International Congress of Interdisciplinary Studies in Science and Engineering
December 2017 - Iran, Tehran
کنگره بین المللی مطالعات میان رشته ای در علوم پایه و مهندسی
آذرماه ۱۳۹۶ - تهران

مروری بر چالش‌ها و راه‌کارها در مدیریت داده‌های بزرگ

ناصر اصفهانی پور[۱]، مهدی جوانمرد[۲]

[۱] دانشجوی کارشناسی ارشد، گروه مهندسی کامپیوتر و فناوری اطلاعات، دانشگاه پیام نور، صندوق پستی ۳۶۹۷-۱۹۳۹۵ تهران، ایران
esfahani.naser@gmail.com
[۲] استادیار, گروه مهندسی کامپیوتر و فناوری اطلاعات، دانشگاه پیام نور، صندوق پستی ۳۶۹۷-۱۹۳۹۵ تهران، ایران javanmard@pnu.ac.ir

چکیده

در حال حاضر، افزایش خیلی زیاد داده‌ها در بسیاری از سازمان‌ها در سرتاسر دنیا مشاهده می‌شود. تحلیل‌گران صنعتی و کسب و کارها به دنبال داده‌های بزرگ به عنوان عامل بزرگ بعدی هستند که فرصت‌ها، بینش‌ها، راه‌کارها و روشی جدید برای افزایش سود در کسب و کار ارائه می‌دهند. داده‌های بزرگ از سایت‌های شبکه‌های اجتماعی گرفته تا سوابق بیماران در بیمارستان نقش مهمی در ارتقای کسب و کار و نوآوری داشته‌اند. کسب و کارها تلاش می‌کنند اطلاعات کافی کسب کنند و آن‌ها را برای تجزیه و تحلیل داده‌ها و اهداف کسب و کار بازیابی کنند. اگرچه داده‌های بزرگ از منابع متعددی بدست می‌آیند، اما مسائل و چالش-های متعددی وجود دارند که شرکت‌ها در هنگام ذخیره‌سازی و مدیریت داده‌های بزرگ با آن مواجه هستند. شیوه‌های مناسب مدیریت داده‌ها، تکنیک‌ها، فن-آوری‌ها و زیرساخت‌های داده‌ها می‌توانند برای غلبه بر این چالش‌ها، مشکلات و مسائل کمک کنند. در این مقاله مروری بر مسائل و چالش‌های مدیریت داده‌های بزرگ و همچنین با راه‌کارها و شیوه‌های مقابله با آن‌ها ارائه می‌شود.

کلمات کلیدی – داده‌های بزرگ، مسائل، چالش‌ها، مدیریت داده‌های بزرگ.

مقدمه

امروزه، سازمان‌ها و شرکت‌ها داده‌های بسیار زیادی تولید می‌کنند در عین‌حال، مقدار زیادی از داده‌ها از منابع مختلف بدست آمده و ذخیره می-شوند. فرآیند مدیریت، کنترل و نگهداری از مقادیر زیاد داده‌ها تحت عنوان مدیریت داده‌های بزرگ شناخته می‌شود. داده‌های بزرگ نه تنها با توجه به حجمشان بلکه توسط ویژگی‌های دیگری مانند اندازه، سرعت داده‌ها، ساختار و کیفیتشان نیز تعریف می‌شوند. بخاطر اینکه داده‌های بزرگ ویژگی‌های بسیار زیادی دارند و از نظر ماهیت متفاوت هستند، به همین دلیل مدیریت و ذخیره‌سازی آن‌ها کاری بسیار مهمی است. حجم داده‌ها در سال‌های اخیر به شدت در حال افزایش است و بسیاری از سازمان‌ها نیاز به تکنولوژی و تکنیک‌هایی برای کار و مدیریت روی داده‌های در حال افزایش هستند.

امروزه، داده ها از طیف گسترده‌ای از منابع مختلف مانند پایگاه‌داده، اینترنت، وب‌سایت‌ها و غیره بدست می‌آیند. قبل از ذخیره‌سازی، داده‌ها با کمک الگوریتم‌های مختلف تحلیلی پردازش و پاک‌سازی می‌شوند. اما بیش‌تر مواقع، سازمان‌ها با توجه به ماهیت متنوع داده‌های بزرگ با مسائل و چالش‌هایی مواجه می‌شوند. داده‌های بزرگ بدست آمده ممکن است پیچیده یا ساده، ساختاریافته یا بدون ساختار، ایمن یا دارای ایمنی حداقل باشند. این کار باعث دشوار شدن کار روی داده‌ها شده است]. این مسائل و چالش‌ها باید طوری حل شوند که بتوان اطلاعات ذخیره‌سازی شده را براحتی برای اتخاذ تصمیمات مناسب کسب و کار در آینده بازیابی کرد. هدف اصلی از مدیریت داده‌های بزرگ، ذخیره‌سازی و مدیریت داده‌های موجود بصورت ساده و قابل درک و با بازیابی آسان و انعطاف‌پذیر است. این کار باعث می‌شود تا کسب و کارها بتوانند مسائل مختلفی را درک کرده و راه‌حل‌هایی با داده‌های موجود پیدا کنند. ممکن است الگوهای جدیدی در داده‌های بدست آمده کشف شوند که بتوان از آن‌ها برای حل مسائل در کسب و کارهای مختلف مانند بیمارستان‌ها، بانک‌ها، سایت‌های شبکه‌های اجتماعی، بخش‌های تولیدی و غیره با مدیریت موثر داده‌های بزرگ بهره برد. بطور کلی، راه‌حل‌های مدیریت داده‌های بزرگ این امکان را برای شرکت‌ها فراهم می‌کنند تا با اتخاذ تصمیمات درست و راهبردی برای بهبود کسب و کار خود به هدف تصمیم‌گیری درست دست یابند.

مسائل و چالش‌های موجود در مدیریت داده‌های بزرگ

الف) حفظ حریم خصوصی و امنیت داده‌ها

بخاطر اینکه داده‌های بزرگ تکنولوژی جدیدی هستند لذا امنیت اهمیت بسیار زیادی دارد و ممکن است همه‌ی شرکت‌ها درک درستی از آن نداشته باشند. از آنجاییکه بیش‌تر داده‌های موجود در مجموعه داده‌ها مهم هستند، تضمین امنیت داده‌ها در مقابل نقص‌های امنیتی از اولویت

بالایی برخوردار است. برای مثال، فرض کنید داده‌ها از یک سرویس مبتنی بر مکان گرفته شده‌اند. این سرویس از کاربر می‌خواهد تا مکان فعلی خود را با ارائه‌دهنده‌ی خدمات به اشتراک بگذارد. اگر نقص امنیتی رخ دهد، ممکن است نگرانی‌های امنیتی جدی پیش بیاید زیرا مکان کاربر به خطر افتاده است. پنهان کردن هویت کاربر تنها نگرانی نیست، زیرا هویت کاربر ممکن است توسط مسیرهای متعدد مکانی او حاصل شود. استفاده از تکنیک‌هایی مانند رمزگذاری و ورود (لاگینگ) برای ایجاد ایمنی در داده‌های بزرگ ضروری است.

ب) ناهمگنی و ناکاملی

بر خلاف انسان‌ها، دستگاه‌ها فقط اطلاعات همگن را درک می‌کنند. دستگاه نمی‌تواند تفاوت‌های ظریف را مانند انسان درک کند. از این‌رو، داده‌ها با ساختار دقیق برای تجزیه و تحلیل کارآمد و دقیق داده‌ها ضروری هستند. داده‌های ناقص نیز ممکن است به تجزیه و تحلیل نادرست داده‌ها منجر شوند. پایگاه داده‌ی ثبت سلامت را در نظر بگیرید که تاریخ تولد، شغل و گروه خون را برای هر بیمار ثبت می‌کند. بسیاری مواقع بیماران همه‌ی اطلاعات را ارائه نمی‌دهند و از این‌رو این داده‌ها ارزشی ندارند. تحلیلی که بیماران را بر حسب شغل طبقه‌بندی می‌کند باید بیمارانی را هم در نظر بگیرد که اطلاعات او مشخص نیست و نمی‌توان از مجموعه‌داده‌های نامشخص چشم‌پوشی کرد. می‌توان فرض کرد که مقادیر نامشخص بصورت آماری همانند مقادیر مشخص هستند. با این‌حال، این امر به نتیجه‌ی نادرستی منجر خواهد شد.

ج) افزایش

با توجه به اینکه حجم داده‌ها سریع‌تر از منابع محاسباتی افزایش می‌یابند و سرعت CPUها به علت محدودیت‌های توان ثابت است، روش نوآورانه-ای برای مدیریت داده‌های بزرگ لازم است.

اولا، طی پنج سال گذشته سرعت ساعتی پردازش‌گرها تا حد زیادی ثابت مانده است و پردازش‌گرها با تعداد زیادی از هسته‌ها ساخته می‌شوند. در گذشته، سیستم‌های پردازش داده‌های بزرگ در مورد توازی گره‌ها در خوشه نگران بودند؛ در حالیکه باید روی توازی در یک گره منفرد کار کرد [۲]. تکنیک‌های پردازش موازی داده‌ها که از آن‌ها در توازی بین گره‌ها استفاده می‌شود متفاوت از توازی درون گره‌ای، چراکه معماری آن‌ها به نظر بسیار متفاوت هستند. این تغییرات بی‌سابقه نیاز به بازنگری در طراحی، ساخت و عملیات اجزای پردازش داده‌ها است.

دوما، حرکت در راستای محاسبات ابری، که در حال حاضر چند بار کاری مجزا و مختلف با اهداف عملکردی متفاوت را در خوشه‌های بسیار بزرگ تجمیع می‌کند. این میزان به اشتراک‌گذاری منابع در خوشه‌های گران‌قیمت و بزرگ نیازمند روش‌های جدیدی برای تعیین نحوه‌ی اجرای وظایف پردازش داده‌هاست بطوریکه بتوان اهداف هر بار کاری را بصورت مقرون‌بصرفه‌ای تامین کرد و با نقص‌های سیستمی مقابله کرد که اغلب هنگام کار بر روی خوشه‌های بسیار بزرگ‌تر رخ می‌دهد.

مورد سوم تغییر زیرسیستم I/O متعارف است. بجای درایوهای هارد دیسک (HDDها) از درایوهای حالت جامد و سایر فن‌آوری‌ها مانند حافظه‌ی تغییر فاز استفاده می‌شود. پیامدهای این تغییر زیرسیستم ذخیره‌سازی بطور بالقوه‌ای با هر جنبه‌ای از پردازش داده‌ها از جمله الگوریتم‌های پردازش جستار، برنامه‌ریزی زمانی جستار، طراحی پایگاه‌داده، روش‌های کنترل همزمان و روش‌های بازیابی ارتباط دارد

د) برنامه‌ریزی زمانی

زمان‌بندی یعنی داده‌ها بسیار سریع‌تر از زمان واقعی در دسترس قرار دارند. دیجیتالی کردن تقریبا «همه چیز» در حال حاضر انواع جدیدی از داده‌های بزرگ و زمان-واقعی را در طیف گسترده‌ای از صنایع ایجاد می‌کند، بنابراین مقدار زیادی از داده را باید پردازش کرد و تجزیه و تحلیل آن مدت زمان بیش‌تری طول می‌کشد. نیاز به طراحی سیستمی داریم که همه‌ی این پارامترها را در نظر خواهد گرفت. در دنیایی که با سرعت در حال تغییر است به نتیجه‌ی تجزیه و تحلیل در کسری از ثانیه نیاز داریم. به عنوان مثال، اگر تراکنش کارت اعتباری جعلی مشکوک باشد، بطور ایده‌آل باید قبل از تکمیل تراکنش این کار تشخیص داده شود و بطور بالقوه‌ای باید کلا از انجام این تراکنش جلوگیری شود. تحلیل کامل تاریخچه‌ی خرید کاربر در زمان-واقعی غیر ممکن است. در عوض، باید از قبل نتایج جزئی را پیش ببریم بطوریکه بتوان از مقدار کمی از محاسبات افزایشی با داده-های جدید برای تصمیم‌گیری سریع استفاده کرد.

د) چالش‌های کسب و کار/شرکتی

داده‌های داخلی و خارجی که برای پیاده‌سازی کسب و کار و رقابت در انواع بازارهای جهانی مورد استفاده قرار می‌گیرند، بطور پیوسته‌ای از نظر حجم، ارزش و اهمیت در حال افزایش هستند. استخراج ارزش این داده‌ها تا حد امکان بطور موثرتر و کاراتر یک چالش واقعی محسوب می‌شود. برای دستیابی به هوش کسب و کار (BI) جدید از طریق تجزیه و تحلیل‌های غنی، باید بتوان با استفاده از شیوه‌های جدید و بصورت مشخصی مجموعه داده‌های نامرتبط را بصورت متقابل ارجاع داد. اگرچه داده‌های اصلی موجود هستند و از آن‌ها برای مدیریت شرکت استفاده می‌شود، هر گروه کسب و کار داده‌های خود را مدیریت می‌کند. این داده‌ها تنها زمانیکه توسط گروه‌های دیگری درخواست داده می‌شوند به اشتراک گذاشته می‌شوند و این کار اغلب باعث تکرار پرهزینه‌ی تلاش‌ها می‌شود. اهمیت و ارزش بالقوه‌ی داده‌ها برای گروه دیگر اغلب توسط تولیدکننده‌ی مجموعه‌ای از داده‌های خاص غیر قابل تشخیص می‌شود. همچنین هیچ انگیزه‌ای برای به اشتراک گذاشتن اطلاعات برای گروه‌های کسب و کار وجود ندارد. بنابراین، اطلاعات کمی در مورد داده‌های پردازش شده توسط آن‌ها با گروه درخواست‌کننده به اشتراک گذاشته می‌شود.

گروه‌های کسب و کار که از سامانه‌های تراکنشی یا سایر منابع داده‌ای استفاده می‌کنند بندرت اطلاعات را در فرمت‌های مورد نیاز در اختیار سایر گروه‌های کسب و کار قرار داده‌اند. تغییرات در قراردادهای نام‌گذاری محصولات، مشتریان و سایر اطلاعات و نیز استفاده از نام‌های فیلد غیر استاندارد و رمزی باعث شده است تا یکپارچه‌سازی و استفاده از داده‌ها از چند منبع دشوار شود. کپی‌های یکسان از روی داده‌های مشابه در تلاش برای استخراج داده‌ها از یک منبع و تبدیل آن به فرمت مورد نیاز مشتری انجام شده است و این کار در تعیین اطلاعات نهایی برای مشتری بعدی که بطور مکرر به داده‌ها دسترسی داشته است مشکلی ایجاد می‌کند.

با توجه به افزایش میزان داده‌ها، هزینه‌ی ذخیره‌سازی داده‌ها و پردازش داده‌ها بطور قابل توجهی افزایش یافته است. همچنین، با توجه به پیدایش فن‌آوری‌های جدید مانند سیستم‌های داده‌های بزرگ بسیاری از گروه‌های کسب و کار درخواست دسترسی به فن‌آوری‌های گران‌قیمت را برای کار بر روی حجم زیادی از داده‌ها می‌دهند حتی اگر داده‌های آن‌ها واقعا چنین راه‌حلی را توجیه نکند.

راه‌کارها/شیوه‌هایی برای مدیریت داده‌های بزرگ

الف) حجم، سرعت، تنوع

حجم یعنی میزان داده‌های گردآوری شده. حجم پرونده‌های مختلف با افزایش حجم کاهش می‌یابد. سرعت در این دنیای چالش‌برانگیز نقش بسیار مهمی را ایفا می‌کند. سرعت بالا برای ایجاد، ذخیره‌سازی، و انتقال داده‌ها لازم هستند. سرعت بر تأخیر زمانی –تاخیر بین زمان ایجاد یا ضبط داده‌ها - و هنگامیکه در دسترس هستند تاثیر می‌گذارد. امروزه، اطلاعات در فرم‌های مختلف و با تغییرات قابل‌توجهی پدید می‌آیند. تغییرات یعنی مدیریت انواع داده‌های پیچیده از جمله داده‌های ساختار یافته و بدون ساختار. بعضی از راه‌حل‌های بسیار خوب برای کار روی حجم، سرعت و تنوع داده‌ها عبارتند از:

RDBMS (سیستم مدیریت ارتباطی پایگاه‌داده) یا پایگاه داده‌های متعارف SQL (زبان جستار ساختاریافته) باید هماهنگی و یکپارچگی داده‌ها را همراه با مقیاس‌پذیری حفظ کنند. به عنوان مثال، در تراکنش‌های بانکی، اگر مشتری پول نقد از بانک بگیرد، پایگاه داده باید قبل از انجام تراکنش‌های دیگر، از تغییر در حساب مشتری اطمینان حاصل کند. در غیر این صورت، ممکن است برداشت پول از حساب بیش‌تر از موجودی باشد.

۱. NOSQL (نه تنها زبان جستار ساختار یافته): NOSQL پایگاه داده‌های ستون‌گرایی دارد. داده‌ها توسط ستون‌ها ذخیره می‌شوند و میزان داده‌های ذخیره شده را می‌توان با فشرده‌سازی یا بیت‌مپینگ ذخیره کرد. سرعت عملکرد جستار فقط با بازگردانی ستون‌های درخواستی و فشرده-سازی داده‌ها حاصل می‌شود. در این طرح‌های داده‌ای، سرعت ثابت نیست و در نتیجه می‌توانیم گره‌های داده‌ای مورد نظر را حذف و اضافه کنیم. در نتیجه الاستیسیته حاصل می‌شود. ساز و کار تکرار داده‌های داخلی باعث می‌شود داده‌ها حتی اگر نقص سخت‌افزاری رخ دهد در دسترس قرارگیرند و این قابلیت تحمل‌پذیری خطا را ارائه می‌دهد. برای نمونه، جدول گوگل وب (Google Web) برای جدول بزرگ گوگل بکار برده می-شود. جدول گوگل وب یک کپی کامل از بخش از اینترنت را در یک جدول تکی ذخیره‌سازی می‌کند که توسط گوگل با گراف کاملی از لینک‌ها بین صفحات نمایه‌سازی شده است. چهار مدل داده‌های از NOSQL وجود دارد:

الف) *سیستم‌های ارزش کلیدی:* این مدل جداول هش را با کلیدی منحصربفرد و نشان‌گری در آیتم داده ذخیره می‌کنند. این مدل داده‌ها در رشته ذخیره می‌کند و این پایگاه‌داده‌ها تایپ نمی‌شوند، اما ویژگی ACID (تجزیه‌پذیری، سازگاری، انزوا و پایایی) و تحمل‌پذیری خطا را ندارد. مثال‌ها عبارتند از Example Memcached، Dynamo و Voldemort.

ب) *سیستم‌های ستونی:* از این مدل برای ذخیره‌ی داده‌ها روی سطر و پردازش حجم زیادی از داده‌های توزیعی استفاده می‌شود. می‌توانیم ستون‌ها را براحتی اضافه کنیم، در نتیجه این مدل ویژگی‌های مقیاس‌پذیری، انعطاف‌پذیری و عملکرد را در پی دارد

ج) *پایگاه‌داده‌های سند:* این مدل به سیستم ارزش کلیدی شباهت دارد، اما داده‌های مرتبط با کلید بصورت سند نیمه‌ساختاری هستند. هرگاه لازم باشد تعداد بسیاری زیادی از گزارش‌ها تولید شوند، این مدل کاراترین مدل است، و از عناصری که بطور مداوم در حال تغییر هستند گردآوری می-شود. مثال؛ MongoDB.

د) *سیستم‌های پایگاه‌داده‌ی گراف:* این مدل زمانی موثر واقع می‌شود که داده‌ها بسیار بهم‌پیوسته باشند و ساختار اساسی تحت عنوان «رابطه‌ی گره» شناخته می‌شود. مثال Neo4j.

برخی از مهم‌ترین راه‌کارهای NOSQL عبارتند از HBase، Cassandra، Hypertable، Voldemort، Membase، MongoDB، CouchDB.

۲. *کاساندرا:* برای هر تراکنشی، کاساندرا سطح مناسبی از سازگاری را به ازای مقیاس‌پذیری میسر می‌سازد. می‌توان دسترسی به مهم‌ترین داده‌ها را توسط حافظه‌ی نهان در حافظه تسریع کرد که در سرورها روی خوشه‌ها قرار می‌گیرد.

۳. تمایز بین کاساندرا، نسخه‌های قبلی RDBMS و سایر نسخه‌های NOSQL عبارتند از: /ویژگی‌ها:
این معماری مقیاس‌پذیر مقادیر پتابایت از داده‌ها و هزاران عملیات همزمان را به قدری براحتی مدیریت می‌کند که همانند مدیریت ترافیک داده‌ها و کاربر بسیار کوچک است هیچ نقصی در هیچ پردازش یا کارکرد پایگاه‌داده به دلیل طراحی همتا به همتا وجود ندارد. بهره‌های عملیات خطی برای هر دو عمل خواندن و نوشتن توسط افزودن آنلاین ظرفیت تحویل داده می‌شود. روش مستقل از مکان حقیقی ذخیره‌سازی و دسترسی به اطلاعات برابر با قابلیت‌های استقلال مکان است (می‌توان داده‌ها را در هر جایی خواند و نوشت).

۱) پایایی و حفاظت مانند RDBMS توسط سازگاری داده‌های قابل تنظیم ارائه می‌شوند.
۲) تمامی فرمت‌های اپلیکیشن داده‌های بزرگ از جمله داده‌های ساختاریافته، نیمه‌ساختاریافته، و بدون ساختار را می‌توان بخاطر طرح انعطاف‌پذیر تطبیق داد.
۳) افزونگی داده‌ها با تکرار ساده‌سازی شده انجام می‌شود و کاساندرا قابلیت تبدیل به مرکز چند داده‌ای یا از نظر ماهیتی به ابر را دارد.
۴) اثر داده‌های خام بخاطر فشردگی داده‌ها کاهش می‌یابد.
۵) استفاده از زبان SQL-مانند باعث کاهش منحنی یادگیری برای ایجادکننده و مدیرانی می‌شود که از دنیای RDBMS می‌آیند.
۶) نیازی به هیچ تجهیزات خاصی وجود ندارد.
۷) کاساندرا یک مدل داده‌ی غنی‌تری را ارائه می‌دهد که المان‌های داده‌های جدولی در پایگاه‌داده‌های SQL و انعطاف‌پذیری نوعی ذخیره‌سازی-های ارزش کلیدی را ترکیب می‌کند.

1. روش‌های جایگزین کاساندرا

الف) *Amazon DynamoDB*

ویژگی‌ها:
ترکیب خواندن/نوشتن قابل تنظیم و کل ظرفیت ذخیره‌سازی این امکان را در این روش بوجود می‌آورد تا نحوه‌ی پیکربندی داده‌ها را مشخص کرد.
این روش بصورت خاصی برای استفاده از توان عملیاتی بالا و پایایی دستگاه ذخیره‌سازی SSD فلش طراحی شده است.
این روش خدمات مدیریت شده‌ای است. مشتریان مجبور نیستند از هر چیزی در مورد پیاده‌سازی یا مدیریت خوشه‌ی پایگاه داده اطلاع داشته باشند.

ب) HBase:

این روش یک پایگاه‌داده‌ی توزیعی است که بخش از پشته‌ی هادوپ است و تا حد زیادی خواندن/نوشتن سازگار را در مقابل ذخیره‌سازی سازگار و نهایی داده‌ها فراهم می‌کند این روش برای پردازش و تحلیل داده‌های خیلی زیاد مناسب است.

ب) راه‌کار منبع باز برای مدیریت داده‌های بزرگ

۱) هادوپ

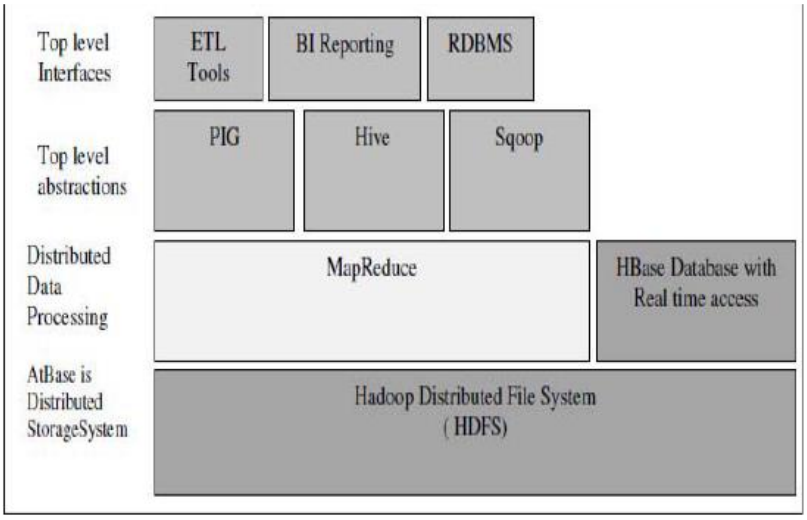
امروزه،حجم زیادی از داده‌ها در سازمان‌ها دریافت و تولید می‌شوند که باید پردازش شوند و زمان زیادی برای تحلیل آن‌ها لازم است. به همین دلیل، سازمان‌ها به فن‌آوری پیشرفته‌تری برای رقابت در این دنیای در حال پیشرفت نیاز دارند. راه‌کار کنونی می‌تواند مسائل مرتبط با داده‌های بزرگ را حل کند اما برای پیاده‌سازی بهتر باید هزینه‌ی ساخت این ابزارها را در نظر بگیریم. هادوپ راه‌کارهایی برای این مساله ارائه می‌دهد چراکه منبع باز است اما هزینه‌ی جواز ندارد در هادوپ، محاسبات همزمان، نرخ جریان و مدیریت پیکربندی از طریق سیستم فایل، ذخیره‌سازی داده‌ها، و پلت‌فرم‌های تحلیلی و یک لایه کنترل می‌شود. هادوپ یک نوع معماری است که روی سخت‌افزار جامعه اجرا می‌شود و برای ذخیره‌سازی و پردازش داده‌های مقیاس‌پذیر، سیستم عامل شبکه‌ی مجازی مقاوم در برابر خطا بکار برده می‌شود. معماری ذخیره‌سازی خوشه‌بندی شده با پهنای باند زیاد و مقاوم در برابر خطا که توسط هادوپ بکار برده می‌شود تحت عنوان سیستم فایل توزیعی هادوپ (HDFS) شناخته می‌شود. HDFS برای پردازش داده‌های توزیعی و کارهایی که با داده‌های ساختار یافته و بدون ساختار انجام می‌شود *MApReduce* را اجرا می‌کند.

هادوپ اکوسیستم هادوپ را دارد که شامل موارد زیر است:

۱. HDFS: در خوشه‌ی هادوپ، HDFS روی گره‌ها و برای تبدیل سیستم‌های فایلی به یک سیستم فایلی بزرگ اجرا می‌شود که به سیستم‌های فایلی روی تعداد بسیار زیادی از گره‌های داده‌های ورودی و خروجی متصل می‌شود.
۲. MapReduce: ابزارهای لازم برای توزیع تصادفی وظیفه‌ی پردازشی روی خوشه‌ی بزرگی از کامپیوترها توسط MapReduce فراهم می‌شود.
۳. HBase: این مورد روی HDFS ایجاد می‌شود که یک راه‌کار NoSQL هادوپ است.
۴. Pig: این مورد برای تحلیل محاسبات داده‌هاست و زبانی با سطح بالا محسوب می‌شود. این مورد روی MapReduce ساخته می‌شود و پیچیدگی را پنهان می‌کند. در Yahoo، اجرای ۳۰٪ از وظایف هادوپ در حقیقت pig هستند.
۵. Hive: این مورد چیزی مانند مدل دسترسی و رابطه‌ای برای SQL فراهم می‌کند.
۶. Sqoop: این مورد داده‌های را انتقال داده و بین پایگاه‌داده‌ی رابطه‌ای و هادوپ جابجا می‌کند.
۷. Oozie: مدیریت یک مجموعه و جریان کار برای وظایف وابسته‌ی هادوپ.

هادوپ سه بخش اساسی HDFS، مدل MapReduce، هادوپ مشترک دارد. در نمونه‌ی هادوپ، هر گرهی یک گره‌نام[1] دارد که درخت دایرکتوری تمامی فایل‌ها را در سیستم فایلی دارد و هسته‌ی سیستم فایل HDFS محسوب می‌شود. از این گره‌نام برای ردیابی در بین خوشه‌ها برای مکان‌یابی فایل داده‌ها استفاده می‌شود. در HDFS، گره‌داده‌ها[2] داده‌ها را ذخیره می‌کنند. برای سیستم ارتباطی، از پروتکل لایه‌ی TCP/IP استفاده می‌شود در حالیکه مشتری از تماس روش از راه دور (RPC)[3] بهره می‌گیرد. نیازی به ذخیره‌سازی RAID روی میزبان‌ها وجود ندارد زیرا داده‌ها روی چند میزان تکرار می‌شوند. داده‌ها روی سه گره ذخیره می‌شوند: دو گره روی یک قفسه (رک) و گره دیگر روی قفسه‌ای دیگر با ارزش تکرار پیش‌فرض. HDFS تنها دستگاه‌های جدید را در هر لحظه از زمان روی خوشه اضافه می‌کند، در نتیجه تقریبا قابلیت افزایش به ظرفیت ذخیره‌سازی نامحدود را دارد.

[1] Namenode
[2] dataNode
[3] Remote Procedure Call



شکل ۱. ابزار معماری هادوپ

۲. *MapReduce*: در این روش، پردازش هر داده‌ای به گام‌های نقشه و کاهش تقسیم می‌شوند. در گام نقشه، پردازش نقشه وظایف را به زیروظایف مستقل تقسیم می‌کند که بصورت همزمان روی سرور مجزای اجرا می‌شوند. در گام کاهش، نتایج پردازش‌های نقشه بصورت بازگشتی با پردازش‌های کاهشی ترکیب می‌شوند که تا زمان محاسبه‌ی نتیجه‌ی نهایی بصورت همزمان و روی سرورهای مجزا اجرا می‌شوند. به عنوان مثال، فرض کنید بخواهیم تعداد رخدادهای برخی از رشته‌های معین را در پایگاه‌داده‌ی NoSQL شمارش کنیم. فرایند تقسیم مساله را به داده‌های محتوای خودکار با استفاده از معیارهای خاصی تقسیم می‌کند. در این مورد، داده‌ها جملات هستند هر فرایند نقشه تعداد رخدادهای هر رشته را در یک گره داده‌ی تکی شمارش می‌کند. هر فرآیند کاهش خروجی‌های فرایندهای نقشه را برای رشته‌ی معینی اضافه می‌کند.

الف) ذخیره‌سازی داده‌ها و پردازش داده‌ها:

فرآیند نقشه روی سرورهاییاجرا می‌شود که داده‌ها روی آن‌ها ذخیره می‌شوند و فرایند کاهش به فرایند نقشه نزدیک است تا ورودی آن‌ها را تامین کند برای ایجاد این محیط بهینه، سیستم مدیریت ذخیره‌سازی داده‌ها باید ارتباط نزدیکی با چارچوب MapReduce داشته باشد.

ب) الزامات ساخت:

برای ایجاد MapReduce، طراحان چارچوب باید به روش‌های جدیدی فکر کنند. اگر وظیفه آسان باشد، فرآیندMapReduce بطور مستقیم انجام می‌شود، اما اگر این وظیفه پیچیده باشد و نتوان آن را همزمان انجام داد، این وظیفه به دنباله‌ای از زیروظایف تقسیم می‌شود که خود آن‌ها نیز باید بصورت کارآمدی به فرآیندهای نقشه و کاهش تقسیم شوند که کار دشواری خواهد بود. بنابراین، نیاز به افراد آموزش دیده است.

۳. ابر و داده‌های بزرگ:

برای تجزیه و تحلیل داده‌های بزرگ در هر دو حالت عمومی و خصوصی، ابر طیف گسترده‌ای از روش‌ها را ارائه می‌دهد ابر راه‌کارهای ارزان‌قیمتی را برای مدیریت و دسترسی به مجموعه داده‌های بسیار بزرگ و همچنین پشتیبانی از عناصر زیرساخت قدرتمند فراهم می‌کند.

معماری‌های ابری شامل آرایه‌هایی از دستگاه‌های مجازی است که برای پردازش مجموعه داده‌های بسیار بزرگ ایده‌آل هستند بطوریکه می‌توان پردازش را به فرآیندهای موازی متعددی تقسیم کرد. ماهیت تطبیق‌پذیر، مجازی، انعطاف‌پذیر و قدرتمند باعث می‌شود تا ابر کاربرد خوبی برای داده‌های بزرگ داشته باشد. در اینجا چند نمونه به شرح زیر ارائه می‌شوند:

الف) **IaaS** (زیرساخت بصورت یک سرویس) در ابر عمومی: در IaaS برای خدمات داده‌های بزرگ، شرکت‌ها می‌توانند بدون استفاده از زیرساخت‌های فیزیکی خود از زیرساخت‌های ارائه‌دهنده ابر عمومی بهره ببرند. می‌توان دستگاه‌های مجازی را تقریبا با قدرت محاسباتی و ذخیره-سازی بی‌نهایت توسط IaaS فراهم کرد.

ب) **PaaS** (پلت‌فرم به عنوان یک سرویس) در ابر خصوصی: در محیط ابر عمومی یا خصوصی این زیرساخت بسته‌بندی شده است که می‌توان از آن برای طراحی، پیاده‌سازی و اجرای اپلیکیشن‌ها و خدمات استفاده کرد.

ج) **SaaS** (نرم افزار به عنوان یک سرویس) در ابر مرکب: SaaS پلت‌فرمی برای داده‌های تجزیه و تحلیل و همچنین رسانه‌های اجتماعی ارائه می‌دهد. علاوه بر این، کسب و کارها ممکن است از داده‌های CRM شرکتی خود در محیط ابر خصوصی برای ورود به تجزیه و تحلیل استفاده کنند.

ج) راه‌کارهایی برای چالش کسب و کار

به منظور مدیریت موثر داده‌ها و داده‌های پردازشی، سازمان‌ها باید تیم مدیریت داده‌های هوش کسب و کار IT همانند تیم مدیریت داده‌های هوش کسب و کار IT اینتل تشکیل دهند. با همکاری گروه‌های کسب و کار و اتخاذ دیدگاه‌های مشتری می‌توان داده‌های شرکتی را در بالاترین سطح کیفی در دسترس قرارداده و کاربردی کرد.

۱. نسخه تکی حقیقت:

استانداردسازی نام محصولات و شناسه‌های بخش مهمی از سرویس تک منبعی است. این استانداردسازی در هنگام جستجو و مقایسه‌ی داده‌ها باعث صرفه‌جویی زمانی می‌شود. تیم مدیریت داده‌های هوش کسب و کار (BIDM) با تیم‌هایی از بخش‌های توانمندسازی که مسئول ارائه‌ی راه-کارهای وب، اپلیکیشن و BI مطابق بر اولویت‌های عملیاتی IT هستند کار می‌کنند تا مشخص شود که کدام داده‌های اصلی لازم هستند و قراردادهایی برای نام‌گذاری ایجاد کنند. لیستی از اسامی مورد تایید محصول، اسامی مشتری‌ها و سایر شناسه‌هایی که به قابلیت افزودن داده‌ها کمک می‌کنند در نتیجه‌ی این همکاری تولید می‌شوند. می‌توان تناقض‌ها در اشکال عددی و سایر داده‌ها را از دو گروه مختلف کسب و کار حذف کرد و بنابراین با استفاده از این نسخه تکی حقیقت میزان دقت BI بهبود می‌یابد.

۲. متناسب کردن اندازه‌ی راه‌حل برای داده‌ها

می‌توان تمامی داده‌های سازمانی ساختاریافته از انبار داده‌ها را با در اختیار قرار دادن داده‌های درست برای مشتریان مناسب در زمان مناسب در دسترس قرار داد. برای کاربردی کردن داده‌ها در گروه‌های کسب و کار، می‌توان داده‌ها را وارد انبار داده کرد، توسط زمینه‌ی موضوعی مرتب و سپس بصورت مناسبی ایجاد کرد. اما این رویکرد با چند مساله مواجه است:

الف) پلت‌فرمی متفاوت از انبار داده‌های قدیمی برای استفاده از داده‌های بدون ساختار یا داده‌های بزرگ برای BI نیاز است

ب) جابجایی مشتریانی که هم اکنون از داده‌هایی با کیفیت متغیر استفاده می‌کنند و از قراردادهای نام‌گذاری مشابهی استفاده نمی کنند برای ایجاد نسخه‌های بعدی ممکن است به معنای از کار انداختن انبار داده‌های آن‌ها و بازگرداندن اطلاعاتشان باشد که کاری وقت‌گیر، تخریبی و پر هزینه است.

ج) حفظ اطلاعاتی که در حال حاضر در یک کانتینر نزدیک به سیستم تراکنش فعلی و کاربران داده‌ها قرار دارد اغلب مواقع مفیدتر است که باعث کمینه‌سازی مسائل مرتبط با تاخیر و نقاط نقص بالقوه می‌شود که با هر حرکت اضافی داده‌ها پیش می‌آید.

برای پرداختن بر مسائل مربوط به مدیریت داده‌ها، برنامه‌ای برای موارد زیر طراحی شده است:

الف) نسخه‌ی تکی داده‌های حقیقی باید در کاتالوگ داده‌ها نگهداری شوند.

ب) دسته‌بندی پلت‌فرم‌های داده‌ها از پایگاه‌های رابطه‌ای و راه‌حل‌های انبار داده‌ها تا پلت‌فرم‌های داده‌های بزرگ مانند هادوپ آپاچی باید پشتیبانی شوند.

ج) برای پشتیبانی از هر کاربرد موردی BI خاص، یک فرآیند تعامل با مشتری ایجاد کنید تا بتوان بر فرآیند، ارزش، معماری، کانتینر داده‌ها، محل کانتینر داده‌ها و میزان سرمایه‌گذاری کسب و کار پرداخت

از طریق فرایند مشارکت مشتری با ارائه سطح بالایی از کیفیت خدمات و پاسخگویی، می‌توان نیازهای مشتریان را بصورت مناسبی تطبیق داد. با حفظ تیم‌های کامل که بطور موثری از تخصص زمینه و کسب و کار استفاده می‌کنند می‌توان در طول زمان به سودی دست یافت که عنصر کلیدی این راهبرد است. تخصیص مدیر خدمات به هر گروه کسب وکار، همانند فروش و بازاریابی، زنجیره‌ی تامین یا طراحی با فرآیند مشارکت

مشتری آغاز می‌شود مدیران محصول نیز انتصاب می‌شوند که برای استفاده از داده‌ها بر نرمالیزه کردن آن‌ها تمرکز دارند و هر کدام از این مدیران محصول در یک زمینه‌ی موضوعی مانند مکان، موجودی، مالی، توزیع فروش و آیتم تخصص دارند.

۳. تامین نیازهای مشتریان:

مدیران خدمات و مدیران تولید محصولات زمینه‌ی موضوعی با یکدیگر همکاری می‌کنند تا اطمینان حاصل شود که نیازهای داده‌های سازمانی هر یک از مشتریان IT از طریق مدل‌سازی مناسب داده‌ها برآورده می‌شود. این به معنای درک نیازهای مشتری، و سپس ساختن داده‌ها و ارائه‌ی این داده‌ها در شکل قابل استفاده است. به عنوان مثال، مدیر محصولات زمینه‌ی موضوعی ممکن است با مدیر خدمات برای ساخت داده‌های موجودی به منظور یکپارچه‌سازی آسان و همبستگی با سایر اطلاعات،] مانند داده‌های فروش و بازاریابی برای حصول به بینشی در مورد موجودی سازمانی با پیش‌بینی‌های فروش فعلی کار کند.

۴. راه‌حل‌های هوش کسب و کار خود-سرو:[4]

دسترسی مستقیم به اطلاعات روشی برای بهبود سرعت است که مشتریان بتوانند به BI دست یابند. دسترسی سریع‌تر به ابزارهای داده و استقلال بیش‌تر در استخراج ارزش و حصول به بینش‌های بیش‌تر و جدید با استفاده از راه‌حل‌های مصورسازی توسط راه‌کارهای خود-سرو ارائه می‌شود. مشتریان IT با اطلاعات سازمانی با کم‌ترین پشتیبانی از IT و بدون دستیابی به مجموعه مهارت‌های فنی از طریق راه‌کارهای BI خود-سرو و با طراحی مناسب کار می‌کنند همچنین، مشتریان می‌توانند تغییرات خود را برای پاسخ سریع‌تر به تغییرات کسب و کار اعمال کنند. تیم مدیریت داده‌های هوش کسب و کار IT نیز می‌توانند مزایای استفاده از راه‌حل‌های خود-سرو را داشته باشند. می‌توانیم از منابع بیش‌تری برای پرداختن بر درخواست‌هایی استفاده کنیم که الزامات پیچیده‌تر و همچنین بررسی خدمات مدیریت داده‌های جدید را در بر می‌گیرد که ممکن است با راه‌حل-های خدمات خود-سرو ارزش جدیدی برای سازمان به ارمغان بیاورد

۵. استفاده از لایه‌های معنایی:

حصول به زمینه‌های موضوعی داده‌ها و بسته‌بندی کردن آن‌ها در لایه‌های معنایی هدف در گروه‌های کسب و کار المانی مهم از راهبرد خود-سرو است. داده‌های شرکتی که برای دسترسی مشتریان IT به داده‌ها بصورت مستقل و با استفاده از شرایط عمومی کسب و کار همانند دستورات فروش بدون آگاهی از جزئیات مهم پیاده‌سازی طراحی شده‌اند توسط این لایه‌های معنایی بیان می‌شوند از لایه‌های معنایی برای ایجاد گزارش‌ها، داشبوردها یا سایر راه‌حل‌های BI مورد نیاز از جمله اشکال مختلفی از مصورسازی داده‌ها توسط بسیاری از گروه‌های کسب و کار استفاده می‌شود.

د) مدیریت داده‌های اصلی در داده‌های بزرگ

سازمان‌ها بارهای داده‌ای را تولید می‌کنند که باید در قالب معناداری مرتب و سازمان‌دهی شوند. مدیریت داده‌های اصلی یکی از راه‌حل‌های موثر برای مدیریت داده‌ها است چراکه فن‌آوری، فرآیندها و افرادی را به هم پیوند می‌دهد که یکنواختی، دقت و منحصر به فرد بودن نهادهای داده‌ای را تضمین می‌کند که توسط چند واحد کسب و کار در یک سازمان به اشتراک‌گذاری شده‌اند. داده‌های اصلی، مخزن مرکزی و قابل اعتمادترین منبع داده‌ها است که در سرتاسر سازمان به اشتراک گذاشته شده است. داده‌های اصلی همراه با داده‌های بزرگ می‌توانند به شرکت‌ها کمک کنند تا بینش خوبی در مورد نهادهای بسیار مهم داده‌ای در سازمان مانند مشتریان، محصولات و غیره بدست بیاورند. با ترکیب داده‌های اصلی و داده‌های بزرگ، شرکت می‌تواند دید ۳۶۰ درجه‌ای از داده‌های در حال استفاده داشته باشد.

1. سیستم MDM:

داده‌های بزرگ و MDM (مدیریت داده‌ها اصلی) با هم می‌توانند در حصول به نتایج کسب و کار موثر به سازمان‌ها کمک کنند. داده‌های بزرگ بدست آمده از منابع مختلف از دو نوع هستند:

الف) داده‌های سطح ویژگی

ب) داده‌های سطح تراکنش

---

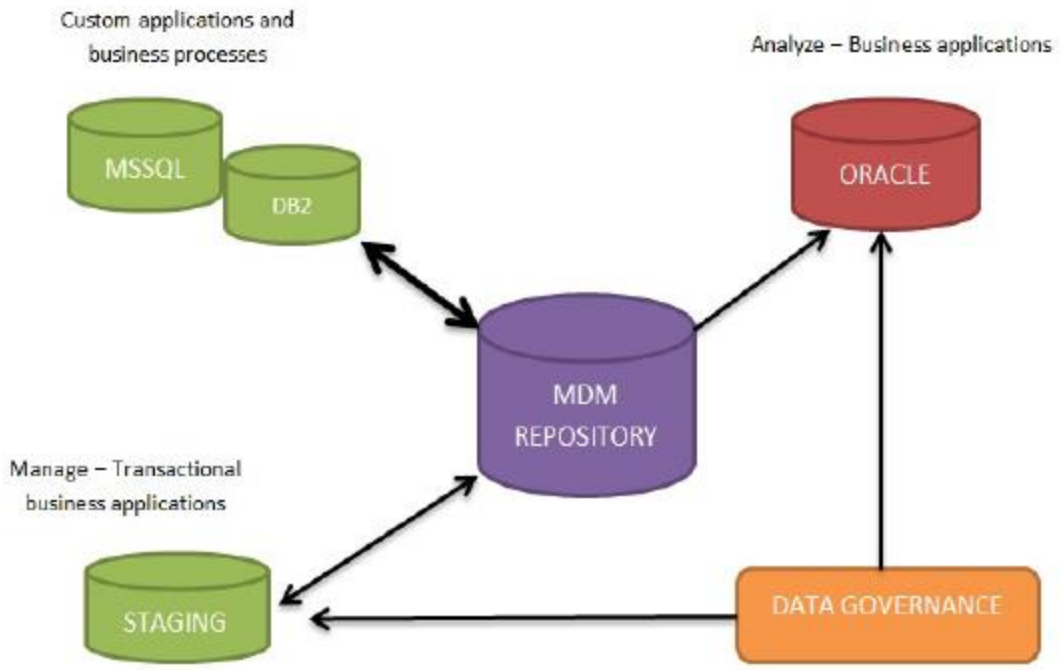
[4] Self-service

داده‌های سطح ویژگی داده‌هایی هستند که نهادها را تعیین می‌کنند. داده‌های سطح تراکنش داده‌هایی هستند که توسط هر یک از نهادها تولید می‌شوند. داده‌های اصلی مخزنی است که هر دو نوع از داده‌ها را ذخیره می‌کند. لایه‌ی استخراج داده‌های بزرگ که در سیستم وجود دارد از سطح ویژگی به مخزن مرکزی انتقال می‌یابد و داده‌های سطح تراکنش به مرکز داده‌ی بزرگ منتقل می‌شوند. سیستم MDM کپی‌هایی از این داده‌های سطح نهاد را پاک، استانداردسازی و حذف می‌کند و آثار و دنباله‌های حسابرسی را حفظ می‌کند. این سیستم نه داده‌های سطح تراکنش را ذخیره می‌کند و نه داده‌های بزرگ را تجزیه و تحلیل می‌کند.

2. MDM با تجزیه و تحلیل داده‌های بزرگ:

شرطی وجود دارد که بر مبنای آن سیستم‌های MD باید حجم زیادی از داده‌های بزرگ را در آینده کنترل و مدیریت کنند. برای این منظور، برخی از تغییرات اساسی در سیستم‌های MDM موجود باید ایجاد شوند:

ویژگی‌های خاص داده‌های بزرگ مانند ترجیح محصول‌ها، شناسه‌های رسانه‌های اجتماعی، قصد خرید، انتخاب، امتیازهای حمایت از برند، سلسله مراتب شبکه‌های اجتماعی، و غیره باید در یک مدل داده ذخیره شوند. برخی از این ویژگی‌ها از سیستم‌های تحلیل داده‌های بزرگ وارد سیستم-های MDM خواهند شد و برخی دیگر از ویژگی‌ها نیز بطور مستقیم از منبع تغذیه می‌شوند. برای کار روی سرعت داده‌ها، معماری‌های بهبود یافته‌ی سیستم MDM لازم خواهد بود. با استفاده از فرآیندهای پیشرفته‌ی تطبیق و ترکیب در داده‌های بزرگ، به ویژه برای تجزیه و تحلیل داده‌های اجتماعی، این فرآیندها برای رفع چالش مهم شناسایی پروفایل‌های اجتماعی کاربر که در سازمان موجود است لازم هستند. این امر بطور خاص در جایی صدق می‌کند که در زمان جستجوی اسامی مشترک چند نتیجه حاصل می‌شود. زمانیکه مشتریان یا کاربران بجای نام واقعی از نام مستعار در پروفایل خود استفاده می‌کنند، چالش شناسایی پیچیده‌تر می‌شود.



شکل ۲. مدیریت داده‌های اصلی

3. داده‌های بزرگ که باید کنترل شوند

روش لازم برای کنترل داده‌های بزرگ عبارتند از:

- طبقه‌بندی داده‌های بدون ساختار با حداکثر دقت با استفاده از فراداده‌های مشترک بین روش‌های MDM و داده‌های بزرگ.
- تصمیم‌گیری در مورد سطوح سرعت و کیفیت مناسب برای سازمان‌ها و استفاده از داده‌های بزرگ برای تامین این سطوح؛

- اعمال خط مشی‌های یکسان حریم خصوصی و امنیت داده‌های اصلی بر داده‌های بزرگ که با داده‌های اصلی مرتبط کرده‌اید؛
- تطبیق قابل اعتماد داده‌های موجود با داده‌های اصلی

ی) شیوه‌های سازمانی

تکامل داده‌های بزرگ بخوبی در حال انجام است. ارزش داده‌های بزرگ با کاهش تاخیر بشدت افزایش می‌یابد و داده‌ها بسرعت تحویل داده می-شوند. بهبود ده و صد برابری در عملکرد منجر به فرصت‌های مختلف تحلیلی کیفی می‌شود که اغلب بصورت افزایش درآمد و سود بیان می‌گردد.

۱. مالکیت و حمایت مالی از داده‌های بزرگ

بسیاری از بخش‌ها و گروه‌ها دارای پلت‌فرم‌های داده‌های بزرگ خود هستند. یک روند در راستای پلت‌فرم‌های داده‌های بزرگ است که توسط IT به عنوان زیرساخت ارائه می‌شود. زیرساخت IT در حال گسترش است تا پلت‌فرم‌های داده‌های بزرگ را در بر بگیرد همان کاری که قبلا در ذخیره-سازی داده‌ها از طریق سیستم‌های SAN (شبکه‌ی منطقه‌ی ذخیره‌سازی) NAS (ذخیره‌سازی الحاقی شبکه) انجام داد. این رویکردی خوب برای سازمان‌هایی است که امیدوارند از سیلوهای داده‌های بزرگ اجتناب کنند که زمانی پیش می‌آید که بخش‌ها (دپارتمان‌ها) از پلت‌فرم‌های خود برای مدیریت داده‌های بزرگ استفاده می‌کنند.

۲. عناوین شغلی و ساختارهای تیمی برای مدیریت داده‌های بزرگ

بسیاری از عناوین شغلی و تیم‌های مختلف در حال حاضر روی مدیریت داده‌های بزرگ کار می‌کنند. انواع افرادی که داده‌های بزرگ را مدیریت و از آن‌ها در کار خود استفاده می‌کنند بطور شگفت‌انگیزی متنوع هستند. معماران مختلف پیشرو در مدیریت داده‌های بزرگ هستند بویژه معمارانی که تمرکزشان روی داده‌هاست. تحلیل‌گران نیز درBDM پیشرو هستند، به خصوص تحلیل‌گرانی که بر داده‌ها تمرکز دارند، چراکه برای تحلیل‌گر داده‌ها و تحلیل‌گر کسب و کار، مدیریت داده‌های بزرگ تنها برای متخصصان داده نیست. در واقع، در حال حاضر عقیده‌ای در تیم‌های اپلیکیشن وب وجود دارد که مدیریت داده‌های بزرگ را انجام می‌دهند، و این دقیقا همان جایی است که هادوپ و دیگر فن‌آوری‌ها و شیوه‌های داده‌های بزرگ از آنجا می‌آیند. بسیاری از کاربران نهایی باید داده‌های بزرگ را مدیریت کنند زیرا این کار در کسب و کار آن‌ها ضروریست.

۳. شیوه‌های مشارکتی در مدیریت داده‌های بزرگ

مدیریت داده‌ها روشی جامع است که تعدادی از رشته‌ها، از جمله انبارداری داده‌ها، یکپارچه‌سازی داده‌ها، کیفیت داده‌ها، کنترل داده‌ها، مدیریت محتوا، پردازش رویداد، مدیریت پایگاه‌داده، و غیره را در بر می‌گیرد [۳۲]. BI/DW (هوش کسب و کار/انبار داده‌ها) و رشته‌های مرتبط بیش‌ترین دخالت را در BDM (مدیریت داده‌های بزرگ) همراه با زمینه‌های بسیار نزدیک یکپارچه‌سازی داده‌ها دارند. در حال حاضر، کیفیت داده‌ها و کنترل داده‌ها فقط رابطه‌ی متوسطی با مدیریت داده‌های بزرگ دارند. همانند تمام داده‌های سازمانی، داده‌های بزرگ مشکلات و فرصت‌هایی در ارتباط با کیفیت و کنترل دارند.

۴. کنترل داده‌ها

کنترل داده‌ها عبارتست از مشخص کردن حقوق تصمیم و چارچوب مسئولیت‌پذیری برای ترغیب رفتار مطلوب در ارزیابی، ایجاد، ذخیره‌سازی، استفاده، بایگانی و حذف داده‌ها و اطلاعات. کنترل داده‌ها شامل فرآیندها، نقش‌ها، استانداردها و معیارهای ارائه شده برای اطمینان از استفاده‌ی مؤثر و کارآمد از داده‌ها و اطلاعات در سازمان به منظور دستیابی به اهداف آن است کنترل داده‌ها یکی از مهم‌ترین اجزای مدیریت اطلاعات سازمانی است و با تمام رشته‌های دیگر در کارکردهای مدیریت داده‌ها ارتباط دارد. استفاده از کنترل داده‌ها یعنی یافتن مقدار مناسب و سطح درستی از کنترل.

۵. روش‌های فنی برای مدیریت داده‌های بزرگ

الف) BDM برای انواع و ساختارهای مختلفی از داده‌ها

امروزه، داده‌های ساختار یافته بیش از هر نوع داده‌ی دیگری مدیریت می‌شوند. اکثر این داده‌های ساختار یافته در واقع رابطه‌ای هستند و در نتیجه داده‌های رابطه‌ای هنوز بسیار مهم هستند. لاگ‌های وب و جریان‌های کلیک خیلی کم مدیریت می‌شوند در حالیکه از داده‌های رسانه‌های اجتماعی

استفاده‌ی زیادی می‌شود. تقریبا نیمی از سازمان‌های کاربر امروزه داده‌های حس‌گر، داده‌های ماشین و فرم های داده‌های جغرافیایی را مدیریت کرده و از آن‌ها استفاده می‌کنند و تا حدودی فرمت‌های داده‌های ثانویه‌ی برجسته‌ای هستند. داده‌های علمی و داده‌های نظارتی بعد از آن‌ها قرار دارند

ب) راهبردهای ذخیره‌سازی برای مدیریت داده‌های بزرگ:
اکثر داده‌های بزرگ در درایوهای قدیمی ذخیره می‌شوند، اما درایوهای حالت جامد و کارکردهای داخل حافظه در حال رشد هستند. یکی از بهترین روش‌ها پیکربندی سرور یا دستگاه با ترکیب انواع درایوها است. بخش عمده‌ای از داده‌های بزرگ روی درایوهای قدیمی با قیمت‌گذاری کالا ذخیره می‌شوند، در حالی‌که داده‌هایی که بطور مرتب توسط اپلیکیشن‌هایی با ارزش بالا به آن‌ها دسترسی می‌شود و در نتیجه نیاز به بهترین عملکرد جستار دارد به روی درایوهای حالت جامد گران‌قیمت منتقل می‌شوند.

ج) حجم داده‌های بزرگی که مدیریت می‌شوند
امروزه، منابع داده‌های بزرگ به حجم بالاتری از داده‌های بزرگ نیاز دارند. بسیاری از شرکت‌ها پیش‌بینی می‌کنند که در عرض سه سال آینده به حد یک پتابایت می‌رسند.

کاربردهای موردی
مدیریت داده‌های بزرگ برای سازمان‌ها مهم است، زیرا به آن‌ها کمک می‌کند تا داده‌ها را سازمان‌دهی و از آن‌ها برای اهداف آتی خود استفاده کنند. می‌توان از داده‌های بدست آمده پس از مدیریت برای تجزیه و تحلیل داده‌های بزرگ استفاده کرد و ارزش کسب و کار بسیار بالایی را می‌توان از روی آن‌ها تولید کرد. شرکت‌های موجود از داده‌های بزرگ برای افزایش فروش خود و بدست آوردن تعداد بیشتری از مشتریان در کسب و کار خود استفاده می‌کنند. داده‌های بزرگ مدیریت‌شده به شرکت‌ها کمک می‌کنند تا مدل‌های کسب و کار راهبردی جدیدی را برای جذب مشتریان در سراسر جهان تولید کنند.

۱. طراح مد و خرده فروش
Burberry که یکی از معروف‌ترین و موفق‌ترین مارک‌های مد در جهان است در حدود ۲۰ میلیون ارتباط در سامانه‌های مختلف شبکه‌های اجتماعی از طریق بازاریابی دیجیتالی و اجتماعی ایجاد کرد. بسیاری از شرکت‌ها میلیون‌ها طرفدار اما سود کمی دارند. با این حال، Burberry سود خود را با ایجاد روشی جدید برای خرید و همچنین برقراری ارتباط با علاقه‌مندان به مد افزایش داده است. Burberry این کار را با ارسال عکس‌ها و فیلم‌هایی از مجموعه‌های خود روی صفحات فیسبوک و توییتر حتی قبل از نمایش آن‌ها در فشن‌شوها انجام می‌دهد.. به همین دلیل، تعقیب‌کنندگان (فالوورهای) او می‌توانند با توجه به لایک‌ها، دیسلایک‌ها، موردها مورد علاقه، و روندها براحتی با شرکت تعامل داشته باشند. این شرکت از تجزیه و تحلیل پیش‌بینی برای تحلیل فعالیت‌های اجتماعی در پیش‌بینی ترجیح‌های مشتریان استفاده می‌کند. این شرکت داده‌ها را برای ایجاد تجربه‌ی یکنواخت بین ارتباطات اجتماعی، دیجیتالی و موبایلی و فروشگاه‌های واقعی کاوش می‌کند. هرچند Burberry مدل‌های قدیمی کسب و کار برای افزایش سود دارد، این شرکت مدل کسب و کار جدیدی را برای داشتن ۲۰ ملیون طرفدار در سراسر جهان با استفاده از روش‌های مدیریت داده‌های بزرگ روی داده‌های مشتری به راه انداخته است.

۲. نوشیدنی‌ساز
می‌توان از مدیریت داده‌های بزرگ در صنعت نوشیدنی نیز استفاده کرد. یک تولیدکننده‌ی نوشیدنی را در نظر بگیرید که صادرات و فروش نوشیدنی‌ها را به ایالات‌های مختلفی در داخل کشور ارسال می‌کند. اگر تولیدکننده سیستم توزیع خوبی نداشته باشد، کسب و کار او با فروش کمی مواجه خواهد شد. به منظور افزایش خرید، این شرکت می‌تواند از تکنیک‌های مدیریت داده‌های بزرگ استفاده کند. این شرکت می‌تواند از بازاریابی رسانه‌های اجتماعی، تبلیغات خیابانی و بازاریابی در پلت‌فرم‌های شبکه‌ی اجتماعی استفاده کند. از طریق داده‌کاوی و مدیریت داده‌ها، شرکت می‌تواند از انتخاب مشتریان آگاهی پیدا کند. همچنین شرکت می‌تواند علایق مصرف‌کنندگان را با استفاده از برنامه‌ی اندروید یا iOS ردیابی کند و نظرات آن‌ها را بداند. شرکت می‌تواند از این نظرات و علایق مصرف‌کنندگان برای بهبود محصولات استفاده کند که به شرکت کمک می‌کند تا فروش خود را در آینده افزایش دهند.

نتیجه‌گیری

پیدایش اینترنت منجر به ظهور داده‌های بزرگ شده است. حجم زیادی از داده‌ها توسط کسب و کارهای مختلف در سرتاسر جهان تولید می‌شود. امروزه، بسیاری از مسائل و چالش‌ها با توجه به ماهیت متنوع داده‌های بزرگ بوجود می‌آیند. کار روی داده‌ها در بین کسب و کارها نیز چالشی دیگر محسوب می‌شود. از طریق راه‌حل‌های منبع باز مانند هادوپ، به قابلیت تحمل خطا، مقیاس‌پذیری و تامین مزایای اجرا روی سخت‌افزار جامعه دست می‌یابیم. یکی دیگر از راه‌حل‌ها، یعنی مدیریت داده‌های اصلی یک دید ۳۶۰ درجه از داده‌ها در سراسر سازمان ارائه می‌دهد. با استفاده از شیوه‌های فنی و تحلیلی مناسب در مدیریت داده‌های بزرگ، شرکت‌ها می‌توانند ارزش کسب و کار را از روی داده‌ها تولید کنند.

مدیریت داده‌های بزرگ نقش بسیار مهمی در رشد امروز کسب و کارها بازی می‌کنند. داده‌های بزرگ قدرت تبدیل کسب و کارهای قدیمی به کسب و کارهای آنلاین موفق و افزایش درآمد برای شرکت‌های کنونی را دارند. با توجه به مدیریت داده‌های بزرگ، شرکت‌ها می‌توانند به ایده‌ها و مسیرهای جدید تحقیقاتی دست یابند. مدیریت داده‌های بزرگ ممکن است به تجزیه و تحلیل داده‌های بزرگ موثر کمک کند. اگرچه امروزه افزایش بسیار زیادی در داده‌هاو ایراداتی در راه‌حل‌های موجود مدیریت داده‌های بزرگ وجود دارد، با توجه به رقابت شدید در بازار، بسیاری از شرکت‌ها دائما در تلاش برای طراحی و راه‌اندازی راه‌حل‌های جدید برای مقابله با مسائل مربوط به داده‌های بزرگ هستند. سطح بالای کیفیت داده‌ها و دسترسی‌پذیری هوش کسب و کار هدف اصلی در مدیریت داده‌های بزرگ است.

مفهوم داده‌های بزرگ باقی می‌ماند. داده‌های بزرگ مسیر جدیدی برای کسب و کار در این عصر اطلاعات ارائه می‌دهند و امیدواریم که این مقاله تاثیر مثبتی داشته باشد و بتواند در کاهش چالش‌هایی که در مدیریت داده‌ها با آن مواجه هستیم کمک کند.

منابع

1. Challenges and Opportunities with Big Data, pg. 8-10.
2. Stephen Kaisler, Frank Armour, J. Alberto Espinosa, William Money, Big Data: Issues and Challenges Moving Forward, 2013 46th Hawaii International Conference on System Sciences, pg. 995 – 999.
3. Big Data Meets Big Data Analytics, SAS, pg. 4 – 7,
4. Oracle: Big Data for the Enterprise An Oracle White Paper, June 2013,pg. 3-12,
5. Datastax Corporation White Paper ,Big Data: Beyond the Hype Why Big Data Matters to You, October 2013,pg. 3-16
6. Doug Laney, Meta Group: Application Delivery Strategies, 6th February 2013, pg 1-4.
7. From The Editors, Big Data And Management , Academy Of Management Journal 2014, Vol.57, No.2, 321 – 326
8. Open Source Solutions for Big Data Management – Atos, pg.3-6,
9. Quantum White Paper, Big Data : Managing Explosive Growth The Importance Of Tiered Storage, pg. 3-7
10. Datastax corporation: Big Data: Beyond the hype Why big data matters to you. October2013.pp-11-12
11. Jaseena K.U., Julie M. David: Issues, Challenges, And Solutions: Big Data Mining.pp 7-9
12. Oguntimilehin. , Ademola E.O. : A Review of Big Data Management, Benefits and Challenges, Journal of Emerging Trends in Computing andInformation Sciences, Cis 6 June 2014.pp 3-4
13. Intel White Paper - Intel IT, Enabling Big Data Solutions with Centralized Data Management, January 2013, pg.1-7
14. Michael Feldman, The Big Data Challenge: Intelligent Tiered Storage at Scale White Paper, November 2013, pg.7-8
15. Capgemini White Paper, Mastering Big Data –Why taking control of the little things matters when looking at the big picture, pg. 3-11.
16. Reshmi Nath, Cognizant White Paper, Mastering Big Data: The Next Big Leap for Master Data Management, May 2013, pg.1-6.
17. IBM White Paper, Master Data Management: The Key to leveraging big data, November 2012, pg. 1-8.
18. Gautam Vemuganti, Metadata Management in Big Data, Infosys Labs Briefings, Vol 11 No 1, 2013, pg. 3-8.
19. Oracle White Paper on Enterprise Architecture, Enterprise Information Management: Best Practices in Data Governance, May 2011, pg. 3-8.
20. Lekha R. Nair, Sujala D. Shetty, Research in Big Data and Analytics: An Overview, International Journal of Computer Applications (0975- 8887), Volume 108 –No 14, December 2014, pg. 19-22.
21. Surajit Chaudhuri, What Next? A Half-Dozen Data Management Research Goals for Big Data and The Cloud, Microsoft Research, pg.1- 2.
22. Michael Schroeck, Rebecca Shockley, Dr. Janet Smart, Professor Dolores Romero-Morales and Professor Peter Tufano, IBM Institute of Business Value, Analytics: The real world use of big data, pg.2-10.