تحلیل دیتاهای بزرگ: چالش‌ها و راه‌حل‌های آن

پوهنیار احمدضیا بهرامی[۶]
تقریظ‌دهنده: پوهندوی عبدالرحمن مجددی

مجـلهٔ علمی-تحقیقی حـوزهٔ علوم طـبیعی پـوهـنتون کـابـل، ۲ (٤) ۱٤۰۰

چکیده

امروزه در عصـر دیتاهای بزرگ؛ دیتا بسـیار سریع و با حجم عظیم و با انواع گوناگون از منابع مختلف تولید و جمع‌آوری می‌شود. بسیاری از دولت‌ها و سازمان‌های خصوصی عصر دیتاهای بزرگ را یک فرصت برای رقابت و پیشرفت دانسته و با اسـتفاده از تکنیک‌های تحلیل دیتاهای بزرگ از میان انبوه دیتا به دانش و بینشـی ارزش‌مند دسـت یافته و از آن در تصـمیم‌گیری و پلان‌گذاری خویش بهره برده اسـت. تجزیه و تحلیل دیتاهای بزرگ به دلیل حجم، تنوع و سرعت تولید آن با تجزیه و تحلیل دیتاهای عادی تفاوت داشته، در این زمینه چالش‌های زیاد وجود دارد. این مقاله به مفاهیم اساسی تجزیه و تحلیل دیتاهای بزرگ پرداخته، مسـائل و چالش‌های تحلیل دیتاهای بزرگ را بررسـی و راه‌حل‌های مناسـب را برای حل چالش‌ها پیشنهاد می‌نماید.

اصطلاحات کلیدی: دیتاهای بزرگ؛ تحلیل دیتا؛ چالش‌ها؛ راه‌حل‌ها؛ کلود کامپیوتینگ

Big Data Analytic: Challenges & Solutions

Jr. Teaching Asstt. Ahmad Zia Bahrami

Abstract

In today's Big Data era, Data is generated and collected very quickly and with huge volume and with different types from different sources. Many governments and private organizations perceive the big data era as an opportunity to compete and thrive. Using big data analytic techniques, they have gained valuable knowledge and insight from the mass of data and have utilized it in their decision-making and planning. Big data analytics differs from normal data analytic due to its volume, variety and production speed, and there are many challenges in this field. This article deals with the basic concepts of big data analytic and examines the problems and challenges of big data analytic and suggests appropriate solutions to overcome the challenges.

Keywords: Big Data; Data Analytics; Challenges; Solutions; Cloud Computing

ارجاع

بهرامی، احمدضیا. (۱٤۰۰). تحلیل دیتاهای بزرگ: چالش‌ها و راه‌حل‌های آن. مجـلهٔ علمی- تحـقیقی حـوزهٔ علـوم طبیعی پوهـنتون کـابل، شماره ۲ (٤)، صص ۷۵ – ۸۵.

---

[۶] استاد پوهنځی کمپیوترساینس، پوهنتون کابل

مقدمه

در سال‌های اخیر، توانایی تولید و جمع‌آوری دیتا به‌طور تصاعدی افزایش یافته است. در واقع، در عصر اینترنت اشیا، مقدار زیادی دیتا دیجیتالی توسط منابع مختلف (مانند حسگرها، دوربین‌های امنیتی، کنتورهای هوشمند، دستگاه‌های تلفن موبایل، برنامه‌های وب و سرویس‌ها) تولید و جمع‌آوری می‌شود. حجم عظیم دیتاهای تولید شده، اعم از نظر سرعت تولید آن و اعم از نظر ناهمگنی فارمت آن (به عنوان مثال، ویدیو، متن، xml، ایمیل)، چالشی برای ذخیره‌سازی، پردازش و تجزیه و تحلیل است. به ویژه، با رشد شبکه‌های اجتماعی (به عنوان مثال، فیسبوک، توییتر، اینستاگرام، و غیره)، انتشار وسیع تلفن‌های موبایل و استفاده زیاد از سرویس‌های مبتنی بر مکان، هر روز میلیون‌ها نفر به خدمات شبکه‌های اجتماعی دسترسی پیدا می‌کنند و از طریق آن علایق و فعالیت‌های خویش را به اشتراک می‌گذارند و حجم عظیم از دیتا را تولید می‌کنند. از این حجم دیتاها که معمولاً به عنوان دیتاهای بزرگ (Big Data) یاد می‌شود، می‌توان برای استخراج معلومات مفید و تولید دانش مؤثر برای ساینس، صنعت، خدمات عمومی و به طور کلی برای پیشرفت زندگی بشر بهره برد (۱).

تجزیه و تحلیل دیتاهای بزرگ، چالش‌های دیتاهایی را نشان می‌دهد که بیش از حد گسترده، ساختارناپذیر و بسیار سریع هستند و نمی‌توانند با روش‌های سنتی مدیریت شوند. از تجارت و مؤسسات تحقیقاتی گرفته تا دولت‌ها و سازمان‌ها اکنون به طور معمول دیتاهایی گسترده و باپیچیدگی بی‌سابقه تولید می‌کنند. جمع‌آوری معلومات معنادار و با مزایای رقابتی از انبوه دیتا برای سازمان‌ها در سطح جهان اهمیت بیشتری پیدا کرده است. تلاش برای استخراج بینش معنادار و کارآمد از چنین منابع دیتای سریع و آسان چالش‌برانگیز است. بنابراین، تجزیه و تحلیل برای درک ارزش کامل دیتاهای بزرگ برای بهبود عملکرد تجاری و افزایش سهم بازار، به طور جدایی‌ناپذیری حیاتی شده است. ابزارهای موجود برای کنترل حجم، سرعت و تنوع دیتاهای بزرگ در سال‌های اخیر بسیار بهبود یافته است. به طور کلی، این تکنالوژی‌ها گران نیستند و بسیاری از نرم‌افزارها منبع باز هستند. Hadoop، معمول‌ترین چارچوب است که وسایل سخت‌افزاری را با نرم‌افزار منبع باز ترکیب می‌کند. Hadoop، جریان‌های ورودی دیتاها را می‌گیرد و آن‌ها را بر روی دیسک‌های ارزان قیمت توزیع می‌کند. هم‌چنین ابزارهایی برای تجزیه و تحلیل داده‌ها فراهم می‌کند. با این حال، این تکنالوژی‌ها به مجموعه مهارتی نیاز دارند که برای بیشتر بخش‌های تکنالوژی معلوماتی جدید است و برای ادغام

تمام منابع داخلی و خارجی مربوط به دیتاها، باید سخت کار کنند. گرچه تنها توجه به تکنالوژی کافی نیست، اما همیشه یک جز ضروری از یک استراتژی دیتاهای بزرگ است (۲).

در این مقاله مفاهیم اساسی دیتاهای بزرگ و تجزیه و تحلیل دیتاهای بزرگ بحث می‌شود. چالش‌های اساسی تجزیه و تحلیل دیتاهای بزرگ شناسایی و راحل‌های مناسب پیشنهاد می‌گردد.

پیشینه‌ی تحقیق

Big Data یک روش تجزیه و تحلیل دیتا است که با پیشرفت‌های اخیر تکنالوژی‌هایی که از ثبت، ذخیره‌سازی و تجزیه و تحلیل دیتا با سرعت بالا پشتیبانی می‌کنند، امکان‌پذیر است. منابع دیتا فراتر از دیتابیس‌های سنتی شرکت‌ها هستند و شامل ایمیل‌ها، خروجی‌های موبایل‌های هوشمند و دیتاهای تولید شده توسط سنسور هستند که در آن دیتاها دیگر محدود به سوابق دیتابیس ساختاریافته نیستند، بلکه دیتاهای بدون ساختار و بدون قالب‌بندی استاندارد هستند (۳). از آن‌جا که Big Data و Analytics عبارتی نسبتاً جدید و در حال تحول است. بنابراین، تعریف واحدی وجود ندارد. ذینفعان مختلف تعاریف متنوع و گاه متناقضی ارائه داده اند. یکی از اولین تعاریفی که از دیتاهای بزرگ به طور گسترده نقل شده است، ناشی از گزارش گارتنر در سال ۲۰۰۱ است. گارتنر پیشنهاد کرد که، دیتاهای بزرگ توسط مدل 3V حجم (Volume)، سرعت (Velocity) و تنوع (Variety) تعریف می‌شوند. گارتنر تعریف خود را در سال ۲۰۱۲ گسترش داد و صحت(Veracity) را نیز شامل تعریف ساخت، تا صحت نشان‌دهنده‌ی الزامات مربوط به اعتماد و عدم اطمینان مربوط به دیتاها و نتیجه تجزیه و تحلیل دیتاها باشد. در یک گزارش سال ۲۰۱۰، IDC مدل 5^{th} V را تعریف کرد که در این مدل ارزش (Value) را نیز بخش از تعریف دیتاهای بزرگ دانسته، یعنی تطبیق اپلیکشن‌های دیتاهای بزرگ در سازمان‌ها باعث افزایش ارزش در سازمان می‌گردد. تجزیه و تحلیل دیتاها به‌صورت عموم معلومات‌های غیر ساختاریافته را که از ثبت تماس‌های تلفن، ترانسکشن‌های بانکداری موبایل و از محتویات صفحات مجازی مانند فیسبوک، تویتر، و بلاگ‌ها، جست و جو انلاین و عکس‌ها به‌دست می‌آید بااستفاده از تکنیک پیشرفته، تجزیه و تحلیل نموده و ارزش‌های نهفته در آن را برای تجارت اشکار و قابل استفاده می‌سازد (۲).

بعد دیگر تعریف Big Data شامل تکنالوژی است. دیتاهای بزرگ نه تنها بزرگ و پیچیده است، بلکه برای تجزیه و تحلیل و پردازش آن به تکنالوژی جدید نیاز است. در سال ۲۰۱۳، کارگروه Big Dataمؤسسه استاندارد و فناوری (NIST) تعریف زیر را برای دیتاهای بزرگ ارائه داده است که بر کاربرد تکنالوژی جدید تأکید دارد. Big Data؛ از ظرفیت یا توانایی روش‌ها و سیستم‌های فعلی یا

متداول فراتر رفته و با اســتفاده از روش‌های رایج یا متداول، رویکردهای جدیدی را برای ســؤالات کشف نشده که قبلاً غیرقابل دست‌رسی یا غیر عملی بودند، امکان‌پذیر می‌سازد. چالش‌های تجاری به ندرت از طریق مشکلات دیتا نشان داده می‌شوند و حتی وقتی دیتاها بسیار زیاد باشد، متخصصان برای ادغام آن در تصـمیم‌گیری پیچیده‌ی خود که ارزش تجاری را ا ضافه می‌کند، مشـکل دارند. در سـال ۲۰۱۲، McKinsey & Company یک نظرسـنجی از ۱۴۶۹ مدیر در مناطق مختلف، صـنایع و شرکت‌های مختلف انجام داد که در آن ۴۹ درصــد از پاســخ‌دهندگان گفته اند که شرکت‌های آن‌ها تلاش می‌کنند دیتاهای بزرگ را بر روی بینش مشــتری، تقســیم‌بندی وهدف‌گذاری برای بهبود عملکرد کلی قرار دهند (۴)، حتی تعداد بیشــتری از پاسـخ‌دهندگان ۶۰ درصــد گفتند که شرکت‌های شـان باید تلاش خود را برای استفاده از دیتاها و تجزیه و تحلیل برای ایجاد این بینش متمرکز کنند. با این حال، فقط یک پنجم گفته‌اند که ســازمان‌های آن‌ها به طور کامل دیتاها و تجزیه و تحلیل‌ها را برای ایجاد بینش در یک واحد تجاری مستقر کرده اند و تنها ۱۳ درصد از آن‌ها برای تولید بینش در سراسر شرکت از دیتاها اسـتفاده می‌کنند. همان‌طور که این نتایج نظرسـنجی نشـان می‌دهد، دیگر این سؤال مطرح نیست که آیا دیتاهای بزرگ می‌توانند به تجارت کمک کنند، بلکه سؤال این است که چگونه سازمان‌ها و شرکت‌ها می‌تواند حداکثر نتیجه را از دیتاهای بزرگ کسب کند.

اهمیت تجزیه و تحلیل دیتاهای بزرگ

تحقیقات نشـان داده اسـت که فرصـت‌های ناشـی از تجزیه و تحلیل دیتاهای بزرگ برای دولت‌ها و سـازمان‌ها یک امر محوری می باشـد، با تطبیق تکنالوژی‌های دیتاهای بزرگ، سـازمان‌ها انتظار دارند در بســیاری از حوزه‌ها، از جمله تجارت الکترونیکی، دولت الکترونیکی، تحصــیلات، صــحت و امنیت، مزایایی را به دست آورند. مزایایی که سازمان‌ها به عنوان ارزش مـــــی‌خواهند از دیتاهـــای بزرگ به‌دست آورند، بستگی به اهداف استراتژیک سازمان در چگونگی تطبیق و مدیریت دیتاهای بزرگ دارد (۱).

تطبیق تکنالوژی و مدیریت دیتاهای بزرگ برای ســازمان‌ها و کشــورها دارای ارزش مهم اجتماعی و اقتصـادی در نظر گرفته می‌شـود. ارزش اجتماعی شـامل سـلامتی اجتماعی در زمینه‌هایی مانند تعلیم، مراقبت‌های صــحی، ســلامت عمومی و امنیت می‌گردد. به عنوان مثال؛ دولت می‌تواند برای تقویت شــفافیت، افزایش مشــارکت شــهروندان در روابط عمومی، جلوگیری از کلاه‌برداری و جرم، بهبود امنیت ملی و حمایت از رفاه مردم از طریق آموزش بهتر از دیتاهای بزرگ اســتفاده کند. هم‌چنان،

مزایایی اجتماعی دیتاهای بزرگ در برگیرنده‌ی منافع اجتماعی از جمله رشد اشتغال، بهره‌وری و جلوگیری از مصرف اضافی در اجتماع می‌گردد.

ارزش اقتصادی را می‌توان با افزایش سود، رشد تجاری و مزیت رقابتی یک سازمان ناشی از تطبیق دیتاهای بزرگ سنجید. ارزش اقتصادی اغلب منافع پولی را شامل می‌شود که توسط سازمان‌ها تخصیص می‌یابد. به عنوان مثال؛ از سازمان‌هایی که برای هدایت استراتژی‌های سازمانی و عملیات روزانه‌ی خویش به دیتاهای بزرگ متکی هستند، انتظار می‌رود عملکرد مالی بهتری نسبت به سازمان‌هایی که انجام نمی‌دهند، داشته باشند (۵).

به طور کلی، دیتاهای بزرگ به عنوان منبع محصولات نوین، خدمات و فرصت‌های شغلی درک می‌شوند. علاوه براین، اعتقاد بر این است که دیتاهای بزرگ منجر به عملیات مؤثرتر و کاراتر می‌شوند، به طور مثال به موارد بهترسازی جریان‌های تولیدات زنجیره‌یی، تعیین سودآورترین قیمت محصولات و خدمات، انتخاب افراد مناسب برای بعضی کارها و شغل‌های تخصصی، به حد اقل رساندن خطاها و مشکلات کیفیت و بهبود روابط مشتری درسازمان‌ها می‌توان اشاره نمود. علاوه براین، ارزش اقتصادی واجتماعی بیشتر را از دیتاهای بزرگ می‌توان از طریق تصمیم‌گیری‌های دقیق‌تر و پلان‌گذاری مؤثرتر و مدیریت دیتاهای بزرگ به دست آورد.

تجزیه و تحلیل دیتاهای بزرگ

تجزیه و تحلیل دیتاهای بزرگ، عبارت از تکنیک‌های پیشرفته‌ی تحلیلی است که روی مجموعه دیتاهای بزرگ کار می‌کنند. بنابراین، تجزیه و تحلیل دیتاهای بزرگ به‌صورت واقعی مربوط به دو مفهوم است یکی دیتاهای بزرگ و دیگری تحلیل و تجزیه؛ طوری‌که این دو مفهوم باهم ادغام شده اند تا یکی از عمیق‌ترین روندهای هوش تجاری (BI) امروز را ایجاد کنند (٦). امروزه، دیتاهایی که باید تجزیه و تحلیل شوند، تنها حجم شان بزرگ نیست، بلکه از انواع مختلف دیتا و حتی شامل جریان دیتاها نیز می‌شوند (٦). از آن‌جا که دیتاهای بزرگ دارای ویژگی‌های منحصر به فرد "حجم زیاد، ابعاد مختلف، ناهمگن، پیچیده، بدون ساختار و ناقص" است که ممکن است روی‌کردهای احصایوی و تجزیه و تحلیل دیتاها را تغییر دهد (۷). اگرچه برای یافتن معلومات مفیدتر، دیتاهای بزرگ امکان جمع‌آوری دیتاهای بیشتر را برای ما امکان‌پذیر می‌کند، اما حقیقت این است که دیتاهای بیشتر لزوماً به معنای اطلاعات مفیدتر نیستند. دیتاهای بیشتر ممکن است حاوی دیتاهای مبهم یا غیرعادی باشد. به عنوان مثال، یک کاربر ممکن است چندین حساب داشته باشد، یا یک حساب کاربری توسط چندین کاربر استفاده شود که ممکن است دقت نتایج استخراج را کاهش

دهد(۸). بنابراین، در این صورت چندین مورد جدید برای تجزیه و تحلیل دیتاها مانند حریم خصوصی، امنیت، ذخیره‌سازی، تحمل خطا و کیفیت دیتاها مطرح می‌شود (۱).

دیتاهای بزرگ ممکن است توسط وسایل موبایل، شبکه اجتماعی، اینترنت اشیا، چندرسانه‌یی و بسیاری دیگر از برنامه‌های جدید ایجاد شود که همه دارای ویژگی‌های حجم، سرعت و تنوع هستند. اما در کل، تمامی تجزیه و تحلیل دیتاها باید از دیدگاه‌های زیر بررسی شود:

از منظر حجم: از منظر حجم، حجم عظیمی ورودی اولین چیزی است که باید با آن روبرو شویم زیرا ممکن است حجم عظیم از ورودی پروسه‌ی تجزیه و تحلیل دیتاها را فلج کند.

از منظر سرعت: از منظر سرعت، دیتاها زمان واقعی یا جریان دیتا مشکل دیتاهای ورودی را برای تحلیل و تجزیه به‌وجود می‌آورد که وسایل و سیستم نمی‌تواند در مدت زمان کم مقدار زیادی دیتاهای ورودی را کنترل و تحلیل و تجزیه کنند.

از نظر تنوع: از آن‌جا که دیتاهای ورودی ممکن است از انواع مختلفی استفاده‌کنند یا دیتا ناقصی داشته باشند، نحوه‌ی مدیریت آن‌ها نیز مسأله‌ی دیگری را برای اپراتورهای ورودی تجزیه و تحلیل دیتاها به وجود می‌آورد (۹).

چالش‌های تجزیه و تحلیل دیتاهای بزرگ

زنجیره‌ی اپلیکشن دیتاهای بزرگ به طور کلی شامل چهار مرحله است: تولید دیتا، مدیریت دیتا، تجزیه و تحلیل دیتاها و اپلیکشن‌ها. تجزیه و تحلیل دیتاهای بزرگ، که مهم‌ترین مرحله در کل زنجیره محسوب می‌شود، به روند کشف الگوها از دیتاها اشاره دارد. در این مرحله، شش چالش اصلی وجود دارد. این باعث می‌شود که تجزیه و تحلیل دیتاهای بزرگ را بسیار دشوارتر و پیچیده‌تر از تجزیه و تحلیل دیتاهای اندازه طبیعی می‌کند.

نمایش دیتاهای پیچیده: چگونگی نمایش یک‌نواخت انواع مختلف ویژگی‌ها، چالش بزرگی برای دیتاهای بزرگ با چند بعد است. ما باید دیتاها را در یک چارچوب یک‌سان پردازش کنیم. نمایش یک‌نواخت/ساختار یافته‌ی دیتاها اولین مرحله‌ی پردازش دیتاها است. ضروری است اما به دلیل چندشکل بودن دیتاهای بزرگ، نمایش یک‌نواخت انواع مختلف داده‌ها بسیار دشوار است. این بدان معنا است که استفاده از روش‌های موجود برای مدیریت دیتاهای بزرگ تقریباً غیرممکن است. این اولین چالش تجزیه و تحلیل دیتاهای بزرگ است.

ابعاد فوق‌العاده بالا: دیتاهای بزرگ در حوزه‌های خاص، به ویژه در بیوانفورماتیک یا محاسبات علوم بیولوژیکی، اغلب دارای ابعادی بسیار بالا هستند. مسأله این است که الگوریتم‌های موجود

برای دیتاهای با ابعاد بالا قابل مقیاس نیستند. معمولاً با افزایش ابعاد دیتا، مقدار زمان یا حافظه مورد نیاز به صورت تصاعدی افزایش می‌یابد. طوری‌که ژایی و همکاران شرح مفصلی از تغییر سریع ابعاد مجموعه دیتاها در زمینه‌ی تحقیقات علمی طی ۲۵ سال گذشته داده است. بسیاری از الگوریتم‌های یادگیری ماشین و داده‌کاوی بر اساس اندازه‌گیری فاصله در یک فضای متریک طراحی شده اند، به عنوان مثال، k-nearest neighbor یکی از معروف‌ترین این الگوریتم‌ها است. مطالعات (۷) نشان می‌دهد که، در یک فضای با ابعاد بالا، اندازه‌گیری فاصله یک پدیده‌ی غیر متجانس است. یعنی برخی از نقاط ثابت نزدیک‌ترین همسایه در هر مورد در فضا هستند. به آن hubness گفته می‌شود که نشان می‌دهد فرمول فاصله بی‌اثر و نامعتبر است.

کلاس‌های عظیم: در عصر دیتاهای بزرگ طبقه‌بندی کلاس‌ها یکی از کارهای بسیار مهم و اساسی است که توسط این کار ما باید با هزاران کلاس از قبیل مسأله تشخیص مقیاس بزرگ توسط دستوری‌های طبقه‌بندی کنار بیایم. به نظر می‌رسد الگوریتم‌های طبقه‌بندی موجود از نگاه طبقه‌بندی کلاس‌ها مشکل ندارد، اما از نگاه اجرا آن‌ها به طور جدی تنزیل می‌یابد. مطالعه (۱۰) مقیاس را به طور واضیح توصیف می‌کند.

رابطه‌ی ضعیف: یک رابطه عمومی‌تر از نگاشت (Mapping) است و یافتن رابطه‌ی دشوارتر از یافتن نگاشتی در هنگام انجام تجزیه و تحلیل دیتاهای بزرگ است. به عنوان مثال، دراجرای طبقه‌بندی ممکن برچسب‌ها فراموش شده باشند، یا مواردی اشتباه برچسب‌گذاری شوند. در این مورد هزینه زیاد برای برچسب‌گذاری منجر به مشکل نظارت ضعیف می‌شود. به طور سنتی، ما باید نگاشتی را از مجموعه موارد به مجموعه دیگری پیدا کنیم. در بیشتر شرایط در یک محیط داده بزرگ، فقط باید رابطه‌یی بین دو زیر مجموعه از موارد پیدا کنیم. این بدان دلیل است که گاهی اوقات در یک محیط دیتاهای بزرگ، ممکن است نیازی به نگاشتی دقیق نداشته باشیم و غالباً یافتن چنین نگاشتی دقیق غیرممکن است.

توانایی محاسبه‌ناپذیر: توانایی محاسباتی فعلی برای مسأله دیتاهای بزرگ قابل افزایش نیست. الگوریتم‌های یادگیری موجود نمی‌توانند خود را به خوبی با محیط جدید دیتاهای بزرگ سازگار کنند. این بدان معناست که هم پیچیدگی مسأله و هم توانایی محاسباتی به طور چشم‌گیری در عصر دیتاهای بزرگ افزایش می‌یابد، اما افزایش توانایی محاسباتی با افزایش پیچیدگی مسأله مطابقت ندارد. هنگامی‌که یک مجموعه دیتا از یک اندازه‌ی معمولی به یک اندازه‌ی بزرگ با بسیاری از ویژگی‌ها تغییر می‌کند، برخی از الگوریتم‌های داده‌کاوی و یادگیری ماشین که اغلب استفاده

می‌شوند، مانند Support Vector Machine، Neural Network، Decision Tree، C-means و C-modes به خوبی کار نخواهد کرد.

در بسیاری از حوزه‌ها، الگوریتم یادگیری یا استخراج به عنوان مؤثر برای دیتاهای بزرگ شناخته شده است. تنها درصورتی که پیچیدگی آن خطی یا شبه خطی باشد.

عدم اطمینان در هر مرحله: عدم اطمینان در هر مرحله‌ی از یادگیری دیتاهای بزرگ وجود دارد. به عنوان مثال، دیتاهای بزرگ غالباً اختلال زیادی دارند و اکثراً، مقادیر مشخصه‌ی یک مورد در دیتاهای بزرگ وجود ندارد (به عنوان مثال در شبکه‌های اجتماعی ۸۰٪-۹۰٪ لینک‌های نامعلوم و غیز قابل اطمینان است و در کلینیک و زمینه‌های صحت ۹۰٪ مقادیر مشخصه برای تشخیص دکتر وجود ندارد.) به‌طور وا ضح برخی از الگوریتم‌های یادگیری سنتی برای پردازش دیتاها با ۹۰٪ مقادیر از دست رفته معتبر نبوده اند، بنابراین، نحوه‌ی طراحی الگوریتم یادگیری جدید برای مقابله با دیتاهای گم‌شده در مقیاس بزرگ دشوار است. علاوه براین، مدل‌های بسیاری وجود دارد که می‌توانند برای پردازش دیتاهای بزرگ انتخاب شوند. با توجه به عدم اطمینان روزافزون موجود در پروسه‌ی انتخاب، انتخاب یک مدل مناسب بر اساس عدم اطمینان فرمول‌بندی شده یک چالش بزرگ دیگر است. سومین مشکل این است که چگونه عدم اطمینان دیتاها را به خوبی نشان دهیم و چگونه آن‌ها را در پروسه‌ی استخراج در مرحله‌ی تجزیه و تحلیل دیتاها نشان دهیم. آیا از دیتاهای با اندازه طبیعی گرفته تا دیتاهای بزرگ، عدم اطمینان افزایش یا کاهش می‌یابد؟ بستگی دارد. به عنوان مثال، برای میانگین یک متغیر تصادفی، عدم اطمینان به دلیل قضیه اعداد زیاد کاهش می‌یابد، اما برای مسأله‌ی انتخاب مدل، افزایش می‌یابد (۱۱).

راه‌حل

در ابتدای پدیده Big Data، فقط شرکت‌های بزرگ فناوری اطلاعات، مانند فیسبوک، یاهو، توییتر، آمازون، LinkedIn، منابع زیادی را در توسعه پروژه‌های اختصاصی برای مقابله با مشکلات تجزیه و تحلیل دیتاهای بزرگ سرمایه‌گذاری کردند. اما امروزه، تجزیه و تحلیل Big Data برای سازمان‌های تجاری کوچک و متوسط بسیار قابل توجه و مفید در نظر گرفته می‌شود. برای پاسخ‌گویی به این تقا ضای درحال افزایش، فروشندگان بزرگ شروع به راه‌اندازی پلت فرم‌های توزیع شده برای تجزیه و تحلیل داده‌های بزرگ کرد. در میان پروژه‌های منبع با، Apache Hadoop پیشتازترین پلت فرم پردازش دیتا منبع باز است که توسط غول‌های فناوری اطلاعات مانند فیسبوک و یاهو به کارگرفته شده است.

از سال ۲۰۰۸، چندین شرکت، مانند Cloudera، MapR و Hortonworks، با تلاش بسیار برای بهبود عملکرد Hadoop از نظر ذخیره‌سازی و پردازش دیتاهای مقیاس‌پذیر، شروع به ارائه پلت فرم‌های بزرگ کردند. در عوض، IBM و Pivotal شروع به راه‌اندازی Hadoop اختصاصی که مربوط شرکت شان است، کردند. شرکت‌های بزرگ دیگر تصمیم گرفتند فقط نرم‌افزار و پشتیبانی جانبی را برای پلت فرم Hadoop که توسط تهیه‌کنندگان خارجی توسعه یافته راه‌اندازی کنند: به عنوان مثال، مایکروسافت تصمیم گرفت پیشنهاد خود را بر پایه‌ی پلت فرم Hortonworks قرار دهد، در حالی که Oracle تصمیم گرفت تا پلت فرم Cloudera را مجدداً بفروشد. با این حال Hadoop تنها راه‌حل برای تجزیه و تحلیل دیتاهای بزرگ نیست. خارج از پلت فرم Hadoop راه‌حل‌های دیگر نیز برای تجزیه و تحلیل دیتاهای بزرگ در حال ظهور است. به طور خاص، تجزیه و تحلیل در حافظه به یک روند گسترده تبدیل شده است، به طوری که شرکت‌ها شروع به ارائه ابزارها و خدمات برای تجزیه و تحلیل سریع‌تر در حافظه می‌کنند، مانند SAP، که با پلت فرم Hana3 خود پیشرو در نظر گرفته می‌شود. سایر شرکت‌ها و سازمان‌ها، از جمله HP، Teradata و Actian، ابزارهای دیتابیس تحلیلی را با قابلیت تجزیه و تحلیل حافظه توسعه دادند (۱۲). علاوه براین، برخی از فروشندگان، مانند مایکروسافت، IBM، Oracle و SAP، با ارائه راه‌حل کامل برای تجزیه و تحلیل دیتاها، از سایر فروشندگان متمایز هستند. علاوه براین، بسیاری از فروشندگان تصمیم گرفتند که بیشتر روی Cloud متمرکز کنند. از این میان خدمات وب آمازون (AWS) و 1010DATA می‌توان نام برد. به طور خاص، AWS طیف گسترده‌یی از خدمات و محصولات را در Cloud برای تجزیه و تحلیل داده‌های بزرگ، از جمله سیستم‌های دیتابیس مقیاس‌پذیر و راه‌حل‌های پشتیبانی تصمیم ارائه می‌دهد. سایر فروشندگان کوچک‌تر، از جمله Actian، InfiniDB، HP Vertica، Infobright و Kognitio، پیشنهاد دیتاهای بزرگ خود را فقط برای تجزیه‌وتحلیل بر روی سیستم‌های مدیریت دیتابیس متمرکز کردند.

در واقع، با وجود این‌که Cloud برای بسیاری از کاربران یک راه‌حل مقرون به صرفه است، تعداد راه‌حل‌های تجزیه و تحلیل دیتا بزرگ بسیار محدود است. اکثر راه‌حل‌های موجود امروزه براساس چارچوب‌های منبع باز مانند Hadoop و Spark ساخته شده اند، اما برخی از راه‌حل‌های اختصاصی مانند راه‌حل‌های پیشنهادی IBM، EMC یا Kognitio نیز وجود دارد. تجزیه و تحلیل Big Data یک زمینه به طور مداوم در حال رشد است، بنابراین، راه‌حل‌های جدید و کارآمد (به عنوان مثال، از نظر سیستم عامل، ابزار برنامه‌نویسی، چارچوب‌ها و الگوریتم‌های داده‌کاوی) هر روز برای مقابله با دامنه‌ی رو به رشد علاقمندی به Big Data ظاهر می‌شوند.

نتیجه‌گیری

طوری‌که در این مقاله دیده می‌شود، عصر دیتاهای بزرگ فرصت‌ها و چالش‌های گوناگونه را برای دولت‌ها و سازمان‌ها به‌وجود آورده است. در نتیجه، گفته می‌توانیم بسیاری از دولت‌ها و سازمان‌های تجاری و صنایع از تجزیه و تحلیل دیتاهای بزرگ برای رقابت و پیشرفت بهره برده و از میان انبوه دیتاها به بینش و ارزش قابل اعتماد و ارزش‌مند دسته یافته و آز آن برای تصمیم‌گیری، پلان انکشافی، سیاست و تحقیقات علمی خویش استفاده بهینه نموده اند.

ویژگی‌های دیتاهای بزرگ نشان می‌دهد ضمن این‌که تجزیه و تحلیل دیتاهای بزرگ برای دولت‌ها و سازمان‌ها فرصت خلق کرده است، چالیش‌های جدی را نیز به‌وجود آورده است. این مقاله اول مفاهیم و تکنیک‌های تجزیه و تحلیل دیتاهای بزرگ را مورد مطالعه قرار داده و در نتیجه به شش چالش اساسی دیتاهای بزرگ اشاره و راه‌حل‌های ممکن را پیشنهاد نموده است.

بسیار از تکنالوژی‌ها و تکنیک‌ها برای مقابله با چالش‌های تجزیه و تحلیل دیتاهای بزرگ ایجاد شده است و سازمان‌های مختلف برای غلبه با این چالش از تکنالوژی و تکنیک‌های مختلف تجزیه و تحلیل استفاده کرده است اما واقعیت این است که تاهنوز معایب و جود دارد و راه‌حل کلی توافق نشده است و تحقیقات برای بهبود ویژگی‌های تکنالوژی و قابلیت‌های تکنیک‌های تجزیه و تحلیل دیتاهای بزرگ جریان دارد.

منابع

(1) Katal A, Wazid M, Goudar RH. Big data: Issues, challenges, tools and Good practices. 2013 6th Int Conf Contemp Comput IC3 2013. 2013; pp. 404-409.

(2) 2Kaisler S, Armour F, Espinosa JA, Money W. Big data: Issues and challenges moving forward. Proc Annu Hawaii Int Conf Syst Sci. 2013; pp. 995–1004.

(3) Villars RL, Olofson CW, Eastwood M. Big Data: What It is and Why You Should Care. IDC White Pap [Internet]. 2011; pp. 7-8. Available from: http://www.tracemyflows.com/uploads/big_data/IDC_AMD_Big_Data_Whitepaper.pdf

(4) Sujitparapitaya S, Shirani A, Roldan M. Issues in Information Systems. Issues Inf Syst. 2012; 13(2), pp. 112-22.

(5) Sukumar SR. Open Research Challenges with Big Data - A Data - Scientist' s Perspective. 2015, pp. 1272-8.

(6) Russom P. Big data analytics - tdwi best practices reporT Introduction to Big Data Analytics. TDWI best Pract report, fourth Quart [Internet]. 2011; 19(4), pp. 1-34. Available from: https://vivomente.com/wp-content/uploads/2016/04/big-data-analytics-white-paper.pdf

(7) Ma C, Zhang HH, Wang X. Machine learning for Big Data analytics in plants. Trends Plant Sci [Internet]. 2014; 19(12), pp. 798-808. Available from: http://dx.doi.org/10.1016/j.tplants.2014.08.004

(8) Boyd D, Crawford K. Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. Inf Commun Soc. 2012; 15(5), pp. 662-79.

(9) Tsai CW, Lai CF, Chao HC, Vasilakos A V. Big data analytics: a survey. J Big Data. 2015, pp. 1-32.

(10) Wang X, Yang LT, Member S, Liu H, Deen MJ. A Big Data-as-a-Service Framework : State-of-the-art and Perspectives. 2017; 7790(c), pp. 1-17.

(11) Wang X, He Y. Learning from Uncertainty for Big Data. Ieee Syst Man Cybern Mag [Internet]. 2016; (August). Available from: http://www.hebmlc.org/UploadFiles/20161121203535376.pdf

(12) Wozniak JM, Wilde M, Foster IT. Language features for scalable distributed-memory dataflow computing. Proc - 2014 4th Work Data-Flow Exec Model Extrem Scale Comput DFM 2014. 2014; 2(Vdl), pp. 50-53.