بررسی تطبیقی، کاربردها و چالش‌های فناوری‌های تحلیل بزرگ داده

یاسر قاسمی‌نژاد*
دانشگاه امام‌حسین (ع)، تهران، ایران
yaserghn@gmail.com

سیدعباسعلی کتابچی**
دانشگاه آزاد واحد تهران مرکزی، تهران، ایران
ketabchi@iran.ir

تاریخ دریافت: ۱۳۹۸/۰۲/۱۸
تاریخ اصلاحات: ۱۳۹۸/۰۵/۲۳
تاریخ پذیرش: ۱۳۹۸/۰۶/۱۳

چکیده

امروزه سازمان‌ها، با به‌کارگیری فناوری بزرگ داده، از طریق دریافت و به اشتراک‌گذاری ساده‌تر و ارزان‌تر اطلاعات، قادر بـه اداره حجم زیادی داده‌ها، با سرعت و تنوع زیاد شده‌اند. فناوری داده‌های عظیم، در صورت حل صحیح مشکلات مـرتبط، فرصـت‌هـای زیادی را فراهم می‌کنند. فناوری‌های گذشته، در پردازش داده‌های موجود برای مواجهه بـا مقـادیر زیـاد داده‌هـای تولیـد شـده، مناسب نیستند. درصورتیکه قالب‌های پیشنهادی برای کاربردهای بزرگ داده، به ذخیـره، تجزیـه و تحلیـل و پـردازش داده‌هـای عظیم کمک می‌کنند. در این تحقیق، ابتدا تعاریف و چالش‌های بزرگ داده، بررسی شده و سپس تعدادی از چارچوب‌های بـزرگ دادهٔ موجود (هادوپ، فلینک، استورم، اسپارک و سمزا)، مورد مطالعه و مقایسـه تطبیقـی قـرار گرفتـه اسـت. چـارچوب بـزرگ داده‌های مورد مطالعه، به طور کلی در دو دسته طبقه‌بندی می‌شود: (۱) حالت دسته‌ای و (۲) حالت جریانی. چارچوب هـادوپ، داده‌ها را در حالت دسته‌ای پردازش می‌کند، در حالی که چارچوب‌های دیگر، اجازهٔ پردازش جریانی یا بلادرنگ را می‌دهند. نهایتاً مهم‌ترین کاربردهای فناوری بزرگ داده تشریح شده اسـت. مهـم‌تـرین کاربردهـای تحلیـل بـزرگ داده عبارتنـد از: کاربردهـای برنامه‌های بهداشتی، سیستم‌های توصیه‌گر، شهر هوشمند و تحلیل شبکه‌های اجتماعی. با توجه به رشد دستگاه‌هـای متصـل بـه اینترنت، داده‌های شبکه‌های اجتماعی به طور گسترده در حال رشد بوده و نیاز بیشتری به فناوری بـزرگ داده دارنـد. همچنـین مهم‌ترین چالش‌های کاربرد بزرگ داده‌ها، شامل محرمانگی در سیستم‌های ذخیره‌سـازی، کمبودهـای نـرم افـزاری و محـدودیت ابزارها و امکانات سخت‌افزاری موجود، لزوم سرمایه‌گذاری بزرگ اولیه و فقدان مهارت‌های تکنیکی و نیروی کار خبره می‌باشد.

واژگان کلیدی

فناوری؛ بزرگ داده؛ بررسی تطبیقی چارچوب‌ها؛ کاربرد بزرگ داده؛ چالش‌ها.

۱- مقدمه

پیشرفت‌های فناورانهٔ فناوری اطلاعات منجر به ذخیره اطلاعات بیشـتر در هزینه پایین‌تر و نرخ انتقال بسیار زیاد شده است. حسگرها، موتورهـای هواپیما، معاملات آنلاین، ایمیل‌ها، فیلم‌هـا، فایـل‌هـای دیـداری، تصـاویر، جریان کلیـک، لاگ‌هـا، پسـت‌هـا، جسـتجوگرها، پرونـده‌هـای سـلامتی، شبکه‌های اجتماعی تعاملاتی، داده‌های علمی و تلفـن‌هـای همـراه، همـه این‌ها و نیز برنامه‌های کاربردی آن‌ها شروع به تولید داده‌های حجم زیادی در سرعتی بالا کرده‌اند و ذخیره و پردازش آن‌ها با فناوری‌ها و پارادایم‌های برنامه‌نویسی کلاسیک غیرممکن مـی‌باشـد. ایـن نـوع از داده‌هـا، «بـزرگ داده[1]»، نامیده می‌شوند [۶].

سرعت تولید داده‌ها به صورت روزافزون افـزایش مـی‌یابـد. بـرآورد شـده است که داده‌های جهان دو برابر و یا سه برابر خود در هـر سـال شـود. ایـن فرصت فوق‌العاده‌ای جهت تحلیل و استفاده از داده‌های عظـیم را در اختیـار می‌دهد. افزایش حجم داده‌ها باعث دستیابی بـه فنـاوری‌هـای جدیـد شـده است. بسیاری از شرکت‌هـا بـرای مـدت زمـان طـولانی از داده‌هـای بـزرگ استفاده کرده‌اند و در آن سرمایه‌گذاری کرده‌اند. گوگل حجم عظیمی از داده را پردازش می‌کند، فیس‌بوک با صدها ترابایت محتـوا سـروکار دارد، هـر روز ده‌ها ترابایت فیلم ویدیویی در یوتیوب آپلود می‌شود. به‌عنوان مثـال بـه‌طـور متوسط ۷۲ ساعت ویدئو در یک دقیقه آپلود می‌شود، تـوییتر بـیش از ۵۵۰ میلیون کاربر فعال دارد و آن‌ها ۹۱۰۰ توییت در هر ثانیه تولید مـی‌کننـد، ۳ میلیارد بخش در فیس‌بوک هر روز تولید می‌شود، یک موتـور جـت بوئینـگ می‌تواند اطلاعات عملیاتی ۱۰ ترابایت را در ۳۰ دقیقه تولید کند [۶].

در دهه‌های اخیر، مقـادیر زیـادی از داده‌هـا از منـابع مختلفـی تولیـد می‌شوند. اندازه داده‌های تولید شده در هـر روز در اینترنـت از دو میلیـارد گیگابایت تجاوز کرده است. در یک دقیقه، ۷۲ سـاعت ویـدیو در یوتیـوب

1. Big Data

* نویسنده مسئول- استادیار دانشگاه جامع امام حسین (ع)، دانشکده مدیریت و برنامه‌ریزی
** دانشجوی کارشناسی‌ارشد مدیریت فناوری اطلاعات، دانشگاه آزاد تهران مرکزی

آپلود می‌شود، در حدود ۳۰/۰۰۰ پست جدید بر روی وبلاگ تامبلر[1] ایجاد می‌شوند، بیش از ۱۰۰/۰۰۰ توئیت در توئیتر به اشتراک گذاشته می‌شوند و بیش از ۲۰۰/۰۰۰ تصویر در فیس‌بوک ارسال می‌شوند [۷]. مشکلات بزرگ داده منجر به ایجاد چند سؤال تحقیقاتی مانند (۱) چگونگی طراحی محیط‌های مقیاس‌پذیر، (۲) نحوه ارائه تحمل خطا و (۳) نحوه طراحی راه‌حل‌های مؤثر، شده است. اغلب ابزارهای موجود برای ذخیره‌سازی، پردازش و تجزیه و تحلیل داده‌ها، برای حجم عظیمی از داده‌های ناهمگن کافی نیستند. در نتیجه، نیاز مبرمی به راه‌حل‌های پیشرفته و کافی برای تحلیل بزرگ داده وجود دارد [۸]. بزرگ داده به‌طور بالقوه از دسترس‌پذیری فزاینده داده‌های موجود در زمینه حجم، تنوع، و سرعت که از ویژگی‌های بزرگ داده هستند، بهره می‌گیرند. بررسی هزینه توسط شرکت بین‌المللی داده ها نشان می‌دهد که درآمد جهانی از بزرگ داده و تجزیه و تحلیل کسب و کار می‌تواند تقریباً از ۱۲۲ میلیارد دلار در سال ۲۰۱۵ به ۱۸۷ میلیارد دلار در سال ۲۰۱۹ رشد کند. شرکت‌های بزرگ و بسیار بزرگ به احتمال زیاد نقش مهمی در این هزینه‌ها ایفا می‌کنند، که باید بیشتر هزینه‌های مرتبط با خدمات بزرگ داده را به خود اختصاص دهند [۹]. در این مقاله، ابتدا مروری بر مهم‌ترین تعاریف بزرگ داده و نیز چالش‌های به‌کارگیری فناوری‌های بزرگ داده داشته و در ادامه، برخی از مهم‌ترین چارچوب‌های تحلیل کلان داده را مورد بررسی و مقایسه تطبیقی قرار می‌دهیم. نهایتاً به بررسی و جمع‌بندی کاربردهای فناوری‌های بزرگ داده می‌پردازیم.

۲- ادبیات نظری

۲-۱- تعاریف بزرگ داده

محققانی، بزرگ داده را به‌عنوان یک مجموعه داده، شبیه به یک دارایی داده نشان می‌دهند، که دارای ویژگی‌های ذیل می‌باشد: الف- یک حجم بسیار زیاد داده، اما نه به شدت تعریف شده و ساختارمند ب- سرعت زیاد، به معنای سرعت ایجاد اطلاعات و ج- تنوع در شکل انواع داده و منابع. چنین الگویی در حال تکامل است. مانند صحت، که به‌عنوان کیفیت مجموعه داده و یا مقدار موجود در داخل داده‌ها، می‌توان در نظر گرفت.

محققانی، سیستم بزرگ داده را به‌عنوان "یک رویکرد جامع برای مدیریت، پردازش و تحلیل ۵ مقایسه (یعنی حجم، تنوع، صحت و ارزش) به منظور ایجاد بینش عملی برای تحویل ارزش پایدار، ارزیابی عملکرد و ایجاد مزایای رقابتی" در نظر گرفته‌اند. آن‌ها این داده‌ها را با استفاده از یک رویکرد فناوری برای پیگیری بینش نوآورانه از حجم چالش برانگیز داده‌های ساختاریافته و غیرساختاری ادغام می‌کنند، و برای دستیابی بهتر به مزیت رقابتی، به‌کار می‌گیرند. محققانی دیگر، بزرگ داده را به‌عنوان دارایی اطلاعاتی مشخص‌شده با حجم، سرعت و تنوع بالا که با روش‌های خاص فناورانه و تحلیلی برای تبدیل آن‌ها به ارزش، با هم ترکیب می‌کنند. آن‌ها تعاریف فناورانه بزرگ داده را از یک دارایی داده به دارایی اطلاعاتی گسترش می‌دهند: این انطباق با توجه به نیاز هر موردی برای کیفیت استخراج اطلاعات مفید است و نه صرفاً کشف داده‌ها بدون یک هدف تجاری دقیق [۱۰].

در تحقیقاتی دیگر، تعاریف مربوط به بزرگ داده را در سه گروه (ویژگی، مقایسه‌ای و معماری) دسته‌بندی کرده‌اند: از نظر ویژگی، بزرگ داده‌ها، دارای سه ویژگی اصلی هستند: حجم، سرعت و تنوع[2]. از نظر مقایسه‌ای، بزرگ داده با دادهٔ سنتی، مورد مقایسه قرار می‌گیرد. از نظر معماری، بر مقیاس‌گذاری افقیِ بزرگ داده، برای پردازش مؤثر تأکید می‌شود [۶].

ویژگی حجم، به معنای تولید و جمع‌آوری حجم عظیمی از داده‌هایی است، که نمی‌توان آن‌ها را در پایگاه داده‌های سنتی ذخیره، مدیریت و تجزیه و تحلیل نمود.

ویژگی سرعت به معنای تولید و جمع‌آوری داده‌ها با سرعت بالا (زمان کوتاه) بوده و تأکید بر این نکته که پردازش و تجزیه و تحلیل داده‌ها باید به موقع باشد، مثل سیل؛

و ویژگی تنوع به انواع مختلف داده‌ها، عمدتاً بدون ساختار و نیمه‌ساختاری مانند لاگ‌ها، متون، فیلم‌ها، موزیک‌ها، صفحات وب و ... اشاره دارد.

در بزرگ داده‌ها، نه تنها حجم زیادی از داده‌ها، بلکه انواع داده‌هایی که پیش از این ترکیبشان غیرقابل تصور بود، کنار هم قرار می‌گیرند. مدیریت و آنالیز داده‌های ساختاریافته، آسان‌تر از داده‌های بدون ساختار است. برای مثال استفاده از توابع خوشه‌ای روی داده‌های ساختاری آسان‌تر است. به تازگی، ویژگی چهارم و پنجم به ویژگی‌های بزرگ داده اضافه شده است. ویژگی چهارم ارزش[3] است، که به معنای کشف مقادیر از داده‌های بزرگ است [۱۰]. پیشرفت در فناوری ذخیره‌سازی و پردازش، امکان جمع‌آوری، تجزیه و تحلیل حجم منحصر به فرد داده‌ها و به‌دست آوردن بینش ارزشمندی و عملی را در اختیار قرار می‌دهد. ارزش انباشت حجم زیادی از داده، نه تنها در مقدار، بلکه در بینش‌های جدید منجر به دگرگونی تصمیم‌ها و اقدامات اقتصاد و جامعه می‌شود. ویژگی پنجم صحت[4] را بیان می‌کند که به عدم قطعیت، سوگیری، اختلال و غیرطبیعی‌بودن اشاره دارد. صحت در تجزیه و تحلیل داده‌ها یکی از بزرگ‌ترین چالش‌ها در مقایسه با چیزهایی مانند حجم و سرعت است. قابلیت مشاهده[5] را می‌توان به‌عنوان ویژگی ششم اضافه کرد، که به درستی ارائه و نمایش انتزاع داده‌ها، به منظور تصمیم‌گیری آموزنده اشاره دارد [۶].

1. Tumblr

2. Volume, Velocity, and Variety
3. Value
4. Veracity
5. Visibility

تعاریف زیادی از کلان داده در تحقیقات مختلف پیشنهاد شده است. اکثر آن‌ها بر این عقیده توافق دارند که بزرگ داده، چهار ویژگی اصلی را به اشتراک می‌گذارند، که تحت عنوان چهار V (حجم[1]، تنوع[2]، صحت[3] و سرعت[4]) بیان می‌شوند [۸،۱۱]. حجم، به اندازه مجموعه داده‌های موجود اشاره دارد که معمولاً نیاز به ذخیره‌سازی و پردازش توزیع شده دارند. تنوع به این حقیقت اشاره دارد که داده‌های بزرگ از انواع مختلفی از داده‌ها از قبیل متن، صدا، تصویر و ویدیو تشکیل شده است. درستی به سوگیری[5]، اختلال و ناهنجاری در داده اشاره دارد. سرعت مربوط به مکانی مکانی است که در آن داده‌ها از منابع مختلف مانند شبکه‌های اجتماعی، دستگاه‌های همراه و اینترنت‌اشیاء (IOT[6]) جریان پیدا می‌کند [۸].

همانطور که تعاریف مختلف نشان می‌دهند، اکثر آن‌ها حائز چهار ویژگی عمده، حجم، سرعت، تنوع و صحت (درستی) هستند. بنابراین به‌طور خلاصه می‌توان گفت، بزرگ داده‌ها شامل حجم عظیمی از داده هستند که با سرعت و تنوع زیاد (متن، صوت، تصویر و ...) در حال تولید بوده و به لحاظ صحت (درستی محتوا) هم دچار اختلال می‌باشند. لذا جهت ذخیره‌سازی این حجم زیاد داده، نیازمند چارچوبی با ظرفیت بالا، و نیز قدرت پردازش بالا، جهت تحلیل داده‌های متنوع و زیاد هستیم.

۲-۲- چالش‌های کاربرد بزرگ داده

اگرچه بزرگ داده‌ها، فرصت‌های زیادی را ارائه می‌دهند، کسانی‌که می‌خواهند از آن بهره‌مند گردند، باید با چالش‌های آن نیز مقابله کنند. سیستم‌های پایگاه داده سنتی، اغلب بر سیستم مدیریت پایگاه داده رابطه‌ای (RDBMS) مبتنی بوده و از داده‌های ساختارمند بهره می‌گیرند که ساده‌تر از کاربرد داده‌های غیرساختاری و نیمه ساختاریافته است. علاوه بر این، RDBMS در طول سال‌ها تکامل یافته است. اما نیازمندی‌های سخت‌افزاری برای RDBMS، به خاطر انتظارات عملکرد روزافزون و مجموعه داده‌های در حال رشد گران‌تر می‌شود. به نظر می‌رسد RDBMS از نظر ساختار داده، حجم و ناهمگونی داده‌های بزرگ محدود شده است. از آنجایی‌که پایگاه‌های داده سنتی محدود شده و به حد بالای فناوری سخت‌افزار نیازمند است، دانشگاهیان و صنعت به دنبال یافتن روش‌ها و پارادایم‌های جدید برای رفع نیازهای بزرگ داده‌ها هستند. استفاده از پتانسیل کامل بزرگ داده‌ها، نیاز به برخی موارد مانند سیاست‌های داده، فناوری و تکنیک‌ها، تغییرات سازمانی و استعداد، دسترسی به داده‌ها و ساختار صنعت دارد [۶]. برخی از محققان بیان می‌کنند، چالش‌های بزرگ داده می‌تواند به شرح زیر ذکر شود [۶،۱۲]:

- نمایش داده
- کاهش افزونگی و فشرده‌سازی داده‌ها
- مدیریت چرخه حیات اطلاعات
- مکانیسم تحلیلی
- محرمانه‌بودن اطلاعات
- مدیریت انرژی
- کارایی و مقیاس‌پذیری

یکی از چالش‌های کاربرد بزرگ داده‌ها، محرمانگی در سیستم‌های ذخیره‌سازی است [۲،۳،۶]. مسأله نقض حریم خصوصی افراد، با تجسس در مورد اطلاعات و مشخصات آن‌ها که در سامانه‌های اطلاعاتی متعدد درج می‌گردد، موضوعی است که می‌بایست به درستی مدیریت گردد. این مقوله به ویژه در مورد سازمان‌های بین‌المللی که کارکنانی که از نقاط مختلف دنیا در اختیار دارند، از اهمیت ویژه‌ای برخوردار است [۲]. محققانی، رمزنگاری را برای امن‌کردن داده‌ها پیشنهاد می‌کنند [۳].

همچنین موارد زیر مشکلات و ضعف‌های متفاوت در رابطه با بحث بزرگ داده‌ها هستند که توسط چند تن از محققان دیگر، مطرح شده است [۹]:

۱. ابهام. تعریف فعلی بزرگ داده با مفاهیم مبهم به تصویر کشیده می‌شود و حتی کلمه "بزرگ" گمراه‌کننده است، چرا که اغلب به آستانه‌های حجم متفاوتی اشاره دارد. از این‌رو، برخی از محققان، به‌عنوان مثال جورج و همکاران[7] (۲۰۱۴) ادعا می‌کنند که ارزش بزرگ داده در هوشمندی خودشان (یعنی میزان بینش قابل استخراج از حجم داده‌های جدید) است. از آنجا که این ابهام است، شرکت‌هایی که مایل به استفاده از فناوری‌های بزرگ داده در بستر ICT خود هستند، تلاش می‌کنند تا این مفهوم را بهتر درک کنند و بنابراین ارزش کسب و کار از بزرگ داده را به دست آورند.
۲. کمبود تمرکز مدیریتی. ارتباط شفافی بین بزرگ داده با معیارهای مالی یا نتایج مشابه وجود ندارد. یکی از دلایل عدم وضوح می‌تواند این باشد که "خروجی ابزار طراحی‌شده برای اکثر بزرگ داده‌ها، قابل اعتماد و اطمینان برای تجزیه و تحلیل علمی نیست" [۱۳]. در نتیجه، ارزیابی‌های جاری در مورد بازده بالقوه سرمایه‌گذاری‌های بزرگ داده ممکن است غیرقابل اعتماد و بی‌اعتبار و یا حداقل زودگذر باشد و تحقیقات بیشتری ضروری است.
۳. اهمیت جزئی. روند رو به افزایش حجم داده‌ها، تنوع و سرعت، همواره موجب تحول فناوری اطلاعاتی و ارتباطی شده است و نباید به‌عنوان چیزی غیرمنتظره در نظر گرفته شود. چیزی که واقعاً در حال تغییر است سرعت تولید بزرگ داده است، که شدت آن

1. Volume
2. Variety
3. Veracity
4. Velocity
5. Biases
6. Internet of Things
7. George et al.

مدیریت جدید و مشکلات بهره‌برداری از داده را ایجاد کرده است. مسأله اصلی این است که رویکردهای آماری معمول، به دلیل حجم زیاد داده‌ها، ناکارآمد می‌شوند [۱۴]، علاوه بر این، تفاسیر متفاوتی از حجم داده‌ای که در حال تلاش برای استفاده از آن هستیم؛ وجود دارد. لذا به یک رویکرد اکتشافی‌تر و آزمایشات بیشتری برای تولید ارزش نیاز داریم [۱۵].

از نظر برخی از محققان، این سه نقص، می‌تواند شفافیت درباره نقش بالقوهٔ بزرگ داده را محدود کرده و به احتمال زیاد تأثیر زیان آوری بر انتخاب شرکت‌ها داشته باشد. به منظور همسویی با اقدامات مدیریتی و بهره‌برداری از بزرگ داده، باید توجه داشت که رویکردهای کاملاً فناورانه به بزرگ داده، ممکن است گمراه‌کننده باشند. بدون شک، بزرگ داده شامل استفاده از فناوری می‌شود، اما اگر یک شرکت قصد دارد بر یک تصمیم‌گیری پیشرفته مبتنی بر اطلاعات تکیه کند، یک رویکرد جامع برای تولید، جمع‌آوری، و ارزیابی منابع داده جدید، نه تنها در سطح راهبردی، بلکه با رویکرد تاکتیکی و عملیاتی، می‌تواند مناسب‌تر باشد [۸]. از نظر محققی دیگر، چالش‌های مهمی نیز برای استفاده عملی از بزرگ داده‌ها وجود دارد که برخی از مهم‌ترین آن‌ها عبارتند از [۳]:

کمبودهای نرم‌افزاری و محدودیت ابزارها و امکانات موجود: روش‌های به‌کار رفته در فناوری‌های قدیمی پاسخگوی نیازمندی‌های گسترده برای مدیریت بزرگ داده‌ها نیستند. لذا نیازمند معماری، فناوری، نرم‌افزارها و الگوریتم‌های جدید برای به‌کارگیری و مدیریت هوشمندانه بزرگ داده‌ها هستیم که در حال حاضر به صورت کامل در دسترس نیستند.

برخی دیگر از محققان، چالش‌ها و مسائل سخت‌افزاری و نرم‌افزاری در مواجهه با تحلیل بزرگ داده‌ها را، اعمال محدودیت‌هایی از نوع زمان استفاده از CPU، میزان حافظه مصرفی، تعداد پرس و جوها، تعداد محاسبات و پردازش‌ها، بیان می‌کنند [۳]:

فقدان مهارت‌های تکنیکی و نیروی کار خبره: مهارت و خبرگی در استفاده از موضع جدیدی نظیر بزرگ داده‌ها و چگونگی رفتار با بزرگ داده‌ها یکی دیگر از چالش‌های پیش‌رو است. به‌عنوان مثال در گزارش اکونومیست بیان شده است که در حدود یک سوم از سازمان‌ها تخصص و مهارت کافی برای بهره‌برداری از داده‌هایشان را نداشته‌اند و لذا از یک سوم از داده‌هایشان هیچ استفاده‌ای نکرده‌اند.

لزوم سرمایه‌گذاری بزرگ اولیه و نبود بودجه کافی برای ورود دولت‌ها به این حوزه: برای در اختیار گرفتن داده‌های بزرگ و ارائه سرویس به مشتریان، نیاز به سرمایه‌گذاری نسبتاً زیادی در این حوزه وجود دارد [۳].

۲-۳- پیشینه تحقیق

برخی محققان، چندین الگوی پیاده‌سازی مپ‌ردیوس[1]، مانند هدوپ[2]، تویستر[3] و لمو- ام آر[4] را در بسیاری از حجم‌های کاری مقایسه کردند. به‌خصوص، عملکرد و مقیاس‌پذیری چارچوب‌های مطالعه شده مورد ارزیابی قرار گرفته است [۸]. برخی پژوهش‌های دیگر تلاش کردند تا اصول کلان داده را برجسته کنند. آن‌ها چالش‌های مربوط به کاربردهای کلان داده را مورد بحث قرار داده و ویژگی‌های اصلی برخی از چارچوب‌های پردازش بزرگ داده را ارائه نمودند [۱۲،۱۶]. شی و همکاران[5] (۲۰۱۵)، یک مطالعه تجربی بر روی اسپارک و هادوپ انجام دادند. آن‌ها دو ابزار را توسعه دادند: (۱) یک مطالعه از کاربرد منابع برای مپ‌ردیوس و اسپارک؛ (۲) تفکیک زمان اجرای کار برای تحلیل عمیق. آزمایش‌های انجام‌شده نشان داد که توانایی اسپارک برای شمارش کلمات، میانگین‌ها، و حجم‌های کاری، به ترتیب حدود ۲/۵ برابر، ۵ برابر، و ۵ برابر سریع‌تر از مپ‌ردیوس، است. [۱۸]، به مدیریت بزرگ داده و چارچوب‌های پردازش پرداخته‌اند. نویسندگان بررسی چندین مورد در مدیریت دادهٔ حافظه و پردازش و سیستم‌ها، از جمله سیستم‌های ذخیره‌سازی داده و چارچوب‌های پردازش داده را ارائه کردند. آن‌ها همچنین برخی عوامل کلیدی را ارائه دادند که باید به منظور دستیابی به مدیریت کارآمد و پردازش داده‌های حجیم حافظه در نظر گرفته شوند، مانند RDD[6] (مجموعه داده توزیعی انعطاف‌پذیر) برای تداوم داده‌های حافظه، مجموعهٔ تغییرناپذیری از اشیاء برای بهبود زمان پاسخگویی، و بهینه‌سازی محل داده [۱۸].

ویگا و همکاران[7] (۲۰۱۶)، یک مطالعه تجربی بر روی اسپارک[8]، هادوپ و و فلینک[9] انجام داده‌اند. در این تحقیق، تأثیر برخی از پارامترهای پیکربندی از از چارچوب‌های مطالعه شده (به عنوان مثال، تعداد رشته‌ها و کاهش‌دهنده‌ها[10] در هدوپ، تعداد رشته‌ها در مورد اسپارک و فلینک) در زمان اجرا و اجرای چندین حجم کاری مورد مطالعه قرار گرفت [۱۹]. گارسیاگیل و همکاران[11] (۲۰۱۷)، اسپارک و فلینک را از نقطه‌نظر تئوری و یا تجربی مقایسه کرده‌اند. در این تحقیق، مقیاس‌پذیری و تأثیر اندازه بر روی دیسک، و همچنین عملکرد و قابلیت‌های خاص چارچوب‌های مقایسه شده، در نظر گرفته شده است [۲۰]. اینوبلی و همکاران[12] (۲۰۱۸)، دسته‌بندی‌های مختلفی از چارچوب‌های بزرگ داده را مورد بحث قرار داده و یک مطالعه تطبیقی از چارچوب‌های کلان داده ارائه داده‌اند. در ادامه نیازمندی‌های

1. Mapreduce
2. Hadoop
3. Twister
4. LEMO - MR
5. Shi et al.
6. Resilient Distributed Datasets
7. Veiga et al.
8. Spark
9. Flink
10. Reducers
11. Garcia-Gil et al
12. Inoubli et al.

سخت‌افزاری مرتبط با هر کدام از فناوری‌های بزرگ داده، کابردها و چالش‌های آن را به صورت مختصر، مورد بحث و بررسی قرار داده‌اند [۸].

همچنین تحقیقات محدودی در زمینه بزرگ داده‌ها در داخل کشور صورت گرفته است که برخی از مهم‌ترین تحقیقات مرتبط، شامل موارد ذیل می‌باشد: در پژوهشی، که در ارتباط با مفاهیم نوین رایانش ابری و پردازش داده‌ها در ادبیات مدیریت دانش مطرح شد، هدف اصلی، مفهوم‌پردازی متغیر کلیدی بزرگ داده یا داده‌های عظیم در تحولات امروزی مدیریت دانش بوده است. در این تحقیق بیان شد که زیرساخت رایانش ابری، منابع و مخازن حجیمی را برای ذخیره‌سازی و تحلیل انواع داده‌های ساخت‌یافته و غیر ساخت‌یافته در اختیار می‌دهد [۱]. در پژوهشی دیگر، بهره‌گیری از سامانه‌های اطلاعاتی به ویژه کلان داده در مدیریت منابع انسانی در پژوهش‌ها و مطالعات انجام شده در سازمان‌هایی همچون مکنزی، مورد بررسی قرار گرفته است. محققان دریافتند، در حوزه مدیریت منابع انسانی، جایگزین‌کردن ساختارهای سنتی با نتایج تحلیلی حاصل از کلان داده‌های سامانه‌های اطلاعاتی، تناسب بیشتری برای تعامل و مدیریت نسل کاری آتی خواهد داشت [۲]. محققانی دیگر، ضمن ارائه تعاریف بزرگ داده و ویژگی‌های آن، برخی مسائل مطرح برای مدیریت بزرگ داده‌ها از جمله انتخاب پایگاه داده مناسب، زیرساخت شبکه‌ای لازم و الگوریتم‌های پردازش و جستجوی داده‌های بزرگ و نیز چالش‌های امنیتی و زمینه‌های تحقیقاتی این حوزه را در تحقیق خود مطرح نموده‌اند [۳].

محققانی دیگر، در تحقیق خود به بررسی و معرفی گزینه‌های معماری انباره داده (اعم از انباره داده سنتی، انباره داده سنتی با تغییر در فرایند ETL، انباره داده سنتی مدل ستاره‌ای، انباره داده مبتنی بر هدوپ) برای پردازش داده‌های ساخت‌یافته و غیرساخت‌یافته پرداختند. همچنین در این تحقیق، با بررسی رویکردهای استفاده شده مشخص شد، به دلیل اینکه متدولوژی خاصی برای طراحی انباره داده‌های بزرگ وجود ندارد، اغلب طراحان انباره داده، دانشی که در زمینه طراحی انباره داده سنتی (اغلب مدل داده‌ای ستاره‌ای) دارند، را به‌کار می‌گیرند [۴].

با مرور ادبیات پیشین، مشاهده می‌شود که محققان مختلفی (عمدتاً تحقیقات خارجی)، به بررسی فناوری‌های تحلیل بزرگ داده پرداخته‌اند. اما اکثر آن‌ها تنها به بررسی یک یا دو مورد از فناوری‌های تحلیل بزرگ داده اختصاص یافته است. تعدادی از پژوهش‌های داخلی نیز به بحث تعاریف، مدیریت، کاربردها و چالش‌های بزرگ داده پرداخته‌اند. اما در این مطالعه سعی بر آن است تا علاوه بر مقایسه تطبیقی چارچوب‌های تحلیل بزرگ داده، به کاربردها و موانع و چالش‌های این موضوع به صورت جامع‌تری پرداخته شود.

۳- روش‌شناسی تحقیق

در این پژوهش از روش مطالعه تطبیقی جهت بررسی موضوع پژوهش استفاده گردید. روش تحقیق تطبیقی یا مقایسه‌ای یکی از مهم‌ترین و پرکاربردترین روش‌های پژوهش در حوزه مسائل کلان علوم اجتماعی است، که با پرسش‌های بزرگ سروکار دارد؛ و واحدهای مشاهده شده و تحلیل خود را از سطوح میانی و کلان انتخاب می‌نماید. روش تطبیقی، که مبتنی بر فهم شباهت‌ها و تفاوت‌ها است، یکی از قدیمی‌ترین روش‌ها در اندیشهٔ اجتماعی و علوم اجتماعی است. دورکیم[1] بر مبنای این روش دو نوع تطبیق را از یکدیگر متمایز می‌نماید: «یکی مقایسه براساس مشابهت‌ها، مانند مقایسه جوامعی که دارای ماهیت و ساخت‌های یکسان و مشابهی هستند. و دیگری مقایسه براساس تفاوت‌ها، مانند مقایسه جوامعی که دارای ماهیت و ساخت‌های متفاوتی هستند. در نتیجه روش تطبیقی روش کشف اشتراک‌ها و افتراق‌ها در میان واقعیت‌ها و فرایندهای اجتماعی است.» به نظر اسملسر[2]، اسملسر[2]، هر قدر تعداد مشاهدات مناسب کاهش می‌یابد، امکان آزمون آماری استدلال‌ها کاسته شده و باید از روش‌های دیگر بهره برد، بر این اساس وقتی تعداد موارد مناسب کم باشد، باید از این آزمون استفاده کرد [۵].

مطابق با نظر اسملسر، در کتاب روش‌های تطبیقی در علوم اجتماعی، در این تحقیق نیز برای انتخاب واحدهای تحلیل (چارچوب‌های تحلیل بزرگ داده) پنج ملاک مطرح می‌باشد: «الف) واحدهای تحلیل متناسب با نوع مسأله نظری مطرح شده تحقیق می‌باشند. ب) واحدهای تحلیل مرتبط با پدیده مورد مطالعه هستند. ج) واحدهای تحلیل در ارتباط با معیارهای طبقه‌بندی‌شان ثابت و پایدار هستند. د) میزان دسترسی به داده‌های مربوط به هر واحد تحقیق بیان می‌شود. هـ) تصمیم به انتخاب و طبقه‌بندی واحدهای تحلیل مبتنی بر رویه‌های استاندارد و قابل قبول می‌باشد» [۵].

۴- یافته‌های تحقیق

۴-۱- چارچوب‌های بزرگ داده

در این بخش ما برخی از چارچوب‌های بزرگ داده را بررسی می‌کنیم و آن‌ها را با توجه به ویژگی‌های اصلی آن‌ها دسته‌بندی می‌کنیم. این ویژگی‌های کلیدی شامل (۱) مدل برنامه‌نویسی (۲) زبان‌های برنامه‌نویسی پشتیبانی‌شده (۳) نوع منابع داده و (۴) قابلیت اجازه پردازش داده‌های تکراری، (۵) سازگاری چارچوب با کتابخانه‌های یادگیری ماشین موجود، و (۶) راهبرد تحمل خطا[3] است [۸]. راهبرد تحمل خطا، ویژگی سیستم در تداوم عملیات به‌طور صحیح در مواقع ایجاد خطا (شکست) در یک یا چند جزء می‌باشد [۱۲].

۴-۱-۱- سیستم هادوپ

هادوپ یک پروژه آپاچی[4] است که در سال ۲۰۰۸ توسط داگ کاتینگ[5] در یاهو و مایک کافرلا[1] در دانشگاه میشیگان تأسیس شد [۲۱].

1. Durkheim
2. Smelser
3. Fault Tolerance
4. Apache
5. Doug Cutting

[۲۱]. هادوپ شامل دو جزء اصلی است: (۱) هادوپ سیستم فایل توزیعی[2] (HDFS) برای ذخیره‌سازی داده و (۲) هادوپ مپ‌ردیوس[3]، یک نمونه از پیاده‌سازی مدل برنامه‌نویسی مپ‌ردیوس [۲۲].

- **سیستم مپ‌ردیوس**

مپ‌ردیوس یک مدل برنامه‌نویسی است که برای پردازش موازی مجموعه داده‌های بزرگ طراحی شده است. مپ‌ردیوس در سال ۲۰۰۴ به‌عنوان یک مدل انتزاعی توسط گوگل پیشنهاد شد [۲۲]. که اجازه انجام محاسبات ساده را می‌دهد درحالی‌که جزئیات موازی‌سازی، ذخیره‌سازی توزیع بار، توازن بار و تحمل خطا را مخفی می‌سازد. مشخصه‌های مرکزی مدل برنامه‌نویسی مپ‌ردیوس دو تابع هستند که توسط یک کاربر نوشته شده: نگاشت و کاهش. تابع نگاشت ارزش واحد را به‌عنوان ورودی می‌گیرد و لیستی از مقدار میانی را تولید می‌کند. مقادیر میانی مرتبط، در کنار هم قرار داده می‌شوند و به تابع کاهش منتقل می‌شوند. تابع کاهش، این مقادیر را به هم ترکیب می‌کنند تا مجموعه کوچک‌تری از مقادیر را تشکیل دهند [۸].

محققانی دیگر نیز بیان می‌کنند، گوگل مدل برنامه‌نویسی مپ‌ردیوس را برای پردازش داده‌های چند سازه‌ای بزرگ ایجاد کرد. از نظر این محققان نیز، مدل مپ‌ردیوس شامل دو عملکرد است: نگاشت و کاهش. بارگذاری توابع نگاشت، تجزیه، تبدیل داده‌ها درحالی‌که عملکرد را کاهش می‌دهد، شامل خروجی عملکرد نگاشت است. هدف از مپ‌ردیوس، شمارش حروف است. سیستم ورودی‌ها را می‌گیرد و آن‌ها را با توجه به گره‌های کاری و داده موجود در چند قطعه تقسیم می‌کند. این تکه‌ها به گره‌های چندگانه بزرگ‌تر، تقسیم می‌شوند. هر تابع نگاشت بر روی گره‌های خود اجرا می‌شود و در نتیجه، محاسبات به صورت موازی اتفاق می‌افتد، بنابراین زمان پردازش را کاهش می‌دهد. تابع نگاشت، گره‌ها را با مرتب‌سازی و سپس سیستم خروجی هر تابع نگاشت را می‌گیرد و نتایج را ادغام می‌کند. تابع کاهش نتایج را به‌دست می‌آورد و تعداد کل هر حرف را محاسبه می‌کند. مپ‌ردیوس می‌تواند حجم زیادی از داده‌ها را پردازش و تجزیه و تحلیل کند. نمایه‌سازی، جستجو، مرتب‌کردن، تجزیه و تحلیل نمودار و متن، و یادگیری ماشین، نمونه‌هایی از برنامه‌های مپ‌ردیوس است. با بهبود مپ‌ردیوس و برطرف‌نمودن برخی از نقاط ضعف آن، اسپارک به‌عنوان یک جایگزین قوی برای مپ‌ردیوس مطرح شد و بر سه مفهوم پایهٔ مجموعه داده‌های توزیع‌شده RDD، تبدیل‌ها و اقدامات متکی گردید [۵].

- **سیستم HDFS**

HDFS یک پیاده‌سازی منبع آزاد از سیستم فایل گوگل است (GFS). این سیستم یک سیستم فایل توزیعی را برای ذخیره‌سازی فایل‌های بزرگ بر روی ماشین‌های توزیع‌شده به روشی قابل‌اعتماد و کارآمد فراهم می‌کند. محققانی دیگر نیز بیان می‌کنند، سیستم فایل توزیع‌شده هادوپ (HDFS)، یک سیستم فایل توزیع شده است که برای ذخیره داده‌ها در میان خوشه‌ای از ماشین‌های مناسب استفاده می‌شود، درحالیکه قابلیت دسترسی و تحمل خطای بالایی را فراهم می‌کند [۱۲].

۴-۱-۲- سیستم اسپارک

آپاچی اسپارک، یک چارچوب پردازش قدرتمند است که ابزاری برای تجزیه و تحلیل کارآمد داده‌های ناهمگن فراهم می‌کند. این چارچوب، در ابتدا در دانشگاه برکلی در سال ۲۰۰۹، توسعه یافت [۲۳]. اسپارک در مقایسه با دیگر چارچوب‌های بزرگ داده مانند هادوپ و استورم، چندین مزیت دارد. این سیستم، توسط بسیاری از شرکت‌ها از قبیل یاهو، بایدو[4] و تنسنت[5]، استفاده می‌شود. ویژگی کلیدی اسپارک، مجموعه داده توزیعی انعطاف‌پذیر (RDDs) است. RDD، یک مجموعهٔ تغییرناپذیر از اشیاء، در یک خوشه از اسپارک است. در اسپارک، دو نوع عملیات بر روی RDDs وجود دارد: (۱) تبدیل و (۲) کنش. تبدیل شامل ایجاد RDDs جدید از RDDs موجود با استفاده از توابع شبیه نقشه، فیلتر، اتحاد و الحاق هستند. کنش‌ها شامل نتیجه نهایی محاسبات RDD هستند [۸].

۴-۱-۳- سیستم استورم

استورم یک چارچوب منبع باز برای پردازش داده‌های ساختاریافته و بدون ساختار بزرگ، به صورت بلادرنگ است. استورم یک چارچوب تحمل خطا است که برای تحلیل داده‌های بلادرنگ، یادگیری ماشین، محاسبات متوالی و تکرارشونده مناسب است. پس از بررسی تطبیقی استورم و هادوپ، مشاهده می‌کنیم که اولی برای کاربردهای بلادرنگ به‌کار گرفته می‌شود درحالیکه دومین مورد برای کاربردهای دسته‌ای مؤثر است [۸].

۴-۱-۴- سیستم سمزا

آپاچی سمزا[6]، یک چارچوب پردازش توزیعی است که توسط شبکه اجتماعی لینکدین[7]، برای حل انواع مختلف الزامات پردازش جریانی مانند ردیابی داده‌ها، ورود به سیستم خدمات و ایجاد خطوط انتقال داده برای خدمات بلادرنگ ایجاد شده است. از آن زمان به بعد، این سیستم در چندین پروژه به‌کار گرفته شد. سمزا برای رسیدگی به پیام‌های بزرگ طراحی شده است و تداوم سیستم پرونده‌ای را برای آن‌ها، فراهم می‌کند. این سیستم، از آپاچی کافکا[8]، به‌عنوان یک واسطه توزیعی برای پیام‌رسانی استفاده می‌کند [۸].

1. Mike Cafarella
2. Hadoop Distributed File System (HDFS)
3. Hadoop MapReduce
4. Baidu
5. Tencent
6. Apache Samza
7. LinkedIn
8. Kafka

۴-۱-۵- سیستم فلینک

فلینک، یک چارچوب منبع باز برای پردازش داده در حالت بلادرنگ و پردازش دسته‌ای است. این روش مزایای متعددی مانند تحمل خطا و محاسبات در مقیاس بزرگ را فراهم می‌کند. مدل برنامه‌نویسی در این سیستم، مشابه مپ‌ردیوس است. فلینک در مقابل مپ‌ردیوس، توابع سطح بالای اضافی مانند پیوستن، فیلتر و تجمیع را ارائه می‌دهد. فلینک اجازه پردازش مکرر و محاسبه زمان حقیقی روی داده‌های جریان جمع‌آوری‌شده توسط ابزارهای مختلف مانند فلوم[1] و کافکا را می‌دهد. این سیستم، چندین API[2] (رابط برنامه‌نویسی کاربردی) را در سطح انتزاعی‌تری ارائه می‌دهد که به کاربر اجازه می‌دهد که محاسبه توزیعی را به روش ساده و آسان راه‌اندازی کند. یادگیری ماشین در این سیستم، شامل کتابخانه‌هایی است که طیف وسیعی از الگوریتم‌های یادگیری را فراهم می‌کند تا کاربردهای سریع و مقیاس‌پذیر داده را ایجاد کند [۲۴].

۴-۲- بررسی تطبیقی چارچوب‌های تحلیل بزرگ داده

اینوبلی و همکاران (۲۰۱۸)، در یک تحلیلی جامع، چارچوب‌های تحلیل بزرگ داده را از زوایای مختلف ذیل، مطابق با جدول ۱، مورد بررسی و مقایسه تطبیقی قرار داده‌اند [۸]: زبان‌های برنامه‌نویسی پشتیبانی‌شده، حالت پردازش، سازگاری یادگیری ماشین و اینکه آیا این چارچوب اجازه محاسبه تکراری را می‌دهد یا نه.

جدول ۱- خلاصه بررسی تطبیقی چارچوب‌های تحلیل بزرگ داده [۸]

	هادوپ	اسپارک	استورم	فلینک	سمزا
حالت پردازش	دسته‌ای	دسته‌ای و جریانی	جریانی	دسته‌ای و جریانی	جریانی
زبان برنامه‌نویسی پشتیبانی شده	جاوا	جاوا، اسکالا، پایتون	جاوا	جاوا	جاوا
توضیحات	اطلاعات ذخیره‌شده در HDFS	توسعه برنامه کاربردی تعاملی به چندین API	مناسب برای کاربرد بلادرنگ	یک دنباله از مپ‌ردیوس با روش‌های گراف	توسعه بر روی هادوپ و کافکا
محاسبات تکراری	بله	بله	بله	بله	
سازگاری با یادگیری ماشین	Mahout	SparkMLlib	سازگار با API	FlinkML	سازگار با API

هادوپ، فلینک و استورم از فرمت مقدار کلیدی برای نشان‌دادن داده استفاده می‌کنند. فرمت مقدار کلیدی اجازه دسترسی به داده‌های ناهمگن را می‌دهد. برای اسپارک، هم RDD و هم مدل مقدار کلیدی برای دسترسی سریع به داده‌ها مورد استفاده قرار می‌گیرند. همچنین چارچوب بزرگ داده‌های مورد مطالعه، به دو دسته طبقه‌بندی می‌شود: (۱) حالت دسته‌ای و (۲) حالت جریانی. هادوپ‌ها داده‌ها را در حالت دسته‌ای پردازش می‌کند، درحالیکه چارچوب‌های دیگر، اجازهٔ پردازش جریانی را می‌دهند. تمامی چارچوب‌های ارائه‌شده، قابلیت ارتباط با زبان برنامه‌نویسی جاوا، را دارند. اما اسپارک، قابلیت ارتباط با چندین زبان برنامه‌نویسی دیگر، مثل اسکالا و پایتون را دارد. هر چارچوب مجموعه‌ای از توابع انتزاعی را فراهم می‌کند که برای تعریف محاسبات مورد نظر به‌کار می‌روند. اسپارک و فلینک، کتابخانه‌های یادگیری ماشینِ خود را فراهم می‌کنند، درحالیکه سمزا دارای سازگاری با سایر ابزارها مانند samoa و هادوپ دارای سازگاری با Mahout هستند [۸].

هادوپ در حال حاضر یکی از پیشرفته‌ترین راه‌حل‌های پردازش موازی است. هادوپ به‌طور گسترده در مدیریت خوشه‌های بزرگ به‌کار می‌رود و انتخاب مناسبی برای پیکربندی راه‌حل‌های کلان داده در چندین گره[3] می‌باشد. برای مثال، هادوپ به‌وسیله یاهو استفاده می‌شود، تا ۲۴ هزار گره را مدیریت کند. علاوه بر این، ثابت شده است که هادوپ بهترین انتخاب برای انجام وظایف پردازش متن است [۲۵]. هادوپ می‌تواند چندین کار مپ‌ردیوس را اجرا کند تا از محاسبات تکراری پشتیبانی کند، اما این کار را خوب انجام نمی‌دهد زیرا نمی‌تواند داده‌های میانی حافظه را برای عملکرد سریع‌ترِ حافظه، نهان کند [۸].

اهمیت اسپارک در ویژگی‌های حافظه و قابلیت‌های پردازش گروهی، به خصوص در پردازش مکرر و افزایشی وجود دارد. علاوه بر این، اسپارک یک ابزار تعاملی به نام اسپارک‌شل[4] ارائه می‌کند که اجازه بهره‌برداری از خوشه اسپارک را به صورت بلادرنگ می‌دهد. در برخی از انواع کاربردها به دلیل استفاده از یک مفهوم خاص[5] و مدل برنامه‌نویسی، بسیار سریع شناخته می‌شود.

فلینک، شباهت‌ها و مشخصه‌های مربوط به اسپارک را به اشتراک می‌گذارد. این سیستم، عملکرد پردازش خوبی را در هنگام برخورد با ساختارهای کلان داده مانند نمودارها، ارائه می‌دهد. اگرچه راه‌حل‌های دیگری برای پردازش گراف در مقیاس بزرگ وجود دارد، فلینک و اسپارک با APIهای خاص و ابزار یادگیری ماشین، آنالیز پیشگویانه و تحلیل جریان نموداری، غنی هستند [۸،۲۶].

۴-۳- کاربرد برنامه‌های تحلیل بزرگ داده در دنیای واقعی

1. Flume
2. Application Programming Interfaces
3. Node
4. SparkShell
5. DAG- Based

ورودی‌های جستجو، پست‌ها، سوابق چت، سنسورها، فیلم‌ها، کلیک‌ها در وب‌سایت، داده‌های تجارت الکترونیک، اینترنت‌اشیاء (IOT)، داده‌های تحقیق علمی، همگی نمونه‌هایی از منابع بزرگ داده هستند. این مجموعه داده‌ها در مقیاس وسیع توزیع و تولید می‌شوند و به خودی خود بی‌معنی هستند مگر اینکه در یک مخزن اطلاعات بزرگ انباشته شده و مورد بهره‌برداری قرار گیرند. علاوه بر این، این مجموعه داده‌ها در زیرساخت‌های سنتی فناوری اطلاعات متناسب نیستند و نیاز به ظرفیت محاسباتی بیشتری دارند [۶].

در این بخش، استفاده از چارچوب‌های مطالعه شده در چندین کاربرد در دنیای واقعی از جمله کاربردهای برنامه‌های بهداشتی، سیستم‌های توصیه‌گر، تحلیل شبکه‌های اجتماعی و شهر هوشمند را به تفصیل، مورد بحث قرار می‌دهیم:

۴-۳-۱- برنامه‌های مراقبت بهداشتی

برنامه‌های علمی سلامت، مثل شبکه فضای بدن، قابلیت پایش را برای تصمیم‌گیری در مورد وضعیت سلامتی فرد، فراهم می‌کند. این امر نیازمند استقرار صدها سنسور مرتبط با بدن انسان برای جمع‌آوری داده‌های مختلف از جمله تنفس، قلبی عروقی، انسولین، خون، گلوکز و دمای بدن است [۲۷]. خدمات مراقبت بهداشتی (مانند کمک از راه دور بیماران) نیازمند تصمیم‌گیری و حصول نتیجه در زمان کوتاهی در عرض چند میلی ثانیه می‌باشد. در چنین مواردی، کاربرد چارچوب پردازش فوری (بلادرنگ)، مانند استورم، توصیه می‌شود. ترکیب نقاط قوت چارچوب‌های مذکور، نیز در اکوسیستم‌های هوشمند بین حوزه‌ای، تحت عنوان خدمات بزرگ، مفید واقع می‌شوند [۲۸].

۴-۳-۲- سیستم‌های پیشنهاد

سیستم‌های پیشنهاددهنده، حوزهٔ دیگری است که به خصوص با تغییرات مداوم و جریان‌های رو به رشد کاربران، توجه بیشتری را به خود جلب می‌کند. بر خلاف رویکردهای پیشنهادی سنتی که تنها به صورت ایستا با داده‌های کاربر سر و کار دارند، سیستم‌های توصیه‌گر جدید باید با حجم بالای اطلاعات مورد نظر و جریان بزرگ درجه‌بندی کاربر و سلایق، انطباق پیدا کنند. در این حالت، سیستم‌های توصیه‌گر باید قادر به پردازش جریان بزرگ داده‌ها باشند. به‌عنوان مثال، موارد خبری، با درجه بالایی از تغییر مشخص می‌شوند و علایق کاربر در طول زمان تغییر می‌کنند که نیازمند تعدیل پیوستهٔ سیستم توصیه‌گر است. در این حالت، چارچوب‌ها مانند هادوپ، قادر به برخورد با جریان سریع داده‌ها (به‌عنوان مثال درجه‌بندی و نظرات کاربر)، نیستند که ممکن است بر ارزیابی واقعی آیتم‌های موجود (به‌عنوان مثال محصول و یا اخبار) اثر بگذارد. در چنین وضعیتی، اتخاذ قالب‌های مؤثر پردازش جریانی به منظور اجتناب از تغییر یا ترکیب داده‌های مرتبط با کاربر/ مورد به سیستم توصیه‌گر، توصیه می‌شود. ابزارهایی مانند ماهوت[1] در هادوپ، یادگیری ماشین در فلینک[2] و اسپارک[3] شامل الگوریتم‌های فیلترینگ مشترک هستند که ممکن است برای اهداف تجارت الکترونیک و در برخی خدمات شبکه اجتماعی برای پیشنهاد آیتم‌های مناسب به کاربران به‌کار روند [۲۹].

۴-۳-۳- داده سازمانی

داده‌های سازمانی دامنه برجسته‌ای از داده‌های بزرگ است. در سال ۲۰۱۱ برآورد شد که حجم داده‌های کسب و کار در هر ۱/۲ سال دو برابر می‌شود. داده‌های سازمانی عمدتاً داده‌های ساختاری و مدیریت‌شده توسط سیستم مدیریت پایگاه داده رابطه‌ای (RDBMS[4]) است و هر اقدام قابل ضبط و فعالیت یک شرکت را نگه می‌دارد مانند CRM، ERP، فروش، مالی و تولید. به‌عنوان مثال، ۶۰۰۰ فروشگاه والمارت[5] در سراسر جهان حدود ۲۶۷ میلیون داده تراکنش را تولید می‌کنند. والمارت به منظور بهره‌گیری از این اطلاعات عظیم، انبار داده‌ای با مقیاس پتابایت[6] (یک میلیون گیگابایت) را ایجاد کرد [۲۰]. همانطور که ذخیره‌سازی، پردازش، شبکه و فناوری مدیریت داده‌ها، حجم داده، تنوع و پیچیدگی را افزایش می‌دهد. این افزایش نیاز به تکنیک‌های تجزیه و تحلیل مؤثر و پیشرفته برای اطلاعات تصمیم‌گیری است [۶].

محققانی بیان می‌کنند، تولید فراوان داده‌هایی که در حوزه‌های سازمانی ایجاد می‌شوند، به مثابه قطعات طلایی هستند که از لجن‌های حاصل از فرآوری مس، به‌دست می‌آیند. بنابراین، در دنیای امروز داده‌ها و اطلاعات اولیه که از اهمیت زیادی برخوردار نیستند؛ اغلب به صورت گسترده و توزیعی، در اختیار همه است. اما فرآوری و بازپروری آن و تولید دانش و استخراج گزاره‌های آن در فرایند غنی‌سازی داده‌ها از اهمیت بالایی برخوردار است. بنابراین سازمان‌ها باید در دانش و مهارت کافی مرتبط با تحلیل بزرگ داده‌ها تسلط داشته باشند [۲].

۴-۳-۴- داده‌های شبکه و رسانه اجتماعی

شبکه، زیرساخت پایه برای انتقال و به اشتراک‌گذاری اطلاعات است و تقریباً مردم در هر جنبه‌ای از زندگی، از طریق اینترنت یا از طریق سایر شبکه‌های خصوصی مانند بی‌سیم، اتصالات سیمی یا تلفن‌همراه، ارتباط دارند. موضوع شبکه در مرکز داده، حتی در سیستم‌های محاسبه‌ای توزیع شده و ذخیره‌سازی، مورد استفاده می‌باشد.

جستجو، خدمات شبکه‌های اجتماعی(SNS[7])، وب‌سایت‌ها، جریان‌های کلیک و غیره، می‌تواند به‌عنوان منابع بزرگ دادهٔ شبکه در نظر گرفته شود [۶،۳۰]. داده‌های منابع شبکه، با سرعت بسیار بالا تولید می‌شوند. برای مثال، بین

1. Mahout

2. Flinkml
3. Sparkmllib
4. Relational Database Management System
5. Walmart
6. Petabyte
7. Social Networking Services

سال‌های ۲۰۰۲ و ۲۰۰۹ ترافیک داده‌ها ۵۶ برابر شد، درحالیکه قدرت محاسبات تنها ۱۶ برابر شد. استفاده از گوشی‌های هوشمند برای چند سال گذشته به شدت افزایش یافته است. بیش از نیمی از جمعیت جهان از یک تلفن‌همراه استفاده می‌کنند و بسیاری از آن‌ها تمام روز به اینترنت متصل می‌شوند. ماشین‌ها به مانند انسان به یکدیگر متصل می‌شوند. با توجه به رشد دستگاه‌های متصل به اینترنت/ شبکه، به نظر می‌رسد که داده‌های شبکه به‌طور گسترده رشد می‌کنند و نیاز به فناوری شبکه‌ای بهتر (فناوری بزرگ داده) دارند [۶].

رسانه اجتماعی یک منبع داده نمایشی دیگر برای بزرگ داده‌ها هستند که نیازمند نتایج و پردازش بلادرنگ هستند. محتوای این رسانه‌ها از طریق حوزه وسیعی از کاربردهای اینترنت و وب‌سایت‌هایی شامل شبکه‌های اجتماعی و کسب و کار محور (مانند فیس‌بوک و لینکدین) و خدمات اشتراک تصویر و ویدئوی موبایلی آنلاین (مانند اینستاگرام، یوتیوب و فلیکر) تولید می‌شوند. این حجم عظیم از داده‌های اجتماعی نیازمند مجموعه‌ای از روش‌ها و الگوریتم‌های مربوط به، تحلیل متن، انتشار اطلاعات، ترکیب اطلاعات، ردیابی جامعه و تحلیل شبکه است که ممکن است برای تجزیه و تحلیل و پردازش اطلاعات از منابع مبتنی بر جامعه مورد بهره‌برداری قرار گیرد [۳۱]. این امر همچنین نیازمند پردازش مکرر و قابلیت‌های یادگیری است و مستلزم اتخاذ چارچوب‌های جریانی مانند استورم و فلینک همراه با کتابخانه‌های غنی آن‌ها می‌باشد [۸].

پژوهش‌ها نشان می‌دهند، افراد زمان قابل توجهی را در شبکه‌های اجتماعی می‌گذرانند. این موضوع کنار آسیب‌هایی که به همراه دارد، در بر گیرندهٔ حجم عظیمی از داده‌ها و اطلاعات است که نشان از گرایش‌ها و تمایلات سطوح مختلف افراد جامعه دارد. این سامانه‌های اطلاعاتی، علاوه بر تسریع و تسهیل در امور، مزیت دیگری را با خود به همراه دارند و آن ایجاد بانک اطلاعاتی بزرگی است که شامل تمام داده‌های ثبت‌شده مربوط به افراد و سازمان‌هاست. بهره‌برداری درست از این حجم عظیم داده‌های به ظاهر دست و پاگیر، می‌تواند منافع متعدد و غیرقابل انکاری داشته باشد. بزرگ داده، دریچه‌ای به زندگی حرفه‌ای کارکنان است که باعث افزایش بینش در ساختار منابع انسانی سازمان می‌شود و نوعی تعامل مجازی میان فرد و سازمان به‌شمار می‌رود. کسب اطلاعات از میزان کارآیی و توانمندی‌های هر فرد و گرایش‌ها و تمایلات وی، می‌تواند نقش مهمی در تصمیم‌گیری‌های منابع انسانی ایفاء نموده و بستری برای مدیریت استعدادهای درون سازمان فراهم نماید. لاند و همکاران[1] (۲۰۱۶)، بیان می‌کنند، داده‌های نهفته در حساب کاربری در شبکه‌های اجتماعی، بانک‌های اطلاعاتی حاوی رزومه افراد و حتی نتایج بازی‌های راهبردی که در اینترنت منتشر می‌شوند نیز از جمله داده‌هایی هستند که می‌توانند در دقیق‌سازی اطلاعات مربوط به کارکنان مفید باشند [۲].

۴-۳-۵- مدیریت ارتباط با مشتری

یکی از جدیدترین چالش‌های مدیریت ارتباط با مشتری[2]، تلاش برای برای مهار منابع اطلاعاتی ناهمگن به‌عنوان مثال ترسیم داده‌های مشتری از شبکه‌های اجتماعی برای ایجاد پیشنهاد نوآورانه می‌باشد. با این وجود، بهره‌برداری از منابع بزرگ داده در شرکت‌ها می‌تواند نیاز به تغییرات مربوط به عوامل مدیریتی مرتبط با فعالیت‌های کسب و کار و فناوری اطلاعات و ارتباطات (ICT) داشته باشد [۹]. برخی از محققان، بر این باور هستند که استخراج داده‌های خرده‌فروشی می‌تواند کمک کند تا الگوها و تمایلات خرید و نیازهای مشتری برای طرح‌ریزی مؤثر تولید محصول و جذب مشتریان، بیشتر و افزایش سود شناسایی شوند [۳].

تجزیه و تحلیل شبکه‌ای بزرگ داده به‌طور فزاینده‌ای در حال جمع‌آوری مقادیر زیادی از داده‌های مشتری، مانند رفتار خرید مشتریان، برای یک تصمیم‌گیری بلادرنگ است. شرکت‌ها با داده‌های مشتریان در میان انبوهی از منابع داده فزاینده‌ای که اغلب خارجی و ساختاریافته نیستند، مواجهند و ارزش بالقوه داده‌ها را برای ایجاد بینش در مورد رفتار مشتریان، محاسبه می‌کنند [۳۲]. به‌عنوان مثال هولدینگ سیِرز[3] به صورت هوشمندانه و دقیق با استفاده از بزرگ داده‌هایی که از چندین انبار داده مربوط به مارک‌های خود استخراج کرده، جهت ارائه تبلیغات شخصی به موقع استفاده می‌کند. همچنین شرکتی دیگر[4] از داده‌های برنامه طراحی شده، برای مشتریان دائمی و یا مشتریانی که به صورت لحظه‌ای مراجعه می‌کنند، استفاده می‌کند. اغلب این داده‌ها جهت بهبود درک مشتری و کاهش زمان انتظار برای مشتری می‌باشد، که به صورت نمایش لحظه‌ای از طریق دستگاه‌های همراه انتخاب شده است [۹].

۴-۳-۶- شهرهای هوشمند

شهر هوشمند، مفهومی‌گسترده است که شامل اقتصاد، حاکمیت، حمل و نقل، مردم، محیط‌زیست و زندگی است [۳۳]. این شهر، به استفاده از فناوری اطلاعات برای افزایش کیفیت، عملکرد و تعامل خدمات شهری در یک شهر اشاره می‌کند. همچنین با هدف اتصال چندین شهر دور افتاده از لحاظ جغرافیایی شکل می‌گیرد. داده‌ها در یک شهر هوشمند، از حسگرهای نصب‌شده بر روی تیرهای برق، خطوط آب، اتوبوس‌ها، قطارها و چراغ‌های راهنمایی جمع‌آوری می‌شوند. شبکه‌سازی تجهیزات سخت‌افزاری و سنسورها به‌عنوان اینترنت‌اشیاء (IOT) شناخته‌شده و منبع مهمی برای بزرگ داده‌ها هستند [۳۴]. چن و ژانگ (۲۰۱۴)، بیان می‌کنند که در حوزه IOT بسیاری از سنسورهای ماشین که در شبکه گسترده‌ای مستقر هستند، بسته به عملکرد آن، داده‌های مختلفی را در مراحل مختلف، تولید می‌کنند. ماشین لباسشویی، روشنایی، زنگ هشدار،

1. Lund et al.

2. CRM
3. Sears
4. Caesars Entertainment

GPS، تلفن‌همراه، یخچال و غیره می‌تواند به‌عنوان مثال منابع IOT داده شود. اعتقاد بر این است که داده‌های IOT بخش بزرگی از بزرگ داده را در آینده‌ای نزدیک تشکیل می‌دهند [۶].

فناوری‌های بزرگ داده، برای چندین هدف در یک شهر هوشمند از جمله آمار ترافیکی، کشاورزی هوشمند، بهداشت، حمل و نقل و بسیاری دیگر استفاده می‌شوند [۳۴]. برای مثال، دستگاه‌های حمل و نقل در شرکت‌های لجستیکی مجهز به سنسورهای عملیاتی و دستگاه‌های GPS هستند که به ترتیب حالات موتور و موقعیت خود را گزارش می‌کنند. این داده‌ها برای پیش‌بینی خرابی‌ها و ردیابی موقعیت‌های وسایل نقلیه مورد استفاده قرار می‌گیرند. ترافیک شهری همچنین مقادیر زیادی از داده‌هایی را فراهم می‌کند که از سنسورهای مختلف می‌آیند (به‌عنوان مثال، GPSs، کارت‌های هوشمند حمل و نقل عمومی، دستگاه‌های شرایط آب و هوایی و دوربین‌های ترافیکی). برای درک این رفتار ترافیک، آشکارسازی اطلاعات پنهان و با ارزش، از جریان بزرگ داده‌ها مهم است. هنوز یافتن مدل صحیح برنامه‌نویسی، به خاطر مقادیر در حال رشد و متنوع داده، به‌عنوان یک چالش می‌باشد [۳۵]. در حقیقت، برخی موارد کاربردی، مانند برنامه‌ریزی شهری و کنترل ترافیک، اغلب کند هستند. پردازش دادهٔ شهری، در یک چارچوب دسته‌ای کوچک، برای مثال در مورد خدمات اجرایی عمومی و دولت الکترونیک ممکن بوده و بنابراین اتخاذ چارچوبی دسته‌ای مانند هادوپ، کافی است [۳۶].

۴-۳-۷- داده‌های پژوهشی علمی

دامنه تحقیقات علمی (از جمله فیزیک ذرات، ستاره‌شناسی، بیوانفورماتیک، علوم‌زمین، شبیه‌سازی اجتماعی، پزشکی، ژنومیک، زیست‌شناسی، بیوگرافی شیمی، علوم جوی و غیره)، تحت تأثیر پدیده‌های بزرگ داده، می‌باشند. برخی از محققان، در بحث فیزیک ذرات بیان می‌کند، برای تحلیل نتایج و ذخیره‌سازی داده‌های تولیدشده توسط سازمان تحقیقات هسته‌ای اروپا از سال ۲۰۰۹ تا سال ۲۰۱۰، که ۱۳ پتابایت داده را تولید کرده است، قدرت محاسبه‌ای کافی، مورد نیاز است. در نجوم، دانشمندان با گرفتن تصاویر تلاش می‌کنند فضا را مجازی‌سازی کنند و عمدتاً در این فضای مجازی به جای فضای واقعی کار می‌کنند. در علوم‌اجتماعی، داده‌های بزرگ برای تجزیه و تحلیل، پردازش و ذخیره داده‌های اجتماعی و رفتاری مورد نیاز است [۶].

۵- نتیجه‌گیری

در این مقاله، بعد از مرور تعاریف بزرگ داده، برخی از مهم‌ترین چارچوب‌های تحلیل بزرگ داده و ویژگی‌های آن‌ها مورد بررسی و مقایسه تطبیقی قرار گرفت. بررسی‌ها نشان داد، هادوپ، فلینک و استورم از فرمت مقدار کلیدی برای نشان‌دادن داده استفاده می‌کنند. فرمت مقدار کلیدی اجازه دسترسی به داده‌های ناهمگن را می‌دهد. برای اسپارک، هم RDD و هم مدل مقدار کلیدی برای دسترسی سریع به داده‌ها مورد استفاده قرار می‌گیرند. همچنین چارچوب بزرگ داده‌های مورد مطالعه، به‌طور کلی در دو دسته طبقه‌بندی می‌شود: (۱) حالت دسته‌ای و (۲) حالت جریانی. هادوپ‌ها داده‌ها را در حالت دسته‌ای پردازش می‌کند، درحالیکه چارچوب‌های دیگر، اجازهٔ پردازش جریانی یا بلادرنگ را می‌دهند. چارچوب اسپارک، قابلیت ارتباط با چندین زبان برنامه‌نویسی مثل جاوا، اسکالا و پایتون را دارد.

هادوپ به‌عنوان یکی از بهترین راه‌حل‌ها، در مدیریت حجم عظیمی از داده‌ها و نیز انجام وظایف پردازش متن به‌کار می‌رود. همانطور که در این تحقیق بررسی شد، هادوپ شامل دو جز اصلی است: (۱) هادوپ سیستم فایل توزیعی (HDFS) برای ذخیره‌سازی داده و (۲) هادوپ مپردیوس، یک نمونه از پیاده‌سازی مدل برنامه‌نویسی مپردیوس می‌باشد. هادوپ می‌تواند چندین کار مپردیوس را به صورت موازی اجرا کند تا از محاسبات تکراری پشتیبانی نماید، اما برای پردازش بلادرنگ داده‌ها مناسب نیست. چارچوب استورم برای تحلیل داده‌های بلادرنگ، یادگیری ماشین، محاسبات متوالی و تکرارشونده مناسب است. اسپارک نیز، قابلیت‌های پردازش گروهی، به خصوص در پردازش مکرر و افزایشی را به صورت بلادرنگ دارد. فلینک، شباهت‌ها و مشخصه‌های مربوط به اسپارک را به اشتراک گذاشته و پردازش خوبی را در هنگام برخورد با ساختارهای بزرگ داده‌ها مانند نمودارها، ارائه می‌دهد. اگرچه راه‌حل‌های دیگری برای پردازش گراف در مقیاس بزرگ وجود دارد، فلینک و اسپارک با APIهای خاص و ابزار یادگیری ماشین، از حیث آنالیز پیشگویانه و تحلیل جریان نموداری، غنی هستند. سمزا نیز، یک چارچوب پردازش توزیعی است که توسط شبکه اجتماعی لینکدین برای رسیدگی به پیام‌های بزرگ و برای حل انواع مختلف الزامات پردازش جریانی مانند ردیابی داده‌ها، ورود به سیستم خدمات و ایجاد خطوط انتقال داده برای خدمات بلادرنگ طراحی و ایجاد شده است.

مهم‌ترین کاربردهای تحلیل بزرگ داده از جمله کاربردهای برنامه‌های بهداشتی، سیستم‌های توصیه‌گر، شهر هوشمند و تحلیل شبکه‌های اجتماعی مورد بررسی قرار گرفت. برنامه‌های علمی سلامت، قابلیت پایش را برای تصمیم‌گیری در مورد وضعیت سلامتی فرد، فراهم می‌کند و این خدمات مراقبت بهداشتی (مانند کمک از راه دور بیماران) نیازمند تصمیم‌گیری و حصول نتیجه در زمان کوتاهی در عرض چند میلی ثانیه می‌باشد. این امر نیازمند استقرار صدها سنسور مرتبط با بدن انسان برای جمع‌آوری داده‌های مختلف و تحلیل این کلان داده است. داده‌های سازمانی دامنه وسیعی از داده‌های بزرگ مانند CRM، ERP، فروش، مالی و تولید و ... است. چنانکه ذخیره‌سازی، پردازش، شبکه و فناوری مدیریت داده‌ها، حجم داده، تنوع و پیچیدگی را افزایش می‌دهد. این افزایش نیاز به تکنیک‌های تجزیه و تحلیل مؤثر و پیشرفته برای اطلاعات تصمیم‌گیری است. بر خلاف رویکردهای پیشنهادی سنتی که تنها به صورت ایستا با داده‌های کاربر سر و کار دارند، سیستم‌های توصیه‌گر جدید باید با حجم بالای اطلاعات مورد نظر و جریان بزرگ درجه‌بندی کاربر و سلایق، انطباق پیدا کنند. به‌عنوان مثال، موارد خبری، با درجه بالایی از تغییر مشخص

می‌شوند و علایق کاربر در طول زمان تغییر می‌کنند که نیازمند تعدیل پیوستهٔ سیستم توصیه‌گر است. شهر هوشمند، با استفاده از فناوری اطلاعات برای افزایش کیفیت، عملکرد و تعامل خدمات شهری در یک شهر اشاره می‌کند. در حوزه IOT بسیاری از سنسورهای ماشین که در شبکه گسترده‌ای مستقر هستند، بسته به عملکرد آن، داده‌های مختلفی را در مراحل مختلف، تولید می‌کنند. ماشین لباسشویی، روشنایی، زنگ هشدار، GPS، تلفن‌همراه، یخچال و غیره می‌تواند به‌عنوان مثال منابع IOT داده در شهر هوشمند باشد.

با توجه به رشد دستگاه‌های متصل به اینترنت/ شبکه، داده‌های شبکه به‌طور گسترده رشد می‌کنند و نیاز به فناوری شبکه‌ای بهتر (فناوری بزرگ داده) دارند. رسانه اجتماعی از طریق حوزه وسیعی از کاربردهای اینترنت و وب‌سایت‌هایی شامل شبکه‌های اجتماعی و کسب و کار محور و خدمات اشتراک تصویر و ویدئوی موبایلی آنلاین تولید می‌شوند. افراد زمان قابل توجهی را در شبکه‌های اجتماعی می‌گذرانند. این موضوع در بر گیرندهٔ حجم عظیمی از داده‌ها و اطلاعات است که نشان از گرایش‌ها و تمایلات سطوح مختلف افراد جامعه دارد. این سامانه‌های اطلاعاتی، علاوه بر تسریع و تسهیل در امور، حاوی بانک اطلاعاتی بزرگی است که شامل تمام داده‌های ثبت شده مربوط به افراد و سازمان‌هاست. همچنین تجزیه و تحلیل شبکه‌ای بزرگ داده، در حال جمع‌آوری مقادیر زیادی از داده‌های مشتری، مانند رفتار خرید مشتریان، برای یک تصمیم‌گیری بلادرنگ است. سازمان‌ها با داده‌های مشتریان در میان انبوهی از منابع داده فزاینده‌ای که اغلب خارجی و ساختاریافته نیستند، مواجهند و ارزش بالقوه داده‌ها را برای ایجاد بینش در مورد رفتار مشتریان، محاسبه می‌کنند. این حجم عظیم از داده‌های اجتماعی، نیازمند مجموعه‌ای از روش‌ها و الگوریتم‌های مربوط به تحلیل متن، انتشار اطلاعات، ترکیب اطلاعات، ردیابی جامعه و تحلیل شبکه است.

در مورد چالش‌های کاربرد بزرگ داده‌ها، می‌توان به‌طور خلاصه به موارد ذیل اشاره نمود:

- محرمانگی در سیستم‌های ذخیره‌سازی
- کمبودهای نرم‌افزاری و محدودیت ابزارها و امکانات سخت‌افزاری موجود
- لزوم سرمایه‌گذاری بزرگ اولیه

۶- مراجع

۱- حزباوی، سنا؛ دوستی، پریسا؛ رستمی نوروزآباد، مجتبی؛ شیخ اسماعیلی، سامان. مفهوم‌پردازی بزرگ داده‌ها در مدیریت دانش؛ با تأکید بر رایانش ابری، هفتمین کنفرانس ملی و اولین کنفرانس بین‌المللی مدیریت دانش، ایران، تهران، ۱۳۹۳.

۲- ملک‌زاده، غلامرضا و صادقی، صدیقه. راهبرد مدیریت منابع انسانی در عصر دیجیتال با تکیه بر کلان داده، فصلنامه رشد فناوری، شماره ۵۱، ۶۲-۷۰، ۱۳۹۶.

۳- همتی، مهدی و شیرازی، سحر. مدیریت داده‌های بزرگ، سومین کنفرانس بین‌المللی مدیریت و مهندسی صنایع، ایران، تهران، ۱۳۹۶.

۴- عباسی‌مهر، حسین و پورسلیمان صومعه دل، یوسف. بررسی گزینه‌های معماری اطلاعات سازمان‌ها با ظهور داده‌های بزرگ، چهارمین کنفرانس ملی محاسبات توزیعی و پردازش داده‌های بزرگ، ایران، تبریز، ۱۳۹۷.

۵- غفاری، غلامرضا. منطق پژوهش تطبیقی. مجله مطالعات اجتماعی ایران، دوره ۳، شماره ۴، ۱۳۸۹.

6- Sirin, Erkan, and Hacer Karacan. "A Review on Business Intelligence and Big Data." International Journal of Intelligent Systems and Applications in Engineering 5, no. 4 (2017): 206-215.

7- Gandomi, A., & Haider, M. "Beyond the hype: Big data concepts, methods, and analytics." International Journal of Information Management, 35(2), (2015): 137-144.

8- Inoubli, Wissem, Sabeur Aridhi, Haithem Mezni, Mondher Maddouri, and Engelbert Mephu Nguifo. "An experimental survey on big data frameworks." Future Generation Computer Systems 86 (2018): 546-564.

9- Zerbino, Pierluigi, Davide Aloini, Riccardo Dulmin, and Valeria Mininno. "Big Data-enabled customer relationship management: A holistic approach." Information Processing & Management 54, no. 5 (2018): 818-846.

10- Assuncao, M. D., Calheiros, R. N., Bianchi, S., Netto, M. A., & Buyya, R. Big Data computing and clouds: Trends and future directions. Journal of Parallel and Distributed Computing, 79, (2015): 3-15.

11- Oguntimilehin, A., and E. O. Ademola. "A review of big data management, benefits and challenges." A Review of Big Data Management, Benefits and Challenges 5, no. 6 (2014): 1-7.

12- Singh, D. & Reddy, C.K. A survey on platforms for big data analytics. Journal of Big Data (2015) 2: 8.

13- Lazer, David, Ryan Kennedy, Gary King, and Alessandro Vespignani. "The parable of Google Flu: traps in big data analysis." Science 343, no. 6176 (2014): 1203-1205.

14- Fan, J., Han, F., & Liu, H. (2014). Challenges of big data analysis. National science review, 1(2), 293-314.

15- Jukić, Nenad, Abhishek Sharma, Svetlozar Nestorov, and Boris Jukić. "Augmenting data warehouses with big data." Information Systems Management 32, no. 3 (2015): 200-209.

16- Chen, CL Philip, and Chun-Yang Zhang. "Data-intensive applications, challenges, techniques and technologies: A survey on Big Data." Information sciences 275 (2014): 314-347.

17- Shi, Juwei, Yunjie Qiu, Umar Farooq Minhas, Limei Jiao, Chen Wang, Berthold Reinwald, and Fatma Özcan. "Clash of the titans: Mapreduce vs. spark for large scale data analytics." Proceedings of the VLDB Endowment 8, no. 13 (2015): 2110-2121.

18- Zhang, Fan, Junwei Cao, Samee U. Khan, Keqin Li, and Kai Hwang. "A task-level adaptive MapReduce framework for real-time streaming data in healthcare applications." Future generation computer systems 43 (2015): 149-160.

19- Veiga, Jorge, Roberto R. Expósito, Xoán C. Pardo, Guillermo L. Taboada, and Juan Tourifio. "Performance evaluation of big data frameworks for large-scale data analytics." In 2016 IEEE International Conference on Big Data (Big Data), pp. 424-431. IEEE, 2016.

20- García-Gil, Diego, Sergio Ramírez-Gallego, Salvador García, and Francisco Herrera. "A comparison on scalability for batch big data processing on Apache Spark and Apache Flink." Big Data Analytics 2, no. 1 (2017): 1.

21- Polato, Ivanilton, Reginaldo Ré, Alfredo Goldman, and Fabio Kon. "A comprehensive view of Hadoop research—A systematic literature review." Journal of Network and Computer Applications 46 (2014): 1-25.

22- Dean, Jeffrey, and Sanjay Ghemawat. "MapReduce: simplified data processing on large clusters." Communications of the ACM 51, no. 1 (2008): 107-113.

23- Zaharia, Matei, Mosharaf Chowdhury, Michael J. Franklin, Scott Shenker, and Ion Stoica. "Spark: Cluster computing with working sets." HotCloud 10, no. 10-10 (2010): 95.
24- Alexandrov, Alexander, Rico Bergmann, Stephan Ewen, Johann-Christoph Freytag, Fabian Hueske, Arvid Heise, Odej Kao et al. "The stratosphere platform for big data analytics." The VLDB Journal—The International Journal on Very Large Data Bases 23, no. 6 (2014): 939-964.
25- Lin, Jimmy, and Chris Dyer. "Data-intensive text processing with MapReduce." Synthesis Lectures on Human Language Technologies 3, no. 1 (2010): 1-177.
26- Bajaber, Fuad, Radwa Elshawi, Omar Batarfi, Abdulrahman Altalhi, Ahmed Barnawi, and Sherif Sakr. "Big data 2.0 processing systems: Taxonomy and open challenges." Journal of Grid Computing 14, no. 3 (2016): 379-405.
27- Zhang, Fan, Junwei Cao, Samee U. Khan, Keqin Li, and Kai Hwang. "A task-level adaptive MapReduce framework for real-time streaming data in healthcare applications." Future generation computer systems 43 (2015): 149-160.
28- Xu, Xiaofei, Quan Z. Sheng, Liang-Jie Zhang, Yushun Fan, and Schahram Dustdar. "From big data to big service." Computer 7 (2015): 80-83.
29- Domann, Jaschar, Jens Meiners, Lea Helmers, and Andreas Lommatzsch. "Real-time News Recommendations using Apache Spark." In CLEF (Working Notes), pp. 628-641. 2016.
30- Hu, Han, Yonggang Wen, Tat-Seng Chua, and Xuelong Li. "Toward scalable systems for big data analytics: A technology tutorial." IEEE access 2 (2014): 652-687.
31- Bello-Orgaz, Gema, Jason J. Jung, and David Camacho. "Social big data: Recent achievements and new challenges." Information Fusion 28 (2016): 45-59.
32- Phillips-Wren, Gloria, and Angela Hoskisson. "An analytical journey towards big data." Journal of Decision Systems 24, no. 1 (2015): 87-102.
33- Yin, ChuanTao, Zhang Xiong, Hui Chen, JingYuan Wang, Daven Cooper, and Bertrand David. "A literature survey on smart cities." Science China Information Sciences 58, no. 10 (2015): 1-18.
34- Stimmel, Carol L. Building smart cities: analytics, ICT, and design thinking. Auerbach Publications, 2015.
35- Piro, Giuseppe, Ilaria Cianci, Luigi Alfredo Grieco, Gennaro Boggia, and Pietro Camarda. "Information centric services in smart cities." Journal of Systems and Software 88 (2014): 169-188.
36- Xu, Xiaofei, Quan Z. Sheng, Liang-Jie Zhang, Yushun Fan, and Schahram Dustdar. "From big data to big service." Computer 7 (2015): 80-83.