مروری بر اهمیت وچالشهای کلان داده

سارا سکوت[1]، علیرضا نوروزی*[2]

۱- **دانشجوی کارشناسی ارشد نرم افزار، دانشگاه آزاد اسلامی، واحد اصفهان(خوراسگان)، اصفهان، ایران**

sarasokoot@gmail.com

۲- **عضو هیات علمی، استادیار ، دانشگاه آزاد اسلامی واحد شهر مجلسی، گروه کامپیوتر**

norouzi.arz@gmail.com

چکیده

در سال‌های اخیر، گسترش سریع اینترنت، اینترنت اشیاء و پردازش ابری منجر به رشد انفجاری داده‌ها تقریباً در هر فضای صنعتی و کسب‌وکار شده است. کلان‌داده به سرعت تبدیل به موضوعی داغ شده است که توجهی فراوان را از سمت دانشگاه، صنعت و دولت‌ها در سرتاسر جهان به سمت خود کشیده است. در این مقاله در ابتدا به شکلی خلاصه ایده و مفهوم کلان‌داده را به واسطه تعاریف، ویژگی‌ها و ارزشی که دارد، معرفی می‌کند. سپس از جانب دیدگاه‌های متفاوت اهمیتی که کلان‌داده در اختیارمان قرار می‌دهد را بررسی نموده و در نهایت چالش های پیش روی آن را جهت ذخیره سازی و تجزیه و تحلیل داده ها بررسی خواهد کرد.

کلمات کلیدی: کلان‌داده، اهمیت، تجزیه و تحلیل داده، چالش ها و چارچوب کلان داده.

۱–مقدمه

افزایش مداوم حجم و جزئیات داده های ثبت شده توسط سازمان ها، از قبیل طلوع رسانه های اجتماعی، اینترنت اشیاء، و رسانه های جمعی، منجر به ایجاد جریان قریب الوقوع داده ها هم در فرمت ساخت یافته، هم در فرمت غیر ساختیافته شده است. داده ها که از این پس، آن ها را داده های بزرگ می خوانیم با سرعت بی سابقه ای، ایجاد می شوند و روند رو به رشدی را دارند. چالش اصلی محققان و شاغلان این حوزه، این است که این نرخ رشد، از توانایی آن ها برای طراحی چارچوب مناسب به منظور انجام محاسبات برای تحلیل داده ها و به روز رسانی حجم کار فشرده، جلو زده است [1, 2] .

در سال‌های اخیر، کلان‌داده به سرعت تبدیل به نقطه‌ای کانونی گشته است به گونه‌ای که توجهی ویژه و فراوان را از سمت دانشگاه‌ها، صنعت و حتی دولت‌ها از سرتاسر دنیا به سمت خود می‌کشاند [3-5] . مک‌کینزی[1]، یک شرکت معروف در زمینه مدیریت و مشاوره، بر این باور است که کلان‌داده به درون هر فضای صنعتی و توابع کسب وکار نفوذ کرده است و تبدیل به عامل مهم در تولید شده است[6]. استفاده و استخراج کلان‌داده خبر از موج جدیدی از رشد بهره‌وری و انگیزه‌های مصرفی می‌دهد. بعضی حتی می‌گویند که کلان‌داده می‌تواند به مثابه نفت خام تازه‌ای در نظر گرفته شود که به اقتصاد اطلاعاتی آینده قدرت و توان می‌بخشد.

[1] McKinsey

۲- تعریف و اهمیت کلان داده

تا کنون هیچ تعریفی وجود ندارد که به شکلی جهانی پذیرفته شده باشد. در ویکی‌پدیا، کلان‌داده به شکل "اصطلاحی کاملاً جامع برای مجموعه‌ای از هر دسته‌های اطلاعاتی تعریف می‌شود که بسیار بزرگ و پیچیده‌اند که انجام پردازش با استفاده از ابزار پردازش صنعتی بر رویشان کاریست بسیار سخت"[7]. از دیدگاه کلان، کلان‌داده را می‌توان به عنوان قید یا پیوستگی در نظر گرفت که به شکلی دقیق دنیای فیزیکی، جامعه انسانی و فضای سایبری را به هم متصل می‌کند. اینجا دنیای فیزیکی بازتابی در فضای سایبری دارد، به واسطه اینترنت، اینترنت اشیاء و دیگر تکنولوژی‌های اطلاعاتی به شکل کلان‌داده تجسم می‌یابد، در حالیکه جامعه انسانی کلان‌داده خود را مبتنی بر انگاشت در فضای سایبری به واسطه مکانیسم‌هایی مانند ارتباط انسان-کامپیوتر، ارتباط مغز-ماشین و اینترنت موبایل تولید می‌کند [8-10] . در این معنا، کلان‌داده می‌تواند اساساً در دو دسته جای گیرد که عبارتند از داده‌هایی از دنیای فیزیکی که معمولاً به واسطه سنسورها، آزمایشات و مشاهدات علمی (مانند داده‌های بیولوژیکی، داده‌های عصبی، داده‌های نجومی و داده‌های سنجش از راه دور) و داده‌هایی از جامعه انسانی که اغلب حاصل چنین منابع یا دامنه‌هایی مانند شبکه‌های اجتماعی، اینترنت، سلامت، امور مالی، اقتصاد و حمل‌ونقل هستند، به دست می‌آیند.

کلان داده، بر اساس اهمیت و ارزش بسیار زیادی که دارد، اساساً شیوه زندگی، کار و تفکرمان را تغییر می‌دهد و دگرگون می‌سازد [3] . در ادامه جزئیات کامل اهمیت کلان‌داده به لحاظ دیدگاه‌های مختلف توصیف شده است:

۲-۱- اهمیتی که برای رشد شهروندان دارد

اکنون، دنیا به شکلی کامل وارد عصر اطلاعات شده است. استفاده گسترده از اینترنت، اینترنت اشیاء، پردازش ابری و اشکال مختلفی از ظهور تکنولوژی‌های IT باعث افزایش منابع مختلف اطلاعات در اندازه و مقیاسی بی‌سابقه شده است در حالیکه ساخت و ایجاد ساختارها و گونه‌های مختلف داده‌ای هر روز، بیش از پیش، پیچیده‌تر می‌شود. تجزیه و تحلیل و به کارگیری دقیق و عمیق کلان‌داده نقشی مهم را در ارتقاء و گسترش رشد پایدار اقتصادی کشورها بازی خواهد کرد و توانایی رقابتی شرکت‌ها را افزایش خواهد داد.

در آینده، کلان‌داده به نقطه جدیدی در رشد اقتصادی تبدیل خواهد شد. با کلان‌داده، شرکت‌ها ارتقاء پیدا خواهند کرد که به موجب آن فناوری اطلاعات و دیگر صنایع را تغییر می‌دهند. در چنین زمینه‌ای، غول‌های جهانی صنعت IT (از جمله IBM، Google، Microsoft و Oracle) پیش از این شروع به برنامه‌ریزی در راستای گسترش تکنیکی در عصر کلان‌داده نموده‌اند.

در سطح ملی، ظرفیت انباشت، پردازش و به کارگیری حجم بسیار زیادی از داده‌ها به یک نشان اختصاصی جدید در توانایی و قدرت خاص کشورها تبدیل خواهد شد. پادشاهی و سلطه اطلاعاتی یک کشور در فضای سایبری فضایی دیگر در بازی قدرت در کنار، زمین، دریا، هوا و فضاهایی از این دست خواهد بود[11].

به طور کلی، کشورهای غربی به نمایندگی ایالات متحده، تحت زمان‌بندی ملی‌شان به سمت مدرنیته سازی قدرت و توان ملی‌شان به واسطه پژوهش‌های مربوط به کلان‌داده و نرم‌افزارهای کاربردی مربوط به آن پیشروی می‌کنند. پیش‌بینی می‌شود که رقابت‌های اقتصادی و سیاسی آینده در این کشورها مبتنی بر بهره‌وری از قابلیت‌ها و پتانسیل کلان‌داده در میان دیگر جنبه‌های سنتی، خواهد بود. به شکلی خلاصه، این پژوهش‌های مربوط به کلان‌داده و به کارگیری آن اهمیتی استراتژیک برای بهبود قدرت و توانایی رقابتی هر کشور خواهد داشت.

۲-۲-اهمیتی که برای به روزرسانی‌های صنعتی دارد

پژوهش بر روی مشکلات و مسائل مشترک کلان‌داده، به ویژه بر روی از میان برداشتن موانع پیش روی فنّاوری‌های هسته‌ای، صنایع را قادر خواهد ساخت پیچدگی را مهار سازند که به هم پیوستگی اطلاعاتی را شامل می‌شود و بر عدم

قطعیت‌هایی تسلط یابند که حاصل افزونگی و یا کسری اطلاعات هستند. هر کسی خواهان رسیدن به اطلاعات مبتنی بر تقاضا، دانش و حتی آگاهی به واسطه کلان داده است و امیدوار است تا در نهایت بتواند به سودمندی کاملی برسد که حاصل ارزش کلان کلان‌داده‌هاست. چنین موردی بدین معناست که داده‌ها دیگر محصول جانبی و ثانویه بخش صنعتی نیستند اما تبدیل به یک پیوندی کلیدی در تمامی جنبه‌ها شده‌اند. در این معنا، مطالعه و پژوهش بر روی مشکلات مشترک و فنّاوری‌های هسته کلان‌داده بر روی نسل جدیدی از IT و کاربردهایش متمرکز خواهد شد. چنین موردی تنها ماشین و اسباب جدیدی برای نگهداری رشد سریع صنعت اطلاعات نیست بلکه ابزار جدیدی برای صنایع برای بهبود قابلیت‌های رقابتی‌شان است.

۳-۲- اهمیتی که برای پژوهش‌های علمی به دنبال دارد

به خوبی آشکار است که اولی پژوهش‌های علمی در تاریخ بشر مبتنی بر تجربیات و آزمایشات بوده است. بعدها، دانش نظری رخ نمود که به واسطه مطالعه بر روی قوانین و قواعد مختلف شناخته می‌شود. با این حال، به این دلیل که تجزیه و تحلیل نظری بسیار پیچیده است و حل کردن مشکلات عملی امکان‌پذیر نیست، انسان جستجوی برای روش‌های مبتنی بر شبیه‌سازی را آغاز کرد که به دانش محاسباتی منتهی شد.

پیدایش ناگهانی کلان‌داده پارادایم پژوهشی جدیدی را به دنبال داشت؛ بر این اساس، در کنار کلان داده، پژوهشگران می‌توانند تنها نیازمند یافتن یا جستجو به دنبال اطلاعات، دانش و آگاهی مورد نیاز باشند. آن‌ها حتی نیازی به دسترسی مستقیم به موضوعات مورد پژوهش ندارند. در سال ۲۰۰۷،جیم گری،[1] در آخرین سخنرانی خود پارادایم چهارم از پژوهش علمی متمرکز بر داده را شرح داد که علم متمرکز بر اطلاعات را از دانش محاسباتی جدا کرد[12]. گری بر این باور بود که پارادایم چهارم می‌تواند تنها راه سیستماتیک برای حل بعضی از سخت‌ترین چالش‌های جهانی باشد که امروزه با آن‌ها مواجهیم. در اساس، پارادایم چهارم نه تنها یک تغییر در پژوهش علمی است بلکه تغییر در راهی است که افراد بر آن اساس فکر می‌کنند [3].

۴-۲-اهمیتی که برای پیدایش پژوهش‌های میان‌رشته‌ای دارد

فنّاوری‌های کلان‌داده و مطالعاتی بنیادین وابسته به آن تبدیل به نقطه کانونی مطالعات در حوزه دانشگاهی شده‌اند. یک رشته میان‌رشته‌ای در حال ظهور که به نام دانش اطلاعات [13] خوانده می‌شد، به تدریج وارد میدان شد. این رشته کلان‌داده را به مثابه موضوع پژوهشی خود در نظر گرفت و قصد دارد استخراج و برداشت دانش از داده و اطلاعات را عمومیت ببخشد. این رشته بر روی رشته‌های بسیاری مانند دانش اطلاعات، ریاضیات، علوم اجتماعی، دانش شبکه، دانش سیستم، روان‌شناسی و اقتصاد [15 , 14] پل بسته است. این رشته تکنیک‌ها و تئوری‌های متفاوتی را از بسیاری زمینه‌ها شامل پردازش سیگنال، نظریه احتمالات، فراگیری ماشین، دانش آماری، برنامه‌ریزی کامپیوتر، مهندسی اطلاعات، الگوشناسی، تجسم، مدلسازی عدم قطعیت، انبار داده‌ها و محاسبات با کارای بالا به کار گرفته است.

۵-۲-اهمیتی که برای کمک به درک بهتر شرایط افراد می‌کند

کلان‌داده، به ویژه کلان‌داده تحت شبکه، حاوی قدرت اطلاعات اجتماعی است و از این رو می‌تواند به شکل شبکه‌ای برای نقشه‌برداری اجتماعی در نظر گرفته شود. برای این منظور، با تجزیه و تحلیل کلان‌داده و خلاصه بیشتر و یافتن کلیدها و مدارک و قوانینی که حاوی آن است می‌تواند به ما کمک کند تا درک بهتری از شرایط حاضر داشته باشیم.

برای مثال، دو شاخص نمونه از علایقی که در چین توسعه یافته است استفاده‌ای بزرگ از داده‌هایی را ممکن می‌سازد که به شکلی عمومی از طریق اینترنت در دسترس هستند. مرکز زمینه‌یابی و ارزیابی چین مرتبط با دانشگاه رنمین چین، "شاخص پیشرفت چین" را به صورت سالانه منتشر کرده است. این شاخص، با چهار شاخص جداگانه مربوط به سلامت، تحصیلات، استاندارد زندگی و محیط اجتماعی، قصد دارد وضعیت موجود و حل مسائل مربوط به پیشرفت چین را اندازه‌گیری کند. با

[1] Jim Gray

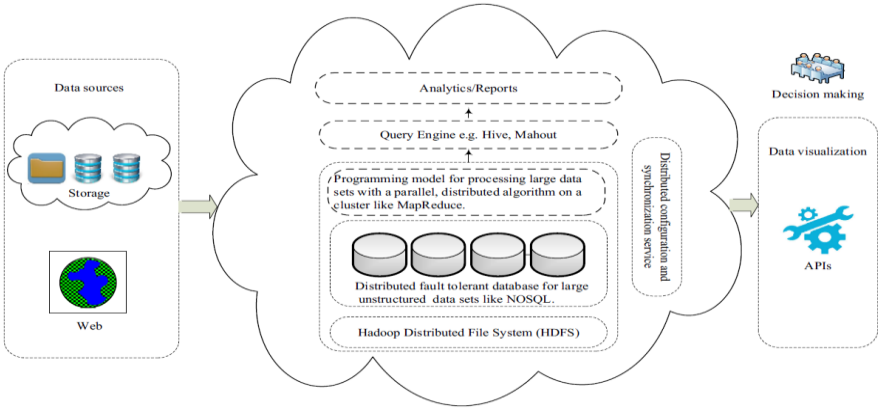
مقایسه و تجزیه و تحلیل شاخص‌های عینی و ذهنی مختلف و ترکیب تجزیه و تحلیل کمّی و کیفی، این شاخص وضعیت‌ها و قوانین پیشرفت حاضر مراکز مالی بین‌المللی را آشکار می‌کند.
استخراج دقیق و کامل اطلاعات موجود در کلان‌داده می‌تواند به افراد در گرفتن تصمیماتی بهتر کمک کند

۶-۲- اهمیتی که برای کمک به پیش‌بینی بهتر افراد در مورد آینده دارد

قابلیت داده بزرگ شبکه به شدت رو به افزایش است و به نحوی موثر در زمینه‌های امنیت و نظامی به کار برده می‌شود [16]. تجزیه و تحلیل پیشگویانه مبتنی بر کلان‌داده برای پیش بردن موضوعات اجتماعی شامل سلامت عمومی و توسعه اقتصادی نیز به کار گرفته می‌شود. گینسبرگ و همکاران دریافتند که اگر حجم سوالاتی با کلمات کلیدی مانند "علائم سرماخوردگی" و "درمان سرماخوردگی" که در Google ثبت می‌شود در یک منطقه خاص رو به افزایش بگذارد، تنها پس از چند هفته، نتیجتاً، تعداد بیماران مبتلا به آنفولانزا در بخش اورژانس بیمارستان‌های مربوط به همان منطقه رو به افزایش خواهد گذاشت [17]. با این کشف، آن‌ها قادر به پیش‌بینی شیوع آنفولانزا و به کاربستن اقدام متقابل در زمانی زودتر از آنچه انتظار می‌رود، خواهند بود. در توسعه اقتصادی، سازمان ملل متحده اخیراً پروژه جدیدی را راه‌اندازی کرده است که با نام Global Pulse [18]. شناخته می‌شود و انتظار بر این است که از کلان‌داده برای ترویج و گسترش توسعه اقتصادی جهانی بهره ببرد. سازمان ملل متحد تجزیه و تحلیلی به اصطلاح عاطفی را هدایت خواهد کرد که از نرم‌افزار پردازش زبان طبیعی برای تجزیه و تحلیل پیام‌های متنی در سایت‌های مربوط به شبکه‌های اجتماعی به منظور پیش‌بینی مسائل اجتماعی مانند نرخ بیکاری، کاهش هزینه‌ها و شیوع بیماری‌ها در ناحیه‌ای خاص بهره می‌برد. هدف کلی چنین برنامه‌ای به کارگیری سیگنال‌ها و پیام‌های هشداردهنده اولیه دیجیتالی برای راهنمایی و پیشبرد پروژه‌های پشتیبانی به منظور جلوگیری از به دام افتادن یک منطقه خاص در دام فقر است.

۳- پردازش داده های بزرگ

داده های بزرگ، برای کاربرانش، توانایی استفاده ی راحت از محاسبات مناسب را برای پردازش پرس و جوهای پراکنده در سرتاسر مجموعه داده های چندگانه و بازگشتی، به شیوه ای زمان بندی شده فراهم کرده است. محاسبات ابری، موتوری متضمن را از طریق استفاده از Hadoop کلاسی از چارچوب های پردازش داده های توزیع شده فراهم می کند. استفاده از محاسبات ابری در داده های بزرگ، در شکل ۱، نشان داده شده است. منابع داده های عظیم از ابر و وب در یک پایگاه داده ی توزیع شده تحمل پذیر خطا ذخیره شده است و از طریق مدل برنامه نویسی برای مجموعه داده های بزرگ، با یک الگوریتم توزیع شده موازی در یک کلاستر پردازش شده است. هدف اصلی بصری سازی داده ها همچنان که در شکل ۱مشاهده می کنید. دیدن نتایج تحلیلی است که با استفاده از نمودارهای مختلف، به منظور تصمیم گیری، به تصویر کشیده شده اند[19].



شکل ۱- استفاده از محاسبات ابری در داده های بزرگ

داده های بزرگ از تکنولوژی ذخیره سازی توزیع شده استفاده می کند که به جای اینکه ذخیره ی محلی متصل به کامپیوتر یا دستگاه الکترونیک، بر اساس محاسبات ابری است. ارزیابی داده های بزرگ، توسط اپلیکیشن های مبتنی بر ابر توسعه یافته با استفاده از تکنولوژی های مجازی شده انجام شده است. بنابراین، محاسبات ابری نه تنها تسهیلاتی برای محاسبات و پردازش داده های بزرگ فراهم می کند، بلکه به عنوان یک مدل خدماتی نیز عمل می کند. جدول۱، مقایسه ی ارائه دهندگان ابری داده های بزرگ متعدد را نشان می دهد.

جدول۱- مقایسه ی چند پلت فرم ابری داده های بزرگ

	Google	Microsoft	Amazon	Cloudera
ذخیره ی داده های بزرگ	خدمات ابر Google	Azure	S3	
Map Reduce	App Engine	Hadoop on Azure	Elastic Map Reduce (Hadoop)	Map Reduce YARN
تحلیل داده های بزرگ	Big Query	Hadoop on Azure	Elastic Map Reduce (Hadoop)	Elastic Map Reduce (Hadoop)
پایگاه داده ی رابطه ای	Cloud SQL	SQL Azure	MySQL or Oracle	MySQL, Oracle, PostgreSQL
پایگاهNoSQL داده	AppEngine Data store	ذخیره ی جدول	Dynamo DB	Apache Accumulate
Streaming پردازش	Search API	Stream insight	هیچ چیزی از قبل بسته بندی نشده	Apache Spark
یادگیری ماشینی	پیش بینیAPI	Hadoopþ Mahout	HadoopþMahout	Hadoop þ Oryx
Data import	شبکه	شبکه	شبکه	شبکه
منابع داده ها	چند مجموعه داده ی نمونه	Windows Azure marketplace	مجموعه داده های عمومی	مجموعه داده های عمومی
قابلیت دسترسی	چند سرویس در بتای خصوصی	چند سرویس در بتای خصوصی	تولید عمومی	صنایع

تالیا در باره ی پیچیدگی و تنوع انواع داده ها و قدرت پردازش برای انجام تحلیل و آنالیز بر روی مجموعه داده های عظیم، به بحث و گفتگو می نشیند. نویسنده، اظهار داشت که زیرساخت محاسبات ابری می تواند به عنوان چارچوب ی برای حل و فصل مشکل ذخیره ی داده های مورد نیاز برای انجام تحلیل و آنالیز بر روی داده های بزرگ، عمل کند.

محاسبات ابری با الگوی جدید توشه ی زیرساخت محاسباتی و روش پردازش داده های بزرگ برای همه ی منابع موجود در ابر –از طریق تحلیل داده ها– همبستگی دارد. تکنولوژی های مبتنی بر ابر بسیاری باید خود را با این محیط جدید سازگار کنند زیرا کنارآمدن با داده های بزرگ برای پردازش همروند، به طرز فزاینده ای پیچیده شده است[20].

Map Reduce مثال خوبی از پردازش دادهای بزرگ در یک محیط ابری است؛ Map Reduce، پردازش میزان زیادی از مجموعه داده های ذخیره شده به صورت موازی در کلاستر را مجاز می داند. محاسبات کلاستر، عملکرد خوبی در محیط های سیستمی پراکنده – از قبیل قدرت، ذخیره، و شبکه ی ارتباطی کامپیوتر– نشان داده است. به همین ترتیب، بولی یر و فایرستون بر توانایی محاسبات کلاستر برای ارائه ی بستر مهمان نوازی برای رشد داده ها، تاکید کرده اند. اما میلر در این باره بحث می کند که فقدان در دسترس بودن داده ها، گران تمام می شود، زیر کاربران تصمیمات بیشتری برای روش های تحلیلی offload کردند؛ استفاده ی نادرست از روش ها یا ضعف های ذاتی در روش ها، می تواند منجر به تصمیم های غلط یا هزینه بر شود. DBMS ها به عنوان بخشی از معماری محاسبات ابری جاری در نظر گرفته می شود و در حصول اطمینان از اینکه تبدیل اپلیکیشن ها از زیرساخت های قدیمی شرکت ها به سوی معماری های زیرساخت های ابری جدید به آسانی

انجام می شـود، نقش مهمی ایفا می کند. فشـار سـازمان ها برای انطباق سریع خود با شرایط جدید و پیاده سازی تکنولوژی های جدید، از قبیل محاسـبات کامپیوتری برای روبه رو شـدن با چالش ذخیره و پردازش داده های بزرگ، عواقب و ریسـک های غیرمنتظره ی طولانی را می طلبد[21].

۴- چالش های کلان داده

اگرچه محاسبه ابری بطرز گسترده‌ای توسط بسیاری از سازمان ها پذیرفته شده است، تحقیق در مورد داده های زیاد در ابر در مراحل اولیه ی خود باقی مانده اسـت. چندین موضـوع موجود بطرز کاملی کنترل و هدایت نشـده اند. به علاوه، چالش های جدیدی از برنامه‌های کاربردی توسط سازمان ها پدیدار شده اند. در بخش‌های بعدی، در مورد برخی از چالش های تحقیقاتی کلیدی، نظیر مقیاس‌پذیر بودن، دسـترس پذیری، یکپارچگی داده‌ها، تبدیل داده‌ها، کیفیت داده‌ها، ناهمگونی داده ها، مسـائل پوشیدگی و قانونی، و کنترل باقاعده، بحث شده است.

۴-۱- مقیاس‌پذیری

مقیاس‌پذیری توانایی ذخیره برای کنترل مقادیر افزایش یافته ی داده ها به شـیوه ای مناسـب است. سیستم های ذخیره داده های توزیع شـده مقیاس‌پذیری به بخشـی حیاتی از زیرسـاختار محاسبات ابری تبدیل شده است. فقدان ویژگی‌های محاسبه ابری برای پشـتیبانی از RDBMS های مرتبط با راه‌حل های مشارکتی جذابیتRDBMSها را برای توسعه ی برنامه‌های کاربردی مقیاس بزرگ در ابر کمتر کرده است.این نقص باعث محبوبیت NoSQL شده است[22].

یک پایگاه داده‌ی NoSQL مکانیسمی برای ذخیره و بازیابی حجم زیاد داده های توزیع شده ارائه می‌کند. ویژگی‌های پایگاه داده‌ی SQL شامل حالت های عاری از طرح (شما/الگو)، پشتیانی از تکثیر آسان، API ساده، و حالت سازگار و انعطاف پذیر می‌باشـد. انواع مختلفی از پایگاه داده های NoSQL، نظیر مقدار کلید، متمایل به سـتون، و متمایل به سـند، از داده های بزرگ پشتیبانی میکند[23].

۴-۲- دسترس‌پذیری

دسـترس‌پذیری اشـاره به منابع سیستم قابل دسترس توسط فرد مجاز دارد. در یک محیط ابری، یکی از موضوعات اصلی در رابطه با ارائه دهندگان خدمات ابری دسترس پذیری داده های ذخیره شده در ابر است.

با تعداد رو به رشـد کاربران ابری، ارائه دهندگان سرویس ابری باید موضوع در دسترس قرار دادن داده های موردنیاز کاربران را برای ارائه ی خدمات با کیفیت بالای خود حل کنند[24].

۴-۳- یکپارچگی داده‌ها

جنبه ی کلیدی امنیت داده ها یکپارچگی اسـت. یکپارچگی بدین معناست که داده ها می توانند تنها توسط اشخاص مجاز یا مالک داده ها دستکاری شوند تا از سوء استفاده ی احتمالی از داده ها جلوگیری به عمل آید. تکثیر برنامه های کاربردی ابری به کاربران این امکان را می دهد تا بتوانند داده هایشان را در مراکز داده ای ابری ذخیره و مدیریت نمایند. چنین برنامه هایی باید از یکپارچگ داده ها اطمینان حاصل کنند. اما، یکی از چالش های اصلی که باید هدایت و مدیریت شود اطمینان یافتن از صحت داده های کاربر در ابر است[25].

۴-۴- تبدیل (تغییر)

تبدیل داده ها به شکل مناسبی برای تحلیل مانع پذیرش داده های بزرگ است.

در مورد داده ای ساختاریافته، داده ها قبل از ذخیره شدن در پایگاه داده های منطقه پیش پردازش شده اند تا محدودیت های شمای در حال چاپ مرتفع شوند. داده ها پس از آن می توانند برای تحلیل بازیابی گردند. اما، در داده های ساختارنیافته، داده ها باید ابتدا در پایگاه داده های توزیع شده ذخیره شوند، نظیر HBase، و این کار باید قبل از پردازش آنها برای تحلیل انجام گیرد[26].

۴-۵- کیفیت داده ها

در گذشته، کار پردازش داده ها معمولاً در مجموعه های داده ای روشنی از منابع مشهور و محدود شده انجام می گرفت. بنابراین، نتایج دقیق بودند. اما، با پیدایش داده های عظیم، داده ها از منابع بسیار مختلفی نشئت می گیرند؛ تمام این منابه شناخه شده یا قابل تأیید نیستند. کیفیت داده ای ضعیف به مسأله ای جدی برای ارائه دهندگان خدمات ابری تبدیل شده است چون داده ها اغلب از منابع مختلفی جمع آوری شده اند. به عناون مثال، مقادیر زیادی از داده ها از تلفن های هوشمند تولید شده اند، که در آن ها قالب های داده ای متناقضی می توانند تولید شوند و این نتیجه ی منابع ناهمگون است. مسأله ی کیفیت داده ها معمولاً بعنوان هر صعبی که در مورد یک یا چند بعد کیفی با آن مواجه می شویم تعریف شده که داده ها را برای استفاده منتقل می کند. بنابراین، بدست آوردن داده های با کیفیت از مجموعه های گسترده از منابع داده ای خود یک چالش است[27].

۶-۴- ناهمگنی

گوناگونی یکی از جنبه های اصلی داده نویسی وسیع است که در نتیجه رشد در منابع مختلف داده های واقعا نامحدود است . این رشد موجب نا همگنی در ماهیت داده وسیع میشود . داده هایی از منابع گوناگون عموما در انواع مختلف و به شکل های نمایشی وعمدتا بهم پیوسته . در شکل هایی نا سازگار و به طور ناهماهنگی نمایش داده میشوند . چالش بر سر چگونگی اداره کردن انواع منبع داده های گوناگون است [28].

۷-۴- حریم

اهمیت حریم مانع کاربرانی میشود که داده های خصوصیشان را در اینترنت پخش می کنند. این نگرانی با گسترش استخراج داده های وسیع وتحلیل هایی که نیازمند اطلاعات شخصی برای تولید نتایج مربوط مانند سرویس های شخصی و موقعیت بنیاد جدی تر شده است [28].

۵- نتیجه گیری

کلان‌داده اثری بسیار قوی، تقریباً بر روی هر بخش و صنعتی در دنیای امروز به جای گذاشته است. روند تولید سریع کلان داده یک نوع پیشی گرفتن از رقبا درنظر گرفته می‌شود و اگر یک تجارتی قادر به تجزیه و تحلیل اطلاعات موجود در بخش داده باشد قادر خواهد بود تا مشتریان بیشتری جذب نماید و هزینه هایش را نیز کاهش دهد. با این حال تحلیل کلان داده هنوز چالشی است و نیازمند وظایف نرم افزارهای گران قیمت است و زیرساخت محاسباتی عظیمی را نیاز دارد .

این مقاله به شکلی خلاصه فرصت‌ها و اهمیتی را که کلان‌داده به خود اختصاص می‌دهد بررسی نموده و در کنار آن بعضی چالش‌های بزرگی را مطرح نموده که از جمله می توان به مقیاس‌پذیری، دسترس‌پذیری، یکپارچگی داده‌ها، تبدیل، کیفیت داده ها، ناهمگنی و حریم اشاره نمود.

از بررسی مطالب بالا ابن نتیجه حاصل می‌شود که راه حل‌های زیادی در رابطه با کلان داده ارائه شده است که مرتبط با محاسبات ابری است مانند راه حل‌های زیادی که بخاطر نرخ زیاد تحلیل و کاربران غیرمجرب و تحلیل ترافیکی ارائه گرده‌یده است. تحلیلات نیز می‌توانند بصورت پیش بینی کننده و توصیفی باشند. همچنین واضح است که پردازش امری پیچیده است که نیازمند افرادی با زمینه تخصصی تمیزکردن داده و فهم و انتخاب متدهای مناسب و تحلیل نتایج باشد پس ابزارها برای رسیدن به این اهداف تعبیه شده اند.

موضوعات کلیدی که در فضای اینترنتی کلان داده تاکید شده، چالش های چشمگیری دارد که باید در آینده توسط هیات های علمی و صنعتی بررسی و حل شود. محققان و دانشمندان علوم اجتماعی نیز باید در اعتماد سازی بلند مدت در دستیابی به مدیریت داده در فضای فناوری اینترنتی با هم کار کنند و بدنبال راه های های جدیدی برای جستجو باشند.

۶- مراجع

1.R.L .Villars, C.W. Olofson, M. Eastwood, Big data: what it is and why you should care, White Paper, IDC, 2011, MA, USA.

2.R. Cumbley , P. Church, Is Big Data creepy Compute. Law Secur. Rev. 29 (2013) 601–609.

3.V. Mayer-Schonberger , K. Cukier , Big Data: A Revolution That Will Transform How We Live, Work, and Think, Houghton Mifflin Harcourt, 2013.

4.R. Thomson, C. Lebiere , S. Bennati , Human, model and machine: a complemen-tary approach to big data, in: Proceedings of the 2014 Workshop on Human Centered Big Data Research, HCBDR '14, 2014.

5.A. Cuzzocrea , Privacy and security of big data: current challenges and future research perspectives, in: Proceedings of the First International Workshop on Privacy and Security of Big Data, PSBD '14, 2014.

6.J. Manyika, M. Chui, B. Brown, J. Bughin, R. Dobbs, C. Roxburgh, A. Hung, Big data: the next frontier for innovation, competition, and productivity, Tech. rep., McKinsey Global Institute, 2011, available at:http://www.mckinsey.com/insights/business_technology/big_data_the_next_frontier_for_innovation.

7.Big data, http://en.wikipedia.org/wiki/Big_data, 2014.

8. G. Li, X. Cheng, Research status and scientific thinking of big data, Bull. Chin. Acad. Sci. 27(6) (2012) 647–657.

9. Y. Wang, X. Jin Xueqi, Network big data: present and future, Chinese J. Comput. 36(6) (2013) 1125-1138.

10. X.-Q. Cheng, X. Jin, Y. Wang, J. Guo, T. Zhang, G. Li, Survey on big data system and analytic technology, J. Softw. 25(9) (2014) 1889–1908.

11. T. Kalil, Big data is a big deal, available at: http://www.whitehouse.gov/blog/2012/03/29/big-data-big-deal, 2012.

12.T. Hey, S. Tansley, K. Tolle (Eds.), The Fourth Paradigm: Data-Intensive Scientific Discovery, Microsoft Corporation, 2009.

13.Data science, http://en.wikipedia.org/wiki/Data_science, 2014.

14.M. Loukides, What Is Data Science?, O'Reilly Media, Inc., 2011.

15.M.A. Stokes, China's nuclear warhead storage and handling system, Tech. rep., 2049 Project Institute, March 2010.

16. I.M. Easton, L.R. Hsiao, The Chinese people's liberation army's unmanned aerial vehicle project: organizational capacities and operational capabilities, Tech. rep., 2049 Project Institute, March 2013.

17. J. Ginsberg, M.H. Mohebbi, R.S. Patel, L. Brammer, M.S. Smolinski, L. Brilliant, Detecting influenza epidemics using search engine query data, Nature 7232 (2009) 1012–1014.

18. Big data for development: challenges & opportunities, available at: http://www.unglobalpulse.org/projects/BigDataforDevelopment, May 2012.

19. D. Talia,Cloudsforscalablebigdataanalytics,Computer46(2013) 98–101.

20.C.Ji,Y.Li,W.Qiu,U.Awada,K.Li,Bigdataprocessingincloud computing environments, Pervasive Systems, Algorithm sand Net- works(ISPAN),2012,in:Proceedingsofthe12th International Symposium on, IEEE, 2012, pp.17–23.

21. J. Dean, S. Ghemawat, Map Reduce: simplified data processing on large clusters, Common. ACM51(2008)107–113.

22. P. Mell, T. Grance, The NIST definition of cloud computing (draft), NIST Spec. Publ. 800 (2011) 7.

23. R.Cattell,Scalable SQL and NoSQL datastores, ACMSIGMOD Record, 39(4), ACM NewYork, NY, USA, 2011, 12–27.

24. D. Zissis, D.Lekkas, Addressing cloud computing security issues, Futur. Gener. Comput. Syst. 28(2012)583–592.

25. R. Sravan Kumar, A. Saxena, Data integrity proofsin cloud storage, in: ProceedingsoftheThirdInternationalConferenceonCommu- nication Systemsand Networks(COMSNETS), 2011,pp.1–4.

26. R. Akerkar,BigDataComputing,CRCPress,2013.

27. T.C.Redman,A.Blanton,DataQualityfortheInformationAge, Artech House,Inc.,Norwood,MA,USA,1997.

28.D. Che, M.Safran,Z.Peng,Frombigdatatobigdatamining: challenges, issues,andopportunities,in:B.Hong,X.Meng,L.Chen, W.Winiwarter, W.Song(Eds.), DatabaseSystemsforAdvanced Applications, Springer,BerlinHeidelberg,2013,pp.1–15.