ISSN: 2745-3673
http://www.Rcsj.ir

مجله‌ی علمی پژوهش در علوم رایانه (سال دوم)
شماره ۴ / زمستان ۱۳۹۵ / ص ۷- ۱

فناوری‌ها، چالش‌ها و چشم‌انداز آینده کلان داده‌ها

مرضیه فلاح[۱]، احمد فراهی[۲]

[۱] دانشجو کارشناسی ارشد، گروه مهندسی کامپیوتر و فناوری اطلاعات، دانشگاه پیام نور
[۲] استادیار، گروه مهندسی کامپیوتر و فناوری اطلاعات، دانشگاه پیام نور

نام و نشانی ایمیل نویسنده مسئول:
مرضیه فلاح
fallah.system@gmail.com

چکیده

حجم داده‌های ذخیره‌شده ناشی از فعالیت‌های کاربران در شبکه‌های اجتماعی، تراکنش‌ها، داده‌های مربوط به سنسورهای آب‌وهوا، تصاویر و ویدئوهای دیجیتال، سیگنال‌های GPS و غیره در بستر اینترنت و به‌واسطه به‌کارگیری آن‌ها از تجهیزات و ابزارهای مختلف با سرعت خیره‌کننده‌ای در حال افزایش است. به مجموعه‌ای از این داده‌ها که نرخ رشد آن‌ها بسیار بالاست و در مدت‌زمان کوتاهی، شامل چنان حجمی از اطلاعات می‌شوند که کار با آن‌ها با ابزارهای مدیریت داده موجود غیرقابل انجام خواهد بود، «داده‌های عظیم» می‌گویند. حجم عظیم این داده‌های پیچیده می‌تواند چشم‌اندازها و واقعیت‌های بسیاری را به‌صورت پنهان در خود داشته باشد. ازاین‌رو در این مقاله، ضمن معرفی کلان داده‌ها و زیرساخت‌ها، به کاربردهای آن در فناوری و همچنین چالش‌هایی که به دلیل حجم، سرعت تولید و تنوع داده‌ها به وجود می‌آید و نیز چشم‌اندازهای آینده آن را بیان می‌کنیم.

واژگان کلیدی: کلان داده- علم داده- چالش- فناوری

مجلـه‌ی علمی پژوهش در علوم رایانه ، شماره ۴ ،زمستان ۱۳۹۵، ص ۷ – ۱

ISSN: 2745-3673
http://www.Rcsj.ir

مقدمه

در دنیای امروز، تجزیه‌وتحلیل داده‌های عظیم بسیار حیاتی است. بر این مبنا شرکت‌ها و دانشمندان توجه زیادی به این مقوله دارند. شرکت‌ها و سازمان‌ها حجم عظیمی از داده‌های مربوط به مشتریان، تأمین‌کنندگان، معاملات کسب‌وکار، سنسورهایی که در مکان‌های مختلف تعبیه می‌شوند را که بیش از تریلیون۱ بایت را در زمان واقعی۲ تولید می‌کنند. علاوه بر این میلیاردها نفر در سراسر جهان به‌صورت روزانه در شبکه‌های اجتماعی و برنامه‌های کاربردی به افزایش حجم و در دسترس بودن داده‌ها کمک می‌کنند.[3]

چنین داده‌های عظیم یک فرصت عالی برای تجزیه‌وتحلیل و استفاده ازآن‌جهت تصمیم‌گیری‌ها است؛ که سازمان‌ها می‌توانند در مورد کسب‌وکارشان اطلاعات بیشتری کسب کنند و فرآیندهای تصمیم‌گیری و عملکردشان را بهبود ببخشند. بااین‌حال وسعت هر مجموعه چالش‌هایی را نیز دارد؛ مثلاً: ظرفیت ذخیره‌سازی، مدیریت، سازمان‌دهی پردازش و تجزیه‌وتحلیل داده‌ها است.[4]

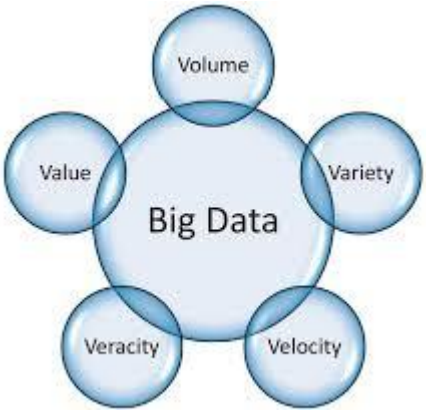
رویکرد ما در این مقاله بر سه محور بناشده است. در محور اول مروری بر کلان داده‌ها، کاربردها و تاریخچه آن خواهیم داشت و در رویکرد دوم بررسی چالش‌های موجود در داده‌های عظیم و رویکرد سوم چشم‌انداز آینده کلان داده‌هاست.

۱- کلان داده‌ها و تاریخچه آن

کلان داده‌ها معمولاً به مجموعه‌ای از داده‌ها اطلاق می‌شود که اندازه آن‌ها فراتر از حدی است که با نرم‌افزارهای معمولی بتوان آن‌ها را در یک‌زمان معقول اخذ، دقیق سازی، مدیریت و پردازش کرد. مفهوم اندازه در کلان داده‌ها به‌طور مستمر در حال تغییر است و به‌مرور بزرگ‌تر می‌شود.

طی بیست سال گذشته در حوزه‌های مختلف، حجم داده‌ها به مقدار زیادی افزایش‌یافته است. طبق گزارش موسسه بین‌المللی داده[3]، در سال ۲۰۱۱ حجم داده‌های ایجاد و تکثیرشده در سراسر دنیا ۱,۸ زِتابایت ($۱۰^{۲۱}$ بایت) بود که این مقدار در ظرف پنج سال نزدیک به ۹ برابر شده است. با این روند، در آینده‌ی نزدیک حداقل در هر دو سال یک‌بار این مقدار دو برابر خواهد شد.

در سال ۲۰۰۱ موسسه گارتنر[4] تعریف جدیدی را ارائه کرد: «کلان داده‌ها، حجم بالا و تنوع بالایی از اطلاعاتی هستند که نیازمند شکل جدیدی از پردازش هستند تا بتوانند تصمیم‌گیری را غنی‌تر سازند، بینش جدیدی را کشف کنند و نیز فرآیندها را بهینه نمایند.» و همچنین آن را در چند بعد از چالش‌ها و فرصت‌های پیش رو در حوزه رشد داده‌ها را مطرح کرد که عبارت بودند از:



شکل ۱: پنج ویژگی اصلی کلان داده‌ها

- حجم داده‌ها[5]: حجم داده‌ها به‌صورت نمایی در حال رشد است. منابع مختلفی نظیر شبکه‌های اجتماعی، لاگ سرورهای وب، جریان‌های ترافیک، تصاویر ماهواره‌ای، جریان‌های صوتی، تراکنش‌های بانکی، محتوای صفحات وب، اسناد دولتی وجود دارد که حجم داده بسیار زیادی را تولید می‌کنند.

[1] Trillions
[2] Real-time
[3] International Data Corporation
[4] Gartner
[5] Volume

ISSN: 2745-3673
http://www.Rcsj.ir

- سرعت تولید داده‌ها[6]: داده‌ها از طریق برنامه‌های کاربردی و سنسورهای بسیار زیادی که در محیط وجود دارند با سرعت بسیار زیاد و به‌صورت بلادرنگ تولید می‌شوند.
- تنوع داده‌ها[7]: انواع منابع داده و تنوع در نوع داده بسیار زیاد است که درنتیجه ساختارهای داده‌ای بسیار زیادی وجود دارد. مثلاً: در وب افراد از نرم‌افزارها و مرورگرهای مختلفی برای ارسال اطلاعات استفاده می‌کنند. بسیاری از اطلاعات مستقیماً از انسان دریافت می‌شود و بنابراین وجود خطا اجتناب‌ناپذیر است. این تنوع سبب می‌شود جامعیت داده تحت تأثیر قرار بگیرد؛ زیرا هر چه تنوع بیشتری وجود داشته باشد، احتمال بروز خطای بیشتری نیز وجود دارد.
- صحت و اعتبار داده‌ها[8]: این موضوع دلالت بر این دارد که داده‌ها برای تصمیم‌گیری چقدر حائز ارزش هستند و آیا اعتباری دارند؟ و از چه منابعی به‌دست‌آمده‌اند.
- ارزش[9]: ارزش آن را ازنظر تصمیم‌گیری دارد یا نه و ارزش و فایده موردنظر را برای یک سازمان خواهند داشت.

۲- فناوری‌های وابسته

جهت درک عمیق کلان داده‌ها چند فناوری وابسته به کلان داده‌ها را معرفی می‌نماییم که ارتباط زیادی باکلان داده‌ها دارند.

۲-۱- رایانش ابری

رایانش ابری در یک معنای جامع به معنی شیوه‌ی تحویل و استفاده از سرویس‌ها است؛ یعنی سرویس‌های لازم را از طریق اینترنت به‌محض تقاضا یا به‌صورت یک روش قابل‌گسترش به دست می‌آورد. این سرویس ممکن است به نرم‌افزار، اینترنت یا چیزهای دیگر وابسته باشد. توسعه رایانش ابری راهکاری برای ذخیره‌سازی و پردازش کلان داده است. به‌عبارتی‌دیگر پیدایش کلان داده‌ها، توسعه رانش ابری را شتاب داده است. فناوری ذخیره‌سازی توزیع‌شده مبتنی بر رانش ابری، مدیریت مؤثر کلان داده‌ها را ممکن می‌سازد و ظرفیت محاسباتی موازی ایجادشده به‌وسیله رایانش ابری باعث بهبود بهره‌وری، اکتساب و تحلیل کلان داده‌ها می‌شود.[5]

۲-۲- اینترنت اشیا[10]

هدف اصلی IOT اتصال اشیای مختلف در جهان واقعی (اشیایی مانند RFID[11]، قرائت گرهای بارکد[12]، حسگرها، تلفن‌های همراه و غیره) به‌منظور تبادل اطلاعات و همکاری آن‌ها برای تکمیل یک وظیفه مشترک است. در الگوی IOT تعداد بسیار زیادی از حسگرهای شبکه در وسایل دنیای واقعی جاسازی‌شده است. این حسگرها که در زمینه‌های مختلف توسعه‌یافته‌اند ممکن است انواع مختلف داده‌ها را ازجمله داده‌های محیطی، جغرافیایی، نجومی و منطقی جمع‌آوری کنند. تجهیزات سیار، وسایل حمل‌ونقل، امکانات عمومی و وسایل خانگی همه می‌توانند تجهیزات اکتساب داده در IOT باشند.[5]

۲-۳- مرکز داده‌ها

یک مرکز داده‌ها انبوهی از داده‌ها دارد و داده‌ها را طبق هدف اصلی و مسیر توسعه‌ی آن، سازمان‌دهی و مدیریت می‌کند که این باارزش‌تر از داشتن منبع و مکان خوب است. پیدایش کلان داده‌ها، فرصت‌های توسعه و چالش‌های بسیاری را برای مراکز داده‌ها به وجود آورده است.[5]

۲-۴- هادوپ

هادوپ یک فناوری است که به کلان داده‌ها وابستگی نزدیکی دارد و از طریق ذخیره‌سازی داده‌ها، پردازش داده‌ها، مدیریت سیستم و یکپارچه کردن ماژول‌های دیگر یک راهکار قدرتمند اصولی کلان داده‌ها را شکل می‌دهد. هادوپ از دو بخش تشکیل‌شده

[6] Variety
[7] Velocity
[8] Veracity
[9] Value
[10] IOT
[11] Radio-frequency identification
[12] Bar code readers

ISSN: 2745-3673
http://www.Rcsj.ir

است:HDFS[۱۳] و چارچوب توزیع‌شده در حال اجرا روی سخت‌افزار تجاری است و بر اساس مرجع سیستم فایل توزیع‌شده گوگل طراحی‌شده است. در حال حاضر در کاربردهای کلان داده‌ها در صنعت به‌طور مثال: فیلتر کردن هرزنامه[۱۴]، جست‌وجوی شبکه، تحلیل جریان کلیک[۱۵] و توصیه گرهای اجتماعی[۱۶] از هادوپ به‌طور گسترده استفاده می‌شود.[5]

۳- کاربردها

کلان داده‌ها تقاضا برای متخصصان در این حوزه را به‌شدت بالابرده است و شرکت‌هایی چون اوراکل، ای‌بی‌ام، مایکروسافت هزینه‌های بسیاری را برای توسعه نرم‌افزارهای مدیریت و تحلیل داده سرمایه‌گذاری کرده‌اند. کلان داده‌ها نحوه کار سازمان‌ها و افراد را تحت تأثیر قرار می‌دهد. کلان داده‌ها فرهنگی را در سازمان‌ها ایجاد می‌کند که از طریق آن کسب‌وکارها و مدیران فناوری اطلاعات را به سمت استفاده از تمامی ارزش‌های پنهان در داده‌ها سوق می‌دهد. ادراک این ارزش‌ها به همه کارکنان سازمان‌ها این امکان را می‌دهد که یا بینش وسعی‌تری تصمیم‌گیری کنند، نزدیکی بیشتری با مشتریان داشته باشند، فعالیت‌های خود را بهینه کنند، با تهدیدات مقابله کنند و درنهایت سرمایه‌های خود را بر روی منبع جدیدی از سود سرشار پنهان در داده‌ها متمرکز سازند. سازمان‌ها برای رسیدن به این مرحله نیازمند معماری جدید، ابزارهای نو و فعالیت‌ها و تلاش مستمری هستند تا بتوانند از مزیت‌های چارچوب‌های مبتنی بر داده‌های بزرگ بهره‌مند گردند.[6]

۴- نمونه کاربردها

- حمل‌ونقل هوشمند: یکی از مهم‌ترین کاربردهای کلان داده تجزیه‌وتحلیل داده‌ها و فعال کردن دستگاه‌های حمل‌ونقل هوشمند است، سنسورهایی در دوربین جاده‌ها. سنسورهای خودرو داده‌های مربوط به تعداد خودروها در جاده، سرعت رانندگان، مجاورت وسایل نقلیه و غیره جمع‌آوری کرده و وضعیت ترافیک جاده و رفتار رانندگان را به دست می‌آورند.[7]
- معاملات: در هر ثانیه داده‌های مالی بسیاری در معاملات سهام، نرخ ارز، نرخ بهره، قیمت کالا و غیره به‌صورت پویا تولید می‌شوند. سازمان‌ها برای تشخیص فرصت‌ها و فرصت‌ها از آن‌ها استفاده می‌کنند؛ مانند پیش‌بینی کاهش و یا افزایش اوراق بهادار، فعالیت‌های غیرقانونی و تقلب، امنیت سرمایه‌گذاری، پیش‌بینی فرصت‌های مالی و غیره استفاده می‌شود.
- آموزش: کلان داده‌ها در صنعت آموزش می‌تواند به شخصی‌سازی فرآیند یادگیری کمک کند، موضوعی که تا قبل از پیدایش دستگاه‌های یادگیری الکترونیکی و جمع‌آوری داده‌های آموزشی نبود. این شخصی‌سازی به‌نوبه خود می‌تواند باعث شکوفایی استعدادهای دانش آموزان دانشجویان شود و پویایی محیط یادگیری را افزایش دهد.
- تولید: در صنعت تولید استفاده از کلان داده‌ها می‌توان کالا را طب نیازمندی‌های مشتریان تولید نمود و می‌توان خط تولید را به‌صورت بهینه طراحی کرد و عیوب آن را پیش از تولید کشف و برطرف کرد.
- خرده‌فروشی: در صنعت خرده‌فروشی با جمع‌آوری داده‌های فروش و تجارب مشتریان، آن‌ها را بهتر شناخت و تبلیغات و فروش را بهینه کرد.
- بهداشت و درمان: آنا کاوی کلان داده‌ها می‌تواند در صنعت بهداشت و درمان در قالب ارائه خدمات بهتر به عموم مردم کمک کند که این امر منجر به شناسایی روش‌های شخصی‌سازی‌شده برای درمان بیماران می‌شود. این شخصی‌سازی درمان می‌تواند به افزایش سلامت جامعه و کاهش هزینه‌های دولت در این بخش کمک کند.
- دولت: در این بخش استفاده از کلان داده‌ها منجر به کاهش هزینه‌ها، افزایش بهره‌وری و نیز ظهور و بروز نوآوری‌های جدید می‌شود؛ و نیازمند همکاری بخش‌های مختلف همچون وزارتخانه‌ها است. [10]

[13] Hadoop distributed file system
[14] Spam
[15] Clickstream analysis
[16] Social recommendation

ISSN: 2745-3673
http://www.Rcsj.ir

۵- چالش‌ها

در بحث کلان داده، ما نیاز داریم که داده‌ها را به‌منظور استخراج اطلاعات، کشف دانش و درنهایت تصمیم‌گیری در خصوص مسائل مختلف کاربردی به‌صورت صحیح مدیریت کنیم. مدیریت داده‌ها عموماً شامل ۵ فعالیت اصلی است که چالش‌هایی در آن‌ها وجود دارد.

- جمع‌آوری: بی‌گمان ضبط و نگهداری داده‌ها، یک دارایی ارزشمند برای سازمان‌ها به‌حساب می‌آید. درواقع اقتصاد داده مبتنی بر ارزش داده‌هایی است که از طریق تجزیه‌وتحلیل استخراج می‌شوند. بااین‌حال اعتقاد بر این است که برخلاف کالا، ارزش داده به نسبت مساوی از حجم آن رشد نمی‌کند. حتی اگر این دیدگاه اقتصادی در مورد داده درست باشد، برای این قاعده که حجم فزاینده و انواع مختلف داده فرصت‌های بیشتری برای استخراج ارزش‌افزوده فراهم می‌کند موردمحاسبه قرار نمی‌گیرد لذا قابلیت ضبط داده‌های ساختاریافته، شبه ساختاریافته و غیر ساختاریافته موجب تغییر در این فرضیات شده است .کلان داده‌ها، موجب تغییر روش‌های تجزیه‌وتحلیل داده‌ها از داده‌کاوی به تجزیه‌وتحلیل‌های پیشرفته شده‌اند.
- ذخیره‌سازی: با توجه به پویایی، غیر ساخت‌یافته بودن مقادیر و افزایش روزافزون حجم داده‌ها، باید زیرساخت‌های لازم جهت ذخیره‌سازی داده‌ها را فراهم آوریم که این مسئله چالش‌های زیادی را در رابطه با امنیت و هزینه‌ها به بار می‌آورد.
- جست‌وجو: اغلب، تجزیه‌وتحلیل کلان داده‌ها به‌عنوان «پیدا کردن یک سوزن در انبار کاه» معرفی‌شده است. اگرچه تصویر جالبی است ولی می‌تواند بسیار گمراه‌کننده باشد. اول موضوع یافتن سوزن نیست بلکه یافتن الماس است. با توجه به غیر ساخت‌یافتگی اکثریت داده‌ها نیازمند الگوریتم‌های پیچیده جهت این امر هستند.
- به اشتراک‌گذاری داده‌ها: این امر به دلیل حجم انبوه و غیر ساخت‌یافتگی امری مشکل است.
- تحلیل داده‌ها: در این حوزه همواره چالش‌های زیادی به دلیل ماهیت کلان داده‌ها و ویژگی‌های آن مطرح بوده و است. از چالش‌های روز آن می‌توان به تحلیل اطلاعات نیمه ساختاری و بدون ساختار اشاره نمود. یکی از روش‌های تحلیل اطلاعات در داده‌های بدون ساختار متنی استفاده از فراداده[17] است. برای مثال فردی در شبکه اجتماعی پیامی به ای شکل می‌نویسد «من از وضعیت پوشش شبکه تلفن همراه خود راضی نیستم درصورتی‌که در تبلیغات گفته‌شده بود بهترین شبکه را دارد. بهتر است سرویس‌دهنده خود را عوض کنم! » برای آگاه شدن از قصد مشتری شبکه‌های اجتماعی با استفاده از موتور استنتاج خود فراداده‌های کلیدی مانند:«سرویس‌دهنده »،«راضی نیستم »،«رضایت »، «قصد » را نشانه‌گذاری کرده و در لحظه می‌توانند داده‌ها را تحلیل کنند. مشخص است که کلان داده‌ها محدود به متن نبوده و شامل حجم عظیمی از تصاویر، صداها و ویدئو نیز است و همواره مبحث مدیریت فراداده‌ها به‌عنوان یکی از روش‌های تحلیل اطلاعات در کلان داده موضوعی جذاب است که نیاز به پژوهش بیشتری دارد.

۶- چشم‌انداز

پیدایش کلان داده‌ها، فرصت‌های زیادی را فراهم می‌کند. در عصر فناوری اطلاعات[18]، "T" یا فناوری نگرانی اصلی بود، درحالی‌که فناوری، توسعه داده‌ها را هدایت می‌کرد. در اصل کلان داده‌ها با برتری ارزش داده‌ها و پیشرفت در "I" یا اطلاعات، داده‌ها در آینده، پیشرفت فناوری‌ها را هدایت خواهند کرد. کلان داده‌ها نه‌تنها زندگی اجتماعی و اقتصادی را تغییر می‌دهند، بلکه در شیوه‌های زندگی و تفکر هرکسی تأثیر می‌گزارد که این فقط آغاز کار است. ما نمی‌توانیم آینده را پیش‌بینی کنیم اما می‌توانیم برای حوادث احتمالی که در آینده رخ می‌دهد اقدام احتیاطی انجام دهیم.

[17] Meta data
[18] IT

ISSN: 2745-3673
http://www.Rcsj.ir

۷-جمع‌بندی و پیشنهادها

با استفاده و تجزیه‌وتحلیل کلان داده‌ها به شاخه از علم به نام علم داده در حال تکامل است. علم اطلاعات شاخه‌ای از علم است که در مجموعه‌های عظیم داده بدون ساختار و نیمه ساختار به کشف و استنتاج دانش می‌پردازیم. علم داده انقلابی است که در حال تغییر جهان است در صنایع مختلف مانند امور مالی، خرده‌فروشی، بهداشت و درمان، تولید، ورزشی و ارتباطات، جستجو کردن موتور و بازاریابی دیجیتال شرکت‌هایی مانند گوگل، یاهو و بینک، شبکه مانند فیس بوک، توییتر و غیره است.

یکی از مسائل مهم در علم داده بهینه کردن سرعت جست‌وجو و مقادیری است که مورد انتظار کاربر است و مسئله دیگر مربوط به آسیب‌های مربوط به امنیت داده‌هاست که بستر مناسبی برای پژوهش در این زمینه است.

ISSN: 2745-3673
http://www.Rcsj.ir

منابع و مراجع

[۱] فرجامی، ی. مولانا پور، ر. هوش تجاری از ایده تا عمل چاپ اول، انتشارات ناقوس، تهران، ۱۳۹۰.

[۲] روحانی، س. حسینی، س. تحلیل‌های عظیم داده، چاپ اول، انتشارات نیاز دانش، تهران، ۱۳۹۴.

[3] J .Manyika, M .Chui, B .Brown, J .Bughin, R .Dobbs, C .Roxburgh, and A .H .Byers "Big data : The next frontier for innovation, competition, and productivity", McKinsey Global Institute, 2011.

[4] A .McAfee and E .Brynjolfsson "Big data :the management revolution ".Harvard business review, Vol .90, No .10, pp .60-68, 2012.

[5] composable architecture for rack scale big data computing august 2016, chung – sheng Li, Hubertus franke ,colin parris, bulent abali,mukil kesavan , victor chang.

[6] Giuseppe decandia, deniz hastorun, madan jampani, gunavardhan kakulapati, avinash lakshman, alex pilchin, swaminathan sivasubramanian, peter vosshall, and werner vogels. Dynamo: amazon highly available key-value store. In sosp, volume 7, pages 205-220,2007.

[7] A data quality in use model for big data original research article, October 2016, jorgemerino, is mael caballero, bibiano rivas,manuel Serrano, Mario piattini.

[8] M. Ferreira, R. Fernandes, H. Conceição, P. Gomes, P.M. d'Orey, L. Moreira - Matias, J. Gama, F. Lima, and L. Damas, "Vehicular Sensing: Emerge nce of a Massive Urban Scanner," In Sensor Systems and Software, pp. 1-14. Springer Berlin Heidelberg, 2015

[9] A. Mukherjee, P. Diwan, P. Bhattacharjee, D. Mukherjee, and P. Misra, "Capital market surveillance using stream processing," In 2nd International Conference on Computer Technology and Development (ICCTD), pp. 577-582. IEEE, 2014

[10] modeling and management of big data : challeng and opportunities, October 2016, david gil,il-yeol song.

[11] Sivarajah, Kama ,Irani and Weerakkody, "Critical analysis of Big Data challenges and analytical methods", Journal of Business Research , 2017.