



۱۳۶۵

مؤسسه چاپ و نشر

دانشگاه افسری و تربیت پاسدار امام حسین

ذخیره سازی و پردازش داده‌های بزرگ

مؤلفان: بهرام دهقانی - آرش غفوری

www.ketab.ir

سرشناسه: دهقانی، بهراد، ۱۳۶۱
 عنوان و نام پدیدآور: ذخیره‌سازی و پردازش داده‌های بزرگ / مولفان بهراد دهقانی، آرش غفوری؛ ناظر علمی علیرضا حبیبی؛
 ویراستار علمی محمدعلی جوادزاده؛ ویراستار ادبی سعید گودرز؛
 با همکاری معاونت تربیت و آموزش دانشگاه افسری و تربیت پاسداری امام حسین علیه السلام.
 مشخصات نشر: تهران: دانشگاه افسری و تربیت پاسداری امام حسین علیه السلام، ۱۳۹۹
 مشخصات ظاهری: ۱۷۲ ص.
 شابک: ۷-۱۰-۸۹۶۵-۶۰۰-۹۷۸ قیمت: ۳۳۰۰۰۰ ریال
 وضعیت فهرست نویسی: فیا
 ۷۳- یادداشت: کتابنامه: ص. ۷۲
 موضوع: داده‌های کلان
 موضوع Big data
 موضوع: آپاچی هادوپ
 موضوع: Apache Hadoop
 شناسه افزوده: غفوری، آرش، ۱۳۶۹
 شناسه افزوده: جوادزاده، محمدعلی، ۱۳۴۶ - ویراستار
 شناسه افزوده: دانشگاه افسری و تربیت پاسداری امام حسین علیه السلام. معاونت تربیت و آموزش
 شناسه افزوده: دانشگاه افسری و تربیت پاسداری امام حسین علیه السلام
 شناسه افزوده: Guards & Imam Hossein University for Training Officers
 QA/۹ رده بندی کنگره: ۷۶
 ۷۴ / رده بندی دیویی: ۰۰۵
 شماره کتابشناسی ملی: ۶۲۴۰۸۳۱

ذخیره سازی و پردازش داده‌های بزرگ

تألیف: بهراد دهقانی - آرش غفوری

ناظر علمی: علیرضا حبیبی

نوبت چاپ: اول - بهار ۱۴۰۰

قیمت: ۴۲۰۰۰ تومان

شمارگان: ۵۰۰ جلد

شابک: ۷-۱۰-۸۹۶۵-۶۰۰-۹۷۸

نشانی:

بزرگراه شهید بابایی، بعد از پل لشکرک، دانشگاه افسری و تربیت پاسداری امام حسین علیه السلام

موسسه چاپ و انتشارات.

تلفن: ۷۷۷۱۰۴۸۳۴

دورنگار: ۷۷۱۰۴۸۳۶

کلیه حقوق اعم از چاپ و تکثیر و نسخه برداری برای مولف محفوظ است.

نقل و چاپ بخش‌هایی از متن با ذکر منبع آزاد است.

سخن ناشر

کتاب یکی از مهم‌ترین دست‌آوردهای بشری و بهترین ابزار ارتباطی برای انتقال معارف الهی و انسانی است. تألیف آثار علمی برجسته و معتبر ازجمله رهیافت‌های کارآمد و اثربخش برای بهره‌گیری از تجربه‌های ارزشمند و اندوخته‌های دانش صاحب‌نظران هر حوزه از علم است. دانشگاه افسری و تربیت پاسداری امام حسین علیه‌السلام با تدوین کتاب‌های علمی، فرهنگی، انقلابی و نظامی موردنیاز با بهره‌گیری از علوم جدید و بومی‌سازی آن با تجارب متراکم انقلاب اسلامی تلاش می‌نماید تا مرزهای دانش را گسترش دهد و علوم موردنیاز در عرصه پاسداری از انقلاب اسلامی تا ایجاد تمدن نوین اسلامی را عرضه نماید.

امید است این امر جوششی در حرکت و پویایی این دانشگاه در راستای گام دوم انقلاب اسلامی باشد.

مؤسسه چاپ و انتشارات

دانشگاه افسری و تربیت پاسداری امام حسین علیه‌السلام

پیش گفتار

«یرفع الله الذین امنوا منکم والذین اوتوا العلم درجات»

خداوند مقام اهل ایمان و دانشمندان عالم را (در دوجهان) رفیع می‌گرداند.

(سوره مبارکه مجادله. آیه ۱۱)

حقیقت جویی و اشتیاق برای شناخت واقعیات، از جمله مواردی است که در فطرت بشری به صورتی غریزی نهادینه شده است. چنان‌که در اکثریت قریب به اتفاق مکاتب آیینی از جمله ادیان الهی بر اهمیت علم و علم‌آموزی تأکید گردیده است.

لکن به گواه مکتوبات، دین مبین اسلام بیش از هر آیین و مسلک دیگری بر ضرورت و اهمیت دانشجویی و دانش‌آموزی اصرار ورزیده است؛ این تأکید و تکرار، با مطالعه معجزه‌گران سنگ اسلام، قرآن کریم، خود را بیش‌ازپیش می‌نمایاند. کما اینکه لفظ "علم" و "معرفت" به معنای شناخت - و مشتقات این دو مجموعاً ۸۱۱ بار در قرآن کریم، در آیات متعدد و به فراخور موقعیت‌های مختلف به کار رفته‌اند. این تعدد و تکرار (که با تعدد کاربرد لفظ "ایمان" و مشتقات آن برابر است)، نه تنها در لفظ، بلکه در معنای آن اهمیت دانش و هم‌ارزی آن با ایمان از منظر اسلام حکایت می‌کند؛ تا آنجا که کلام الله، ملاک برتری و تمایز افراد به لحاظ علم، و نه فقط مسلمین - را در علم‌اندوزی برمی‌شمارد: "آیا کسانی که می‌دانند و کسانی که نمی‌دانند یکسان‌اند؟" (سوره‌ی (مر: آیه ۹). بی‌سبب نیست که نخستین آیات وحی شده بر پیامبر اعظم ﷺ، او را به خواندن و آموختن امر می‌نمایند: کوی یاری تعالی خود را با "تعلیم و تعلم" به فرستاده‌ی خویش می‌شناساند: "بخوان! بخوان به نام پروردگارت که (جهان را) آفرید... هم او که به وسیله‌ی قلم تعلیم نمود... و به انسان آنچه را که نمی‌دانست آموخت!" (سوره علق: آیات ۱ تا ۵).

به تاسی از کلام خدا، پیامبر ﷺ نیز بر دانشجویی و علم‌آموزی تأکید فرموده‌اند: "اطلبوا العلم بالصین فان طلب العلم فریضة علی کل مسلم". از این حدیث می‌توان دریافت که فرستاده‌ی خدا هیچ قید مکانی و زمانی را برای علم‌آموزی لحاظ نفرموده و آن را به شکل یک فریضة برای پیروان خود مطرح نموده‌اند، که این‌ها همه اشاراتی بر اهمیت علم از منظر اسلام است. در سیره‌ی معصومین ﷺ و روایات نیز توجه ویژه‌ای به مقوله علم و علم‌آموزی گردیده است. به عنوان نمونه، وصی رسول اکرم و امام متقیان حضرت علی علیه السلام می‌فرمایند: ((أَلْعِلْمُ سُلْطَانٌ))؛ که در تفسیر کلام حضرت امیر، امام خامنه‌ای رحمه الله علم و دانش را ثروت آفرین و قدرت آفرین (در حوزه‌های مختلف، اعم از نظامی، اقتصادی، سیاسی و سایر موارد) می‌دانند و آن را به عنوان عامل تعیین‌کننده در تفوق یک ملت معرفی می‌نمایند:

"خود دانش مایه‌ی اقتدار است؛ ثروت آفرین است؛ قدرت نظامی آفرین است؛ قدرت سیاسی آفرین است. می‌فرماید: «أَلْعِلْمُ سُلْطَانٌ» - علم، قدرت است - «مَنْ وَجَدَهُ صَالًا بِهِ وَمَنْ لَمْ يَجِدْهُ صَبِيلَ غَلِيَّةٍ». یعنی مسئله، دو طرف دارد: اگر علم داشتید، می‌توانید سخن برتر را بگوئید، دست برتر را داشته باشید - «صال» یعنی این - اگر نداشتید، حالت میانه ندارد؛ «صیل علیه». پس کسی که علم دارد، او دست برتر را بر روی شما خواهد داشت."

دانایی و علم‌آموزی یکی از ویژگی‌های برجسته جوامع پویا و پیشرو است که همواره مؤثرترین نقش در رشد و تعالی مردمان آن جامعه داشته است. این خصوصیت یگانه که پایه و اساس شکل‌گیری و انسجام تمدن‌ها و فرهنگ‌ها است ارتباطی مستقیم با مقوله کتاب و کتاب‌خوانی دارد و چه نیکو گفته است پیامبر اعظم صلی الله علیه و آله که «در علم را با نوشتن زنجیر کنید». امروزه همان‌طور که قرن‌ها پیش سعدی شاعر بزرگ ایرانی گفته است «توانا بود هر که دانا بود» دانایی و دانش‌آموختگی برای بشر از اهمیت فوق‌العاده‌ای برخوردار شده است و نیاز به کتاب و افزایش دانش به سرعت در حال رشد است.

انتشار نتیجه مطالعات دانشمندان و پژوهشگران یکی از مؤثرترین راه‌های پاسخگویی به این نیاز است. در ایران اسلامی سالیان درازی است که دانش و علم‌آموزی چه به لحاظ دینی و چه از نظر فرهنگی ارزش‌های والایی داشته است. دانشمندان این مرزوبوم همواره در تلاش بوده‌اند که نتایج پژوهش‌ها و تلاش‌های خود را به رشته تحریر درآورند تا نسل‌های آینده بتوانند از دانسته‌ها و یافته‌های آنان در جهت رشد و شکوفایی کشور عزیزمان بهره ببرند. جهت تحقق این امر و گام برداشتن به سمت خودکفایی چاپ کتاب‌های علمی که یکی از پایه‌های توسعه فرهنگی و رشد و شکوفایی استعدادهاست به خصوص در زمینه علوم جدیدی همچون رایانه که توسعه و پیشرفت سایر تکنولوژی‌ها و علوم به آن وابسته است، ضروری به نظر می‌رسد.

آنچه که نویسندگان را بر آن داشت تا دست به تألیف این کتاب ببرند نیاز دانشجویان به منبعی فارسی و بومی جامع بود تا بتوانند نیازهای علاقه‌مندان به این مبحث را تأمین کند. هدف این کتاب در مرحله اول آشنایی با اصول و مفاهیم پایه و تکنیک‌های اساسی کلان داده و در مرحله دوم به‌کارگیری آن است. طبیعی است که در یک کتاب نمی‌توان تمام مسائل یک مبحث وسیع را جای داد. به همین دلیل هم سعی شده که کتاب حجمی بیشتر از معمول پیدا نکند. در نگارش این کتاب سعی شده علاوه بر تأمین نیازهای علمی دانشجویان مقطع کارشناسی و کارشناسی ارشد فناوری اطلاعات، دفاع سایبری، مهندسی رایانه، و دانشجویان سایر رشته‌های دانشگاهی (مخصوصاً دانشجویان رشته‌های فنی و مهندسی) نیز از درک مطالب این کتاب بی‌بهره نمانند.

در پایان، با عنایت به بار علمی کتاب و همچنین تلاش ما در ارائه قالبی روان و خوش‌خوان، امیدواریم توانسته باشیم قدمی هرچند کوچک در راستای اعتلای علمی میهن‌مان و دانشجویان و پژوهشگران این مرز و بوم برداریم. در انتها لازم می‌دانیم از معاونت محترم پژوهش، دانشکده علوم و مهندسی دفاعی و همچنین مسئول مرکز تحقیقات مبین جناب مهندس حبیبی که نظارت بر تألیف این کتاب را به عهده داشته‌اند تشکر نموده و از کلیه استادان گران‌قدر، همکاران که بدون زحمت‌های ایشان امکان انتشار کتاب با کیفیت فعلی وجود نداشت خالصانه تشکر و قدردانی نماییم. همچنین از همسران صبور و گرمی که اندک زمان ماندن در کنار ایشان صرف تألیف این اثر گردید صمیمانه تشکر و قدردانی می‌نمایم.

بدیهی است علیرغم دقت فراوان برای تقدیم کتابی بی‌عیب و نقص، این اثر خالی از اشکال نیست، امید است استادان محترم دانشگاه‌ها، دانشجویان عزیز و خوانندگان گرمی با ارائه نظرات و یادآوری اشکالات، این‌جانبان را مورد لطف خود قرار دهند.

پذیرای نظرات و پیشنهادات ارزشمند شما هستیم. آدرس پست الکترونیکی: dehghani@ihu.ac.ir

بهراد دهقانی

آرش غفوری

فهرست مطالب

مفاهیم و معماری کلان داده‌ها

فصل ۱

۲	۱. مقدمه
۲	۲. تاریخچه کلان داده‌ها
۴	۳. آینده کلان داده‌ها
۵	۳-۱. سناریوی خوش بینانه
۵	۳-۲. سناریوی بدبینانه
۶	۳-۳. سناریوهای میانه
۶	۴. معرفی کلان داده و ویژگی‌های آن
۶	۴-۱. ویژگی‌های اصلی کلان داده
۷	۴-۲. ویژگی‌های فرعی کلان داده
۸	۵. چرخه حیات کلان داده‌ها
۸	۶. انواع داده‌ها (ساخت یافته، غیر ساخت یافته، نیمه ساخت یافته)
۱۰	۶-۱. داده‌های ساخت یافته
۱۰	۶-۲. داده‌های بی ساختار یا غیر ساخت یافته
۱۰	۶-۳. داده‌های نیمه ساخت یافته
۱۱	۶-۴. انواع داده‌ها در یادگیری ماشین
۱۱	۷. رایانش کلان داده‌ها
۱۲	۷-۱. انواع رایانش/پردازش کلان داده‌ها
۱۳	۸. رایانش ابری و کلان داده‌ها
۱۳	۸-۱. ارتباط رایانش ابری با کلان داده‌ها
۱۴	۸-۲. کلان داده‌ها و رایانش ابری از دیدگاه سرویس
۱۵	۹. مدل مرجع کلان داده‌ها
۱۸	۱۰. یک معماری پشته کلان داده نمونه
۱۸	۱۰-۱. معماری پشته کلان داده در قالب مدل لامبدا
۲۲	۱۰-۲. معماری لایه‌ای پشته محاسباتی
۲۴	۱۱. فرصت‌ها و چالش‌های کلان داده
۲۴	۱۱-۱. چالش‌های کلان داده
۲۵	۱۱-۲. راهکارهای رایانش ابری برای چالش‌های حوزه کلان داده‌ها
۲۹	۱۲. خلاصه فصل
۳۰	۱۳. آزمون فصل اول

اکوسیستم هدوپ

فصل ۲

۳۲	۱. مقدمه
۳۲	۲. معرفی هدوپ
۳۳	۱-۲. معماری هدوپ
۳۳	۲-۲. نگاشت کاهش هدوپ
۳۷	۳-۲. سیستم فایل توزیع شده هدوپ
۳۹	۴-۲. یارن هدوپ
۴۵	۵-۲. رویکرد آموزشی در هدوپ
۴۶	۳. اکوسیستم هدوپ
۴۶	۱-۳. اچ بیس
۴۹	۲-۳. کاساندرا
۵۴	۳-۳. اسکوپ
۵۵	۴-۳. اسپارک
۵۶	۵-۳. ماهوت
۵۷	۶-۳. استورم
۵۸	۷-۳. اسپارک استریم
۵۹	۸-۳. هایو
۶۰	۹-۳. پیگ
۷۰	۱۰-۳. سامزا
۷۲	۱۱-۳. فلینک
۷۳	۱۲-۳. تریدنت
۷۳	۱۳-۳. زوکیپر
۷۳	۱۴-۳. آمباری
۷۴	۱۵-۳. اووزی
۷۵	۱۶-۳. کافکا
۸۵	۴. خلاصه فصل
۸۶	۵. آزمون فصل دوم

نصب و راه اندازی هدوپ

فصل ۳

۸۸	۱. مقدمه
۸۸	۲. روش‌های مختلف نصب هدوپ
۸۸	۱-۲. روش اصلی یا نصب کامل هدوپ
۸۸	۲-۲. نصب هدوپ به وسیله استفاده از تصاویر آماده (ایلاینس)
۸۹	۲-۳. نصب هدوپ از طریق داکر
۸۹	۲-۴. اجرای آنلاین هدوپ از بستر وب
۸۹	۳. نصب و راه‌اندازی هدوپ روی سیستم عامل لینوکس ابونتو
۹۵	۴. آموزش نصب آپاچی هدوپ در ویندوز
۹۸	۵. نصب هدوپ به وسیله استفاده از تصویر آماده هورتون ورکز
۹۸	۵-۱. دلیل انتخاب ماشین مجازی و لینوکس
۹۹	۵-۲. دلیل انتخاب هورتون ورکز
۹۹	۵-۳. دانلود ابزار و امکانات لازم
۱۰۲	۶. خلاصه فصل
۱۰۲	۷. آزمون فصل سوم

معرفی بستر جستجوی آپاچی سولار

فصل ۴

۱۰۴	۱. مقدمه
۱۰۴	۲. راه اندازی آپاچی سولار
۱۰۵	۲-۱. پیش نیازهای راه اندازی آپاچی سولار
۱۰۵	۲-۲. اجرای آپاچی سولار بر روی Jetty
۱۰۷	۲-۳. اجرای سولار در سایر کانتینرهای J2EE
۱۰۷	۲-۴. مثال Hello word با آپاچی سولار
۱۰۸	۳. مدیریت سولار
۱۰۹	۴. ناوبری سولار
۱۱۰	۵. مشکلات و راه حل‌های متداول
۱۱۱	۶. معماری آپاچی سولار
۱۱۳	۷. پیکربندی سولار
۱۱۳	۷-۱. ساختار سولار
۱۱۳	۷-۲. تعریف شما در سولار

۱۱۴	۳-۷. فیلدهای سولار
۱۱۶	۴-۷. فیلدهای پویا در سولار
۱۱۶	۵-۷. کپی کردن فیلدها
۱۱۶	۶-۷. پرداختن به انواع فیلد
۱۱۷	۷-۷. پیکربندی فراداده اضافی
۱۱۸	۸-۷. سایر عناصر مهم شمای سولار
۱۱۹	۹-۷. فایل‌های پیکربندی آپاچی سولار
۱۲۲	۸. افزونه سولار
۱۲۳	۹. سایر تنظیمات
۱۲۳	۱۰. بارگذاری داده‌ها در آپاچی سولار
۱۲۴	۱-۱۰. کنترل کننده درخواست استخراج - solr cell
۱۲۵	۲-۱۰. معرفی کنترل کننده‌های ورود داده
۱۲۶	۱۱. تعامل با سولار از طریق SolrJ
۱۲۸	۱۲. کار با اسناد غنی (Apache Tika)
۱۲۸	۱۳. جستجوی اطلاعات در سولار
۱۳۰	۱۴. خلاصه فصل
۱۳۰	۱۵. آزمون فصل چهارم

جستجوی توزیع شده با آپاچی سولار

فصل ۵

۱۳۲	۱. مقدمه
۱۳۲	۲. جستجوی توزیع شده
۱۳۲	۱-۲. الگوهای جستجوی توزیع شده
۱۳۴	۲-۲. جستجوی توزیع شده و آپاچی سولار
۱۳۵	۳. کار با SolrCloud
۱۳۵	۱-۳. زوکیپر چیست؟
۱۳۶	۲-۳. معماری SolrCloud
۱۳۹	۳-۳. ایجاد یک جستجوی توزیع شده‌ی با استفاده از SolrCloud
۱۴۵	۴-۳. ایجاد تکه‌ها، مجموعه‌ها، و کپی‌ها در SolrCloud
۱۴۷	۵-۳. مشکلات رایج و حل آن‌ها
۱۴۸	۶-۳. الگوریتم تکه‌کردن و تحمل خطا
۱۴۸	۷-۳. مسیریابی و تکه‌کردن سند
۱۵۱	۸-۳. تقسیم تکه‌ها

۱۵۲	۳-۹. تعادل بار و تحمل خطا در SolrCloud
۱۵۳	۴. ادغام آپاچی سولار و MongoDB
۱۵۳	۴-۱. ارتباط پایگاه داده‌ی NoSQL و کلان داده
۱۵۴	۴-۲. پایگاه داده MongoDB
۱۵۵	۴-۳. نصب MongoDB
۱۵۷	۴-۴. ایجاد شاخص‌های سولار از MongoDB
۱۶۰	۵. خلاصه
۱۶۰	۶. آزمون فصل پنجم

فصل ۶ جستجوی کلان داده با استفاده از هدوپ

۱۶۲	۱. مقدمه
۱۶۲	۲. استفاده از پایگاه داده‌ی NoSQL
۱۶۳	۳. ادغام کاساندرا و سولار
۱۶۳	۳-۱. پیکر بندی تک گره
۱۶۶	۳-۲. یکپارچه سازی با کاساندرا ی چندگره ی
۱۶۶	۴. مقیاس دهی سولار از طریق استورم
۱۶۷	۴-۱. راه اندازی آپاچی استورم
۱۶۹	۵. خلاصه
۱۷۰	۶. آزمون فصل ششم

پیوست پیوست

۱۷۲	۱. مراجع
۱۷۳	۲. پیوست: آدرس برنامه های کاربردی

مقدمه

با رشد فزاینده فناوری های اطلاعاتی در دو دهه اخیر و تولید فزاینده داده ها، حجم بسیار عظیمی از داده ها با قالب های گوناگون از منابع مختلفی نظیر فرآیندهای تجارت الکترونیکی، اینترنت اشیا، سیستم های مراقبت سلامت، شبکه های اجتماعی و کاربردهای مبتنی بر وب، ایجاد شده است. کلان داده ها (داده های عظیم یا داده های حجیم) مفهوم گسترده ای می باشد که دنیای فناوری را متحول کرده است. به طور کلی می توان گفت کلان داده ها به حجم زیادی از داده ها گفته می شود که پردازش آن ها به وسیله پایگاه داده های سنتی و ابزارهای پردازشی و تحلیل سنتی قابل انجام نیست. از یک جهت ارزش بالقوه نهفته در کلان داده ها آن قدر بالا است که در دهه آتی می توان آن را به مثابه یک چاه جدید نفتی کشف نشده دانست. در سال ۲۰۱۹ هم زمان با تدوین این کتاب، داده های بزرگ به عنوان ۱۰ فناوری برتر در کنار فناوری های نوظهور دیگر مانند اینترنت اشیا قرار گرفت.

چرخه حیات کلان داده شامل چهار مرحله اصلی می باشد که شامل مرحله ذخیره سازی، پردازش، تحلیل و بصری سازی می باشد. در این کتاب، هر چهار موضوع اصلی مورد بررسی قرار گرفته و ابزارهای مربوطه معرفی و پیاده سازی می گردند. در این کتاب سعی کردیم مفاهیم تنوری را در کنار کاربرد و در استفاده از ابزارهای مربوطه در محیط عملیاتی مطرح نماییم.

کتاب ذخیره سازی و پردازش داده های بزرگ بر اساس تنوع موضوعات مختلف تحت پوشش شش فصل مطابق ذیل تقسیم شده است:

فصل اول: مفاهیم و معماری کلان داده ها

فصل دوم: اکوسیستم هدوپ

فصل سوم: نصب و راه اندازی هدوپ

فصل چهارم: معرفی آپاچی سولار

فصل پنجم: جستجوی توزیع شده

فصل ششم: جستجوی کلان داده با استفاده از هدوپ