مطالعات علوم کاربردی در مهندسی

دوره ۸، شماره ۳، پاییز ۱۴۰۱

صفحات ۲۹-۲۱

Online ISSN: ۲۴۷۶-۴۵۰۷

Print ISSN: ۲۵۳۸-۲۰۵۵

www.irijournals.com

کلان‌داده و فناوری هدوپ

سحر جعفری

دانشجوی مهندسی کامپیوتر، دانشگاه صنعتی شیراز، شیراز، ایران

چکیده

کلان‌داده مربوط به داده‌هایی با حجم زیاد است که به صورت نمایی در حال رشد می‌باشد؛ این داده حجیم با یک سرعت زیاد از منابع مختلف و در انواع مختلف ساختاریافته، بدون ساختار و نیمه ساختاریافته تولید می‌شود که ما می‌توانیم اطلاعات ارزشمندی را از آن استخراج کنیم و در تصمیم‌گیری ها از کمک‌های آن بهره‌مند شویم. این پژوهش به منظور آشنایی بیشتر با کلان‌داده و فناوری هدوپ بوده و یافته‌ها حاکی از آن است که به طور کلی کلان‌داده توسط سه ویژگی اساسی خود، یعنی حجم (volume)، سرعت (velocity) و تنوع (variety) مشخص می‌شود که این سه ویژگی باید هم‌زمان وجود داشته باشند در غیر این صورت نمی‌توان درباره داده‌های بزرگ صحبت کرد. برخی از محققان برای بیان بهتر کلان‌داده، ویژگی‌های دیگری را نیز معرفی کرده اند از جمله ارزش (value) و صحت (veracity). تحلیل کلان‌داده با فراهم کردن اطلاعات ارزشمند، می‌تواند در حوزه‌های مختلف پزشکی، تجارت و سیاست بسیار کمک کننده باشد؛ اما استفاده از روش‌های سنتی برای ذخیره سازی و پردازش کلان‌داده کاری وقت‌گیر و هزینه‌بر است به همین خاطر فناوری‌هایی مانند هدوپ از طریق برقرار کردن امکان ذخیره‌سازی هر نوع داده در یک محیط توزیع‌شده و پردازش آن‌ها به صورت موازی به کمک ما آمده اند.آپاچی هدوپ از سه بخش سیستم فایل توزیع‌شده (HDFS)، چهارچوب برنامه‌نویسی نگاشت‌کاهش (MapReduce) و سرویس مدیریت منابع (YARN) تشکیل شده است که به ترتیب به عنوان واحد ذخیره‌سازی، واحد پردازش و واحد مدیریت منابع در هدوپ مورد استفاده قرار می‌گیرند و از این طریق مدیریت کلان‌داده برای ما میسر می‌شود.

واژه‌های کلیدی: کلان‌داده، هدوپ، نگاشت‌کاهش، سیستم‌فایل توزیع‌شده، HDFS، YARN، MapReduce.

مقدمه

داده‌ها با پیشرفت تکنولوژی روز به روز در حال گسترش هستند؛ IBM نشان داد که هر روز ۲.۵ اگزا بایت داده تولید میشود و همچنین ٪۹۰ از داده ها در ۵ سال اخیر تولید شده است [1].گوشی‌های موبایل، حسگرهای محیطی، لاگ نرم‌افزارهای مختلف، دوربین‌ها، ایستگاه‌های هواشناسی، شبکه‌های اجتماعی، اطلاعات پزشکی، اطلاعات سامانه‌های خرید از فروشگاه‌ها و غیره، بخشی از منابع تولید داده در مقیاس کلان هستند که با کمک این داده‌های حجیم می‌توان مسیر حرکت کسب و کار و فرایند چرخش کار سازمان ها را مشخص کرد. بنابراین ذخیره سازی و پردازش این داده ها از اهمیت بالایی در صنعت و تجارت برخوردار است و نیازمند استفاده از سیستم ها و چهارچوب های مدرنی است که با رویکرد های سنتی پردازش داده متفاوت باشد.

بهترین راهکار برای ذخیره سازی کلان‌داده استفاده از محیط توزیع‌شده می‌باشد؛ همچنین استفاده از مدل برنامه‌نویسی نگاشت‌کاهش(MapReduce) روشی بهینه برای پردازش موازی و مدیریت کلان‌داده است که توسط دانشمندان علم داده مورد استفاده قرار می‌گیرد.در این تکنیک داده‌ی ما با توجه به یک سری صفت‌ها به دسته‌ای نگاشت می‌شود، سپس تمام داده‌های نگاشت شده با هم ادغام می‌شوند و خروجی نهایی را تولید میکنند.تمام این امکانات در یک چهارچوبی به اسم هدوپ فراهم شده است که اجازه می‌دهد مدیریت و پردازش کلان‌داده با سرعت و اثرگذاری بیش‌تری انجام شود.

انواع داده

داده‌ها را می‌توان در سه شکل مختلف ساختاریافته، بدون‌ساختار و نیمه ساختاریافته تقسیم بندی نمود.

داده‌های ساختاریافته

هر داده‌ای که قابلیت دخیره شدن، قابلیت دسترسی و پردازش را داشته باشد و به یک فرمت ثابت نیز باشد، به عنوان یک داده ساختار یافته در نظر گرفته می شود [2].داده‌هایی که در پایگاه داده های رابطه ای ذخیره میکنیم از نوع ساختاریافته هستند [3].

داده‌های بدون ساختار

هر داده‌ای با فرمت یا ساختار ناشناخته، به عنوان داده‌های بدون ساختار طبقه‌بندی می‌شود. داده‌های بدون ساختار علاوه بر داشتن اندازه بزرگ، چالش‌های متعددی را از نظر پردازش آن برای استخراج ارزش آن داده‌ها، به وجود می‌آورد [2].انواع مختلف فایل های متنی، تصویری، ویدیویی و صوتی جز داده‌های بدون ساختار هستند [3].

داده‌های نیمه‌ساختاریافته

داده‌های نیمه ساختار یافته می‌توانند هر دو نوع داده (ساختار یافته و غیرساختاریافته) را شامل شوند. ما می‌توانیم داده‌های نیمه ساختار یافته را به صورت ساختاریافته ببینیم اما این داده ی نیمه ساختاریافته، به صورتی که بخواهیم آنها را به عنوان یک جدول رابطه ای در نظر بگیریم، در DBMS تعریف نشده است [2].فایل‌های XML و JSON نمونه ای از داده‌های نیمه ساختاریافته می‌باشند [3].

دوره ۸، شماره ۳، پاییز ۱۴۰۱، صفحات ۲۹-۲۱

کلان‌داده

اگر بخواهیم تعریفی از بیگ‌دیتا یا کلان‌داده ارایه کنیم میتوانیم آن را مجموعه ای از داده های حجیم بدانیم که توسط منابع مختلف مثل شبکه‌های اجتماعی، سنسورها، دستگاه‌های IOT و بسیاری منابع دیگر و در انواع مختلف ساختاریافته، بدون ساختار و نیمه ساختاریافته تولید می‌شوند، به طوری که نتوان آن‌ها را با روش‌های معمول، ذخیره، پردازش و تحلیل کرد [1] [3].

نمونه‌هایی از کاربرد کلان‌داده

در اینجا به برخی از کاربردهای کلان‌داده در سه حوزه‌ی سلامت، تجارت و سیاست اشاره خواهیم کرد.

کاربرد کلان‌داده در بخش سلامت

با تجزیه و تحلیل اطلاعات در مورد یک بیماری می‌توان درکی خوب از پیشروی بیماری به دست آورد و از این طریق می‌توان اقدامات پیشگیرانه‌ای انجام داد همچنین با تجزیه و تحلیل روی جراحی‌های قبلی موجود در پرونده بیماران وترکیب آن با علایم فعلی یا سوابق بیمار می‌توان بهترین روش‌های درمان را بر اساس مشخصات بیمار پیدا کرد [4].

کاربرد کلان‌داده در بخش تجارت

با تجزیه و تحلیل داده‌های بزرگ در بخش تجارت می‌توان رفتار کاربران و علایق آن‌ها را درک کرد؛ همین موضوع باعث افزایش رضایت مشتری، افزایش سود و افزایش رقابت در سازمان‌ها می‌شود به طور کلی تحلیل کلان‌داده، در دستیابی به فرصتهای تجاری و پیش‌بینی رکود، به سازمان‌های تجاری کمک میکند [4].

به عنوان مثال در تحقیقی در سنگاپور با استفاده از ۵۰۰ شرکت‌کننده در صنعت خرده‌فروشی انجام گردید. نتایج مطالعه نشان داد که در میان تحلیل‌های مختلف داده‌های بزرگ مورد استفاده در صنعت خرده‌فروشی سنگاپور، تحلیل‌های رسانه‌های اجتماعی به طور عمده توسط شرکت‌کنندگان پاسخ داده شده است. تأثیر داده‌های بزرگ بر عملکرد سازمان مطابق نتایج این مطالعه شامل افزایش بیشتر فروش و کاهش هزینه می‌باشد، اما بیشتر پاسخ‌ها به نفع فروش بیشتر بوده است و در صورت تأثیر بر خدمات مشتری، رضایت بالای مشتریان مشاهده شد. این مطالعه عمدتاً بر استراتژی در دو راهی قرار گرفتن به عنوان رویکرد اصلی در مدیریت داده‌های بزرگ و الگوی VLDB به عنوان روش‌شناسی اصلی تأکید دارد. از نظر کاهش موانع در زمینه مدیریت داده‌های بزرگ، استخدام منابع انسانی ماهر در تجزیه و تحلیل‌های مختلف بوسیله بیشتر شرکت‌کنندگان انتخاب شده بود. علاوه‌بر این، برای بستر قابلیت‌های احتمالی بر بازاریابی سفارشی یا سفارشی‌سازی انبوه تأکید گردید [5].

کاربرد کلان‌داده در بخش سیاست

با تحلیل اطلاعات دوربین‌های نظارت دولتی و خصوصی، نظرات شهروندان در شبکه‌های اجتماعی، معاملات آنلاین، داده‌های GPS و ارتباطات تلفن همراه می‌توان کنترل و نظارت پیوسته، برای محافظت از شهروندان و کاهش جرایم داشت [4].

دوره ۸، شماره ۳، پاییز ۱۴۰۱، صفحات ۲۹-۲۱

خصوصیات کلان‌داده و چالش‌های آن

یکی از چالشهای بزرگ در حوزه‌های مختلف سیستم محاسباتی، داده‌های بزرگ بوده و این یک موضوع اصلی در کسب و کار هوش داده ای است. دو تکنیک اصلی برای پردازش داده‌های بزرگ وجود دارد: پردازش دسته ای و پردازش جریانی. در پردازش دسته ای، داده‌ها ابتدا در پایگاه داده‌های بزرگی ذخیره میشوند و بعداً پردازش می گردند؛ اینکار معمولاً با مدلهای برنامه نویسی مقیاس پذیری مانند Google's MapReduce انجام میشود. با این حال با اندازه رو به رشد داده ها، هزینه انتقال و ذخیره سازی آنها قابل جلوگیری نیست. بعلاوه در بسیاری از دامنه ها، چیزی که مهم است نگه داشتن داده‌های اولیه نیست بلکه تحلیل آنها در سریعترین زمان ممکن است تا اطلاعات ارزشمندی را ایجاد کنند. سیستم‌های پردازش جریانی برای حل این مسائل، بر واکنش پذیری و تحلیل داده‌ها به محض تولید شدن آنها، تأکید دارند. سالهای اخیر شاهد پیدایش راه حل‌های مختلفی از پردازش جریانی بوده است. بیشتر مطالعات در این خصوص بر روی تأثیر محفظه‌های اجرایی مختلف بر عملکرد یک سیستم پردازش جریانی ارتجاعی تمرکز دارند و در اینگونه تحقیقات سلسله مراتب محفظه‌های اجرایی (ماشین‌ها و فرآیندها) را بررسی میکنند و به این نتیجه رسیده اند که فراهم سازی آنها با هزینه‌های مختلفی انجام می شود و مهمتر اینکه فراهم سازی نوع اشتباهی از محفظه‌ها میتواند سبب افت عملکرد گردد [6].

حجم، سرعت و تنوع سه مورد از ویژگی های اصلی کلان داده می‌باشند که به عنوان 3V's مطرح شدند [1] اما در ادامه ویژگی های بیش‌تری توسط محققان برای کلان داده مطرح شده است:

- حجم(volume): به اندازه داده اشاره دارد که به صورت نمایی در حال رشد است؛ معمولا کلان داده دارای حجمی بیش‌تر از ترابایت و پتابایت می‌باشد [7].

- سرعت(velocity): داده‌های حجیم توسط منابع مختلف با سرعت بسیار زیاد و به صورت بلادرنگ تولید میشوند [7].

- تنوع(variety): دادهای حجیم با سرعت زیاد توسط منابع مختلف و در انواع متفاوت تولید میشوند، بیش‌تر حجم داده‌های دنیا بدون‌ساختار و بسیار متنوع هستند. امروزه بخشی از داده‌ها در بانکهای‌اطلاعاتی، بخشی در صفحات وب و بقیه نیز در فایل‌ها با قالب های متفاوت ذخیره شده اند که پردازش آن‌ها پیچیده می‌باشد [1].

- ارزش(value): باید بتوان از داده حجیم که در انواع مختلف و با سرعت زیاد تولید می‌شود اطلاعات ارزشمندی را استخراج کرد به عبارت دیگر باید دید که این داده از نظر اطلاعاتی برای تصمیم‌گیری چقدر دارای ارزش است [1].

- صحت(veracity): با توجه به اینکه داده ها توسط منابع مختلفی ایجاد می‌شوند؛ این موضوع بر این دلالت دارد که داده ما قابل اطمینان و دارای اطلاعات درست باشد [1].

با توجه به خصوصیات کلان‌داده که در بالا اشاره کردیم؛ در هنگام مواجهه با بیگ‌دیتا با چالش هایی روبه‌رو خواهیم شد. اولین چالش مربوط به ذخیره کردن کلان‌داده می‌باشد با توجه به حجم کلان‌داده که فراتر از ترابایت و پتابایت است هراندازه هم که واحد ذخیره سازی را بزرگ در نظر بگیریم باز هم جوابگوی آن حجم از داده نخواهد بود [3].دومین چالش مربوط به پردازش کلان‌داده می‌باشد که کاری وقت‌گیر و هزینه‌بر است و عملا روش‌های پردازش سریالی قادر به پردازش این حجم از داده نخواهند بود [3].سومین چالش عدم توانایی سیستم ها در پردازش داده‌های بدون‌ساختار می‌باشد با توجه به اینکه بیش‌تر حجم داده‌های دنیا بدون‌ساختار و بسیار متنوع هستند این موضوع بسیار پر اهمیت می‌باشد [3].

این چالش‌ها منجر به ظهور پلتفرم‌های جدیدی مانند Apache Hadoop شده است که می‌تواند مجموعه داده‌های بزرگ را به راحتی مدیریت کند [8].

تعریف هدوپ

با افزایش داده های بزرگ، بنیاد نرم افزار آپاچی در سال ۲۰۰۸ یک چارچوب متن باز به نام Apache Hadoop ایجاد کرد که راه حلی برای تمام مشکلات کلان داده است. هدوپ یک چهارچوب متن باز(open source) می‌باشد، که به زبان جاوا نوشته شده است، برای ذخیره سازی توزیع شده و پردازش حجم عظیمی از مجموعه داده‌ها مورد استفاده قرار میگیرد [3]. متن‌باز بودن به این معنی است که به صورت رایگان در دسترس است و حتی ما می توانیم سورس کد آن را مطابق با نیاز خود تغییر دهیم. هدوپ یک سیستم بسیار مقاوم در برابر خطا و بسیار در دسترس است [3].

مواردی که باعث شده تا هدوپ بعنوان یک ابزار برای مدیریت کلان‌داده مورد استفاده قرار بگیرد، عبارتند از:

- امکان ذخیره سازی هر نوع داده از جمله ساختاریافته، نیمه ساختاریافته و بدون ساختار بدون هیچ محدودیتی [3].
- توانایی انجام پردازش‌های پیچیده و سرعت بخشیدن به پردازش‌ها به این طریق که در هدوپ به جای آنکه داده‌ها به سمت برنامه‌ی درحال اجرا منتقل شوند کدهای برنامه‌ی پردازشی به سمت داده‌ها ارسال می‌شوند در حقیقت کدهای برنامه بین ماشین‌های کلاستر توزیع می‌شوند.کلاستر(خوشه) به معنای گروهی ازسیستم‌ها است که از طریق یک شبکه ارتباطی به هم متصل شده‌اند. این موضوع باعث استفاده بهینه از پهنای باند کلاستر شده و از ایجاد بار اضافی ناشی از انتقال داده‌ها در بین گره‌های کلاستر جلوگیری می‌کند و اینگونه سرعت افزایش میابد [3].
- هدوپ می‌تواند روی یک دستگاه معمولی نصب شود.همین موضوع هدوپ را، هم از نظر اقتصادی به صرفه می‌کند و هم باعث می‌شود دسترسی پذیری بیش‌تری داشته باشد چون وابسته به یک فروشنده خاص نیست [3].

ساختار هدوپ

هدوپ متشکل از سه بخش اصلی HDFS، MapReduce و YARN می‌باشد [9] که در ادامه هر یک از این بخش‌ها را با جزییات بیش‌تر شرح خواهیم داد.

سیستم‌فایل توزیع شده هدوپ(HDFS)

HDFS واحد ذخیره سازی در هدوپ می‌باشد [8]که امکان ذخیره سازی حجم عظیمی از داده را در چندین گره از یک کلاستر فراهم می‌کند، به عبارتی ما داده خود را در یک محیط توزیع‌شده ذخیره می‌کنیم. سیستم فایل توزیع‌شده‌ی هدوپ بصورتی توزیع‌شدگی را ایجاد میکند که از حافظه‌ی اختصاصی هریک از دستگاه‌ها برای ذخیره‌سازی داده‌ها استفاده می‌کند [3]. برای ذخیره‌سازی داده در HDFS، داده‌ی ما ابتدا به چندین بلاک با سایز پیش‌فرض 128 MB تقسیم می‌شود و سپس در گره‌ها ذخیره می‌شوند. HDFS از دو بخش name node و data node تشکیل شده است [8].

Name node فقط فراداده‌ها (داده‌هایی درباره داده) را ذخیره می‌کند [9]؛ در حالی که data node گره‌هایی هستند که شامل داده‌های حقیقی یا بلاک‌های داده هستند [8].

نگاشت‌کاهش(MapReduce)

نگاشت‌کاهش واحد پردازش در هدوپ می‌باشد [8]. نگاشت‌کاهش یک تکنیک برنامه نویسی است که امکان اجرای پردازش موازی بروی مجموعه بزرگی از داده‌ها در یک محیط توزیع‌یافته را می‌دهد [3]. نگاشت/کاهش شامل دو وظیفه مجزا است یعنی نگاشت و کاهش و همانطور که از نام آن مشخص است، بعد از اجرای کامل فاز نگاشت، فاز کاهش اجرا میشود. خروجی

فاز نگاشت، جفت‌های کلید-مقدار می‌باشد که به عنوان ورودی در فاز کاهش در نظر گرفته می‌شوند [8] [3] این ورودی‌ها در تابع کاهش پردازش می‌شوند و سپس به مجموعه‌های کوچک‌تر تبدیل می‌شوند و با یکدیگر ادغام می‌شوند و خروجی نهایی را تولید می‌کنند، خروجی نهایی در HDFS ذخیره می‌شود [3]. با استفاده از این تکنیک درخواست پردازشی ارسال شده توسط کاربر به بخش های مستقلی تقسیم می‌شود که هر کدام از آن‌ها بین گره‌هایی که حاوی داده هستند پخش می‌شوند و به صورت موازی و همزمان اجرا می‌شوند [3].در نتیجه‌ی این کار داده‌های مورد نظر به بخش‌های کوچک‌تر می‌شکنند و یک سری مقادیر به یک سری مقادیر دیگر نگاشت می‌شوند و جفت‌های کلید-مقدار به وجود می‌آیند سپس این جفت‌ها بر اساس کلیدشان طبقه بندی می‌شوند و خروجی‌های این مرحله وارد فاز کاهش می‌شوند و در نهایت یک خروجی تولید می‌شود [8].

سرویس مدیریت منابع(YARN)

YARN واحد مدیریت منابع در هدوپ می‌باشد که زمان‌بندی کارها را نیز انجام می‌دهد [8].در یک کلاستر با چندین گره مدیریت منابع کاری دشوار است به همین خاطر با کمک YARN منابع به شکل کاملا کارآمد مدیریت می‌شوند [3]. برای اینکه YARN بتواند هدف مدیریت منابع را محقق کند احتیاج به دو بخش resource manager و node manager دارد [8] [9]. Resource manager وظیفه مدیریت کل منابع موجود در یک خوشه را بر عهده دارد. Node manager وظیفه مدیریت و کنترل اجرای منابع موجود در یک گره را عهده دار است [8] .هر گره خود شامل دو بخش container و application master می‌باشد [8] [9] که در مدیریت منابع به node manager کمک می‌کنند؛ container مجموعه‌ای است که منابع فیزیکی مانند RAM و CPU را در خود دارد که application master درخواست نیاز به منابع را از طرف node manager به سمت container ارسال می‌کند و در آخر node manager نیز گزارشی از منابع خود را به سمت resource manager می‌فرستد [8].به طور کلی به ازای هر خوشه تنها یک resource manager داریم و به ازای هر گره یک node manager داریم.

به طور کلی هدوپ به سبک MasterSlave (ارباب و برده) کار می‌کند. در یک خوشه هدوپ تنها یک گره ارباب وجود دارد و تعداد زیادی گره برده که این تعداد می‌تواند تا هزاران گره نیز ادامه پیدا کند. گره ارباب ، گره‌های برده را مدیریت، حفظ و کنترل می‌کند در حالی که گره‌های برده ، مولفه‌های واقعی انجام کار هستند [3]. Name node و resource manager در گره‌های master قرار دارند در حالی که data node و node manager در گره‌های slave قرار دارند [8].گره‌های برده هر سه ثانیه سیگنال‌هایی به سمت گره‌های ارباب ارسال می‌کنند و گزارشی از وضعیت خود می‌دهند [8].

نتیجه گیری

همانطور که در گفته شد، کلان‌داده توسط سه ویژگی اساسی خود، یعنی حجم، سرعت و تنوع، مشخص می‌شود که این سه ویژگی باید هم‌زمان وجود داشته باشند البته برخی از محققان ویژگی‌های دیگری همچون "ارزش و صحت" را نیز به این ویژگیها اضافه کرده‌اند. کلان‌داده می‌تواند در حوزه‌های مختلف پزشکی، تجارت و سیاست بسیار مفید باشد؛ به همین دلیل بجای استفاده از روش‌های سنتی برای ذخیره سازی و پردازش کلان‌داده از فناوری‌هایی همچون هدوپ استفاده می‌شود. به طور کلی برای مدیریت کلان‌داده می‌توانیم از چهارچوب هدوپ که تحت بنیاد آپاچی است استفاده کنیم و برای مدیریت

دوره ۸، شماره ۳، پاییز ۱۴۰۱، صفحات ۲۹-۲۱

مطلوب بر کلان‌داده‌ها، آپاچی هدوپ از سه بخش سیستم فایل توزیع‌شده(HDFS)، چهارچوب برنامه‌نویسی نگاشت‌کاهش(MapReduce) و سرویس مدیریت منابع(YARN) استفاده می‌کند.

با استفاده از هدوپ می‌توانیم داده حجیم خود را در یک سیستم توزیع‌شده ذخیره کنیم و آن را به صورت موازی با استفاده از تکنیک MapReduce در گره‌های محاسباتی موجود در خوشه‌ها پردازش کنیم که همه منابع موجود در خوشه‌ها برای انجام کارهای پردازشی و دیگر کارها به شکل کاملا کارآمد، توسط YARN مدیریت می‌شوند.

منابع

۱. و. جانی, "داده های حجیم و نقش آن در بهبود تجارت الکترونیک," در سومین کنفرانس بین‌المللی پژوهش در علوم و تکنولوژی, ۱۳۹۵.

2. "upGrad," [Online]. Available: https://www.upgrad.com/blog/what-is-big-data-types-characteristics-benefits-and-examples/.

۳. س. ص. م. واهبی, "بررسی سرویس پردازش موازی هدوپ در تحلیل کالن داده," در پنجمین کنفرانس بین‌المللی مهندسی برق، کامپیوتر و مکانیک, ۱۳۹۹.

٤. ن. ا. پ. م. جوانمرد, "داده های بزرگ، مروری بر فرصت ها، کاربردها و ابزارهای مورد استفاده," در کنگره بین‌المللی مطالعات میان رشته‌ای در علوم پایه و مهندسی, ۱۳۹۶.

5. S. S. S. A. C. C. J. S. Song Ying, Managing big data in the retail industry of Singapore: Examining the impact on customer satisfaction and organizational performance, European Management Journal, 2020 .

6. V. Marangozova-Martin, Multi-Level Elasticity for Data Stream Processing, Grenoble Alpes, CNRS, LIG, F-38000 Grenoble France: Noël de Palma and Ahmed El Rheddane Univ, 2019 .

7. F.-Z. B. ,. A. A. L. ,. B. Ahmed Oussous, "Big Data technologies," in Journal of King Saud University Computer and Information Sciences, 2017 .

8. "TechVidvan," [Online]. Available: https://techvidvan.com/tutorials/how-hadoop-works-internally/.

9. "TechVidvan2," [Online]. Available: https://techvidvan.com/tutorials/apache-hadoop-tutorials/.

مطالعات علوم کاربردی در مهندسی

دوره ۸، شماره ۳، پاییز ۱۴۰۱، صفحات ۲۹-۲۱

Big Data and Hadoop Technology

Sahar Jafari

Computer engineering student, Shiraz University of Technology, Shiraz, Iran,

Abstract

Big data is related to data with a large volume that is growing exponentially; This massive data is produced at a high speed from different sources and in different types of structured, unstructured and semi-structured, from which we can extract valuable information and benefit from its help in making decisions. In general, big data is characterized by three characteristics. Its basic characteristics, namely volume, velocity and variety, are defined. These three characteristics must exist at the same time, otherwise we cannot talk about big data. In order to better express big data, some researchers include other characteristics have introduced, including value and veracity. By providing valuable information, big data analysis can be very helpful in various fields of medicine, business and politics; But using traditional methods to store and process big data is time-consuming and expensive, that's why technologies like Hadoop have come to our aid by making it possible to store any type of data in a distributed environment and process them in parallel. Apache Hadoop from There are three parts of distributed file system (HDFS), map reduction programming framework (MapReduce) and resource management service (YARN) which are used as storage unit, processing unit and resource management unit respectively in Hadoop and through this management Big data becomes available to us.

Keywords: Big Data, Hadoop, MapReduce, Distributed File System, HDFS, YARN, MapReduce.