

بسم الله الرحمن الرحيم

از کلان داده‌ها

تا

انتخاب اطلاعات

بر بندگان:

مهندس منساب یاقفی زاده

دکتر سینا دامی

عضو هیأت علمی دانشگاه آزاد اسلامی - واحد تهران غرب

انتشارات شهر آشوب

۱۳۹۸

سرشناسه	یا حقی زاده، مهتاب، ۱۳۶۷ - .
عنوان و نام پدیدآور	: از کلان داده‌ها تا انقلاب اطلاعات/ نویسندگان مهتاب یا حقی زاده، سینا دامی.
مشخصات نشر	: تهران: شهر آشوب، ۱۳۹۸.
مشخصات ظاهری	: ۲۲۸ ص.
شابک	: 978-964-7513-94-4
وضعیت فهرست‌نویسی: فیا	
موضوع	: داده‌های کلان
موضوع	: Big data
موضوع	: داده‌کاوی
موضوع	: Data mining
شناسه اثر ده:	: دامی - سینا، ۱۳۶۵ -
رده‌بندی کنگره	: ۰۵۷۶۱۰
رده‌بندی دیویی	: ۰۰۷
شماره کتابشناسی	: ۰۰۱۸۳۲



انتشارات شهر آشوب

تهران، خیابان انقلاب، خیابان فخر رازی، کوچه انرژی، پلاک ۸، طبقه سوم

تلفن: ۰۱۱۲۱۸۳۱۶۴۴-۶۶۴۶۴۸۰۷

عنوان	: از کلان داده‌ها تا انقلاب اطلاعات
ناشر	: انتشارات شهر آشوب
نویسندگان	: مهندس مهتاب یا حقی زاده، دکتر سینا دامی
نوبت چاپ	: اول، ۱۳۹۸
تیراژ	: ۵۰۰ نسخه
قیمت	: ۳۴۰۰۰ تومان
شابک	: ۹۷۸-۹۶۴-۷۵۱۳-۹۴-۴

حق چاپ محفوظ و هر گونه کپی برداری به صورت مکتوب یا الکترونیکی بدون مجوز از ناشر و مؤلف پیگرد قانونی دارد

مقدمه

در فوری‌ی سال ۲۰۱۱، بیش از ۱۳۰۰ نفر جهت شرکت در یک کنفرانس علمی در سانتا کلارای کالیفرنیا گرد هم آمدند.

اگرچه، حوزه‌های سخت‌افزاری، از بیمه گرفته تا رسانه و فناوری‌های پیشرفته در حوزه‌ی مراقبت‌های بهداشتی ارائه شدند اما شرکت کنندگان در کنفرانس با یک نهاد عمومی نوظهور با عنوان دانشمندان داده غافلگیر شدند. کارآفرینی زیرکانه‌ای که مهارت‌های برنامه نویسی را با ریاضی ترکیب می‌کند و دانشمندان داده به عنوان حرفه و سمت جدیدی که به سمت کسب و کارهای مبتنی بر داده سوق داده شده‌اند، نام می‌کنند. این حرفه‌ی جدید بر موج کلان داده سوار است. به لطف شبکه‌های اجتماعی و تلفن، با هوشمند کسب و کار ما در حال ایجاد داده‌های بیشتر است و ما به عنوان مصرف کنندگی منبع جریان‌های عظیم اطلاعات هستیم. در این جنس خام، ارزش‌های بسیاری وجود دارد و بینش در کسب و کار و بازار و همچنین چشم اندازی برای انواع جدید محصولات و خدمات فوق العاده شخصی ایجاد می‌کند. در چندین سال گذشته، تنها کسب و کارهای بزرگ می‌توانستند در کلان داده سود ببرند مانند: Walmart، گوگل و تاجران مالی متخصص.

امروزه به لطف یک پروژه‌ی منبع باز که هادوپ^۱ نامیده شده است، سخت‌افزاری لینوکس و رایانش ابری، قدرتی است که در دسترس همگانی قرار گرفته است. انقلاب داده‌ای رخ داده است که در حال فراگیری کسب و کار، دولت و علم همراه با عواقبی است که تاکنون و در مدت زمان طولانی توسط خودش ایجاد شده است. هر انقلابی باید از یک جایی شروع شود و برای بسیاری این سؤال مطرح می‌شود که "چگونه دانشمندان داده و کلان داده می‌توانند به

ساماندهی من کمک کنند؟" بعد از سال‌ها که ماهیت انتخاب‌های پردازش داده در مسیری درست روبه پیشروی است، در حال حاضر چشم انداز متنوعی برای مذاکره وجود دارد. علاوه بر این، برای سروکار داشتن با داده، شما باید با تغییراتی که فرهنگی و تکنولوژیکی هستند روبرو شوید.

هدف این کتاب، این است که آنچه را که کلان داده گفته می‌شود و نیز اینکه چرا کلان داده مهم است و نقطه‌ی شروع کجاست را کاملاً دریابید و از آن در دنیایی که در جریانی از داده‌ها شناور است استفاده کنید.

از آنجمله از موضوعات بر شرح و توصیف کلان داده به بسیاری از مسائل مرتبط با آن مانند تولید و اکتساب و ذخیره سایر کلان داده، شرح سخت افزار و سیستم‌های توزیع شده، فناوری‌های مرتبط با کلان داده، تجزیه و تحلیل و کاربردهای کلیدی کلان داده، بررسی بازار کلان داده، شرح و معرفی پایگاه داده‌های چالش‌های کلان داده و همچنین طرح بسیاری از مباحثی که باعث درک بیشتر از کلان داده خواهد شد پرداخته‌ایم.

اگر شما با کلان داده‌ها کار می‌کنید، با این کتاب به همکاران یا هیئت رئیسه‌ی خود کمک کنید و جهت ارائه‌ی خدمتی بهینه‌تر موضوعات و مطالب این کتاب را به خوبی مطالعه و درک کنید.

امید بر این است که در برداشتن گامی هرچند کوچک در جهت ارائه‌ی بینش و آگاهی در خصوص علم داده و کلان داده موثر بوده باشیم و تلاش می‌کنیم در جهت تدوین این کتاب برای دستیاران و علاقمندان این علم مسخر و مرقوع شده باشد.

بدیهی است که هراثری کامل و بدون نقص نیست از آنجمله از خوانندگان علاقمند خواهشمندیم که نظرات و بازخورد سازنده‌ی خود را با ما از طریق mahtab.yahaghizadeh@gmail.com به اشتراک بگذارند و در ایندکس را همراهی کنند.

فهرست مطالب

مقدمه.....	۳
فصل ۱: خورد اقتصادی.....	۱۷
۱-۱. فواید داده‌ها، دیجیتالی - سریع.....	۱۸
۲-۱. زنجیره‌ی مین با داده‌ها.....	۱۹
۱-۲-۱. جمع‌آوری داده.....	۱۹
۲-۲-۱. جذب و شفاف سازی.....	۲۰
۳-۲-۱. سخت افزار.....	۲۱
۴-۲-۱. پلت فرم‌ها.....	۲۱
۵-۲-۱. یادگیری ماشین.....	۲۲
۶-۲-۱. کشف انسانی.....	۲۳
۷-۲-۱. ذخیره سازی.....	۲۳
۸-۲-۱. سنجش و جمع‌آوری بازخورد.....	۲۵
۳-۱. جایگزینی همه چیز با داده.....	۲۵
۴-۱. یک بازخورد اقتصادی.....	۲۵
فصل ۲: سخت افزار.....	۲۷
۱-۲. ذخیره سازی (DISK).....	۲۷
۲-۲. واحد پردازش مرکزی.....	۳۰
۳-۲. حافظه.....	۳۲

۴-۲ شبکه ۳۴

فصل ۳: کلان داده چیست ؟ ۳۵

۱-۳ کلان داده ها چگونه به نظر میرسند ؟ ۳۶

۱-۱-۳ ظرفیت ۳۷

۲-۱-۳ سرعت ۳۸

۳-۳-۳ نوع ۴۰

۲-۳-۳ آیا در خانه ؟ ۴۱

۳-۳ کلان داده چیست ۴۲

۴-۳ کلان داده چیست ۴۲

۵-۳ فرهنگ ۴۳

۶-۳ می دانید کجایم خرید کنید ۴۴

فصل ۴: تولید و اکتساب کلان داده ۴۵

۱-۴ تولید داده ۴۵

۱-۱-۴ داده های شرکت ۴۶

۲-۱-۴ داده های اینترنت اشیا ۴۶

۳-۱-۴ داده بیوپزشکی ۴۸

۴-۱-۴ تولید داده از دیگر فیلدها ۴۹

۲-۴ اکتساب کلان داده ۵۰

۱-۲-۴ جمع آوری داده ۵۱

۲-۲-۴ انتقال داده ۵۴

۳-۲-۴ پیش پردازش داده ۵۷

فصل ۵: ذخیره سازی کلان داده ۶۱

۱-۵ سیستم ذخیره سازی برای داده های حجیم ۶۱

۶۳	۲-۵ سیستم ذخیره سازی توزیع شده
۶۵	۳-۵ مکانیزم ذخیره سازی کلان داده
۶۷	فصل ۶: سیستم های توزیع شده
۶۸	۱-۶ محاسبات پایگاه داده
۶۹	۲-۶ محاسبات سیستم فایل
۷۱	۳-۶ ملاحظات
۷۵	فصل ۷: فایزری های مرتبط
۷۵	۱-۷ فایزری های مرتبط با کلان داده
۷۵	۱-۱-۷ ارتباط بین ریانش ابری و کلان داده
۷۷	۲-۱-۷ ارتباط بین اینترنت اشیا و کلان داده
۷۸	۳-۱-۷ مرکز داده
۷۹	۴-۱-۷ آپاچی هادوپ
۸۰	۱-۴-۱-۷ هسته ی هادوپ: Map Reduce
۸۱	۲-۴-۱-۷ سطوح پایین تر هادوپ: HDFS و MapReduce
۸۱	۳-۴-۱-۷ بهبود قابلیت برنامه نویسی: Hive و Pig
۸۲	۴-۴-۱-۷ بهبود دسترسی به داده: Hbase، Zooop و Flume
۸۴	۵-۴-۱-۷ هماهنگی و گردش کار: Zookeeper و Curie
۸۴	۶-۴-۱-۷ مدیریت و استقرار: Ambari و Whirr
۸۵	۷-۴-۱-۷ یادگیری ماشینی: Mahout
۸۵	۸-۴-۱-۷ استفاده از هادوپ
۸۷	فصل ۸: تجزیه و تحلیل کلان داده
۸۷	۱-۸ تجزیه و تحلیل داده های سنتی
۸۹	۲-۸ روش های تحلیل کلان داده

- ۸-۳ معماری تجزیه و تحلیل کلان داده ۹۰
- ۸-۳-۱ تجزیه و تحلیل بلادرنگ در برابر تجزیه و تحلیل آنلاین ۹۱
- ۸-۳-۲ تجزیه و تحلیل در سطوح مختلف ۹۲
- ۸-۳-۳ تجزیه و تحلیل با پیچیدگی متفاوت ۹۳
- ۸-۴ ابزار کاوش و تحلیل کلان داده ۹۳
- فصل ۹: کاربردهای کلیدی کلان داده ۹۷
- ۹-۱ کاربردهای کلان داده ۹۷
- ۹-۱-۱ کاربرد کلان داده در شرکت ها ۹۷
- ۹-۱-۲ کاربرد کلان داده در صنعتی بر IoT ۹۸
- ۹-۱-۳ کاربرد کلان داده های شبکه محور اجتماعی آنلاین ۹۹
- ۹-۱-۴ کاربردهای کلان داده های سلامت بهداشتی و پزشکی ۱۰۲
- ۹-۱-۵ هوش جمعی ۱۰۳
- ۹-۱-۶ شبکه های هوشمند ۱۰۴
- فصل ۱۰: بررسی بازار کلان داده ۱۰۷
- ۱۰-۱ فقط هادوپ ؟ ۱۰۸
- ۱۰-۲ سیستم های یکپارچه ی هادوپ ۱۰۸
- ۱۰-۲-۱ EMC Greenplum ۱۰۹
- ۱۰-۲-۲ IBM ۱۱۰
- ۱۰-۲-۳ مایکروسافت ۱۱۱
- ۱۰-۲-۴ اوراکل ۱۱۲
- ۱۰-۳ دسترسی پذیری ۱۱۳
- ۱۰-۴ پایگاه داده های تحلیلی با اتصال به هادوپ ۱۱۳
- ۱۰-۴-۱ نگاهی اجمالی ۱۱۴
- ۱۰-۵ شرکت های متمرکز بر هادوپ ۱۱۵

۱۱۵	 Cloudera ۱-۵-۱۰
۱۱۵	 Hortonworks ۲-۵-۱۰
۱۱۶	 ۶-۱۰ مروری بر توزیع های هادوپ
۱۱۸	 ۷-۱۰ یادداشت ها
۱۱۹	 فصل ۱۱: برنامه ی میکروسافت برای کلان داده
۱۱۹	 ۱-۱۱ توزیع هادوپ میکروسافت
۱۲۰	 ۲-۱۱ توسعه دهندگان، توسعه دهندگان، توسعه دهندگان
۱۲۱	 ۳-۱۱ داده های جریانی و NoSQL
۱۲۲	 ۴-۱۱ به سمت یک محیط یکپارچه
۱۲۲	 ۵-۱۱ بازار داده
۱۲۳	 ۶-۱۱ خلاصه
۱۲۵	 فصل ۱۲: کلان داده در ابر
۱۲۵	 ۱-۱۲ ۱۸۸۵ و ابرهای خصوصی
۱۲۶	 ۲-۱۲ راه حل های پلت فرم
۱۲۷	 ۳-۱۲ خدمات وب آمازون
۱۲۸	 Elastic Map Reduce ۱-۳-۱۲
۱۲۸	 DynamoDB ۲-۳-۱۲
۱۲۸	 ۴-۱۲ گوگل
۱۲۹	 BigQuery ۱-۴-۱۲
۱۳۰	 Prediction API ۲-۴-۱۲
۱۳۰	 ۵-۱۲ میکروسافت
۱۳۱	 ۶-۱۲ پلت فرم های کلان داده ی مقایسه شده
۱۳۳	 ۷-۱۲ نتیجه گیری و نکات / یادداشت ها
۱۳۳	 نکات / یادداشت ها

۱۳۵.....	فصل ۱۳: مخاطرات رایانش ابری
۱۳۵.....	۱-۱۳ اصول سه گانه‌ی CIA
۱۳۵.....	۱-۱-۱۳ محرمانگی
۱۳۶.....	۲-۱-۱۳ یکپارچگی
۱۳۶.....	۳-۱-۱۳ دسترسی پذیری
۱۳۷.....	۴-۱۳ سایر مفاهیم مهم
۱۳۸.....	۱-۱۳ حریم خصوصی و خطرات پذیرش
۱۳۹.....	۳-۱۳ امنیت داده‌ها در صنعت کارت‌های پرداخت
۱۴۱.....	۴-۱۳ استاندارد (PC, SS)
۱۴۲.....	۵-۱۳ حریم خصوصی، اطلاعات و قوانین حریم خصوصی
۱۴۴.....	۱-۵-۱۳ قوانین حریم خصوصی
۱۴۵.....	۲-۵-۱۳ قانون حساسیتی و حمل و نقل بیمه سلامت (HIPPA)
۱۴۶.....	۳-۵-۱۳ قانون HITECH
۱۴۷.....	۴-۵-۱۳ قانونی برای اولویت‌های حذف حریم خصوصی (P3P)
۱۴۸.....	۵-۵-۱۳ قانون PATRIOT
۱۴۹.....	۶-۵-۱۳ قانون مدیریت امنیت اطلاعات فدرال (FISMA)
۱۵۲.....	۷-۵-۱۳ سایر قوانین مرتبط با حریم خصوصی
۱۵۵.....	۶-۱۳ تهدیدهای مربوط به زیرساخت‌ها، داده‌ها و کنترل دسترسی
۱۵۵.....	۷-۱۳ پنهان سازی مقیاس
۱۵۶.....	۸-۱۳ تهدیدها و آسیب پذیری‌های رایج
۱۵۷.....	۹-۱۳ ابرهای خصوصی در مقابل ابرهای عمومی
۱۵۷.....	۱۰-۱۳ انواع حملات و ورود غیر مجاز به سیستم
۱۵۷.....	۱-۱۰-۱۳ سوء استفاده از ورود به سیستم
۱۵۸.....	۲-۱۰-۱۳ استفاده از سیستم نامناسب
۱۵۸.....	۳-۱۰-۱۳ استراق سمع

۱۵۹.....	۴-۱۰-۱۳ نفوذ به شبکه
۱۵۹.....	۵-۱۰-۱۳ حملات محرومیت از سرویس (DoS)
۱۶۰.....	۶-۱۰-۱۳ حملات ربودن حساب کاربری
۱۶۰.....	۷-۱۰-۱۳ حملات متلاشی کردن
۱۶۱.....	۸-۱۰-۱۳ حملات شماره گیری
۱۶۱.....	۱۱-۱۳ مسائل مربوط به کنترل دسترسی ابری
۱۶۲.....	۱۲-۱۳ مسائل مربوط به یکپارچگی پایگاه داده
۱۶۳.....	۱۳-۱۳ طرهای تأمین کنندگان سرویس های ابری
۱۶۴.....	۱-۳-۱۳ درب پشتی
۱۶۴.....	۲-۳-۱۳ حملات
۱۶۴.....	۳-۳-۱۳ فرستادن رمز
۱۶۵.....	۴-۳-۱۳ پاسخ
۱۶۵.....	۵-۳-۱۳ ربودن TCP
۱۶۵.....	۶-۳-۱۳ مهندسی اجتماعی
۱۶۶.....	۷-۳-۱۳ غوطه وری در زباله ها
۱۶۷.....	۸-۳-۱۳ حدس رمز عبور
۱۶۷.....	۹-۳-۱۳ اسب های تروجان و بدافزار
۱۶۸.....	۱۴-۱۳ خلاصه
۱۶۹.....	فصل ۱۴: برنامه ی مایکروسافت برای رایانش ابری
۱۶۹.....	۱-۱۴ درک مبحث ویندوز آزور
۱۶۹.....	۱-۱-۱۴ ویندوز آزور چیست؟
۱۶۹.....	۲-۱-۱۴ چرا از ابر استفاده می کنیم؟
۱۷۰.....	۱-۲-۱-۱۴ سرعت
۱۷۰.....	۲-۲-۱-۱۴ مقیاس
۱۷۱.....	۳-۲-۱-۱۴ اقتصاد

- ۱۷۲..... ۱۴-۱-۳ ابرمقاوم
- ۱۷۲..... ۱۴-۱-۴ آیا شما برای استفاده از ابرآماده هستید؟
- ۱۷۴..... ۱۴-۱-۵ شبکه‌ی تحویل محتوای ویندوزآزور
- ۱۷۴..... ۱۴-۱-۶ سرویس‌های ویندوزآزور
- ۱۷۵..... ۱۴-۱-۶-۱ سرویس‌های محاسباتی
- ۱۷۵..... ۱۴-۱-۶-۲ سرویس‌های شبکه
- ۱۷۶..... ۴-۱-۶-۳ سرویس‌های داده
- ۱۷۷..... ۴-۱-۶-۴ سرویس‌های برنامه
- ۱۷۸..... ۱۴-۱-۷ مجموعه‌های ویندوزآزور
- ۱۷۹..... ۱۴-۱-۸ راه‌های ویندوزآزور
- ۱۸۰..... ۱۴-۱-۸-۱ زیرساخت
- ۱۸۰..... ۱۴-۱-۸-۲ تلفن همراه
- ۱۸۱..... ۱۴-۱-۸-۳ وب
- ۱۸۱..... ۱۴-۱-۸-۴ رسانه
- ۱۸۱..... ۱۴-۱-۸-۵ یکپارچه‌سازی
- ۱۸۲..... ۱۴-۱-۸-۶ شناسایی و مدیریت دسترسی
- ۱۸۲..... ۱۴-۱-۸-۷ کلان داده
- ۱۸۳..... ۱۴-۱-۸-۸ توسعه آزمایش
- ۱۸۳..... ۱۴-۱-۸-۹ ذخیره‌سازی، پشتیبان‌گیری و بازیابی
- ۱۸۳..... ۱۴-۱-۸-۱۰ مدیریت داده
- ۱۸۴..... ۱۴-۱-۹ ویندوزآزور در عمل
- ۱۸۴..... ۱۴-۱-۹-۱ Cooltrax
- ۱۸۵..... ۱۴-۱-۹-۲ شبکه‌ی دانش‌باز (OKN)
- ۱۸۵..... ۱۴-۱-۱۰ ویندوزآزور و سیستم عامل ابر

فصل ۱۵: بازارهای داده	۱۸۹
۱-۱۵ بازارهای داده چه کاری انجام می‌دهند؟	۱۸۹
Infochimps ۱-۱-۱۵	۱۹۰
Factual ۲-۱-۱۵	۱۹۱
۳-۱-۱۵ بازار داده‌ی ویندوز آזור	۱۹۲
DataMarket ۴-۱-۱۵	۱۹۳
۵-۱-۱۵ مقایسه‌ی بازارهای داده	۱۹۴
۲-۱۵ زیرتأمین کنندگان داده	۱۹۵
فصل ۱۶: جنبش NoSQL	۱۹۷
۱-۱۶ اندازه، پاسخ، استی	۱۹۹
۲-۱۶ تغییر داده و صرفاً جوی در جنبه	۲۰۳
۳-۱۶ گاوهای مقدس	۲۰۸
۴-۱۶ سایر ویژگی‌ها	۲۱۱
۵-۱۶ ملاحظات	۲۱۳
فصل ۱۷: بصری سازی داده	۲۱۵
۱-۱۷ یک عکس گویاتر از ۱۰۰۰ کلمه است	۲۱۵
۲-۱۷ انواع بصری سازی	۲۱۶
۳-۱۷ توضیح و کاوش	۲۱۶
۴-۱۷ مشتریان شما، بسیار تصمیم گیرنده هستند	۲۱۷
۵-۱۷ در حق خودتان لطف کرده و یک طراح استخدام کنید	۲۱۸
فصل ۱۸: چالش‌های کلان داده	۲۱۹
۱-۱۸ چالش‌های کلان داده	۲۱۹
۱-۱-۱۸ نمایش داده	۲۱۹

۲۲۰.....	۱۸-۱-۲ کاهش افزونگی و فشرده سازی داده
۲۲۰.....	۱۸-۱-۳ مدیریت چرخه‌ی عمر داده
۲۲۰.....	۱۸-۱-۴ مکانیزم تحلیلی
۲۲۱.....	۱۸-۱-۵ محرمانگی داده
۲۲۱.....	۱۸-۱-۶ مدیریت انرژی
۲۲۲.....	۱۸-۱-۷ قابلیت توسعه و مقیاس پذیری
۲۲۲.....	۱۸-۱-۸ همکاری
۲۲۳.....	فصل ۱۹: آینده‌ی کلان داده
۲۲۳.....	۱۹-۱ ابزارهای استاندارد برای تحلیل
۲۲۴.....	۱۹-۲ پردازش داده‌ها در جریان
۲۲۵.....	۱۹-۳ پیشرفت بازارهای داده
۲۲۵.....	۱۹-۴ توسعه‌ی ابزارها و گردش کارهای داده‌ها
۲۲۶.....	۱۹-۵ افزایش درک و تقاضا برای بصری سازی
۲۲۷.....	مراجع