داده‌های بزرگ
ویژگی‌ها و کاربردها

مجید حبیبی *
mhabibi@noornet.net

* کارشناس ارشد نرم‌افزار کامپیوتر.

اشاره

در این مقاله، بهترین دستاوردهای داده‌های بزرگ را مورد بررسی قرار داده و سپس، مصارف عمومی داده‌های بزرگ را معرفی کرده، فناوری‌های مربوطه را بررسی می‌نماییم؛ مانند: محاسبات ابری، اینترنت اشیا، مراکز داده‌ای و هادوپ. آنگاه بر چهار فاز زنجیره ارزشی داده‌های بزرگ تمرکز می‌کنیم؛ همچون: تولید داده، اکتساب داده، ذخیره داده و تحلیل داده. پس‌ازاین، چالش‌های فنی را در این زمینه مورد بحث قرار داده، آخرین پیشرفت‌ها را بررسی کرده در نهایت نیز به چندین نمونه از برنامه‌های کاربردی داده‌های بزرگ اشاره می‌نماییم؛ از جمله، مدیریت تجاری، اینترنت اشیا، شبکه‌های اجتماعی آنلاین که همگی می‌توانند در مباحث علوم انسانی و اسلامی مورد استفاده قرار گیرند یا دست‌مایه پژوهش برای دست‌اندرکاران مهندسان و متخصصان این حوزه قرار گیرد.

منبع اصلی نوشتار حاضر، مقاله «BigData» نوشته: Min Chen، Shiwen Mao و Yun-hao Liu می‌باشد که در وب‌گاه Springer (ناشر بین‌المللی منابع علمی، فنی و پزشکی) منتشر گردیده است. گفتنی است، نگارنده علاوه بر ترجمه مقاله مذکور، جهت روشن‌شدن بحث، نکات و مطالب لازم را نیز به متن افزوده است.

کلیدواژگان: داده‌های بزرگ، محاسبات ابری، اینترنت اشیا، مرکز داده، هادوپ، تحلیل‌های داده‌های بزرگ.

مقدمه

امروزه داده‌های حجیم در مرکزِ توجه علوم مدرن و کسب‌وکار است. این داده‌ها از تراکنش‌های آنلاین، ایمیل‌ها، ویدیوها، صوت‌ها، متون، اسناد، تصاویر، کلیک بر روی لینک‌ها، گزارش خطاها، پست‌ها، گزارش‌های جست‌وجو، رکوردهای اطلاعات سلامت، عملیات متقابل در شبکه‌های اجتماعی، داده‌های علمی، حسگرها، سوابق جزئیات تماس، اطلاعات اعتباری و شخصی، شناسایی فرکانس رادیویی، تلفن‌های همراه و نرم‌افزارهای نصب‌شده روی تلفن‌های همراه تولید می‌شوند.

برای شفاف‌شدن بیشتر، به چند مثال علمی در این خصوص اشاره می‌کنیم؛ حجم اطلاعاتی که تا سال ۲۰۰۳ توسط انسان ایجاد شد، تنها ۵ اگزابایت (10^{18}بایت) است؛ اما امروزه این حجم از اطلاعات، تنها در عرض دو روز ایجاد می‌شود. IBM در تحقیقی نشان داد هر روز ۲/۵ اگزابایت داده تولید می‌شود که حدود ٪۹۰ داده‌های موجود، فقط در دو سال اخیر تولید شده است. شرکتی مانند گوگل، بیلیون‌ها سِرور در سطح جهان دارد. حدود ۶ بیلیون مشترک تلفن همراه در جهان همه روزه ۱۰میلیون پیام متنی ارسال و دریافت می‌کنند و تا سال ۲۰۲۰ حدود ۵۰ بیلیون وسیله متصل به اینترنت و شبکه وجود

خواهد داشت. از سال ۲۰۱۲، داده‌های حجیم، به عنوان یک پروژه مهم و جهانی مطرح شد؛ پروژه‌هایی که به جمع‌آوری، بصری‌سازی و آنالیز مقدار زیادی داده می‌پردازند.

تعریف و ویژگی‌های داده‌های بزرگ

برای داده‌های حجیم، تعاریف مختلفی ارائه شده است. داده‌های حجیم را می‌توان داده‌هایی که پردازش آنها خارج از حدّ توان سیستم‌های کنونی است، تعریف نمود و یا داده‌های حجیم را افزایش حجم داده دانست؛ به‌گونه‌ای‌که ذخیره، پردازش و آنالیز آن از طریق فناوری‌های قدیمی دیتابیس‌ها به‌سختی ممکن باشد.

داده‌های بزرگ، شامل یک مفهوم انتزاعی است؛ یعنی مجموعه داده‌هایی که قابل مشاهده، اکتسابی، مدیریت‌شده و فرآوری‌شده توسط IT سنّتی و ابزارهای نرم‌افزاری و سخت‌افزاری نباشند.

Apache Hadoop، داده‌های بزرگ را به عنوان مجموعه داده‌هایی تعریف کرده که نمی‌تواند -cap tured یا گرفته‌شده مدیریت شوند و یا توسط رایانه‌های عمومی به طور قابل قبولی، پردازش شوند.

McKinsey & Company، یک آژانس مشاوره جهانی است که داده‌های بزرگ را به عنوان مرز بعدی نوآوری، رقابت و بهره‌وری بیان کرده است. ازاین‌رو، می‌توان گفت داده‌های بزرگ باید به معنای چنین مجموعه داده‌ای باشد که نمی‌توانند اکتساب‌شده ذخیره شوند و یا توسط نرم‌افزار بانک اطلاعاتیِ کلاسیک مدیریت شوند. این تعریف، شامل دو شرط است: اوّل، حجم پایگاه داده که با استانداردِ داده‌های بزرگ مطابقت دارد، در حال تغییر است و می‌تواند در طول زمان و یا با پیشرفت‌های فناورانه، رشد کند. دوم، نوع پایگاه داده که با استاندارد داده‌های بزرگ فرق دارند.

برای نمونه، به چند مثال کاربردی اشاره می‌کنیم؛ یک فروشگاه اینترنتی را در نظر بگیرید که می‌خواهد با بررسی کلیک بازدیدکنندگان بر روی لینک‌های مختلف وب‌سایت، به علایق و سلیقه مشتریان پی برده، بدین سان، کسب‌وکار خود را بهبود بخشد. یا دولت یک کشور می‌تواند با رصد شبکه‌های اجتماعی، به مقابله و پیشگیری از ناهنجاری‌های جامعه بپردازد. همچنین، در مثال دیگر، فرض نمایید برای انجام یک کار ضروری یا مسافرت از منزل خارج شده‌اید؛ در طول مسیر متوجه می‌شوید که یکی از دستگاه‌های موجود در خانه خاموش نشده است. در این صورت، به جای نگرانی یا بازگشت به منزل، کافی است با مراجعه به رایانه شخصی یا تلفن همراه خویش، وارد فضای شبیه‌سازی‌شده منزل خود شوید و دستگاه مذکور را تنظیم یا خاموش نمایید.

داگ لونی، یک تحلیلگر IT، چالش‌ها و فرصت‌هایی را که توسط داده‌های افزایش یافته، با مدل 3V تعریف می‌کند. در این مدل، سه معیار وجود دارد:

ـ «حجم» که با تولید و جمع‌آوری توده‌ای از داده‌ها، مقیاس داده به طور فزاینده‌ای بزرگ می‌شود؛

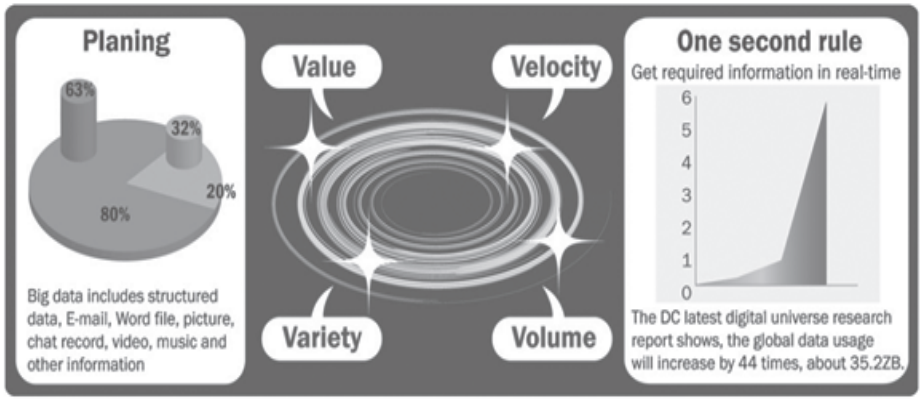
ـ «سرعت»، یعنی خط زمان داده‌های بزرگ که با جمع‌آوری و تحلیل داده‌ها، باید به طور زمانی انجام شود. پس، جریان داده به داده سازمان‌یافته تبدیل می‌شود تا ارزش آن به حداکثر برسد.

ـ «تنوع» که انوع مختلف داده‌ها را نشان می‌دهند و شامل داده‌های غیرساخت‌یافته، نیمه‌ساخت‌یافته و ساخت‌یافته است؛ مانند: صدا، تصویر، صفحات وب و متن. داده‌های ساخت‌یافته در انباره‌های داده، تگ خورده و به‌آسانی ذخیره می‌شوند؛ اما داده‌های غیرساخت‌یافته به صورت تصادفی بوده، آنالیز آنها دشوار است. داده‌های نیمه‌ساخت‌یافته نیز از فیلدهای ثابتی تشکیل نشده‌اند؛ اما تگ‌هایی برای جداسازی عناصر داده دارند.

از آنچه گفته شد، می‌توان به این نکته اشاره داشت که مشخصه‌های داده‌های بزرگ می‌تواند در چهار V خلاصه شود: Volume یا حجم (حجم بالا)، Variety یا تنوع (روش‌های متفاوت)، Veloc-ity یا شتاب و سرعت (تولید سریع) و Value یا مقدار(مقدار بسیار زیاد؛ اما تراکم پایین). (شکل ۱)

چالش‌های داده‌های بزرگ

کاربردهای متفاوت داده‌های بزرگ، می‌تواند بر اساس این فناوری‌های خلاقانه یا پلتفرم‌ها، توسعه



«**شکل ۱**»

یابند؛ اما در این میان، موانع بسیاری در راه توسعه کاربردها و برنامه‌های داده‌های بزرگ وجود دارد که به قرار ذیل‌اند:

ـ **نمایش داده‌ها**: بسیاری از داده‌ها در نوع، ساختار، معناشناسی، دانه‌دانه‌بودن و قابلیت دسترسی، ناهمگن هستند. ازاین‌رو، نمایش داده‌های ناهمگن، بر حجم داده‌ها تأثیر می‌گذارد و حتی مانع تحلیل مؤثر داده‌ها نیز می‌شود. ازاین‌رو، باید گفت نمایش داده‌های مناسب و همگن، باید بر ساختار، کلاس، نوع و همچنین فناوری‌های یکپارچه داده‌ها تأثیرگذار باشد؛ به‌طوری‌که عملیات کارآمدسازی را در مجموعه داده‌های مختلف مقدور سازد.

ـ **کاهش افزونگی و فشرده‌سازی داده‌ها**: عموماً، داده‌ها دارای یک سطح بالایی از افزونگی هستند. کاهش افزونگی و فشرده‌سازی داده‌ها، برای کاهش هزینه غیرمستقیم در کلّ سیستم مؤثر است؛ برای مثال، بیشتر داده‌هایی که با شبکه‌های حسگر تولید می‌شوند، به‌شدت افزونه دارند که می‌توانند فیلتر شده، به ترتیب بزرگی فشرده شوند.

ـ **مکانیزم تحلیلی**: سیستم تحلیلی داده‌های بزرگ باید توده‌هایی از داده‌های ناهمگن را در یک زمان محدود پردازش نمایند. سیستم مدیریت پایگاه داده رابطه‌ای یا RDBMS، غیرقابل گسترش طراحی شدند؛ اما پایگاه داده غیررابطه‌ای، مزیت‌های منحصربه‌فردی را در پردازش داده‌های ساختاربندی‌شده نشان داده است.

ـ **محرمانگی داده‌ها**: صاحبان داده‌های بزرگ، در حال حاضر نمی‌توانند چنین مجموعه داده‌ای بزرگی را به دلیل ظرفیت محدودشان تحلیل کنند. آنها باید به حرفه‌ای‌ها و ابزارهای پیشرفته برای تحلیل چنین داده‌هایی متکی باشند که خطرات بالقوه‌ای را افزایش می‌دهد و امنیت داده‌ها را مخدوش می‌کند.

ـ **مدیریت انرژی**: مصرف انرژی سیستم محاسباتی یک پردازنده، از نظر اقتصادی بسیار مهم است. با افزایش حجم داده‌ها و تقاضاهای تحلیلی، پردازش، ذخیره‌سازی و انتقال داده‌های بزرگ، به‌ناچار انرژی

الکتریکی بسیاری مصرف می‌شود. ازاین‌رو، کنترل مصرف برق سیستم باید برای داده‌های بزرگ انجام شود تا قدرت گسترش و قابلیت دسترسی، تضمین شود.

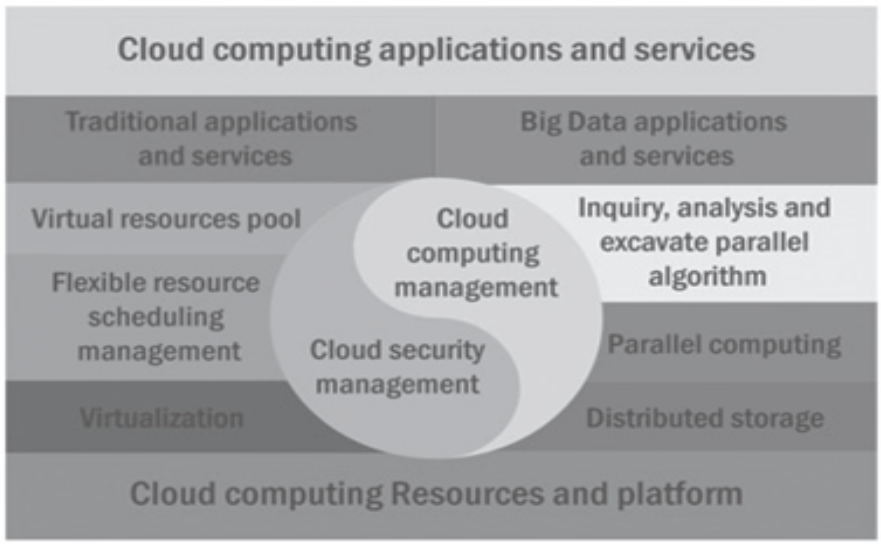
ـ **همکاری و تعاون**: یک معماری شبکه داده‌های بزرگ و جامع، باید برای کمک به پژوهشگران و مهندسان در زمینه‌های مختلف ایجاد شود که بتوانند به انوع مختلفی از داده‌ها دسترسی پیدا کرده، از اطلاعات و دانش خودشان استفاده کنند؛ به‌طوری‌که برای تکمیل اهداف تحلیلی با یکدیگر همکاری نمایند.

فناوری‌های داده‌های بزرگ

الف. ارتباط بین محاسبات ابری و داده‌های بزرگ

رایانش ابری، نوعی فناوری قدرتمند برای اجرای محاسبات پیچیده و سنگین است که نیاز به استفاده از سخت‌افزارهای گران را حذف نموده، فضای محاسباتی و نرم‌افزار مورد نیاز را در اختیار کاربر قرار می‌دهد. همچنین، رایانش ابری، یک الگوی جدید زیرساخت محاسباتی محسوب می‌شود که روشی مناسب برای پردازش داده‌های حجیم در ابر فراهم می‌آورد و توسط همه انواع منابعِ در دسترس، قابل استفاده است.

محاسبات ابری، ارتباط نزدیکی با داده‌های بزرگ دارند؛ جزئیات کلیدی محاسبات ابری، در شکل «۲» نشان داده شده‌اند. داده‌های بزرگ، هدف عملیات محاسبه فشرده داده‌هاست و بر ظرفیت ذخیره‌سازی یک سیستم ابری تأکید دارد؛ درحالی‌که هدف اصلی محاسبات ابری، استفاده از محاسبات بزرگ و ذخیره منابع تحت مدیریت متمرکز می‌باشد؛ به‌طوری‌که برنامه‌های داده‌های بزرگ را با ظرفیت محاسباتی

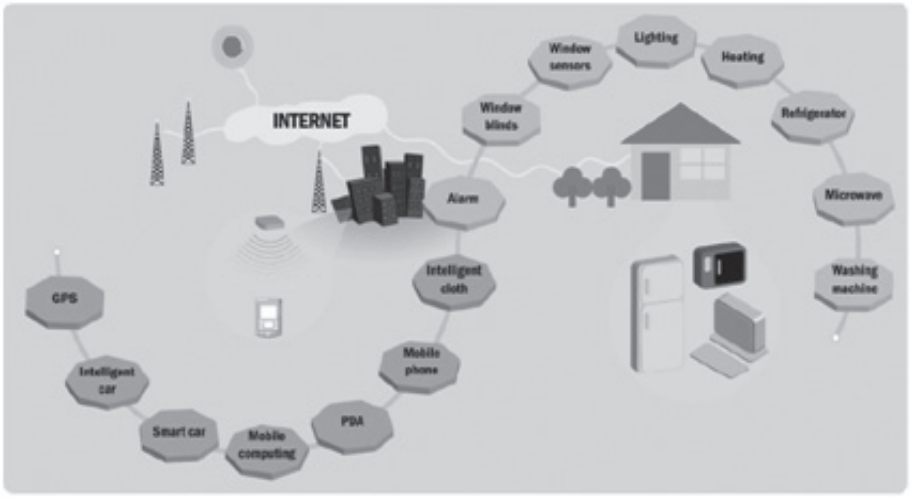


«**شکل ۲**»

جزء، ارائه می‌کند. توسعه محاسبات ابری، راه‌حل‌هایی را برای ذخیره‌سازی و پردازش داده‌های بزرگ ارائه می‌نماید. از سوی دیگر، ظهور داده‌های بزرگ نیز به توسعه محاسبات ابری سرعت می‌بخشد. حتی اگر هر دو مفهوم محاسبات ابری و داده‌های بزرگ در جهات زیادی با یکدیگر هم‌پوشانی داشته باشند، ولی از دو منظر، دارای تفاوت‌هایی هستند: اوّل، محاسبات ابری، معماری IT را تغییر می‌دهد؛ درحالی‌که داده‌های بزرگ بر تصمیم‌گیری تجاری تأثیر می‌گذارند؛ هرچند داده‌های بزرگ به محاسبات ابری وابسته هستند. دوم، داده‌های بزرگ و محاسبات ابری، مشتریان متفاوتی دارند. محاسبات ابری،

فناوری محصول مدیر ارشد فناوری اطلاعات را به عنوان یک راه‌حل IT پیشرفته بررسی می‌کند؛ اما داده‌های بزرگ، هدف مدیر عاملی است که بر عملیات‌های تجاری متمرکز است.

با پیشرفت داده‌های بزرگ و محاسبات ابری، این دو فناوری قطعاً و به طور فزاینده با همدیگر در هم



«**شکل ۳**»

پیچیده‌اند. محاسبات ابری، با توابعی شبیه به آنهایی که در رایانه‌ها و سیستم‌عامل‌ها هستند، منابع سیستم را ارائه می‌کنند و داده‌های بزرگ در سطح بالاتری که توسط محاسبات ابری پشتیبانی می‌شود، عمل می‌کنند و توابعی را شبیه آنهایی که در پایگاه داده‌ها و ظرفیت پردازش داده‌های کارآمد هستند، ارائه می‌کنند. ازاین‌رو، می‌توان گفت برنامه‌های کاربردی داده‌های بزرگ باید بر اساس محاسبات ابری باشد. پس، محاسبات ابری، فقط محاسبات و پردازش داده‌های بزرگ نیست؛ بلکه به‌خودی‌خود نوعی خدمت است که توسعه داده‌های بزرگ را نیز ارتقا می‌دهد.

ب. رابطه بین اینترنت اشیا (IoT) و داده‌های بزرگ

در مبحث IoT، تعداد بسیاری حسگر شبکه‌ای، در دستگاه‌های مختلف سخت‌افزاری تعبیه شده است. چنین حسگرهایی که در زمینه‌های مختلف گسترش یافته‌اند، می‌توانند انوع مختلفی از داده‌ها را جمع‌آوری کنند؛ مانند: داده‌های محیطی، داده‌های جغرافیایی، داده‌های نجومی و داده‌های منطقی یا رقمی. ازاین‌رو، تجهیزات تلفن همراه، وسایل حمل و نقل و کاربردهای خانگی، همگی می‌توانند تجهیزات اکتساب داده در IoT باشند.

بر اساس فرآیندهای اکتساب داده‌ها و انتقال در IoT، معماری شبکه آن می‌تواند در سه لایه تقسیم شود: لایه سنجش، لایه شبکه و لایه کاربردی. لایه سنجش، مسئول کسب داده‌هاست و اساساً شامل شبکه‌های حسگر می‌باشد. لایه شبکه، مسئول انتقال اطلاعات و پردازش آنهاست که انتقال نزدیک، می‌تواند متکی بر شبکه‌های حسگر بوده و انتقال از راه دور باید وابسته به اینترنت باشد. در نهایت، لایه کاربردی، کاربردهای خاص IoT را پشتیبانی می‌کند.

ج. رابطه بین DataCenter و داده‌های بزرگ

در بحث داده‌های بزرگ، مرکز داده، نه فقط یک پلتفرم برای ذخیره‌سازی متمرکز داده‌هاست، بلکه مسئولیت‌های بیشتری را به عهده دارد؛ مانند: اکتساب داده‌ها، مدیریت داده‌ها و سازمان‌دهی داده‌ها.

برای داده‌های حجیم، تعاریف مختلفی ارائه شده است. داده‌های حجیم را می‌توان داده‌هایی که پردازش آنها خارج از حدّ توان سیستم‌های کنونی است، تعریف نمود و یا داده‌های حجیم را افزایش حجم داده دانست؛ به‌گونه‌ای‌که ذخیره، پردازش و آنالیز آن از طریق فناوری‌های قدیمی دیتابیس‌ها به‌سختی ممکن باشد

داده‌های بزرگ، الزام‌های سخت‌گیرانه‌ای در ذخیره‌سازی، پردازش و همچنین انتقال شبکه‌ای دارند. شرکت‌ها باید توسعه مراکز داده‌ای را برای بهبود ظرفیت سریع و مؤثر پردازش داده‌های بزرگ تحت نرخ عملکرد/ قیمت محدود، مورد بررسی قرار دهند. مرکز داده‌ای باید یک زیرساخت با تعداد زیادی گره، یک شبکه داخلی سرعت بالا، دفع مؤثر گرما و داده پشتیبان مؤثر، ارائه کند. فقط زمانی که یک مرکز داده‌ای با کارآمدی بالا، ثابت، امن و گسترش‌پذیر ساخته شود، می‌تواند عملیات عادی کاربردهای داده‌های بزرگ را تضمین نماید.

در بحث داده‌های بزرگ، مرکز داده‌ای نباید فقط با تجهیزات سخت‌افزاری بررسی شود؛ بلکه باید ظرفیت‌های نرم‌افزاری را نیز تقویت کند؛ مانند: ظرفیت‌های اکتساب، پردازش، تجزیه و تحلیل و برنامه‌های داده‌های بزرگ.

رابطه بین هادوپ (Hadoop) و داده‌های بزرگ

در حال حاضر، هادوپ، به طور گسترده در کاربردها و برنامه‌های داده‌های بزرگ در صنعت استفاده می‌شود؛ مانند فیلترینگ اسپم یا هرزنامه، جست‌وجوی شبکه‌ای و تحلیل کلیک‌ها. در واقع،Hadoop یک فناوری نرم‌افزاری رایگان است که مختص پردازش حجم انبوهی از اطلاعات بر روی ذخیره سازها و سِروِرها، طراحی گردیده است.

منابع تولید داده‌های بزرگ

ـ **مدیای اجتماعی:** اطلاعاتی است که از طریق به‌اشتراک‌گذاری و یا تبادل اطلاعات توسط نشانی‌های اینترنتی و یا ارتباطات مجازی و شبکه‌های مجازی به دست می‌آیند؛ نظیر اطلاعاتی که در پروژه‌های اشتراکی، بلاگ‌ها، میکروبلاگ‌ها، فیسبوک و توییتر تولید می‌شوند.

ـ **داده‌های ماشین:** اطلاعاتی است که به صورت خودکار توسط سخت‌افزار و نرم‌افزارهای ابزارهایی نظیر: رایانه‌ها، وسایل پزشکی یا دیگر ماشین‌ها بدون دخالت انسان تولید می‌گردد.

ـ **حسگرها:** وسایل حسگر مختلفی برای اندازه‌گیری کمیت‌های فیزیکی و تبدیل آنها به سیگنال وجود دارد که بخشی از داده‌های حجیم را تولید می‌نماید.

ـ **اینترنت اشیا:** IoT، مجموعه‌ای است از اشیا که به صورت یکتا قابل تعریف هستند و به عنوان بخشی از اینترنت می‌باشند. این اشیا، شامل: تلفن‌های کوچک، دوربین‌های دیجیتال و تبلت‌ها هستند. وقتی این وسایل از طریق اینترنت به یکدیگر متصل می‌شوند، قادرند بیشتر پردازش‌های کوچک

امروزه سازمان‌ها و شرکت‌ها، از جمله مراکز و سازمان‌هایی که در زمینه علوم اسلامی فعالیت دارند، می‌توانند از داده‌های حجیم خود استفاده‌های گوناگونی ببرند. مرکز تحقیقات کامپیوتری علوم اسلامی، در شمار مراکزی است که بعد از گذشت نزدیک به سه دهه از تحقیقات نرم‌افزاری خود، حجم عظیمی از داده‌های علوم و معارف اسلامی را در اختیار دارد؛ گنجینه‌ای که پاسخگوی نیاز جامعه علمی ایران و دستمایه‌ای بزرگ برای تولید علم در کشور به شمار می‌رود

و سرویس‌های پشتیبانی اقتصادی، محیطی و سلامت را فراهم آورند. تعداد زیاد وسایل متصل به اینترنت، انواع مختلفی از سرویس‌ها را فراهم می‌آورند و مقادیر زیادی داده و اطلاعات تولید می‌نمایند.

تولید و اکتساب داده‌های بزرگ

اگر داده‌ها را به عنوان مواد اوّلیه در نظر بگیریم، تولید و کسب داده‌ها، یک فرآیند بهره‌بردای از این مواد خام هستند و ذخیره داده‌ها یک فرآیند ذخیره‌سازی محسوب شده، تجزیه تحلیل داده‌ها یک فرآیند تولید می‌باشد که مواد اوّلیه را برای تولید محصول جدید به‌کار می‌بندد. بنابراین، می‌توان گفت زنجیره مقادیر داده‌های بزرگ، به طور کلی، در چهار فاز دسته‌بندی می‌شود: تولید داده، کسب داده، ذخیره داده و تجزیه و تحلیل داده.

ـ **تولید داده:** اوّلین مرحله داده‌های بزرگ است؛ به عنوان مثال، داده‌های اینترنتی را درنظر بگیرید که در آن حجم عظیمی از داده‌ها، مانند: ورودی‌های جست‌وجو، پست‌های انجمنی، رکوردها و سوابق چت و پیام‌های میکروبلاگ، تولید می‌شوند. این داده‌ها مربوط به زندگی روزمره مردم می‌باشد و ویژگی‌های مشابهی با مقدار بالا و تراکم پایین دارد. چنین داده‌های اینترنتی، ممکن است به‌تنهایی بی‌ارزش باشند؛ اما در بهره‌برداری متراکم داده‌های بزرگ، اطلاعات مفیدی مانند عادات و سرگرمی‌های کاربران را می‌توان از آنها به دست آورد و حتی ممکن است حالات احساسی و رفتاری کاربران را پیش‌بینی کرد.

ـ **جمع‌آوری داده‌ها:** وقتی کسی به طور گسترده از روش جمع‌آوری داده‌ها استفاده می‌کند، logfiles (فایل‌های ثبت رخداد)، فایل‌هایی را که به طور خودکار با سیستم منبع داده تولید شده، ثبت نموده، سپس، فعالیت‌های انجام‌گرفته را در فرمت‌های تعیین‌شده برای تحلیل‌های بعدی ثبت می‌نماید؛ برای مثال، تعداد کلیک‌ها، دفعات بازدید از وب‌سایت‌ها و دیگر سوابق کاربران وب را ثبت می‌کند. برای ضبط فعالیت‌های کاربران در وب‌سایت‌ها، سرورهای وب اساساً شامل سه فرمت فایل ثبت رخداد می‌باشد: فرمت فایل ثبت رخداد عمومی (NCSA)، فرمت ثبت گسترده (W3C) و فرمت ثبت ISS (Microsoft). هر سه نوع فایل‌های ثبت رخداد، در فرمت متن اَسکی (ASCII) هستند.

ـ **سنجش حس‌کردن یا Sensing:** حسگرها معمولاً در زندگی روزمره برای اندازه‌گیری کمیت‌های فیزیکی و تبدیل کمیت‌های فیزیکی به سیگنال‌های دیجیتالیِ قابل خواندن، رایج می‌باشند. داده‌های حسی، می‌توانند به شکل‌هایی چون: امواج صوتی، صدا، لرزش یا ویبره، شیمیایی، جریان، آب و هوا، فشار و یا دما دسته‌بندی شوند. اطلاعات این حسگرها توسط شبکه‌های سیمی و بی‌سیم، به یک نقطه جمع‌آوری داده انتقال می‌یابد.

ـ روش‌های اکتساب داده‌های شبکه: در حال حاضر، اکتساب داده‌های شبکه، با استفاده از Web crawler یا خزنده وب انجام شده است. Web crawler، برنامه‌ای است که با موتورهای جست‌وجو برای دانلود و ذخیره‌سازی صفحات وب استفاده می‌شود.

ـ تجهیزات موبایل: امروزه، دستگاه‌های موبایل به طور گسترده‌تری استفاده می‌شوند. همان‌طورکه توابع و عملکردهای موبایل به طور فزاینده‌ای قوی‌تر می‌شوند، پیچیدگی بیشتر و امکانات متعددتر و همچنین، تنوع بیشتر داده‌ها را نمایان می‌کنند. دستیابی به اطلاعات موقعیت جغرافیایی از طریق سیستم‌های موقعیت‌یاب، دسترسی به اطلاعات صوتی از طریق میکروفن، دستیابی به تصاویر، فیلم‌ها، نماهای خیابان، بارکدهای دوبعدی و دیگر اطلاعات چندرسانه‌ای از طریق دوربین، دسترسی به حرکات کاربر مانند لمس صفحه نمایش توسط انگشت، از ویژگی‌های این دستگاه‌هاست؛ برای مثال، iPhone به‌خودی‌خود یک جاسوس موبایل است و می‌تواند داده‌های بی‌سیم و اطلاعات موقعیت جغرافیایی را جمع‌آوری کند و سپس، آنها را به شرکت Apple، برای پردازش ارسال نماید که البته معمولاً کاربران

از این مسئله بی‌خبرند. همچنین، سیستم‌های‌عامل تلفن‌های هوشمند مانند اندروید گوگل و ویندوز فون مایکروسافت نیز می‌توانند اطلاعات را با همین روش جمع‌آوری کنند.

ـ ذخیره داده‌ها: مکانیزم‌های ذخیره‌سازی داده‌های بزرگ را می‌توان در سه سطح: سیستم‌های فایل، پایگاه‌های داده و مدل‌های برنامه‌نویسی دسته‌بندی نمود. به محض تکمیل جمع‌آوری داده‌های خام، داده‌ها به یک زیرساخت ذخیره‌سازی برای پردازش و تجزیه و تحلیل انتقال می‌یابند که این انتقال می‌تواند در مرکز داده رخ دهد.

ـ تجزیه و تحلیل داده‌ها: به دلیل تنوع وسیع منابع داده‌ای، مجموعه داده‌های جمع‌آوری‌شده با توجه به وجود نویز و یا افزونگی، متفاوت هستند و بی‌شک، ذخیره این گونه داده‌ها، بی‌معناست. به علاوه، بعضی روش‌های تحلیلی، الزام‌های جدی در کیفیت داده‌ها دارند. ازاین‌رو، به منظور فعال‌سازی تجزیه و تحلیل مؤثر، باید داده‌ها را تحت شرایطی برای مرتبط‌کردن آنها با هم، از منابع مختلف پیش‌پردازش کنیم که این کار، فقط هزینه‌های ذخیره‌سازی را کاهش نمی‌دهد؛ بلکه دقت تحلیل را نیز بهبود می‌بخشد. از فنون پیش‌پردازش داده‌ای می‌توان به: یکپارچه‌سازی داده‌ها، پاک‌سازی داده‌ها و حذف افزونگی داده‌ها اشاره نمود.

عملکرد روش‌های رمزگذاری در داده‌های کوچک و متوسط، نمی‌تواند تقاضای داده‌های بزرگ را برآورده کند؛ چراکه باید روش‌های رمزنگاری داده‌های بزرگ کارآمد شود و توسعه یابد. ازاین‌رو، در این خصوص باید طرح‌های مؤثر مدیریت امنیت، کنترل دسترسی و ارتباطات امن برای داده‌های ساخت‌یافته، نیمه‌ساخت‌یافته و غیرساخت‌یافته، مورد بررسی و تحقیق قرار گیرد

ابزارهای کاوش و تحلیل داده‌های بزرگ

ـ R (30.7%): یک زبان برنامه‌نویسی منبع باز و محیط نرم‌افزاری می‌باشد که برای کاوش، تحلیل و حالت بصری داده‌ها طراحی شده است.

ـ Excel (29.8%) : یک کامپوننت اصلی Microsoft Office می‌باشد که پردازش داده‌ها و قابلیت‌های تحلیل آماری قوی را ارائه می‌کند.

ـ Rapid-I Rapidminer (26.7%) : یک نرم‌افزار منبع باز است که برای داده‌کاوی، ماشین یادگیری و تحلیل‌های پیش‌بینی استفاده می‌شود. از جمله برنامه‌های داده‌کاوی و ماشین یادگیری که توسط رپیدماینر پیاده‌سازی می‌شوند، می‌توان به: استخراج، تبدیل و بارگذاری، پیش‌پردازش و بصری‌سازی، مدل‌سازی و گسترش داده اشاره نمود.

ـKNMINE (21.8%): یک پلتفرم کاربرپسند و هوشمند است که برای یکپارچه‌سازی داده‌های منبع باز غنی، پردازش داده‌ها، تحلیل داده‌ها و داده‌کاوی کاربرد دارد.

ـ Weka: Weka/Pentaho (14.8%) ، یک ماشین یادگیری است که عملیاتی چون: پردازش داده، انتخاب ویژگی، کلاس‌بندی، رگراسیون، کلاسترسازی، قانون مشارکت و تجسم‌سازی را ارائه می‌کند. Pentaho، این نرم‌افزار، شامل یک پلتفرم وب‌سرویس و چندین ابزار برای اموری مانند: پشتیبانی گزارش‌گیری، تحلیل، چارت‌بندی، یکپارچه‌سازی داده و داده‌کاوی می‌باشد.

روش‌های تحلیل داده‌ها در داده‌های بزرگ

۱. تحلیل داده‌های ساخت‌یافته

کاربردهای تجاری و پژوهش‌های علمی می‌توانند داده‌های ساخت‌یافته حجیمی را تولید کنند که مدیریت و تحلیل‌ها بر فناوری‌های تجاری – تبلیغاتی متکی باشند؛ برای مثال، ماشین یادگیری آماری مبتنی بر مدل‌های ریاضی دقیق و الگوریتم‌های قوی، برای کشف وضعیت نامطلوب و کنترل انرژی به‌کار گرفته می‌شود که برگرفته از حفظ حریم شخصی در تجارت الکترونیک و دولت الکترونیک و کاربردهای پزشکی و بهداشتی است.

۲. تحلیل داده‌های متن

رایج‌ترین فرمت ذخیره‌سازی اطلاعات، متن است؛ مانند: ایمیل‌ها، اسناد تجاری، صفحات وب و رسانه اجتماعی. معمولاً تحلیل متن، یک فرآیند برای استخراج اطلاعات و دانش مفید از متن ساختارنیافته می‌باشد. متن‌کاوی به طور خاص، بین رشته‌ای، شامل بازیابی اطلاعات، ماشین یادگیری، آمار و ارقام،

زبان‌شناسی محاسباتی و داده‌کاوی می‌باشد. بیشتر سیستم‌های متن‌کاوی، بر اساس توصیف متن و پردازش زبان طبیعی (NLP) با تأکید بیشتر بر حروف می‌باشد.

۳. تحلیل داده‌های وب

تحلیل داده‌های وب، به عنوان یک رشته پژوهشی فعال ظهور پیدا کرد. هدف این تحلیل، بر بازیابی، استخراج و ارزیابی اطلاعات از اسناد و خدمات وب به طور خودکار استوار است که دانش مفید را کشف می‌کند. تحلیل وب، مربوط به چند رشته پژوهشی می‌باشد؛ از جمله: پایگاه داده، بازیابی اطلاعات، NLP و متن‌کاوی. کاوش محتویات وب، فرآیندی برای کشف دانش مفید در صفحات وب می‌باشد که معمولاً شامل چند نوع از داده‌ها، همچون: متن، تصویر، صوت، ویدئو، کد، فراداده و لینک می‌باشد.

٤. تحلیل داده‌های چندرسانه

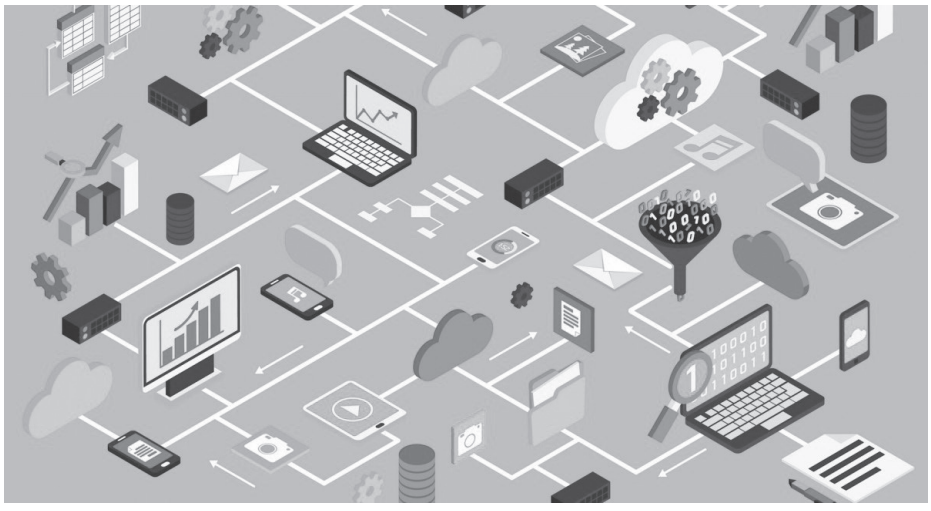
چون داده‌های چندرسانه، ناهمگون و بیشتر شامل اطلاعات غنی‌تر نسبت به داده‌های ساخت‌یافته ساده یا داده‌های متنی هستند، استخراج اطلاعات، با چالش بزرگی از تفاوت‌های معنایی روبه‌رو می‌شود. تحقیق بر تحلیل چندرسانه، بسیاری از رشته‌ها را پوشش می‌دهد. بعضی از اولویت‌های پژوهشی اخیر، عبارت است از: خلاصه‌سازی چندرسانه، حاشیه‌نویسی چندرسانه، نمایه‌سازی و بازیابی چندرسانه‌ای.

۵. تحلیل داده‌های شبکه

تحلیل داده‌های شبکه، از تحلیل مقدار اوّلیه در تحلیل شبکه اجتماعی آنلاین در اوایل قرن ۲۱ میلادی تکامل یافت. بسیاری از خدمات شبکه اجتماعی آنلاین، از قبیل: Twitter, Facebook و Linkedin به طور فزاینده‌ای در این سال‌ها محبوب و مشهور شده‌اند. این شبکه‌های اجتماعی، عموماً شامل داده‌های لینکی حجیم و داده‌های محتوایی می‌باشند. داده‌های لینکی، اساساً در قالب ساختارهای گرافیکی هستند که ارتباطات میان دو موجودیت را توصیف می‌کنند. داده‌های محتوایی شامل: متن، تصویر و دیگر داده‌های چندرسانه‌ای شبکه می‌باشند. محتوای غنی در این شبکه‌ها، چالش‌های بی‌سابقه و نیز فرصت‌هایی برای تحلیل داده‌ها به ارمغان آورده‌اند.

٦. تحلیل داده‌های موبایل

با رشد تعداد کاربران موبایل و بهبود عملکرد آنها، تلفن‌های همراه اکنون برای ساخت یک جامعه کارآمد، مفیدند؛ مانند جامعه‌هایی با موقعیت‌های جغرافیایی و جامعه‌های مبتنی بر پس‌زمینه فرهنگی و علایق. جوامع شبکه سنّتی یا جوامع SNS، در تعامل آنلاین کوتاه‌مدت میان اعضا نقش دارند. این



جوامع، فقط زمانی که اعضا پشت رایانه خود هستند، فعال می‌باشند. در مقابل، تلفن‌های همراه می‌تواند تعامل غنی‌ای در هر زمان و مکان را پشتیبانی نمایند.

کاربردهای مهم داده‌های بزرگ

۱. کاربرد داده‌های بزرگ در سازمان‌ها و شرکت‌ها

کاربرد داده‌های بزرگ در شرکت‌ها، می‌تواند بهره‌وری تولید و رقابت‌پذیری را از جنبه‌های بسیاری بالا ببرد. به طور خاص، در بازاریابی، با تحلیل داده‌های بزرگ، شرکت‌ها می‌توانند با دقت بیشتری رفتار مشتری را پیش‌بینی نمایند و شرایط همکاری بهتری را در تجارت پیدا کنند. شرکت‌ها می‌توانند در برنامه‌ریزی و معرفی طرح‌های فروش، بعد از مقایسه داده‌های حجیم، قیمت کالاهای خودشان را بهینه سازند تا کارآمدی، بهره‌وری و رضایت‌بخشی تجاری، بهینه‌سازی نیروی کار، پیش‌بینی درست و دقیق تخصیص الزام‌های پرسنلی، پرهیز و دوری از ظرفیت تولید اضافی و کاهش هزینه انجام کار را بهبود بخشند. همچنین، این شرکت‌ها می‌توانند در زنجیره تأمین نیازهای اساسی خود و با استفاده از داده‌های بزرگ، بهینه‌سازی موجودی انبار و نیازهای ضروری خویش را برای کاهش توقف بین تأمین و تقاضا، کنترل بودجه و بهبود خدمات اداره کنند.

امروزه سازمان‌ها و شرکت‌ها، از جمله مراکز و سازمان‌هایی که در زمینه علوم اسلامی فعالیت دارند، می‌توانند از داده‌های حجیم خود استفاده‌های گوناگونی ببرند. مرکز تحقیقات کامپیوتری علوم اسلامی، در شمار مراکزی است که بعد از گذشت نزدیک به سه دهه از تحقیقات نرم‌افزاری خود، حجم عظیمی از داده‌های علوم و معارف اسلامی را در اختیار دارد؛ گنجینه‌ای که پاسخگوی نیاز جامعه علمی ایران و دست‌مایه‌ای بزرگ برای تولید علم در کشور به شمار می‌رود.

از دیگر کاربردهای داده‌های بزرگ، تولید محصولات نرم‌افزاری هوشمند است که بر اساس داده‌های موجود یا اطلاعات دریافت‌شده از تعامل با کاربران سامان می‌یابد؛ برای مثال، وقتی یک محقق علوم اسلامی در جست‌وجوی موضوعی خاص است، داده‌های بزرگ در این زمینه او را در دستیابی به مطالب منسجم، جامع و دقیق یاری می‌رسانند و همه منابع و محتواهای معتبر، میان‌رشته‌ای و مرتبط با موضوع را در اختیار او قرار خواهد داد.

۲. کاربرد داده‌های بزرگ مبتنی بر IoT

برای مثال، کامیون‌های UPS به حسگرها، آداپتورهای بی‌سیم و GPS مجهز هستند که دفتر مرکزی می‌تواند موقعیت‌های کامیون را دنبال کند و از خرابی موتور یا مشکلات احتمالی در طول مسیر جلوگیری نماید. در همین حال، این سیستم به UPS کمک می‌کند که کارمندانش را نیز مدیریت و نظارت کند و مسیرهای تحویل را بهینه سازد. مسیرهای تحویل بهینه که مختص به کامیون‌های UPS است، از سابقه تجربه رانندگی آنها استخراج شده است.

همچنین، شهر هوشمند، یک محدوده پژوهشی بر اساس کاربرد داده‌های IoT می‌باشد؛ برای مثال، همکاری پروژه شهر هوشمند بین Miami-Dade در فلوریدا و IBM نزدیک به ۳۵ نوع دپارتمان دولتی را در آنها به هم متصل می‌کند و در نتیجه، دولت، می‌تواند اطلاعات بهتری برای پشتیبانی و تصمیم‌گیری برای مدیریت منابع آبی، کنترل ترافیک و بهبود امنیت عمومی به دست آورد.

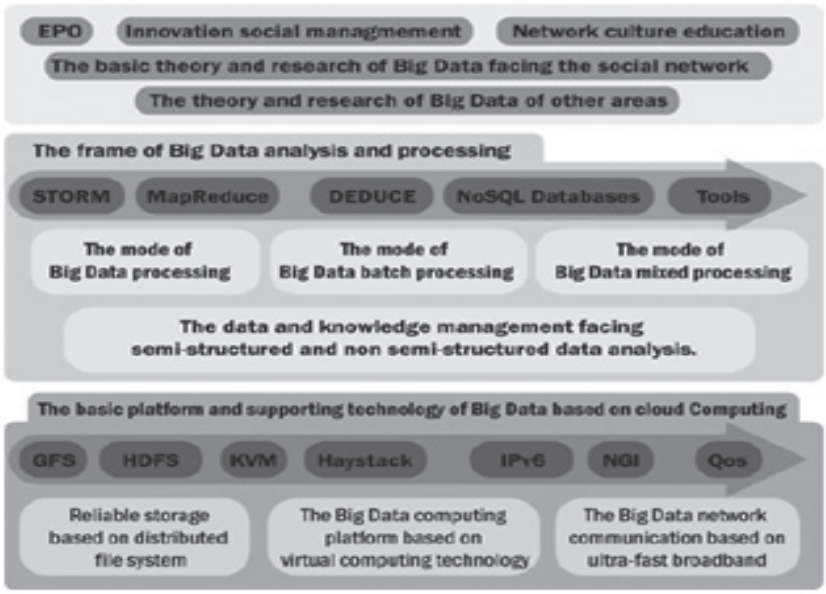
۳. کاربرد داده‌های بزرگ در شبکه‌های اجتماعی آنلاین

SNS آنلاین، یک ساختار اجتماعی تشکیل‌شده از افراد اجتماعی و اتصالات میان افراد بر اساس یک شبکه اطلاعاتی می‌باشد. داده‌های بزرگ SNS آنلاین، از پیام‌های فوری و آنلاین، میکروبلاگ و

یا فضای اشتراک می‌باشد که فعالیت‌های مختلف کاربر را اداره می‌کند. تحلیل‌های داده‌های بزرگ در SNS آنلاین، روش تحلیلی محاسباتی ارائه‌شده را برای فهم روابط جامعه انسانی توسط نظریات و روش‌ها بیان می‌نماید که شامل: ریاضیات، انفورماتیک، جامعه‌شناسی و یا علم مدیریت است و از آنها در سه بُعد: ساختار شبکه، تعامل گروهی و گسترش اطلاعات استفاده می‌کند. این برنامه، شامل توانایی‌هایی همچون: تحلیل هوش شبکه‌ای، بازاریابی اجتماعی، پشتیبانی تصمیم‌گیری‌های دولتی و تحصیل آنلاین می‌باشد. (شکل ۴)

امروزه، فضای مجازی از جمله شبکه‌های اجتماعی، شرایط خوبی را برای گسترش و آموزش علوم و معارف دینی پدید آورده و به‌یقین، داده‌های بزرگ در این فضا می‌تواند بسیاری از علاقه‌مندان این گونه مطالب و مباحث را پاسخ دهد؛ برای مثال، راه‌اندازی پروژه‌های ملی در این زمینه، همچون شبکه ملی اطلاعات و جست‌وجوگر قوی در منابع فارسی فضای مجازی، گامی مؤثر و مفید در این راستا می‌باشد.

امنیت داده‌های بزرگ



«**شکل۴**»

نگران‌کننده‌ترین مسئله دوران کنونی، حریم خصوصی و امنیت اطلاعات می‌باشد. از آنجا که حریم خصوصی برای فرد و انواع داده‌های سازمانی بسیار ضروری است، تبدیل به یک چالش عمده برای کلان‌داده‌ها شده است. تأمین امنیت چنین مجموعه داده‌های بزرگی، از داخل به اندازه بیرون نیز یکی از چالش‌برانگیزترین مسائل کلان داده است. جلوگیری از نشت داده‌ها در هنگام پردازش و دفاع از حملات بیرونی، نیازمند نوعی مدل امنیت داده‌محور قابل اعتماد است. این فناوری، همچنین باید از تهدیدات امنیتی که ممکن است در هنگام ذخیره‌سازی چنین داده‌های بزرگی رخ می‌دهد، مراقبت کند.

در عصر داده‌های بزرگ، همان‌طورکه حجم داده‌ها به‌سرعت رشد می‌کند، خطرات امنیتی شدیدتری وجود دارد؛ درحالی‌که ثابت شده روش‌های حفاظت داده‌های سنّتی، برای داده‌های بزرگ کارآمد نیستند؛ به‌خصوص حریم خصوصی داده‌های بزرگ که با چالش‌های امنیتی زیر مواجه می‌شود:

ـ حفاظت از حریم خصوصی حین کسب داده: علایق و ویژگی‌های شخصی و عادات کاربران می‌تواند به‌راحتی کسب شود و کاربران متوجه نخواهند شد.

ـ داده‌های حریم خصوصی می‌توانند حین ذخیره‌سازی، انتقال و استفاده، نشتی پیدا کنند؛ حتی اگر با تأیید کاربران به‌دست آید. ازاین‌رو، می‌توان نتیجه گرفت که حریم خصوصی کلان‌داده‌ها می‌تواند به وسیله دو رویکرد مختلف حفظ شود: یکی، تحمیل قوانین به فرد و سازمان، و روش دیگر، توسعه حریم خصوصی.

بنابراین، داده‌های بزرگ، چالش‌هایی برای رمزگذاری داده‌های با مقیاس بزرگ و تراکم بالا به ارمغان می‌آورد. عملکرد روش‌های رمزگذاری در داده‌های کوچک و متوسط، نمی‌تواند تقاضای داده‌های بزرگ را برآورده کند؛ چراکه باید روش‌های رمزنگاری داده‌های بزرگ کارآمد شود و توسعه یابد. ازاین‌رو، در این خصوص باید طرح‌های مؤثر مدیریت امنیت، کنترل دسترسی و ارتباطات امن برای داده‌های ساخت‌یافته، نیمه‌ساخت‌یافته و غیرساخت‌یافته، مورد بررسی و تحقیق قرار گیرد.

منابع:

1. (11.03.2013)." An introductory session on Big Data".Available: http://www.humanfaceofbigdata.com

2. A. vailaya, "What's All the Buzz Around "BigData?"," IEEE Women in Engineering Magazine, pp. 24-31, December 2012.

3. C. Eaton, et al., "Understanding Big Data: Analytic for Enterprise Class Hadoop and Streaming Data: Mc Graw-Hill companies", 2012.

4. D. J. Abadi, et al., "Harizopoulos,Column-oriented database systems," Processing VLDB Endow, vol. 2, pp. 1664-1665, 2009.

5. H. Rathod and T. Chauhan, "A Survey on Big Data Analysis Techniques," IJSRD - International Journal for Scientific Research & Development, vol. 1, pp. 1806-1808, 2013.

6. I. A. T. Hashem, et al., "The Rise of "Big Data" on Cloud Computing:Review and Open Research Issues," Information System, vol. 47, pp ,115-98 .Agust 2014.

7. M. Minelli, et al., "Data, Big Data Analytics:Emerging Business Intelligence and Analytic Trend for Today's Businesses": John Wiley & Sons, 2013.

8. Min Chen , Shiwen Mao , Yunhao Liu," Big Data: A Survey",© Springer Science+Business Media New York 2014.

9. P. Neubauer, "Graph databases, NOSQL and Neo4j," 2010.

10. R. D. Schnieder, "Hadoop for Dummies Special" Edition: John Widly&Sons Canada, 2012.

11. S. M. Seeger, Comput.Sci.Media, "Ultra-Large-Sites,Key-Value stores:a practical overview," computer sciences Media, 2009.

12. S. Sagirogolu and D. Sinanc, "Big Data: A Review," IEEE, 2013.

13. S. Singh and N. Singh, "Big Data Analytics," presented at the International Conference on Communication, Information & Computing Tecnology, Mumbai india, 2012.

14. Shrivasta, K., Rizvi, M., & Singh, S. "Big Data Privacy Based on Differential Privacy a Hope for Big Data". Computational Intelligence and Communication Networks (CICN) (pp. 776-781). Bhopal: IEEE, 2014.■