انتشارات دانشگاه کردستان

کلان داده‌ها و اینترنت اشیاء

اصول و ابزارها

سعدون عزیزی
یاسین امینی

کلان داده‌ها و اینترنت اشیاء

اصول و ابزارها

کلان داده‌ها و اینترنت اشیاء

اصول و ابزارها

مؤلفان:

سعدون عزیزی

عضو هیات علمی گروه مهندسی کامپیوتر دانشگاه کردستان

یاسین امینی

فارغ‌التحصیل مهندسی کامپیوتر دانشگاه کردستان

انتشارات دانشگاه کردستان

۱۳۹۹

سرشناسه	:عزیزی، سعدون
عنوان و نام پدیدآور	:کلان داده‌ها و اینترنت اشیاء: اصول و ابزارها/سعدون عزیزی، یاسین امینی.
مشخصات نشر	:سنندج، دانشگاه کردستان، انتشارات، ۱۳۹۹.
مشخصات ظاهری	:۳۱۷ص.، مصور، جدول، نمودار.
شابک	: ۹-۲۴-۶۷۰۲-۶۲۲-۹۷۸
وضعیت فهرست‌نویسی	:فیپا.
یادداشت	:واژه نامه.
یادداشت	:کتاب‌نامه.
موضوع	:کلان داده‌ها، اینترنت اشیاء
شناسه افزوده	:امینی، یاسین.
شناسه افزوده	:دانشگاه کردستان، انتشارات.
رده‌بندی کنگره	:QA۷۶/۹
رده‌بندی دیویی	:۰۰۵/۷
شماره کتابشناسی ملی	:۶۱۸۴۰۰۷

نام کتاب: کلان داده‌ها و اینترنت اشیاء: اصول و ابزارها

مولفین: سعدون عزیزی، یاسین امینی

طراحی جلد: کوروش عنبری

نوبت چاپ: اول

تیراژ: ۵۰۰ جلد

ناشر: انتشارات دانشگاه کردستان – سنندج، بلوار پاسداران، دانشگاه کردستان

قیمت: ۴۵۰۰۰ تومان

تلفن جهت سفارش کتاب: ۰۸۷۳۳۶۲۴۰۰۸ (خانم کاردوسیان)

شابک: ۹-۲۴-۶۷۰۲-۶۲۲-۹۷۸

حق چاپ برای انتشارات دانشگاه کردستان محفوظ است.

پیشگفتار

در حال حاضر داده به عنوان باارزش‌ترین منبع جهان مطرح است و با ظهور اینترنت اشیاء اهمیت آن چند برابر شده است. مجموعه داده‌های نوین به اندازه‌ای پیچیده هستند که نمی‌توان آنها را توسط سخت‌افزارها و نرم‌افزارهای سنتی مدیریت کرد. حجم، سرعت و تنوع سه ویژگی اصلی داده‌های تولید شده‌ی امروزی هستند که باعث ایجاد مفهومی به نام کلان داده‌ها شده است. چنین ویژگی‌هایی روال‌های دریافت، ذخیره‌سازی، پردازش، تحلیل و بصری‌سازی کلان داده‌ها را به یک امر چالش‌برانگیز تبدیل کرده است. در دنیای رقابتی امروز، تحلیل کلان داده‌ها از اهمیت چشمگیری برخوردار است. اهمیت کلان داده‌ها به معنای مقدار داده در اختیار یک شرکت یا سازمان نیست بلکه به چگونگی استفاده از آن داده‌ها بستگی دارد. پردازش و تحلیل داده‌های جمع‌آوری شده به شرکت‌ها و سازمان‌ها کمک می‌کند تا بینش‌های لازم را کسب کرده و از آن در راستای تصمیم‌گیری‌های استراتژیک بهره بگیرند. طی چند سال گذشته، چارچوب‌ها و ابزارهای نوینی برای ذخیره‌سازی، پردازش و تحلیل کلان داده‌ها ارائه شده است که آشنایی و کار با آنها می‌تواند زمینه های پژوهشی و فرصت‌های شغلی متنوعی را پیش‌روی متخصصان این حوزه قرار دهد. با جستجو و مطالعه‌ی منابع مختلف، نویسندگان پی بردند که منبع جامعی برای آشنایی با اصول و ابزارهای کلان داده‌ها و اینترنت اشیاء وجود ندارد. به همین منظور، تألیف چنین کتابی را ضروری دانستند. دانشجویان و علاقمندان به رشته‌های مهندسی کامپیوتر، فناوری اطلاعات، علوم کامپیوتر و آمار در مقاطع مختلف تحصیلی می‌توانند از مطالب این کتاب استفاده کنند. علاوه بر این، شرکت‌ها و سازمان‌های فعال در حوزه کلان داده‌ها و اینترنت اشیاء می‌توانند مخاطب کتاب پیش‌رو باشند. در مورد نحوه‌ی نگارش این کتاب بایستی به این نکته اشاره شود که نویسندگان با تحقیق و مطالعه‌ی منابع معتبر، که در انتهای کتاب ذکر شده‌اند، و همچنین دانش حاصل از تدریس درس کلان داده‌ها و اینترنت اشیاء این اثر را تولید کرده‌اند. امید است تألیف این کتاب بتواند به نوبه‌ی خود سهم برجسته‌ای در افزایش توان علمی جامعه‌ی دانشگاهی و حرفه‌ای کشور داشته باشد. مسلماً این اثر بدون اشکال نیست. لطفا جهت کمک به بهبود و ارتقای بیشتر کیفیت مطالب، پیشنهادات و انتقادات خود را به آدرس iot@uok.ac.ir ارسال نمائید.

سعدون عزیزی

عضو هیات علمی گروه مهندسی کامپیوتر دانشگاه کردستان

تابستان ۱۳۹۹

فهرست مطالب

۱. مقدمه‌ای بر کلان داده‌ها و اینترنت اشیاء ۱

۱-۱ مقدمه ۳

۱-۲ کلان داده‌ها ۴

۱-۲-۱ ویژگی‌های کلان داده‌ها ۵

۱-۲-۲ اهمیت کلان داده‌ها ۱۰

۱-۲-۳ حوزه‌های مرتبط با کلان داده‌ها ۱۳

۱-۳ منابع کلان داده‌ها ۱۸

۱-۴ چالش‌های کلان داده‌ها ۲۰

۱-۵ اینترنت اشیاء ۲۲

۱-۵-۱ تعریف اینترنت اشیاء ۲۳

۱-۵-۲ اهمیت اینترنت اشیاء ۲۵

۱-۵-۳ چالش‌های اینترنت اشیاء ۳۰

۱-۵-۴ معماری اینترنت اشیاء ۳۲

۱-۴-۵-۱ معماری سه لایه ۳۲

۱-۴-۵-۲ معماری چهار لایه ۳۲

۱-۴-۵-۳ معماری پنج لایه ۳۳

۱-۶ ترکیب اینترنت اشیاء با کلان داده‌ها ۳۵

۱-۷ جمع‌بندی و نتیجه‌گیری ۳۶

۲.پایگاه داده‌های غیررابطه‌ای ۳۹

۲-۱ مقدمه ۴۱

۲-۲ تاریخچه ۴۱

۲-۳ مدل داده ACID و BASE ۴۲

۲-۴ دسته‌بندی داده‌ها ۴۳

۲-۵ اصل CAP ۴۵
۲-۶ اهمیت پایگاه داده‌های غیررابطه‌ای ۴۶
۲-۷ مقیاس‌پذیری عمودی و افقی ۴۹
۲-۸ ویژگی‌های پایگاه داده‌های غیررابطه‌ای ۵۰
۲-۹ چالش‌های پایگاه داده‌های غیررابطه‌ای ۵۳
۲-۱۰دسته بندی سیستم‌های مدیریت داده غیررابطه‌ای ۵۵
۲-۱۰-۱پایگاه داده‌های کلید/مقدار ۵۶
۲-۱۰-۲پایگاه داده‌های سندگرا ۵۷
۲-۱۰-۳پایگاه داده‌های ستونی ۶۲
۲-۱۰-۴پایگاه داده‌های مبتنی بر گراف ۶۴
۲-۱۱تفاوت پایگاه داده‌های رابطه‌ای و غیررابطه‌ای ۶۶
۲-۱۲پایگاه داده‌های NewSQL ۷۰
۲-۱۳نتیجه‌گیری ۷۲
۳. داده کاوی در کلان داده‌ها ۷۵
۳-۱ مقدمه ۷۷
۳-۲ داده کاوی ۷۸
۳-۳ مراحل داده کاوی ۷۹
۳-۳-۱ شناسایی منابع اطلاعاتی و تعریف مسأله ۷۹
۳-۳-۲ جمع‌آوری و پیش‌پردازش داده ۸۰
۳-۳-۳ استخراج الگو و مدل از داده‌ها ۸۱
۳-۳-۴ شناسایی موارد کلیدی استخراج شده ۸۲
۳-۳-۵ تفسیر و گزارش نتایج ۸۲
۳-۴ انواع روش‌های داده کاوی ۸۳
۳-۵ یادگیری ماشین ۸۳
۳-۵-۱ یادگیری با نظارت ۸۴

۳-۵-۲ یادگیری بدون نظارت ۸۵
۳-۵-۳ یادگیری نیمه نظارتی ۸۶
۳-۵-۴ یادگیری تقویتی ۸۷
۳-۶ دسته‌بندی ۸۸
۳-۶-۱ رگرسیون ۸۹
۳-۶-۲ درخت تصمیم ۹۱
۳-۶-۳ شبکه‌های عصبی ۹۵
۳-۶-۴ K-نزدیک‌ترین همسایه ۹۹
۳-۷ خوشه‌بندی ۱۰۲
۳-۷-۱ کاربردهای خوشه‌بندی ۱۰۵
۳-۷-۲ الگوریتم‌های خوشه‌بندی ۱۰۶
۳-۷-۲-۱ الگوریتم K-میانگین ۱۰۶
۳-۷-۲-۲ الگوریتم CLARA ۱۰۹
۳-۸ استخراج قواعد انجمنی ۱۱۰
۳-۸-۱ پشتیبان و اطمینان ۱۱۱
۳-۸-۲ الگوریتم‌های استخراج قواعد انجمنی ۱۱۳
۳-۸-۲-۱ الگوریتم AIS ۱۱۴
۳-۸-۳ الگوریتم پیشینار یا Apriori ۱۱۶
۳-۸-۴ الگوریتم AprioriTid ۱۱۷
۳-۸-۵ الگوریتم Apriorihybrid ۱۱۷
۳-۹ ابزارهای داده کاوی ۱۱۷
۳-۹-۱ زبان برنامه‌نویسی پایتون ۱۱۸
۳-۹-۲ زبان برنامه‌نویسی R ۱۱۹
۳-۹-۳ محیط برنامه‌نویسی متلب ۱۲۰
۳-۹-۴ نرم‌افزار رپیدماینر ۱۲۱
۳-۹-۵ نرم‌افزار کلمنتاین ۱۲۲
۳-۹-۶ نرم‌افزار اُرنج ۱۲۳

۳-۹-۷ نرم‌افزار وِکا ۱۲۴
۳-۱۰جمع‌بندی و نتیجه‌گیری ۱۲۵
۴. اکوسیستم هدوپ..... ۱۲۷
۴-۱ مقدمه ۱۲۹
۴-۲ داستانی از هدوپ ۱۳۰
۴-۳ تاریخچه هدوپ ۱۳۲
۴-۴ هدوپ چیست؟ ۱۳۳
۴-۵ هدوپ چگونه کار میکند؟ ۱۳۵
۴-۶ معماری هدوپ ۱۳۶
۴-۷ ویژگی‌ها و مزایای هدوپ ۱۳۷
۴-۸ اکوسیستم هدوپ ۱۴۱
۴-۸-۱ سیستم فایل توزیع‌شده هدوپ (HDFS) ۱۴۲
۴-۸-۲ زمانبند هدوپ (Yarn) ۱۴۷
۴-۸-۳ مدل پردازشی نگاشت/کاهش ۱۴۹
۴-۸-۴ HBASE ۱۵۲
۴-۸-۵ HIVE ۱۵۳
۴-۸-۶ آپاچی پیگ (PIG) ۱۵۴
۴-۸-۷ تز ۱۵۶
۴-۸-۸ اسکوپ (Sqoop) ۱۵۷
۴-۸-۹ آپاچی فلوم (Flume) ۱۵۸
۴-۸-۱۰دریل (Drill) ۱۵۹
۴-۸-۱۱آپاچی آمباری (Ambari) ۱۶۰
۴-۸-۱۲زوکیپر (ZooKeeper) ۱۶۰
۴-۸-۱۳اوزی (Oozie) ۱۶۱
۴-۸-۱۴سامزا (Samza) ۱۶۱
۴-۸-۱۵ماهوت (Mahout) ۱۶۳

۴-۹ توسعه‌دهندگان هدوپ ۱۶۳
۴-۱۰معایب هدوپ ۱۶۴
۴-۱۱نصب هدوپ ۱۶۶
۴-۱۲شروع کار با هدوپ (شمارش کلمات) ۱۶۸
۴-۱۳نتیجه‌گیری و جمع‌بندی ۱۷۰
۵.پردازش درون‌حافظه‌ای با استفاده از گریدگین و اسپارک..... ۱۷۱
۵-۱ مقدمه ۱۷۳
۵-۲ گریدگین ۱۷۳
۵-۲-۱ موتور پردازشی ۱۷۴
۵-۲-۲ فایل سیستم ۱۷۵
۵-۲-۳ ترکیب هدوپ و گریدگین ۱۷۷
۵-۳ آپاچی اسپارک ۱۷۸
۵-۳-۱ مزیت‌های اسپارک..... ۱۷۹
۵-۳-۲ اسپارک در دنیای واقعی..... ۱۸۵
۵-۳-۳ بررسی اجزای اسپارک..... ۱۸۶
۵-۳-۴ چالش‌های اسپارک..... ۱۹۰
۵-۳-۵ نصب اسپارک..... ۱۹۱
۵-۴ شروع کار با اسپارک و زبان برنامه‌نویسی اسکالا ۱۹۳
۵-۵ جمع‌بندی و نتیجه‌گیری ۲۰۲
۶. پردازش جریانی و ابزارهای آن..... ۲۰۵
۶-۱ مقدمه ۲۰۷
۶-۲ پردازش جریانی ۲۰۷
۶-۳ آپاچی استورم ۲۱۲
۶-۳-۱ معماری آپاچی استورم..... ۲۱۳
۶-۳-۲ مزیت‌های آپاچی استورم..... ۲۱۶

۶-۳-۳ نصب آپاچی استورم ۲۱۸
۶-۴ آپاچی فلینک ۲۲۳
۶-۴-۱ معماری آپاچی فلینک ۲۲۵
۶-۴-۲ اجزای آپاچی فلینک ۲۲۷
۶-۴-۳ مراحل اجرای برنامه‌ها در فلینک ۲۲۸
۶-۴-۴ مزایا و ویژگی‌ها ۲۲۹
۶-۴-۵ نصب و پیکربندی ۲۳۱
۶-۵ آپاچی نایفای ۲۳۳
۶-۵-۱ نحوه کار آپاچی نایفای ۲۳۵
۶-۵-۲ ویژگی‌های آپاچی نایفای ۲۳۵
۶-۵-۳ کاربردهای آپاچی نایفای ۲۳۷
۶-۵-۴ برنامه‌نویسی جریانی ۲۳۷
۶-۵-۵ نصب آپاچی نایفای ۲۳۸
۶-۶ نتیجه‌گیری و جمع‌بندی ۲۳۹
۷. سیستم پیامرسانی کافکا ۲۴۱
۷-۱ مقدمه ۲۴۳
۷-۲ سیستم‌های پیامرسانی ۲۴۳
۷-۳ هدف کافکا ۲۴۸
۷-۴ اجزای آپاچی کافکا ۲۵۰
۷-۵ مزایای آپاچی کافکا ۲۵۴
۷-۶ چالش‌های آپاچی کافکا ۲۵۶
۷-۷ انواع سیستم‌های مبتنی بر آپاچی کافکا ۲۵۷
۷-۸ ابزارهای کافکا ۲۵۹
۷-۹ نحوه کار آپاچی کافکا ۲۶۰
۷-۱۰نصب و پیکربندی آپاچی کافکا ۲۶۱

۷-۱۱شروع کار با کافکا ۲۶۳
۷-۱۲نتیجه‌گیری و جمع بندی ۲۶۹
۸. ابزارهای پردزاشی نوین ۲۷۱
۸-۱ مقدمه ۲۷۳
۸-۲ آپاچی بیم ۲۷۳
۸-۲-۱ مزایا و چالش‌ها ۲۷۵
۸-۲-۲ نصب آپاچی بیم ۲۷۶
۸-۳ آپاچی اَپکس ۲۷۸
۸-۳-۱ معماری آپاچی اَپکس ۲۷۹
۸-۳-۲ مدل برنامه‌نویسی ۲۸۰
۸-۴ آپاچی ایگنایت ۲۸۱
۸-۴-۱ نحوه کارکرد آپاچی ایگنایت ۲۸۲
۸-۴-۲ معماری آپاچی ایگنایت ۲۸۴
۸-۴-۳ مزیت‌های استفاده از آپاچی ایگنایت ۲۸۵
۸-۵ نتیجه‌گیری و جمع‌بندی ۲۸۹
واژه‌نامه ۲۹۱
اختصارات ۲۹۶
منابع و مراجع ۲۹۷

فهرست شکل‌ها

شکل ۱-۱ ویژگی‌های اصلی کلان داده‌ها ۶

شکل ۱-۲: در هر دقیقه چه اتفاقی در اینترنت می‌افتد؟ ۱۱

شکل ۱-۳: ارتباط یادگیری ماشین و کلان داده ۱۷

شکل ۱-۴: حوزه‌های کاربردی اینترنت اشیاء ۲۶

شکل ۱-۵: سوسک هوشمند برای نجات زنده‌های زیر آوار ۲۸

شکل ۱-۶: درصد درآمد هر یک از حوزه‌های اینترنت اشیاء تا سال ۲۰۲۵ ..... ۳۰

شکل ۱-۷: لایه‌های مختلف معماری پنج لایه ۳۴

شکل ۲-۱: نمونه‌هایی از سیستم مدیریت داده که در آنها دو مورد از ویژگیهای اصل CAP رعایت شده است ۴۶

شکل ۲-۲: رشده داده‌های غیرساختاریافته ۴۸

شکل ۲-۳: مقیاس‌پذیری عمودی (up-scale) در مقابل مقیاس‌پذیری افقی (out-scale) ۵۰

شکل ۲-۴: رابطه بین جدول‌ها در پایگاه داده‌های رابطه‌ای ۶۰

شکل ۲-۵: پایگاه داده مبتنی بر گراف با ارتباط دوستی یا علاقه‌مندی‌ها ۶۵

شکل ۲-۶: روند روبه رشد استفاده از پایگاه داده‌های غیررابطه‌ای به ویژه Neo4j ۶۶

شکل ۳-۱: انواع روش‌های داده کاوی ۸۳

شکل ۳-۲: یادگیری با نظارت ۸۵

شکل ۳-۳: یادگیری بدون نظارت ۸۶

شکل ۳-۴: یادگیری تقویتی (یادگیری مبتنی بر پاداش و تنبیه) ۸۷

شکل ۳-۵: دسته‌بندی داده‌ها با استفاده از الگوریتم رگرسیون ۹۱

شکل ۳-۶: نمونه‌ای از یک درخت تصمیم برای تشخیص بیماری مراجعه‌کننده ۹۳

شکل ۳-۷: شبکه‌های عصبی ساده که چند ورودی داشته و فقط یک لایه میانی دارد ۹۵

شکل ۳-۸: رسم خط به کمک پرسپترون برای جداسازی سگ از گربه ۹۷

شکل ۳-۹: پرسپترون چند لایه‌ای که دارای چندین لایه مخفی یا میانی است. ۹۸

شکل ۳-۱۰: خوشه‌بندی و تفکیک مشتریان برای سودهی بیشتر ۱۰۳

شکل ۳-۱۱: نمودار داده‌های جدول ۳-۵ ..... ۱۰۸
شکل ۳-۱۲: سبد خریدهای مختلف یک فروشگاه ۱۱۳
شکل ۳-۱۳: محیط ژوپیتر برای برنامه‌نویسی پایتون ۱۱۸
شکل ۳-۱۴: محیط RStudio برای زبان برنامه‌نویسی R ۱۱۹
شکل ۳-۱۵: محیط برنامه‌نویسی متلب ۱۲۰
شکل ۳-۱۶: محیط نرم‌افزار رپیدماینر ۱۲۲
شکل ۳-۱۷: محیط نرم‌افزار spss clementine ۱۲۳
شکل ۳-۱۸: محیط نرم‌افزار اُرنج ۱۲۴
شکل ۳-۱۹: محیط نرم‌افزار وِکا ۱۲۵
شکل ۴-۱: داگ کاتینگ، خالق هدوپ ۱۳۴
شکل ۴-۲: معماری هدوپ ۱۳۷
شکل ۴-۳: اجزای تشکیل دهنده اکوسیستم هدوپ (باغ وحش هدوپ!) ۱۴۲
شکل ۴-۴: تقسیم‌بندی داده در فایل سیستم هدوپ ۱۴۵
شکل ۴-۵: نحوه انجام عمل تکرار در فایل سیستم هدوپ ۱۴۶
شکل ۴-۶: نحوه تکرار داده‌ها در مرکز داده هدوپ ۱۴۶
شکل ۴-۷: نحوه کارکرد آپاچی یارن (زمان بند هدوپ) ۱۴۹
شکل ۴-۸: نحوه کارکرد موتور پردازشی نگاشت/کاهش، برای شمارش کلمات ۱۵۰
شکل ۴-۹: دسترسی ترتیبی در مقابل دسترسی تصادفی ۱۵۳
شکل ۴-۱۰: معماری آپاچی پیگ ۱۵۵
شکل ۴-۱۱: معماری آپاچی تز ۱۵۶
شکل ۴-۱۲: معماری آپاچی اسکوپ ۱۵۸
شکل ۴-۱۳: معماری آپاچی فلوم ۱۵۹
شکل ۴-۱۴: معماری و نحوه کارکرد آپاچی سامزا ۱۶۲
شکل ۴-۱۵: ذخیره نتایج میانی در فایل سیستم هدوپ و کاهش سرعت پردازشی سیستم ۱۶۴
شکل ۵-۱: نحوه کارکرد گریدگین ۱۷۵
شکل ۵-۲: نحوه کارکرد سیستم ذخیره‌سازی گریدگین ۱۷۷

شکل ۵-۳: ترکیب اکوسیستم هدوپ و گریدگین ۱۷۷
شکل ۵-۴: عملیات‌های انجام شده توسط چارچوب اسپارک ۱۸۱
شکل ۵-۵: تفاوت بین ذخیره نتایج میانی در هدوپ و اسپارک ۱۸۲
شکل ۵-۶: معماری آپاچی اسپارک ۱۸۳
شکل ۵-۷: معماری رایانش مه ۱۸۶
شکل ۵-۸: اجزای تشکیل‌دهنده اسپارک ۱۸۷
شکل ۵-۹: پردازش داده‌های جریانی در اسپارک با روش میکرودسته‌ای ۱۸۸
شکل ۵-۱۰: درست نصب شدن اسپارک ۱۹۲
شکل ۶-۱: پنجره ثابت در پردازش‌های جریانی ۲۰۹
شکل ۶-۲: پنجره اسلایدی در پردازش داده‌های جریانی ۲۱۰
شکل ۶-۳: تفاوت بین پنجره‌های داده ۲۱۱
شکل ۶-۴: معماری آپاچی استورم ۲۱۴
شکل ۶-۵: توپولوژی آپاچی استورم ۲۱۵
شکل ۶-۶: نحوه کارکرد و ارتباط بخش‌های متخلف آپاچی استورم ۲۱۶
شکل ۶-۷: لوگوی آپاچی فلینک ۲۲۵
شکل ۶-۸: معماری آپاچی فلینک ۲۲۶
شکل ۶-۹: مراحل اجرای برنامه‌ها در آپاچی فلینک ۲۲۹
شکل ۶-۱۰: داشبورد تحت وب آپاچی فلینک ۲۳۲
شکل ۶-۱۱: نمایش اجرای برنامه‌ها در داشبورد مدیریتی آپاچی فلینک ۲۳۳
شکل ۶-۱۲: محیط توسعه آپاچی نایفای ۲۳۴
شکل ۶-۱۳: محیط یکپارچه پردازشی برای پردازش کلان داده‌ها ۲۳۶
شکل ۶-۱۴: نصب موفق آپاچی نای‌فای ۲۳۹
شکل ۷-۱: سیستم پیامرسانی نقطه به نقطه ۲۴۴
شکل ۷-۲: سیستم پیامرسانی انتشار/اشتراک ۲۴۵
شکل ۷-۳: نحوه عملکرد آپاچی کافکا در دنیای واقعی ۲۴۸

شکل ۷-۴: شبکه استفاده آپاچی کافکا ۲۴۹
شکل ۷-۵: ارتباط مستقیم شبکه پردازشی ۲۴۹
شکل ۷-۶: ارتباط غیرمستقیم شبکه پردازشی با استفاده از آپاچی کافکا ۲۵۰
شکل ۷-۷: اجزای تشکیل‌دهنده آپاچی کافکا ۲۵۰
شکل ۷-۸: کافکا در نقش سیستم انتشار/اشتراک ۲۵۸
شکل ۷-۹: کافکا در نقش سیستم صف ۲۵۸
شکل ۷-۱۰: نحوه کارکرد کافکا استریمز ۲۶۰
شکل ۷-۱۱: نحوه عملکرد آپاچی کافکا ۲۶۱
شکل ۸-۱: آپاچی بیم و نحوه کارکرد آن ۲۷۴
شکل ۸-۲: آپاچی بیم به عنوان یکپارچه‌ساز ابزارهای کلان داده ۲۷۵
شکل ۸-۳: معماری آپاچی اَپکس ۲۸۰
شکل ۸-۴: نحوه کارکرد و اجرای برنامه در آپاچی اَپکس ۲۸۱
شکل ۸-۵: یکپارچه‌سازی ابزارهای ذخیره‌سازی و زبان‌های برنامه‌نویسی در آپاچی ایگنایت ۲۸۴
شکل ۸-۶: نحوه ذخیره داده‌ها در آپاچی ایگنایت ۲۸۵
شکل ۸-۷: معماری پردازش جریانی در آپاچی ایگنایت ۲۸۶
شکل ۸-۸: تغییرات قیمت حافظه اصلی و حافظه جانبی با گذر زمان ۲۸۸
شکل ۸-۹: دایرکتوری‌های لازم برای راه‌اندازی SBT ۱۹۳
شکل ۸-۱۰: نحوه ایجاد دایرکتوری‌های پروژه اسکالا و اسپارک ۱۹۴
شکل ۸-۱۱: موفقیت‌آمیز بودن ایجاد پروژه ۱۹۶
شکل ۸-۱۲: آدرس قرارگیری فایل اجرایی ۱۹۶
شکل ۸-۱۳: خروجی اجرای پروژه ۱۹۷

فهرست جدول‌ها

جدول ۱-۱: تفاوت بین داده‌های سنتی و کلان داده‌ها ۱۳

جدول ۲-۱: ذخیره اطلاعات در جدول پایگاه داده رابطه‌ای ۵۸

جدول ۲-۲: اضافه کردن اطلاعات جدید به پایگاه داده‌های رابطه‌ای ۵۹

جدول ۲-۳: ذخیره اطلاعات در جدول پایگاه داده رابطه‌ای ۶۲

جدول ۲-۴: ذخیره اطلاعات در جدول پایگاه داده غیررابطه‌ای ستون‌گرا ۶۲

جدول ۲-۵: تفاوت بین پایگاه داده‌های رابطه‌ای و غیررابطه‌ای ۷۰

جدول ۳-۱: مجموعه داده‌های آموزشی ۸۹

جدول ۳-۲: مجموعه داده‌های آموزش ۱۰۱

جدول ۳-۳: فاصله ورودی جدید از تک تک نمونه‌ها ۱۰۱

جدول ۳-۴: تعیین برچسب برای نمونه جدید برحسب رأی اکثریت ۱۰۲

جدول ۳-۵: مجموعه داده جهت خوشه‌بندی با روش k-میانگین ۱۰۸

جدول ۳-۶: فاصله داده‌ها از مراکز خوشه ۱۰۹

جدول ۳-۷: سبد خرید برای استخراج قواعد انجمنی ۱۱۲

جدول ۳-۸: داده‌های یک سبد خرید ۱۱۳

جدول ۳-۹: پایگاه داده اصلی ۱۱۴

جدول ۳-۱۰: الگوریتم AIS و اولین پیمایش پایگاه داده برای استخراج محصولات و تعداد آن‌ها ۱۱۵

جدول ۳-۱۱: استخراج قواعد دوتایی با شناسه‌ی تراکنش ۱۱۵

جدول ۳-۱۲: استخراج قواعد سه‌تایی با شناسه‌ی تراکنش ۱۱۶

جدول ۷-۱: مقایسه سیستم‌های پیامرسان نقطه به نقطه و انتشار/اشتراک ۲۴۵

فصل اول

مقدمه ای بر کلان داده ها و اینترنت اشیاء

مطالبی که در این فصل پوشش داده می‌شوند:

- ✓ کلان داده چیست و چه ویژگی‌هایی دارد؟
- ✓ چرا مطالعه کلان داده‌ها اهمیت دارد؟
- ✓ اینترنت اشیاء چیست؟
- ✓ چرا مطالعه اینترنت اشیاء اهمیت دارد؟
- ✓ کلان داده‌ها و اینترنت اشیاء چه حوزه‌ها و فناوری‌هایی را تحت تأثیر قرار می‌دهند؟
- ✓ رابطه کلان داده‌ها و اینترنت اشیاء چیست؟

۱-۱ مقدمه

در دهه‌ی اخیر داده به عنوان باارزش‌ترین منبع جهان مطرح شده است و این ارزش و اهمیت داده روز به روز افزایش می‌یابد تا جایی که در سال ۲۰۱۵ برای اولین بار ارزش داده از نفت پیشی گرفت!

طبق آمارهای منتشر شده، ۹۰ درصد داده‌های ذخیره شده در دنیا در دو سال اخیر تولید شده‌اند! روزانه ۲.۵ کوین‌تلیون[۱] داده در دنیا تولید می‌شود. در سال ۲۰۱۱ یک رکورد جدید ثبت گردید و در هر دو روز به اندازه کل داده‌های تولید شده تا سال ۲۰۰۳، داده تولید شد! این رکورد در سال ۲۰۱۵ به دو ساعت رسید! در سال‌های اخیر حجم داده‌های تولید شده هر دو روز دو برابر می‌شود! در سال ۲۰۰۸ تعداد دستگاه‌های متصل به اینترنت از تعداد آدم‌های روی کره زمین پیشی گرفت و به یک نقطه عطف در هوشمندی دنیا رسید! شرکت‌های بزرگ دنیا هر روز آماری از تعداد دستگاه‌های هوشمند که تا سال ۲۰۲۰ قرار است وجود داشته باشند، ارائه می‌دهند. بر اساس پیش‌بینی شرکت سیسکو، تعداد دستگاه‌های متصل به اینترنت تا سال ۲۰۲۰ به بیش از ۲۶ میلیارد خواهد رسید!

همه عناوین بالا، عناوین اصلی اخباری هستند که ما هر روز در گوشه و کنار مجله‌ها و سایت‌های معتبر مرتبط با علوم کامپیوتر مشاهده می‌کنیم. همین امر ایجاب می‌کند که ما در مورد این فناوری‌ها به اندازه کافی مطالعه داشته باشیم تا بتوانیم دانش خود را به روز نگه داشته و خود را با قطار سریع‌السیر دنیای فناوری، که هر

---

[۱] Quintillion

روز با سرعت بیشتری پیش می‌رود، هم جهت نموده و از این فناوری‌ها در امور کشوری، مدیریت بحران‌ها، ساده‌سازی زندگی افراد جامعه و غیره استفاده نماییم. در این فصل نگاهی به فناوری‌های دنیای امروز، اهمیت آن‌ها و چرایی وجود آن‌ها خواهیم داشت.

۱-۲ کلان داده‌ها

مفهوم کلان داده‌ها[۱] برای اولین بار در دهه اول قرن جاری به علت پیشی گرفتن سرعت و نرخ تولید داده از تحلیل و فضای ذخیره سازی موجود، به طور جدی مطرح گردید. دقیقاً زمانی که اولین جرقه‌های مفهومی فراتر از داده مطرح گردید، سازمان‌ها و شرکت‌های بزرگ با ذخیره‌سازی و تحلیل داده به عنوان چالشی اساسی مواجه شدند.

به دلیل نوپا بودن و وجود سازمان‌های مختلف با حوزه‌های کاری مختلف، تاکنون برای کلان داده‌ها استاندارد واحدی تعریف نشده است. در ادامه، تعریف‌های متعدد و رایجی که برای این مفهوم وجود دارد را ارائه می‌دهیم.

موسسه گارتنر

کلان داده‌ها به داده‌هایی گفته می‌شوند که دارای حجم[۲] بالا، تنوع[۳] بالا و نرخ تولید[۴] بالا باشند.

شرکت اوراکل

کلان داده‌ها به داده‌هایی گفته می‌شوند که دارای چهار ویژگی حجم بالا، تنوع بالا، نرخ تولید بالا و همچنین صحت و درستی باشند.

---

۱ Big Data
۲ Volume
۳ Veriety
۴ Velocity

مک‌کینزی[۱]

مجموعه داده‌هایی که اندازه آنها از قابلیت ابزارهای پایگاه داده سنتی برای دریافت، ذخیره‌سازی، مدیریت و تحلیل خارج است.

موسسه ملی استاندارد و فناوری

مجموعه داده‌هایی هستند که در آن‌ها حجم داده‌ها، سرعت تولید و نمایش داده‌ها، توانایی ما را برای تحلیل کارآمد و موثر با استفاده از روش‌های سنتی محدود می‌سازد و برای تجزیه و تحلیل و پردازش کارآمد این داده‌ها، استفاده از روش‌های نوین و مقیاس‌پذیر ضروری می‌باشد.

تعریف ساده[۲]

کلان داده‌ها به داده‌هایی اطلاق می‌گردد که نتوان آن‌ها را توسط یک سیستم عادی ذخیره‌سازی و پردازش کرد. بدین معنی که یک سیستم عادی توان پردازشی و ذخیره‌سازی لازم برای تجزیه و تحلیل این داده‌ها را نداشته باشد.

۱-۲-۱ ویژگی‌های کلان داده‌ها

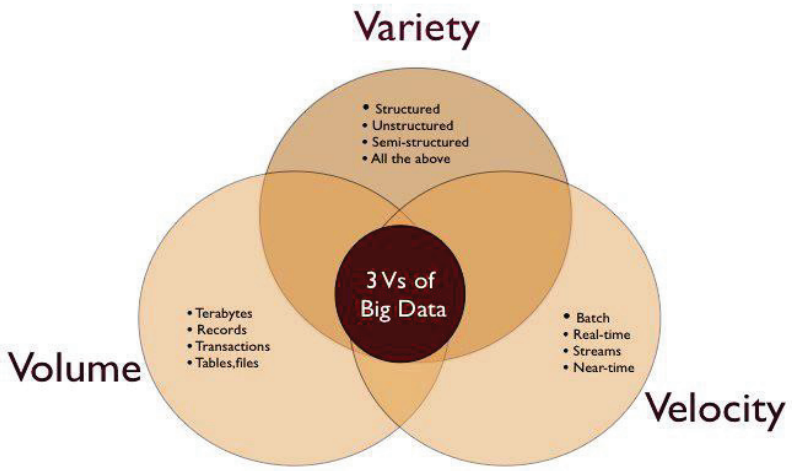
حال این سوال پیش می‌آید که کلان داده‌ها چه ویژگی‌هایی دارند و ما چه مجموعه از داده‌ها را کلان داده می‌نامیم؟ آیا تمام داده‌های دنیای امروز کلان داده محسوب می‌شوند؟ باید بگوییم خیر؛ اینگونه نیست و دانشمندان داده، مجموعه ای از ویژگی‌ها را برای کلان داده‌ها در نظر گرفته‌اند که اگر مجموعه داده ما چنین ویژگی‌هایی را دارا باشد می‌توان آن را کلان داده در نظر گرفت. برای کلان داده سه ویژگی اصلی شامل حجم، تنوع و نرخ تولید در نظر گرفته می‌شود [۱]. در دنیای کلان داده‌ها، این سه ویژگی به یک استاندارد

---

[۱] McKinsey

[۲] این تعریف توسط نویسندگان برداشت شده است.

تبدیل شده است (شکل ۱-۱). ویژگی‌های دیگری همچون صحت[۱]، اعتبار[۲]، بصری‌سازی[۳]، نوسان[۴]، ارزش[۵] و آسیب‌پذیری[۶] هم وجود دارند [۲] که جنبه کاربردی آن‌ها بیشتر مربوط به سازمان‌های خاص می‌باشد و کمتر مورد توجه و اهمیت عام قرار می‌گیرند. در ادامه، به توصیف هر یک از این ویژگی‌ها می‌پردازیم.



شکل ۱-۱ ویژگی‌های اصلی کلان داده‌ها

حجم

حجم یکی از ویژگی‌های بسیار مهم کلان داده‌هاست؛ به ویژه زمانی که گفته می‌شود ۹۰ درصد داده‌های دنیا در دو سال اخیر تولید شده‌اند. اما حجم در کلان داده مفهوم مبهمی محسوب می‌شود و در واقع نمی‌توان چندان به ویژگی حجم برای تشخیص کلان داده بودن یک مجموعه اکتفا نمود. در واقع ما نمی‌دانیم که دقیقاً چه حجمی را باید به عنوان حجم پایه برای کلان داده در نظر بگیریم. به عنوان مثال اگر یک ترابایت داده را حجم پایه‌ای انتخاب نماییم، آیا ۹۹۰ گیگابایت داده، کلان داده محسوب نمی‌شود؟ اگرچه گفته می‌شود که

---

[۱] Veracity
[۲] Validity
[۳] Visualization
[۴] Volatility
[۵] Value
[۶] Vulnerability

حجم کلان داده‌ها در دنیای امروز در حدود ترابایت و پتابایت می‌باشد ولی ما در مورد ویژگی حجم تنها به این نکته بسنده می‌کنیم: *مجموعه داده‌ای کلان داده است که حجم آن توسط پایگاه داده‌های سنتی قابل مدیریت نباشد*. حجم به دو عامل سرعت رشد داده و سرعت توسعه دیسک‌های ذخیره‌سازی بستگی دارد.

تنوع

تنوع باعث می‌شود که کلان داده‌ها واقعا کلان شوند. تنوع در کلان داده‌ها به این معنی است که مجموعه داده‌های موجود در کلان داده شامل انواع داده‌های ساختاریافته[۱] (داده‌هایی که در پایگاه داده‌های رابطه‌ای ذخیره می‌شوند؛ مانند مجموعه نمرات دانشجویان یک کلاس)، داده‌های نیمه‌ساختاریافته[۲] (داده‌هایی که دارای قالب نسبتاً مشخصی هستند ولی امکان دارد که خصوصیت‌های مختلفی داشته باشند؛ مانند فرمت‌های JSON و XML) و داده‌های غیرساختاریافته[۳] (داده‌هایی که هیچ‌گونه ساختار منظمی در آن‌ها وجود ندارد؛ مانند تصویر، فایل‌های چندرسانه‌ای و داده‌های جمع آوری شده از شبکه‌های اجتماعی) است. منابع تولید داده می‌تواند هم ماشین باشد و هم انسان.

نرخ تولید

نرخ تولید یا سرعت تولید تداعی‌کننده این امر است که مجموعه داده‌ها در کلان داده در مدت زمانی کم به میزان بالایی افزایش یابد؛ یعنی منابع تولید داده با سرعت بالایی داده تولید می‌کنند و حجم داده با آهنگ بالایی رشد می‌نماید. داده‌های بلادرنگ[۴] نیز در این دسته قرار می‌گیرند؛ چون در مورد داده‌های بلادرنگ، سرعت تولید و پردازش داده اهمیت بالاتری از حجم داده دارد. به عنوان مثال شرکت فیسبوک روزانه نزدیک به ۵۰۰ میلیون تصویر را در پایگاه داده‌های خود ذخیره می‌نماید و حجم پایگاه داده‌ی آن با یک آهنگ نمائی رشد می‌کند.

---

۱ Structured
۲ Semi-structured
۳ Unstructured
۴ Real-time