۸
شبکه هوشمند.....	۵۹

فصل دوم: فناوری ها و زیر ساخت های مورد استفاده در داده های عظیم

۶۱	فناوری های وابسته
۶۳	رایانش ابری
۶۳	اینترنت اشیا
۶۶	رابطه میان اینترنت اشیا و داده های عظیم
۶۷	مرکز داده ها
۶۸	هدوپ
۶۹	تولید و اکتساب داده های عظیم
۷۳	تولید داده های عظیم
۷۳	داده های اینترنتی
۷۶	داده های بیست پزشکی
۷۶	تولید داده ها از سایر زمینه ها
۷۸	اکتساب داده های عظیم
۷۸	گردآوری داده ها
۷۹	انتقال داده ها
۸۲	پیش پردازش داده ها
۸۴	مجتمع سازی
۸۴	پاکسازی
۸۵	حذف افزونگی
۸۵	ذخیره سازی داده های عظیم
۸۷	سیستم ذخیره سازی برای داده های انبوه
۸۷	سیستم ذخیره سازی توزیع شده
۸۹	مکانیزم های ذخیره سازی برای داده های عظیم
۹۳	فناوری پایگاه داده
۹۳	پایگاه داده های کلید - مقدار
۹۴	پایگاه داده های ستونی
۹۶	پایگاه داده های مبتنی بر سند
۹۹	بلت فرم برای ذخیره سازی جدول جهانی آنی
۱۰۲	فاکتورهای طراحی
۱۰۲	مدل برنامه نویسی پایگاه داده
۱۰۴	MapReduce
۱۰۴	Dryad
۱۰۵	All-Pairs
۱۰۶	

۱۰۷	Pregel
۱۰۸	تحلیل داده‌های عظیم
۱۰۸	تحلیل داده‌های سنتی
۱۱۱	روش‌های تحلیلی داده‌های عظیم
۱۱۴	معماری برای تحلیل داده‌های عظیم
۱۱۴	تحلیل بلادرنگ در برابر آفلاین
۱۱۴	تحلیل در سطوح مختلف
۱۱۵	تحلیل با پیچیدگی‌های مختلف
۱۱۵	ابزار کاوش و تحلیل داده‌های عظیم
۱۱۷	معماری HDFS
۱۱۹	نگاشت کاهش (MapReduce) و YARN
۱۲۱	Zoo keeper
۱۲۲	HBASE
۱۲۳	HIVE
۱۲۴	PIG
۱۲۶	ماهوت
۱۲۷	توسعه‌ی برنامه‌های کاربردی داده‌های عظیم
۱۲۷	موازی سازی
۱۲۸	اسطوره‌ی مقیاس‌پذیری ساده
۱۲۹	چارچوب توسعه‌ی برنامه‌ی کاربردی
۱۳۰	مدل برنامه‌نویسی نگاشت کاهش
۱۳۴	سایر چارچوب‌های توسعه‌ی داده‌های عظیم
۱۳۴	مدل اجرا
۱۳۶	مدیریت داده‌ی NoSQL برای داده‌های عظیم
۱۳۶	NoSQL چیست
۱۳۸	انبارهای کلید مقدار
۱۴۰	انبارهای اسناد
۱۴۱	انبارهای جدولی
۱۴۲	انبارهای شیء داده
۱۴۲	پایگاه داده‌های نموداری
۱۴۲	برداشت‌ها
۱۴۳	تمرینات فکری

فصل سوم: آموزش عملی BIGDATA

۱۴۵	
۱۴۷	مقدمه IBM InfoSphere Biginsights
۱۴۸	ترکیب بندی نرم افزار در دستگاه مجازی
۱۴۸	کاربرد تک گره ای
۱۴۸	شروع استفاده از IBM InfoSphere Biginsights
۱۵۱	تنظیمات IBM InfoSphere Biginsights
۱۵۲	امور مربوطه
۱۵۳	بروز رسانی ترکیب بندی
۱۵۳	بروز رسانی پنجره BIOS
۱۵۴	ترکیب بندی گره های اضافی
۱۵۶	تنظیم ساختار هداوت
۱۵۸	ورود به VM
۱۵۹	وب کنسول Amb ri
۱۶۰	تعاملات اساسی UDI S استفاده از خط دستور
۱۶۵	خلاصه
۱۶۵	Loop Administration
۱۶۶	آماده سازی محیط کاری
۱۶۷	Child image
۱۶۸	Master image
۱۶۹	مراجع

۱. مقدمه

در ۲۰ سال گذشته داده‌ها در مقیاس بزرگ و حوزه‌های مختلف افزایش یافته‌اند. طبق گزارش شرکت سهامی داده‌های بین‌الملل در سال ۲۰۱۱ حجم کلی داده‌های تولید و کپی شده در جهان ۱,۸ ZB (۱۰^{۲۱}) بوده، که این مقدار طی ۵ سال حدوداً نه برابر شده است. در آینده‌ی نزدیک این رقم حداقل هر دو سال یک بار دو برابر می‌شود [۱].

اصطلاح داده‌های عظیم در پی افزایش زیاد داده‌های جهانی پدید آمده و عمدتاً از آن برای توصیف مجموعه داده‌های بزرگ استفاده می‌شود. در مقایسه با مجموعه داده‌های سنتی، داده‌های عظیم عموماً شامل از ره‌ی از داده‌های غیر ساخت یافته می‌شوند که مستلزم به تجزیه و تحلیل بلادرنگ هستند. علاوه بر آن، داده‌های عظیم فرصت کشف مقادیر جدید را فراهم کرده و به ما کمک می‌کنند تا درکی عمیق از مقدار و پهنای پتانسیل داشته باشیم و همچنین چالش‌های جدیدی مثلاً در زمینه‌ی نحوه‌ی سازماندهی و مدیریت مؤثر این داده‌ها ایجاد می‌کنند. در حال حاضر داده‌های عظیم توجه صنایع، نهادهای علمی و دولتی را به خود جلب کرده است. برای نمونه، معمولاً مسائل مربوط به داده‌های عظیم تحت پوشش خبری رسانه‌های جمعی نیویورک تایمز [۴]، اکنومیست [۲ و ۳] و رادیو ملی [۵ و ۶] قرار می‌گیرند. همچنین، دو مجله‌ی علمی برتر Science و Nature ستون‌های ویژه‌ای را به بررسی اهمیت و چالش‌های داده‌های عظیم اختصاص داده‌اند [۷ و ۸]. بیشتر نهادهای دولتی اعلام کرده‌اند که برنامه‌های مهمی برای سرعت بخشیدن به پژوهش در زمینه‌ی داده‌های عظیم و کاربرهای آن دارند [۹] و اخیراً صنعت نیز به پتانسیل بالای داده‌های عظیم علاقمند شده است [۱۲].

بدون شک عصر داده‌های عظیم در حال ظهور است. در حال حاضر رشد سریع داده‌های عظیم از زندگی روزانه‌ی مردم، بویژه بخش‌های مرتبط با شرکت‌های خدمات اینترنتی ناشی می‌شود. مثلاً گوگل صدها پتابایت داده را پردازش نموده و فیس‌بوک بیش از ۱۰۰ پتابایت داده در ماه تولید می‌کند؛ یک شرکت چینی به نام بایدها پتابایت داده و تائوپائو به عنوان یک شرکت فرعی علی‌بابا از طریق معاملات آنلاین روزانه ده‌ها ترابایت داده پردازش می‌کنند. با رشد جدی مقدار مجموعه داده‌های بزرگ مسائل چالشی بسیاری نیز حادث شده‌اند که باید هر چه سریع‌تر راه‌حلی برای آنها پیدا کرد. اولاً، پیشرفت‌های جدید فناوری اطلاعات تولید داده را بسیار آسان نموده است. مثلاً، بطور میانگین در هر دقیقه ۷۲ ساعت ویدئو در یوتیوب بارگذاری می‌شود [۱۳]. بنابراین چالش اصلی جمع‌آوری و مجتمع‌سازی داده‌های انبوه از منابع داده‌هایی است که بطور گسترده‌ای توزیع شده‌اند. ثانیاً، داده‌های جمع‌آوری شده رشد زیادی دارند که با توجه به آن مسئله‌ی نحوه‌ی ذخیره و

مدیریت چنین مجموعه داده‌های عظیم و ناهمگون با نیازمندی‌های خاص در زیرساخت نرم‌افزاری و سخت‌افزاری مطرح می‌شود. ثالثاً، نظر به ناهمگونی، مقیاس پذیری، بلادرنگ بودن، پیچیدگی و محرمانه بودن داده‌های عظیم باید مجموعه داده‌ها را در سطوح مختلف با تحلیل، مدل‌سازی، مصورسازی و تکنیک‌های بهینه‌سازی بطور مؤثری «کاوش» کنیم تا خاصیت ذاتی داده‌های عظیم را نشان داده و قدرت تصمیم‌گیری را افزایش دهیم.

رشد سریع رایانش ابری و اینترنت اشیاء (IoT) رشد داده‌ها را افزایش داده است. رایانش ابری، تأمین امنیت، دسترسی به سایت‌ها و کانال‌ها را برای داده‌های پر ارزش فراهم می‌کند. در الگوی IoT حسگرها در سراسر جهان در حال جمع‌آوری و انتقال داده‌هایی هستند که در یک ابر ذخیره و پردازش خواهد شد. این داده‌ها از نظر کمی و روابط متقابل بسیار فراتر از ظرفیت‌های معماری IT و زیرساخت‌های سازمان‌های موجود خواهند رفت و ملزومات بلادرنگ آن بر ظرفیت محاسبه‌ی موجود تأکید بیشتری خواهد داشت. شکل ۱،۱ نشان دهنده‌ی توسعه‌ی عظیم حجم داده‌های جهانی است.