تجزیه و تحلیل داده های کلان

تبدیل داده های کلان به پول کلان

مولف: فرانک اُهل هورست

مترجمان:

دکتر کیوان رحیمی زاده

عضو هیات علمی دانشگاه یاسوج

دکتر محمدعلی ترکمانی

سرشناسه : اولهورست، فرانک، ۱۹۶۴ - م.
Ohlhorst, Frank
عنوان و نام پدیدآور : تجزیه و تحلیل داده های کلان، تبدیل داده‌های کلان به پول کلان/ اثر فرانک اوهل‌هورست؛ مترجمان کیوان رحیمی زاده، محمدعلی ترکمانی.
مشخصات نشر : مشهد: ارسطو، ۱۳۹۶.
مشخصات ظاهری : ۱۵۰ص.: مصور، جدول، نمودار.
شابک : ۹۷۸-۶۰۰-۴۳۲-۱۷۸-۵
وضعیت فهرست نویسی : فیپا
یادداشت : عنوان اصلی: Big data analytics : turning big data into big money, c2013.
موضوع : هوش تجاری
موضوع : Business intelligence
موضوع : داده‌کاوی
موضوع : Data mining
شناسه افزوده : رحیمی‌زاده، کیوان، ۱۳۵۷ -، مترجم
شناسه افزوده : ترکمانی، محمدعلی، ۱۳۵۴ -، مترجم
رده بندی کنگره : ۱۳۹۶ ۳ ت ۸۳ الف / ۷ / ۳۸ HD
رده بندی دیویی : ۶۵۸/۴۷۲
شماره کتابشناسی ملی : ۴۸۱۷۲۰۳

نام کتاب : تجزیه و تحلیل داده های کلان، تبدیل داده‌های کلان به پول کلان/ اثر فرانک اوهل‌هورست
مترجمان :دکتر کیوان رحیمی زاده(عضو هیات علمی دانشگاه یاسوج)-دکتر محمد علی ترکمانی
ناشر : ارسطو (با همکاری سامانه اطلاع رسانی چاپ و نشر ایران)
صفحه آرایی ، تنظیم و طرح جلد : محمد علی ترکمانی و علی بیات
تیراژ : ۱۰۰۰ جلد
نوبت چاپ : دوم -۱۳۹۸
تعداد صفحات: ۱۸۵ صفحه
چاپ : مدیران
قیمت : ۳۶۰۰۰ تومان
شابک : 978-600-432-178-5
تلفن های مرکز پخش : ۰۵۱۱ ۵۰۹۶۱۴۵ – ۵۰۹۶۱۴۶ ۰۹۱۷۷۱۶۴۹۴۰-۰۵۱۱

فهرست مطالب

مقدمه:

استفان اشتراوس نویسنده‌ی کتاب‌های سرگرم کننده برای کودکان در کتابی با عنوان "بزرگ چقدر بزرگ است؟" توضیح می‌دهد که "بزرگی چیزی است که هیچ‌کس نمی‌تواند آن را مصرف و تمام کند". داده های کلان بیان کننده وضعیتی است که مجموعه داده ها به اندازه ای رشد یافته اند که فناوری های اطلاعاتی مرسوم دیگر بیش از این نمی توانند به طور موثر اندازه یا مقیاس و رشد مجموعه داده ها را کنترل نمایند. به عبارت دیگر، مجموعه داده ها آنقدر رشد یافته است که مدیریت آن مشکل می باشد و حتی سخت تر از این ، کسب ارزش از آن دشوار می باشد.

در عصر داده‌های کلان شاهد افزایش رشد انفجاری داده‌ها در حوزه های مختلف هستیم. به همین علت موضوع داده های کلان یکی از مهمترین موضوعات تحقیقاتی در دوره های تحصیلات تکمیلی و صنعت است. کتاب پیش رو یکی از کتابهای ارزشمند در این حوزه است که توسط آقای فرانک اُهل هورست تالیف شده است و با توجه به حجم مناسبی که دارد، برای تدریس در یک نیمسال تحصیلی در دوره های کارشناسی ارشد و دکتری مناسب می‌باشد. امید است این اثر مورد توجه همکاران و دانشجویان عزیز قرار گیرد. از خوانندگان گرامی تقاضا داریم نقطه نظرات خود را از طریق ایمیل m.a.torkamani@gmail.com با مترجمین در میان بگذارند تا انشالله در ویرایش‌های بعدی کتاب اشکالات یا کاستی‌های احتمالی آن، مورد تجدید نظر قرار گیرد.

در پایان وظیفه خود می‌دانیم از آقای مهندس علی بیات به خاطر طراحی جلد کتاب و مدیریت انتشارات ارسطو و سامانه اطلاع رسانی چاپ و نشر ایران جناب آقای حسین قنبری تشکر و قدردانی نماییم.

مترجمین

پاییز ۱۳۹۶

فصل اول

داده های کلان چیست؟

داده های کلان واقعا چیست؟ در اولین نگاه ممکن است این اصطلاح نسبتاً مبهم به نظر رسیده و به چیزی اشاره کند که بزرگ و سرشار از اطلاعات است. این توصیف در واقع نیاز را برآورده می سازد ولی هیچ اطلاعاتی در رابطه با اینکه داده های کلان واقعا چیست نمی‌دهد. داده های کلان اغلب به عنوان مجموعه داده های بی‌نهایت بزرگی قلمداد می‌شود که به نحوی رشد یافته که قابل مدیریت و آنالیز با ابزار مرسوم داده پردازی نیست. جستجوی وب برای یافتن جواب، یک تعریف کلی را نمایان می‌سازد که توسط اکثریت ارتقا دهندگان این اصطلاح مشترکاً استفاده شده است، که می‌تواند به این صورت خلاصه شود:

داده‌های کلان بیان کننده وضعیتی است که مجموعه داده ها به اندازه ای رشد یافته اند که فناوری های اطلاعاتی مرسوم دیگر بیش از این نمی توانند به طور موثر اندازه یا مقیاس و رشد مجموعه داده ها را کنترل نمایند. به عبارتی دیگر، مجموعه داده ها آنقدر رشد یافته است که مدیریت آن مشکل می باشد و حتی سخت تر از این ، کسب ارزش از آن دشوار می باشد. مشکلات اصلی عبارتند از : اکتساب ، ذخیره سازی ، جستجو ، اشتراک ، تحلیل و تجسم داده ها.

چیزهای بیشتری را می توان در رابطه با اینکه داده های کلان در واقع چیست بیان نمود. مفهوم داده های کلان بوجود آمده که نه تنها شامل اندازه مجموعه داده ها باشد بلکه فرایندهای درگیر در تقویت و بهره گیری از داده ها را نیز شامل شود. داده های کلان حتی مترادف مفاهیم تجاری همچون کسب و کار هوشند ، تجزیه و تحلیل و داده کاوی به کار رفته است. از سوی دیگر، مفهوم داده های کلان خیلی نو نیست. اگرچه مجموعه داده های کلان فقط در دو سال گذشته ایجاد شده است ، داده های کلان ریشه در مجامعی علمی و پزشکی دارد که در آنها تجزیه و تحلیل پیچیده حجم خیلی زیادی از داده ها برای توسعه داروها ، مدلسازی های فیزیکی و دیگر حالات تحقیق و

پژوهش که همه با مجموعه داده ها سروکار داشته اند ، صورت می گرفته است. از اینرو ، ریشه های فراوان این مفهوم است که اینکه داده های کلان چیست ؛ را تغییر داده است.

از راه رسیدن تجزیه و تحلیل

با انجام عملیات تجزیه و تحلیل و پژوهش بر روی مجموعه داده های کلان، محققین به این نتیجه دست یافتند که لفظ "بیشتر" بهتر است. در این زمینه داده های بیشتر تجزیه و تحلیل بیشتر و نتایج بیشتر. محققین شروع به آمیختن مجموعه داده های مرتبط، داده های بدون ساختار، داده های بایگانی شده و داده های بلادرنگ در قالب فرایندها نمودند که منجر به تولد آنچه که ما امروزه داده های کلان می نامیم شد.

در دنیای تجارت، داده های کلان تماماً فرصت می باشد. بنا به گفته IBM هر روزه ما 2.5 کوینتیلیون (2.5×10^{18}) بایت داده تولید می کنیم، آنقدر که 90 درصد داده های موجود در دنیای امروز در دو سال گذشته تولید شده است. این داده ها از هر جایی می آیند: سنسورهایی که به منظور جمع آوری اطلاعات هواشناسی استفاده می شوند، اظهار نظرهای ثبت شده در سایتهای اجتماعی، تصاویر و ویدئو های آنلاین که در وبسایتها گنجانده شده اند، تراکنشهای ثبت شده از خریدهای آنلاین و علائم GPS تلفن های همراه، که اینها فقط اندکی از این داده ها می باشند. این کاتالیزوری برای داده های کلان است همراه با این حقیقت که همه این داده ها دارای ارزش ذاتی هستند که با تجزیه و تحلیل و الگوریتم ها و دیگر تکنیک ها اقتباس می شوند.

داده های کلان پیشتراهمیت و ارزش خودش را در زمینه های متعددی به اثبات رسانده است. سازمانهایی از قبیل سازمان ملی اقیانوسی و جوی (NOAA)، سازمان ملی هوا و فضا (NASA) ، چندین شرکت دارویی و تعدادی شرکت مربوط به انرژی مقادیر بسیار زیادی از داده ها را انباشته نموده اند و اکنون بطور روزانه از فناوریهای مربوط به داده های کلان به عنوان اهرمی یرای استخراج ارزش از داده ها استفاده می نمایند.

سازمان ملی اقیانوسی و جوی (NOAA) از رهیافتهای مربوط به داده های کلان برای کمک گرفتن در مورد شرایط اقلیمی و محیط زیست و آب و هوا و تحقیقات تجاری استفاده می نماید، در حالیکه سازمان هوا و فضا (NASA) از داده های کلان برای تحقیقات هوانوردی و دیگر تحقیقات

استفاده می کند. شرکتهای دارویی و انرژی برای نتایج ملموس تری مانند آزمایش دارویی و تجزیه و تحلیل ژئوفیزیکی از داده های کلان استفاده می نمایند روزنامه نیویورک تایمز از ابزار داده های کلان برای تجزیه و تحلیل داده های متنی و وبکاوی استفاده نموده است، درحالیکه شرکت والت دیزنی از داده های کلان برای درک رفتار مشتری ها در تمام فروشگاه هایش ، پارکهای تفریحی و دارایی های تحت وب خود استفاده می نماید.

داده های کلان نقش دیگری در تجارتهای امروزه دارد: سازمانهای بزرگ بطور فزاینده ای، برای پیروی از مقررات دولتی احتیاج دارند که حجم انبوهی از داده های ساختیافته و غیرساختیافته – از اطلاعات تراکنشی در انبارها تا توییت کارکنان، از داده های ثبت شده تولیدی ها تا فایلهای نظارتی - را نگهداری نمایند. این نیاز حتی با توجه به موارد اخیر دادگاهی که شرکتها را ترغیب نموده به نگهداری مقادیر بزرگ از اسناد، پیام های پست الکترونیکی و دیگر ارتباطات الکترونیکی از قبیل پیام دهی فوری و تلفن اینترنتی که ممکن است برای کشف الکترونیکی تخلفات به کار آید، بیشتر مورد توجه می باشد.

ارزش کجاست؟

استخراج ارزش در گفته بسیار آساتر از انجام است. داده های کلان سرشار از چالش می باشد، اعم از فنی تا مفهومی تا عملیاتی، که هر کدام می تواند مانعی برای استخراج ارزش و استفاده ارزشمند از داده های کلان باشد. شاید بهتر باشدکه داده های کلان را از چندین بعد در نظر بگیریم، که در این زمینه چهار بعد مربوط به جنبه های اصلی داده های کلان را می توان در نظر گرفت. این ابعاد را می توان بصورت زیر معرفی نمود:

۱-حجم. داده های کلان با اندازه بزرگ همراه است. تشکیلات اقتصادی لبریز از داده ها هستند و به آسانی هزاران گیگا بایت و حتی میلیون ها گیگا بایت از اطلاعات را جمع آوری می نمایند.

۲-تنوع. داده های کلان فراتر از داده های ساختیافته توسعه می یابند تا جایی که داده های غیرساختیافته از قبیل متن، صدا، ویدئو، فعالیتهای تحت وب کاربران در قالب کلیک نمودن های لینکها، فایلهای لاگ و حتی بیشتر را شامل می شود.

۳-صحت. حجم انبوه داده های جمع آوری شده برای استفاده به عنوان داده های کلان، می تواند منجر به اشتباهات آماری و سوء تعبیرهای اطلاعاتی شود. خالصی اطلاعات برای ارزش گذاری حیاتی می باشد.

۴-سرعت. اغلب بصورت حساس به زمان، داده های کلان باید استفاده شده و در امور اقتصادی جریان یابد تا ارزش و اعتبارش را برای امور تجاری بیشینه نماید، از طرفی دیگر به همان صورت، دسترسی به این داده ها از منابع بایگانی باید برقرار باشد.

این چهار بعد از داده های کلان، به همراه ارزش ذاتی هر کدام در فرایند کشف ارزش ، مسیر رو به تجزیه و تحلیل را نمابان می سازد.

با این وجود، پیچیدگی داده های کلان تنها با چهار بعد به پایان نمی رسد. عوامل دیگری نیز دخیل هستند: فرایندهایی که داده های کلان آنها را برانگیخته می سازند. این فرایندها، تجمیعی از فناوریها و تجزیه و تحلیل هایی است که برای تعریف ارزش منابع داده ای استفاده می شوند و به عنوان عناصر کاربردی در پیشبرد رو به جلو تجارت معنی می شوند.

تعداد زیادی از آن فناوریها و مفاهیم جدید نیستند اما تحت چتر حمایت داده های کلان قرار می گیرند. به عبارتی بهتر به عنوان دسته های تجزیه و تحلیلی معرفی می شوند. این فناوریها و مفاهیم از قرار زیر می باشند:

- **هوش تجاری سنتی.** این شامل دسته وسیعی از برنامه های کاربردی و فناوریهایی برای جمع آوری، ذخیره سازی، تجزیه و تحلیل، و فراهم نمودن دسترسی به داده ها می باشد. هوش تجاری اطلاعاتی کاربردی را فراهم می نماید که به کاربران تجارت و سرمایه کمک می کند تا تصمیم گیری های تجاری بهتری را با استفاده از سیستم های پشتیبان مبتنی بر واقعیت اتخاذ نمایند. هوش تجاری با استفاده از یک تجزیه و تحلیل عمقی از داده های تجاری دقیق که توسط بانکهای اطلاعاتی، برنامه های کاربردی و دیگر منابع ملموس داده ای فراهم شده است، کار می کند. در برخی از زمینه ها، هوش تجاری می تواند مشاهدات تاریخی، جاری و پیش بینی شونده از عملیات تجاری را فراهم نماید.
- **داده کاوی.** این فرایندی است که در آن داده ها از جنبه های مختلف تجزیه و تحلیل شده و نهایتا به داده های خلاصه شده تبدیل شده که مفید قلمداد می شوند. داده کاوی معمولا با داده های راکد یا داده های بایگانی شده به کار گرفته می شود. فنون داده کاوی

به جای توصیف محض بر روی مدلسازی و کشف دانش برای مقاصد پیش بینی تمرکز دارد – یک فرایند ایده آل برای نمایان سازی الگوهای جدید از مجموعه داده های کلان.

- **برنامه های کاربردی آماری.** اینها با استفاده از الگوریتم های مبتنی بر اصول آماری، به داده ها نگرش داشته و معمولا بر روی مجموعه داده های مرتبط به رای گیری، سرشماری و دیگر مجموعه داده های ایستا تمرکز دارند. برنامه های کاربردی آماری به طور مطلوب مشاهدات نمونه ای را دراختیار می گذارند که می توانند برای مطالعه مجموعه داده های جمعیتی به منظور تخمین، آزمایش و تجزیه و تحلیل پیش بینانه، به کار روند. داده های آزمایشی و تجربی از قبیل نظرسنجی ها و گزارشات آزمایشگاهی، منابع اصلی برای اطلاعات تحلیل پذیر می باشند.
- **تجزیه و تحلیل پیش بینانه.** این زیرمجموعه ای از برنامه های کاربردی آماری می باشد که در آن مجموعه داده ها تحت آزمایش قرار گرفته تا نتایج پیش بینی با توجه به تمایلات و اطلاعات جمع آوری شده از بانکهای اطلاعاتی بدست آید. پس از آنکه عناصر خارجی به مجموعه داده ها اضافه شد، تجزیه و تحلیل پیش بینانه در دنیای علمی و مالی،جایی که تمایلات پیش بینی ها را تحریک می کند ، گرایش به سمت بزرگ شدن دارد. یکی از اهداف اصلی تجزیه و تحلیل پیش بینانه تشخیص مخاطرات و فرصت ها برای فرایند تجارت، بازار و تولیدی ها می باشد.
- **مدلسازی داده ای.** این یک برنامه کاربردی مفهومی از تجزیه و تحلیل می باشد که در آن چندین سناریو "what - if" با استفاده از الگوریتم ها می تواند به چندین مجموعه داده اعمال شود. در حالت مطلوب، اطلاعات مدلسازی شده بر اساس اطلاعاتی که در اختیار الگوریتم ها قرار گرفته، تغییر می نمایند، و سپس یک درکی از تاثیرات این تغییرات بر روی مجموعه داده ها ایجاد می نماید. مدلسازی داده ها دست در دست تجسم سازی داده ها کار میکند، بطوریکه آشکار سازی اطلاعات می تواند به یک تلاش تجاری کمک نماید.

این دسته های تجزیه و تحلیل تنها بخشی از جاهایی است که داده های کلان به آنجا رسیده و تنها بخشی از ارزشهای ذاتی آن برای تجارت می باشد. این ارزش حاصل از تلاش بی پایان برای مزیت رقابتی، سازمانها را وادار میکند که تبدیل به منابع بزرگ از داده های شرکتی و بیرونی به

منظور آشکارسازی تمایلات و آمار و دیگر اطلاعات کاربردی شده تا به آنها کمک نماید که در حرکتهای آتی شان تصمیم گیری نمایند. این باعث شده تا مفهوم داده های کلان به همراه ابزار و چارچوب و تجزیه و تحلیل وابسته به آن، نزد تکنسین ها و مدیران شهرت پیدا کند.

داده های کلان بیشتر از آنچه تصور می شود

حجم و اندازه کلی مجموعه داده تنها بخشی از معادله داده های کلان می باشد. یک اتفاق نظر فزاینده در رابطه با اینکه هر دوی منابع داده ای نیمه ساختیافته و بدون ساختار شامل اطلاعات حیاتی تجاری می باشند و در نتیجه باید در اختیار هوش تجاری (BI) و نیازهای عملیاتی باشند، وجود دارد. همچنین، واضح است که مقدار داده های تجاری بدون ساختار مرتبط نه تنها در حال رشد می باشد بلکه حتی در آینده نیز به رشدش ادامه خواهد داد.

داده ها را می توان به چندین دسته طبقه بندی نمود: داده های ساختیافته، داد های نیمه ساختیافته و داده های بدون ساختار. داده های ساختیافته معمولا در بانکهای اطلاعاتی سنتی و مرسوم (مانند SQL و دیگران) در قالب داده های سازمان یافته درون جداول بر اساس قوانین تجاری تعریفی ، یافت می شوند. داده های ساختیافته معمولا اثبات شده است که ساده ترین نوع داده ها برای کار می باشند، و این به خاطر این است که داده ها تعریف و شاخص گذاری شده و درنتیجه دسترسی به آنها و فیلترسازی آنها ساده تر می باشد.

برعکس، داده های بدون ساختارمعمولا هیچگونه هوش تجاری در پشت آنها نیست. داده های بدون ساختار در قالب جداول سازماندهی نشده اند و از اینرو ذاتاً قابل استفاده توسط برنامه های کاربردی و یا تفسیر توسط بانکهای اطلاعاتی نمی باشند.

یک مثال خوب از داده های بدون ساختار مجموعه ای از فایلهای تصویری باینری خواهد بود. داده های نیمه ساختیافته حدواسط داده های ساختیافته و داده های بدون ساختار می باشد. این داده ها مانند داده های ساختیافته دارای یک ساختار رسمی شبیه بانک اطلاعاتی با جداول و رابطه ها نمی باشد. با این وجود، برخلاف داده های بدون ساختار، داده های نیمه ساختیافته دارای یکسری برچسب یا Tag و یا دیگر نشانه ها می باشند تا عناصر را از هم تفکیک نموده و سلسله مراتبی از رکوردها و فیلدها که داده ها را تعریف می نماید، بوجود آورد.

برخورد با تفاوتهای داده های کلان

برخورد با انواع مختلف داده ها همگرا می باشد، تقدیر از ابزار و برنامه هایی که می توانند مجموعه های داده ای را با استفاده از فرمت های استاندارد XML و استاندارد های داده ای خاص صنعتی XML (مانند ACORD در بیمه و HL7 در مراقبت پزشکی) پردازش نمایند. این فناوریهای XML در حال توسعه انواع داده ای می باشند که می توانند توسط ابزار یکپارچه سازی و تجزیه و تحلیل داده های کلان به کار گرفته شوند. در عین حال، قابلیت تبدیل این فرایندها هنوز تحت فشار ناشی از پیچیدگی و حجم داده ها بوده که منجر به عدم تطابق بین قابلیتهای تبدیل موجود و نیازهای مطرح می گردد. این در را به روی یک نوع جدید از محصول تبدیل داده ای فراگیر می گشاید که اجازه خواهد داد تا تبدیلات برای همه طبقات داده ای(ساختیافته، نیمه ساختیافته و بدون ساختار) تعریف شود بدون اینکه احتیاج به کدنویسی باشد، و این نوع داده ای جدید قادر خواهد بود تا در هر برنامه نرم افزاری یا در هر معماری پلت فرمی به کار گرفته شود.

معنی داده های کلان و انجام تجزیه و تحلیل مربوط به آن هنوز، در یک حالت تغییر و گداختگی می باشد؛ ابزار، فناوریها و رویه ها در حال تکامل می باشند. با این حال، این بدان معنی نیست که کسانی که بدنبال ارزش از مجموعه های داده ای هستند باید منتظر بمانند. داده های کلان برای فرایندهای تجاری خیلی با اهمیت تر از این است که رویکرد صبرکن و ببین اتخاذ شود. فوت و فن واقعی درباره داده های کلان، پیدا کردن بهترین راه برای مواجهه با منابع داده ای گوناگون در عین برآورده شدن اهداف از فرایند تجزیه و تحلیل می باشد. این مستلزم یک رویکرد هوشیارانه است تا سخت افزار، نرم افزار و رویه ها را در یک فرایند قابل مدیریت ادغام نموده که نتایج را در یک چارچوب زمانی قابل قبول بدست آورده و با داده ها شروع می شود.

مخزن ذخیره سازی یک عنصر بحرانی برای داده های کلان می باشد. داده باید جایی ذخیره و نگهداری شود، به سادگی قابل دسترسی باشد و نیز محفوظ باشد. این موضوع تایید شده که یک چالش گران قیمت برای بیشتر سازمانها می باشد چرا که خرید و مدیریت مخازن ذخیره سازی تحت شبکه از قبیل SANS و NAS می تواند خیلی گران باشد.

مخزن ذخیره سازی تکامل یافته تا تبدیل به یکی از عناصر عادی در مراکز داده ای معمول گردد – مهمتر از این، فناوری های ذخیره سازی به بلوغ رسیده و در حال نزدیک شدن به حالت کالا می

باشد. با این اوصاف، شرکتهای امروزی با نیازهای در حال رشدی مواجه می شوند که می تواند فشاری را بر فناوریهای ذخیره سازی وارد آورد. یک مورد آن فشار برای تجزیه و تحلیل داده های کلان است، مفهومی که قابلیتهای هوش تجاری را وارد مجموعه های داده ای بزرگ می کند. فرایند تجزیه و تحلیل داده های کلان متقاضی قابلیتهایی است که معمولا فراتر از استانداردهای ذخیره سازی معمولی است. فناوریهای ذخیره سازی سنتی از قبیل SANS و NAS و غیره نمی توانند ذاتاً با میلیون ها مگابایت و میلیونها گیگا بایت از اطلاعات بدون ساختار ارائه شونده توسط داده های کلان مقابله نمایند. موفقیت در تجزیه و تحلیل داده های کلان نیازمند چیز بیشتری است: یک روش جدید برای مواجهه با حجم زیاد داده ها، یک طرز فکر جدید از چارچوب ذخیره سازی.

ابزار ارمغانی یک Open Source

Hadoop یک پروژه منبع باز که بستری نرم افزاری را جهت کار با داده های کلان معرفی می نماید. اگرچه Hadoop مدت زمان مدیدی است که موجود می باشد ولی تجارتهای خیلی زیادی اینک شروع به استفاده و بهره مندی از قابلیتهایش نموده اند. بستر نرم افزاری Hadoop برای حل مشکلات ناشی از داده های با حجم خیلی زیاد طراحی شده است، علی الخصوص داده هایی که مخلوطی از داده های پیچیده ساختیافته و داده های بدون ساختار بوده و انعطاف لازم برای جای دهی در جداول را ندارند. پروژه Hadoop به خوبی در موقعیت هایی که مستلزم یاری جستن از تجزیه و تحلیل محاسباتی عمیق و بسیط مانند خوشه بندی و هدف گذاری می باشند، عمل می نماید.

برای یک تصمیم گیرنده که بدنبال استفاده از مزایای داده های کلان می باشد، Hadoop شایع ترین مشکلات مربوط به داده های کلان را حل می نماید: ذخیره سازی و دسترسی به مقادیر بزرگ داده ای بطور کارآمد. طراحی داخلی Hadoop این اجازه و قابلیت را به آن می دهد تا به عنوان یک بستر نرم افزاری اجرا گردد که قادر است بر روی تعداد خیلی زیادی از ماشین هایی که حافظه و دیسک اشتراکی ندارند عمل نماید. با این فرض، به آسانی می توان ارزشهای اضافی دیگری که Hadoop ارائه می نماید را دید: مدیران شبکه می توانند به آسانی دسته ای از سرورهای با کاربری خاص را خرید نمایند و آنها را درون کابینها تعبیه نموده و نرم افزار Hadoop را بر روی هر کدام از آنها اجرا نمایند.

نرم افزار Hadoop همچنین کمک می نماید تا خیلی از بارهای اضافی مدیریتی وابسته به مجموعه های داده ای کلان حذف شوند. از نظر عملیاتی، هنگامی که داده های یک سازمان به درون یک بستر نرم افزاری Hadoop بار می شود، نرم افزار داده ها را به تکه های قابل مدیریت می شکند و بطور خودکار آنها را در بین سرورهای مختلف پخش می نماید. ماهیت توزیعی داده به این معنی است که برای دسترسی به داده ها نیاز به مراجعه به یک محل خاص نیست؛ Hadoop داده های مقیم را ردیابی کرده و از داده ها با ایجاد چندین کپی ذخیره شده، محافظت می نماید. قابلیت تاب آوری ارتقا می یابد، زیرا اگر سروری از کار بیافتد یا خاموش گردد، داده ها می تواند بطور خودکار از یک کپی خوب و شناخته شده تکثیر گردد.

مدل نوعی Hadoop هنگام کار با داده ها چندین گام جلوتر می رود. به عنوان مثال، گرفتن محدودیت های وابسته به یک سیستم پایگاه داده ای مرکزی که متشکل است از یک دیسک درایو بزرگ که به سیستمی از رسته سرور با چندین پردازشگر متصل شده است. در این سناریو، تجزیه و تحلیل محدود به کارایی دیسک درایو و نهایتا تعداد پردازشگرهایی است که می توان خرید و دوام آورند.

با داشتن یک خوشه از Hadoop، هر سرور موجود در خوشه با بکارگیری توانایی Hadoop در توزیع کارها و داده ها در سرتاسر خوشه، می تواند در پردازش داده ها شرکت نماید. به عبارتی دیگر، با ارسال یک کد به هر کدام از سرورهای موجود در خوشه، یک کار شاخص گذاری شده صورت می گیرد و در نتیجه هر سروری بر روی تکه داده خودش عمل می نماید. سپس، نتایج بصورت یک نتیجه کلی ارائه می شود. نزد Hadoop، این فرایند بعنوان MapReduce به آن اشاره می شود بطوریکه کد یا پردازنده ها به تمامی سرورها نگاشت می شود و نتایج به یک مجموعه واحد کاهش می یابد.

این فرایند همان است که Hadoop را در مواجهه با مقادیر بزرگ داده ای خوب نشان می دهد. Hadoop داده ها را توزیع و پخش می کند و می تواند با تحت کنترل گرفتن تمامی پردازشگرهای موجود در خوشه بصورت موازی، از عهده سوالات محاسباتی پیچیده برآید.

احتیاط: موانع سر راه

با این حال، ماجراجویی در دنیای Hadoop یک تجربه اتصال و اجرا نیست؛ پیش نیازهای خاصی از قبیل نیازمندیهای سخت افزاری و تنظیمات پیکربندی باید دیده شود تا موفقیت تضمین شود. اولین گام، فهمیدن و تعریف نمودن فرایند تجزیه و تحلیل می باشد. خیلی از مدیران ارشد فناوری اطلاعات با تجزیه و تحلیل تجاری (BA) و یا فرایندهای هوش تجاری (BI) آشنا هستند و میتوانند با رایج ترین لایه فرایند مورد استفاده در ارتباط باشند: لایه استخراج، تبدیل و بارگذاری (ETL) و ایفای نقش حیاتی آن در ساخت راه حلهای هوش تجاری (BI) و یا تجزیه و تحلیل تجاری (BA). تحلیل داده های کلان نیازمند اینست که سازمانها داده های مورد تحلیل را فراهم نمایند، آنها را یکپارچه نموده و سپس متدهای تجمیع را قبل از تحویل آنها به لایه ETL بر روی آنها اعمال نماید، این عملیات باید با حجم بزرگی از داده ها صورت گیرد که می توانند ساختیافته یا بدون ساختار و از منابع متعددی مانند شبکه های اجتماعی، لاگ های داده ای، وبسایت ها، دستگاههای موبایل و سنسورها باشند.

Hadoop این عمل را با ترکیب فرایندهای عملی و ملاحظاتی از قبیل معماری خوشه ای مقاوم در برابر خطا، قابلیت انتقال توان محاسباتی نزدیکتر به داده ها، پردازش موازی یا دسته ای مجموعه های داده ای کلان و یک اکوسیستم باز که از لایه های معماری سازمانی از ذخیره سازی داده گرفته تا فرایندهای تجزیه و تحلیل را پشتیبانی می نماید؛ را محقق می نماید.

همه شرکتهای تجاری به آنچه تجزیه و تحلیل داده های کلان در اختیار می گذارد نیاز ندارند؛ آن شرکتهایی که نیاز دارند باید توانایی Hadoop در مواجهه با چالشها را مد نظر قرار دهند. به هر حال، Hadoop به تنهایی نمی تواند همه کارها را انجام دهد. شرکتهای اقتصادی لازم است اینکه چه اجزای اضافی مورد نیاز Hadoop می باشد تا پروژه Hadoop ایجاد شود را مورد توجه قرار دهند.

به عنوان مثال، یک مجموعه شروع کننده از اجزای Hadoop ممکن است از موارد زیر تشکیل شده باشد: HDFS و HBase برای مدیریت داده ها، MapReduce و OOZIE بعنوان یک چارچوب پردازشی، Pig و Hive به عنوان چارچوبهای توسعه برای بهره وری توسعه دهنده، و Pentaho (بستر مجتمع سازی داده ها) برای هوش تجاری. یک پروژه آزمایشی احتیاج به تعداد زیادی سخت افزاری ندارد. سخت افزار مورد نیاز می تواند بطور ساده شامل موارد زیر باشد: یک

جفت سرور با هسته های چندتایی، ۲۴ گیگا بایت یا بیشتر فضای حافظه ای RAM و ۱۲ عدد یا بیشتر هارد دیسک که هر کدام ۲ ترابایت ظرفیت دارد. این باید برای راه اندازی یک پروژه آزمایشی کافی باشد.

مدیران داده ها باید آگاه باشند که مدیریت موثر و پیاده سازی Hadoop نیاز به مقداری تخصص و تجربه دارد، و اگر تخصص لازم مهیا نباشد مدیریت فناوری اطلاعات باید همکاری با یک موسسه خدماتی که پشتیبانی کامل از پروژه Hadoop را ارائه می نماید، را مد نظر قرار دهد. این تخصصها اهمیت ویژه امنیت را بیان می کند؛ Hadoop، HDFS و HBase امنیت یکپارچه خیلی کمی را ارائه می دهند. به عبارتی دیگر، همچنان لازم است که داده ها را از سرقت و یا مصالحه محفوظ نگاه داشت.

با در نظر گرفتن همه موارد، یک پروژه درون سازمانی Hadoop بهترین گزینه برای آزمایش و بررسی قابلیتهای تجزیه و تحلیل داده های کلان می باشد. بعد از انجام آزمایش، حجم زیادی از راه حلهای تجاری و یا نگهداری شده بوجود می آید که قابل دسترسی توسط کسانی خواهد بود که میخواهند بیشتر پا در عرصه تجزیه و تحلیل داده های کلان بگذارند.