تصورات اشتباه در ارتباط با مدل‌های داده‌ای

با 10 باور نادرست در ارتباط با بزرگ داده‌ها آشنا شوید

برای آنکه اطمینان حاصل کنید برنامه‌های سازمان شما در ارتباط با بزرگ داده‌ها در مسیر درستی قرار گرفته‌اند یا نه بهتر است با 10 باور اشتباه و رایجی که در ارتباط با این فناوری وجود دارند آشنا شوید. به طور مثال، یکی از عجیب‌ترین جملاتی که در ارتباط با بزرگ داده‌ها اغلب می‌شنویم این است که اگر مقدار کمی از داده‌ها خوب هستند، پس حجم بسیار بالایی از داده‌ها ایده‌آل خواهد بود. به نظر شما این حرف تا چه اندازه می‌تواند درست باشد؟ برای آنکه به این پرسش پاسخ روشنی دهیم، اجازه دهید از یک مثال ساده استفاده کنیم. گفتن این جمله درست مصداق این است که بگوییم در یک تابستان گرم یک نسیم خنک حس خوبی را به وجود می‌آورد، اما آیا یک گردباد نیز همان حس را به شما منتقل خواهد کرد؟

این مطلب مقدمه مجموعه مقالات پرونده ویژه «داده‌های بزرگ؛ فردای بزرگ‌تر» شماره 197 ماهنامه شبکه است. علاقه‌مندان می‌توانند کل این پرونده ویژه را از روی سایت شبکه دانلود کنند.

مطلب پیشنهادی

دانلود کنید: پرونده ویژه داده‌های بزرگ؛ فردای بزرگ‌تر

1- بزرگ داده‌ها به‌معنای حجم بسیار زیادی از داده‌ها هستند

در مرکز ثقل این مفهوم، بزرگ داده‌ها توصیف‌گر این نکته هستند که چگونه می‌توان داده‌های ساخت‌یافته و غیر ساخت‌یافته‌ای که از طریق تحلیل شبکه‌های اجتماعی، اینترنت اشیا یا دیگر منابع خارجی به دست می‌آیند را با یکدیگر ترکیب کرد، به طوری که بزرگ داده‌ها درنهایت توصیف‌گر یک داستان بزرگ‌تر شوند. این داستان ممکن است توصیف‌گر یک عملیات سازمانی یا نمایی از یک تصویر بزرگ باشد که از طریق متدهای تحلیل سنتی امکان ترسیم آن وجود نداشته است. همیشه به این نکته توجه داشته باشید که حجم بسیار بالای داده‌ها به‌معنای آن نیست که داده‌ها کارآمد هستند، حال آنکه در بعضی موارد پیچیدگی کارها را دو چندان می‌کند. در اختیار داشتن حجم بالای داده‌ها به‌معنای آن است که بگویید با در اختیار داشتن یک اسب بسیار قدرتمند رسیدن به خط پایانی مسابقه کار سختی نیست، حال آنکه برای برنده شدن در یک مسابقه سوارکاری شما به اسبی نیازی دارید که نه‌تنها قوی باشد، بلکه

به‌خوبی آموزش دیده باشد و مهم‌تر از آن به سوارکار ماهری نیاز دارید که بااستعداد باشد.

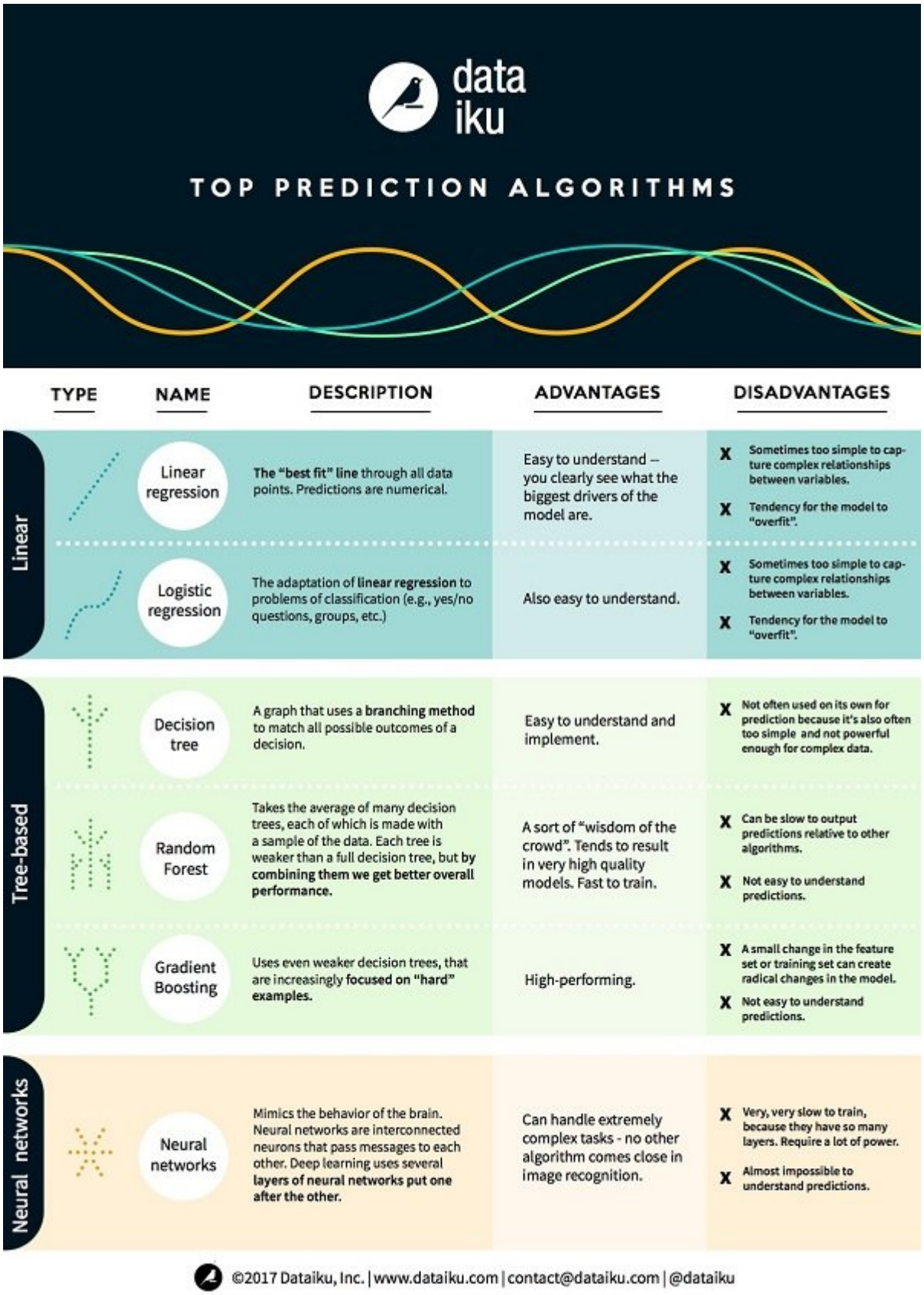
2- بزرگ داده‌ها باید به‌شکل تمیز و درست در اختیار شما قرار داشته باشند

آریجیت سنگوپتا مدیرعامل BeyondCore می‌گوید: «یکی از بزرگ‌ترین افسانه‌هایی که در این زمینه وجود دارد این است که شما باید داده‌های درستی را برای تحلیل در اختیار داشته باشید. اما هیچ‌کس چنین داده‌هایی در اختیار ندارد. این فرضیه از عقل به دور است که من برای آنکه بتوانم فرآیند تحلیل را انجام دهم، در ابتدا باید همه داده‌هایم شفاف و روشن باشند. کاری که شما انجام می‌دهید این است که تجزیه و تحلیل خود را به‌خوبی انجام دهید.» واقعیت این است که شما در اغلب موارد داده‌های خود را به‌شکل کاملاً درهم دریافت می‌کنید و بر مبنای همین داده‌ها باید فرآیند تحلیل را انجام دهید. به همین دلیل است که امروزه ما اعلام می‌داریم با مشکل کیفیت داده‌ها روبه‌رو هستیم. به‌رغم مشکل کیفیت داده‌ها، ما باز هم شاهد شکل‌گیری الگوهای کاملی بودیم که حتی با وجود مشکل کیفیت داده‌ها باز هم به‌خوبی موفق شده‌اند فرآیند تحلیل‌ها را به‌درستی انجام دهند. اما سؤال مهم این است که چگونه می‌توانیم داده‌های خود را شفاف و روشن کنیم؟ روش انجام این کار اجرای برنامه تحلیلگری است که از آن استفاده می‌کنید. برنامه شما باید بتواند نقاط ضعف را در مجموعه داده‌های شما شناسایی کند. یک مرتبه که این ضعف‌ها شناسایی شدند، برنامه تحلیلگر خود را دو مرتبه اجرا کنید تا داده‌های شفاف و روشن را به دست آورید.

واقعیت این است که شما در اغلب موارد داده‌های خود را به‌شکل کاملاً درهم دریافت می‌کنید و بر مبنای همین داده‌ها باید فرآیند تحلیل را انجام دهید

3- همه تحلیلگران انسانی باید با الگوریتم‌ها جایگزین شوند

توصیه‌هایی که از سوی متخصصان علم داده‌ها مطرح می‌شود، غالباً از سوی مدیران کسب و کار چندان رعایت نمی‌شود. آریجیت در مقاله‌ای که در سایت TechRepublic منتشر شده در این ارتباط گفته است: «پیشنهادها اغلب مشکل‌تر از آن چیزی هستند که بتوان آن‌ها را در پروژه‌های علمی به کار گرفت. با این حال، اعتماد بیش از اندازه به الگوریتم‌های یادگیری ماشینی ممکن است به همان اندازه چالش‌برانگیز باشد. الگوریتم‌های یادگیری ماشینی به شما می‌گویند چه کاری را انجام دهید، اما آن‌ها توضیح نمی‌دهند چرا باید این کار را انجام دهید. همین موضوع باعث می‌شود تا ادغام‌سازی تحلیل‌ها با برنامه‌ریزی‌های استراتژیک سازمان به‌سختی امکان‌پذیر باشد.» (شکل 1)



شکل 1- محدوده الگوریتم های پیش‌بینی کننده از الگوریتم های نسبتاً ساده شروع می‌شود و به الگوریتم های پیچیده‌تر درختی و درنهایت به شبکه‌ها ی عمیق عصبی می‌رسد.

مطلب پیشنهادی

چالش بزرگ بعدی اکوسیستم داده‌محور چیست؟
انتقال سریع و هوشمندانه داده‌ها، معدن طلای عصر جدید

4- دریاچه داده و انباره داده دو مفهوم یکسان هستند

جیم آدلر متخصص علم داده‌ها در بخش تحقیقات تویوتا می‌گوید: «مخازن ذخیره‌ساز بزرگی که تعدادی از مدیران فناوری اطلاعات در آرزوی دستیابی به آن هستند، پذیرای حجم بسیار بالایی از داده‌های ساخت‌یافته و غیر ساخت‌یافته‌ای هستند که به این سادگی‌ها به دست نمی‌آیند.» در یک انباره داده‌ها که به طور ویژه به‌منظور تحلیل و گزارش‌گیری‌های مدیریتی پیاده‌سازی می‌شود، اطلاعات ورودی پردازش شده و به یک ساختار هماهنگ تبدیل و ذخیره‌سازی می‌شوند، حال آنکه در یک دریاچه داده‌ها اطلاعات ورودی به همان شکلی که هستند ذخیره‌سازی می‌شوند. یک دریاچه داده در اصل یک مخزن ذخیره‌سازی است که پذیرای حجم بسیار زیادی از اطلاعات ساخت‌یافته و بدون ساختار است. داده‌های درون یک دریاچه داده اغلب در متن بزرگ داده‌ها به کار گرفته می‌شوند. در اغلب موارد سازمان‌ها بدون آنکه نظارت دقیقی بر داده‌ها داشته باشند همه داده‌ها را به یک باره به درون دریاچه‌ها وارد می‌کنند. در یک انباره داده فرآیند پردازش روی داده‌های آماده شده انجام می‌شود، اما در یک دریاچه داده ما بر حسب نیاز ابتدا داده‌ها را سازمان‌دهی می‌کنیم و پس از آن پردازش می‌کنیم. رویکردی که در ارتباط با دریاچه داده‌ها دنبال می‌شود نه‌تنها باعث شفاف‌سازی می‌شود، بلکه به‌راحتی پاسخ‌گوی نیازهای سازمانی است و به‌راحتی با اهداف حاکمیتی سازمان منطبق می‌شود.

مطلب پیشنهادی

برای حرکت به‌سوی SDDC چه اقداماتی لازم است؟
پنج استراتژی برای موفقیت مرکز داده نرم‌افزار محور

5- الگوریتم‌های پیش‌بینی‌کننده از خطا مصون هستند

زمانی نه‌چندان دور، الگوریتم هوشمند ساخته شده از سوی گوگل موسوم به Google Flu Trends ادعا کرد که می‌تواند مکانی که شیوع آنفلوآنزا در آن جا سریع‌تر خواهد بود را پیش‌بینی کند و به مراکز کنترل بیماری‌ها در ایالات متحده و دیگر سرویس‌های فعال در زمینه بهداشت و درمان توصیف‌های لازم را دهد. بسیاری تصور می‌کردند که این رویکرد به‌سادگی پیش‌بینی اوضاع جوی است که با استناد به دمای محلی پیش‌بینی‌هایی را ارائه می‌کند. در نتیجه می‌توان بر اساس جست‌وجوی افرادی که در شرایط مشابه در گذشته به آنفلوآنزای خوکی گرفتار شده‌اند این موضوع را نیز پیش‌بینی کرد. اما درنهایت الگوریتم هوشمند پیش‌بینی‌کننده گوگل در یک تله اطلاعاتی گرفتار شد. زمانی که فرآیند داده‌کاوی روی مجموعه گسترده‌ای از داده‌ها انجام شود، به همان نسبت ضریب خطا افزایش پیدا می‌کند. به‌واسطه آنکه ممکن است روابط اطلاعاتی بی‌موردی که تنها به‌لحاظ آماری قابل توجه هستند مورد توجه قرار گیرند.

6- شما نمی‌توانید برنامه‌های مربوط به بزرگ داده‌ها را روی زیرساخت‌های مجازی اجرا کنید

زمانی که بزرگ داده‌ها به‌شکل عمومی مورد توجه مردم قرار گرفت، نزدیک به ده سال پیش بود که تقریباً مترادف با ظهور آپاچی هادوپ بود. جاستین موری از شرکت VMware در 12 می 2017 میلادی در مقاله‌ای تحت عنوان Inside Big Data به این موضوع اشاره کرد و اعلام داشت که این واژه اکنون با فناوری‌های رایج از (NoSQL MongoDB، Apache Cassandra) گرفته تا آپاچی اسپارک احاطه شده است. منتقدان در گذشته درباره عملکرد هادوپ روی ماشین‌های مجازی پرسش‌هایی را مطرح کرده بودند، اما موری به این موضوع اشاره می‌کند که عملکرد

هادوپ روی ماشین‌های مجازی با عملکرد هادوپ روی ماشین‌های فیزیکی قابل مقایسه است. موری همچنین به این نکته اشاره کرده است که بسیاری بر این باورند که ویژگی‌های اساسی VM به SAN (سرنام Storage Area Network) نیاز دارند، اما این موضوع به هیچ عنوان صحت ندارد. با این حال، فروشندگان اغلب توصیه می‌کنند از ذخیره‌سازهای DAS استفاده شود، به‌واسطه آنکه عملکرد بهتری دارند و هزینه‌های پایین‌تری را تحمیل می‌کنند.

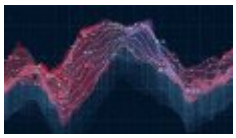
مطلب پیشنهادی

رویای دیروز، واقعیت امروز
اینترنت اشیا با کلان داده‌ها عجین شده است

7- یادگیری ماشینی مترادف با هوش مصنوعی است

شکاف میان الگوریتمی که قادر است الگوها را در حجم عظیمی از داده‌ها شناسایی کند و الگوریتمی که قادر است یک نتیجه منطقی را بر اساس الگوهای داده‌ای نشان دهد، بسیار زیاد است. وینیت جی از ITProPortal در مقاله‌ای که در ارتباط با یادگیری ماشینی منتشر کرده بود، به این موضوع اشاره کرد که یادگیری ماشینی از آمارهای تفسیری برای تولید مدل‌های پیش‌بینی‌کننده استفاده می‌کند. این فناوری در پس‌زمینه الگوریتم‌هایی قرار دارد که پیش‌بینی می‌کنند یک مصرف‌کننده بر مبنای تاریخچه خریدهای خود در گذشته احتمال دارد چه محصولی را خریداری کند یا به چه آهنگ‌هایی بر مبنای تاریخچه گذشته خود گوش فرا دهد. آن چنان که الگوریتم‌ها به‌سمت هوشمندی پیش می‌روند، ممکن است از دستیابی به اهداف هوش مصنوعی که به‌منظور الگوبرداری از تصمیمات انسانی بود دور شوند. پیش‌بینی‌های مبتنی بر آمار فاقد استدلال، قضاوت و تخیل انسانی هستند. حتی پیشرفت‌های حاصل شده در سامانه‌های هوش مصنوعی امروزی همچون واتسون آی‌بی‌ام نیز قادر نیستند در بسیاری از موارد بینشی که دانشمندان علم داده مطرح می‌کنند را ارائه دهند.

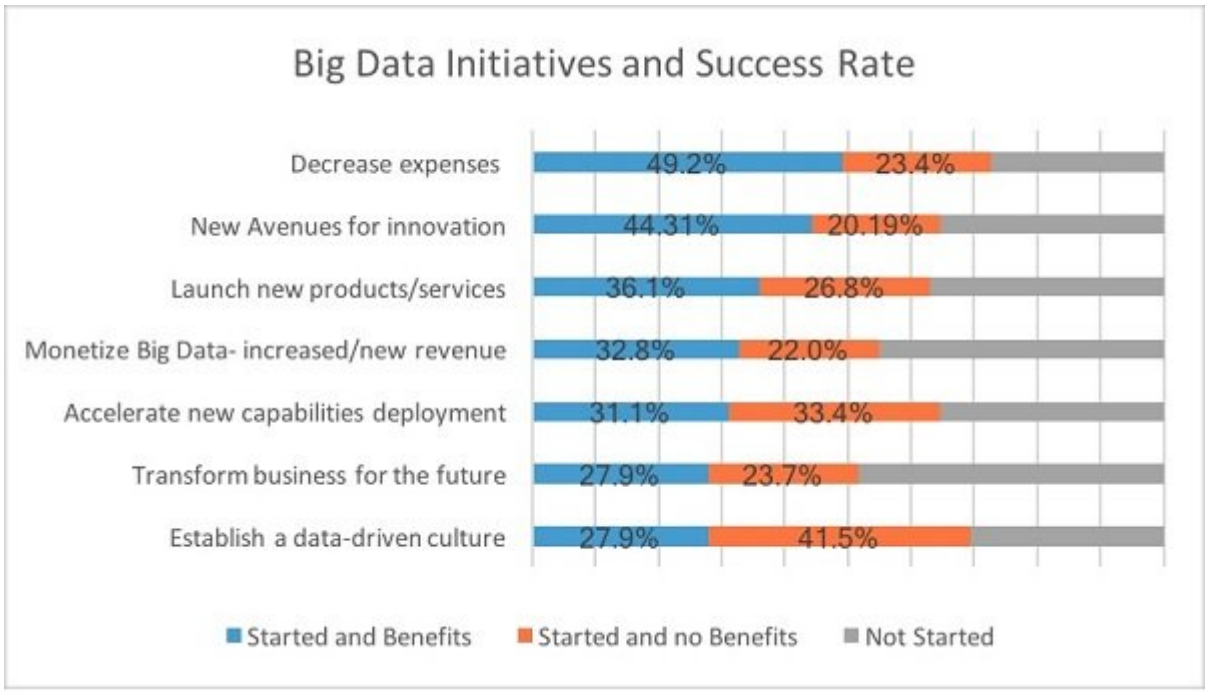
مطلب پیشنهادی



بزرگ داده‌ها بازیگر اصلی فناوری‌های فردا
بزرگ داده‌ها چه هستند و چرا باید به آن‌ها اهمیت دهیم؟

8- اکثر پروژه‌های بزرگ داده‌ها حداقل به نیمی از اهداف از پیش تعیین شده دست پیدا می‌کنند

مدیران فناوری اطلاعات به‌خوبی از این موضوع اطلاع دارند که هیچ پروژه تحلیل داده‌ای به‌طور صددرصد موفقیت‌آمیز نخواهد بود. (شکل 2) زمانی که پروژه‌ای درگیر بزرگ داده‌ها می‌شود، نرخ موفقیت‌آمیز بودن آن ممکن است بالا نباشد. آن چنان که نتایج به دست آمده از نظرسنجیNewVantage Partners این موضوع را به‌خوبی نشان می‌دهد. در نظرسنجی فوق که 95 درصد از رهبران کسب و کار در آن حضور داشتند، بسیاری بر این باور بودند که شرکت‌های مطبوع آن‌ها در 5 سال گذشته تنها در 48.4 درصد از پروژه‌های بزرگ داده‌های خود موفق بوده و به چشم‌اندازهای از پیش تعیین شده در ارتباط با این پروژه‌ها دست پیدا کرده است. بر اساس پژوهشی که سال گذشته میلادی از سوی گارتنر انجام و نتایج آن منتشر شد، پروژه‌های بزرگ داده‌ها به‌ندرت از مرحله آزمایشی سربلند خارج می‌شوند. نظرسنجی گارتنر نشان داد تنها 15 درصد از پروژه‌های بزرگ داده‌ها با موفقیت به مرحله استقرار و استفاده عملی می‌رسند.



شکل 2- نظرسنجی بزرگ انجام شده از سوی مؤسسه NewVantage Partners نشان داد کمتر از نیمی از پروژه‌های بزرگ داده‌ها به اهداف خود رسیده‌اند. همچنین، تغییرات فرهنگی در این زمینه به‌سختی به سرانجام می‌رسد.

حقوق سالانه مهندسان داده به طور میانگین حدود 130 هزار تا 196 هزار دلار است، در حالی که دستمزد متخصصان علم داده‌ها به‌طور میانگین در محدوده 116 هزار تا 163 هزار دلار قرار دارد و همچنین دستمزد تحلیلگران هوش تجاری نیز به طور میانگین در محدوده 118 هزار دلار تا 138 هزار دلار قرار دارد

مطلب پیشنهادی

داده‌های بزرگ؛ فردای بزرگ‌تر
دنیایی که در آن داده‌ها حرف اول و آخر را می‌زنند

9- قیام بزرگ داده‌ها میزان تقاضا برای مهندسان داده‌ها را کاهش خواهد داد

اگر هدف سازمان شما از به‌کارگیری پروژه‌های بزرگ داده‌ها این است که وابستگی خود به مهندسان داده را کم کند، باید بدانید که به‌دنبال یک سراب هستید.
Robert Half Technology Salary Guide در سال جاری میلادی نشان داد حقوق سالانه مهندسان داده به طور میانگین حدود 130 هزار تا 196 هزار دلار است، در حالی که دستمزد متخصصان علم داده‌ها به‌طور میانگین در محدوده 116 هزار تا 163 هزار دلار قرار دارد و همچنین دستمزد تحلیلگران هوش تجاری نیز به طور میانگین در محدوده 118 هزار دلار تا 138 هزار دلار قرار دارد.

10- مدیران میانی و کارکنان با آغوش باز بزرگ داده‌ها را خواهند پذیرفت

نظرسنجی NewVantage Partners نشان داد 85 درصد از شرکت‌ها متعهد شده‌اند تا یک بستر فرهنگی داده‌محور را پیاده‌سازی کنند. با این حال، میزان موفقیت کلی این طرح تنها 37 درصد بوده است. سه مانعی که اغلب این شرکت‌ها در عدم پیاده‌سازی موفقیت‌آمیز این طرح به آن اشاره کرده‌اند، عدم سازگاری سازمانی (42 درصد)، عدم پذیرش و درک مدیریتی (41 درصد) و مقاومت تجاری یا عدم درک صحیح (41 درصد) است.
آینده ممکن است به بزرگ داده‌ها تعلق داشته باشد، اما تحقق مزایای این فناوری به تلاش بیشتر در حوزه تجاری و همچنین مشارکت گسترده عامل انسانی بستگی دارد.

تاریخ انتشار:
05 دی 1396

نشانی منبع:
https://www.shabakeh-mag.com/cover-story/10810/%D8%A8%D8%A7-10-%D8%A8%D8%A7%D9%88%D8%B1-%D9%86%D8%A7%D8%AF%D8%B1%D8%B3%D8%AA-%D8%AF%D8%B1-%D8%A7%D8%B1%D8%AA%D8%A8%D8%A7%D8%B7-%D8%A8%D8%A7-%D8%A8%D8%B2%D8%B1%DA%AF-%D8%AF%D8%A7%D8%AF%D9%87%E2%80%8C%D9%87%D8%A7-%D8%A2%D8%B4%D9%86%D8%A7-%D8%B4%D9%88%DB%8C%D8%AF