دانستنی‌هایی درباره کلان داده

رقیه دهستانی

امروزه درباره کلان داده چیزهای بسیاری می‌شنویم. برای بسیاری از مردم مفهوم کلان داده به معنی سیل عظیمی از داده‌ها می‌باشد. اما حقیقتا کلان داده به چه معنا و مفهوم است؟ چه تفاوتی بین دو واژه کلان داده و مقدار عظیمی از اطلاعات وجود دارد؟ بر طبق نظریه گارتنر اطلاعات زمانی به کلان داده تبدیل می‌شود که حجم داده‌ها توسط ابزارهای پایگاه‌های معمولی قابل مدیریت نباشد. این نوشته به حقایق و یافته‌هایی جالب درباره کلان داده با نگاهی نو پرداخته است.

داده‌ها همه جا هستند.

تنها طی ۶ سال دنیای دیجیتال از ۳. ۲ زتا بایت[۱] تا ۴۰ زتا بایت رشد کرده است. هر روز ما ۲. ۵ کوینتیلیون[۲] بایت داده تولید میکنیم. بیش از ۹۰ درصد داده‌ها در دنیای امروز تنها در ۲ سال اخیر ایجاد شده است. این دادها از هر کجایی آمده‌اند :حسگرهایی که برای جمع‌آوری اطلاعات آب و هوا بکار می‌روند، پست‌هایی که به سایت‌های رسانه‌های اجتماعی ارسال میکنیم، ویدئوها و عکس‌های دیجیتالی، خریداری گزارشات مالی، سیگنال‌های جی پی اس تلفن همراه و غیره. . حجم اطلاعاتی که توسط شرکت‌های آمریکایی در هر سال تولید می‌شود تنها کافیست تا ۱۰۰۰۰ کتابخانه کنگره را پر کند.

زاکر برگ[۳] خاطر نشان می‌کند که ۱ بیلیون جزء از محتوا از طریق گراف‌های آزاد فیس بوک به اشتراک گذاشته می‌شوند. فیس بوک بیش از ۱۰ میلیون عکس در هر ساعت روی صفحه میگذارد و تقریبا ۳ بیلیون لایک هر روز به سمت فیس بوک هدایت می‌شود. گوگل هر روز بیش از ۲۴ پتا بایت[۴] از داده‌ها را مورد پردازش قرار میدهد. هر دقیقه ۴۸ ساعت ویدئو در یوتیوب آپلود می‌شود این مقدار برابر با ۸ سال محتوا در هر روز می‌شود.

۷۰ درصد داده‌ها توسط افراد ایجاد می‌شود اما شرکت‌ها مسئول ذخیره و مدیریت ۸۰ درصد داده‌ها می‌باشند.

۸۸ درصد داده‌ها نادیده گرفته می‌شوند.

بر طبق مطالعه تحقیقاتی فارستر[۵]، اغلب شرکت‌ها صرفا ۱۲ در صد داده‌ها را مورد تجزیه و تحلیل قرار می‌دهند و اغلب ۸۸ درصد داده‌ها نادیده گرفته می‌شود. کمبود ابزارهای تحلیلی و همچنین انبوه سیلوهای داده‌های واپس زده دو عامل مهمی است که باعث می‌شود شرکت‌های بزرگ حجم بزرگی از داده هایشان را در نظر نگیرند. فارستر خاطر نشان می‌کند حقیقت ساده اینست که چه اطلاعاتی ارزشمند است و چه اطلاعاتی بهتر است صرف نظر شود.

داده‌های ساختار یافته در برابر داده‌های غیر ساختاری

---

[۱] zettabytes

[۲] quintillion

[۳] Zuckerberg

[۴] petabytes

[۵] Forrester

رده بندی داده‌ها، همانطور که خدمات مشاوره‌ای TCS محدود کرده است بر این اساس است که چه تعداد از داده‌های شرکت‌ها ساختار یافته و یا غیر ساختاری هستند و یا چه مقدار از داده‌ها درون سازمان بوجود آمده است و چه مقدار بیرون سازمان. بر اساس نتایج تحقیقات ۵۱ درصد داده‌ها ساختار یافته و ۲۷ درصد غیر ساختار ی و ۲۱ درصد هم نیمه ساختاری هستند. کمتر از یک چهارم داده‌ها نیز برون سازمانی هستند.

رونق و شکوفایی مشاغل، اما کاهش مهارت‌ها

در سال ۲۰۱۵، ۴. ۴ میلیون از مشاغل تکنولوژی اینترنت بصورت جهانی بر اساس کلان داده ایجاد شده است و ۱. ۹ میلیون از این مشاغل IT در ایالات متحده شکل گرفته است. هر عملکرد وابسته به کلان داده ۳ شغل دیگر برای ۳ نفر در بیرون از حوزه IT ایجاد می‌کند بنابراین ظرف ۴ سال آینده، ۶ میلیون شغل توسط اقتصاد اطلاعات ایجاد خواهد شد. اما چالش چیست؟ مهارت وتخصص کافی در این صنعت وجود ندارد. در سال ۲۰۱۸ آمریکا به تنهایی توانست با این کاهش مقابله کند. ۱۴۰۰۰۰تا ۱۹۰۰۰۰ افراد با مهارت های تحلیلی عمیق به همان خوبی ۱. ۵ میلیون مدیر و تحلیل گر کلان داده‌ها توانستند در تصمیم گیریهای موثر شرکت کنند.

طبق آمار سایت شغل‌یابی مربوط به تکنولوژی (دیک[۶]) متخصصان متقاضی برای حرفه‌های مهندسی و تکنولوژی و تخصص‌های NoSQL سال به سال در حال رشد ۵۴ درصدی است و مشاغل مربوط به مهارت کلان داده تنها در ماه آوریل رشد ۴۶ درصدی داشته است. مشابه این جریان تخصص‌های حرفه‌ای هادوپ[۷] و پایتون[۸] به ترتیب بیش از ۴۶ درصد و ۱۶ درصد بوده است.

کلان داده به معنی کلان پول است.

طبق آمار مدیس[۹] (فراهم آورنده جهانی خدمات پرسنلی تکنولوژی اینترنت)، دانشمندان داده، بسیار مورد تقاضا هستند اما تعداد کمی را شامل می‌شوند و در نتیجه حقوق شش رقمی سخاوتمندانه‌ای برای افراد با مدرک دکترا و تجربه کاری در حوزه کلان داده در نظر گرفته می‌شود.

طبق آمار رسمی شرکت بورک ورکز[۱۰] (این شرکت افراد با تخصص‌های داده بالا را به مشاغل مهم پیوند می‌زند)، حقوق پایه برای کارمند متخصص کلان داده ۱۲۰۰۰۰ دلار و برای مدیر در این زمینه ۱۶۰۰۰۰ دلار در ماه است. این ارقام بر اساس مصاحبه‌ها با بیش از ۱۷۰ متخصص داده در پایگاه استخدامی شرکت بورک ورکز می‌باشد. اگرچه واحد سنجش حقوق متخصصان داده نشان می‌دهد که حقوق این افراد در آسیا و اروپا بطور قابل ملاحظه‌ای پایین می‌باشد.

کیفیت داده

بیش از نیمی از رهبران تکنولوژی اینترنت (۵۷ درصد) و متخصصان این حوزه (۵۲ درصد) گزارش می‌دهند که آنها همیشه نمی‌دانند که چه کسی منبع داده اطلاعاتی می‌باشد. اگر یک نفر نداند که منبع اصلی داده چیست کیفیت آن

---

[۶] dice

[۷] Hadoop

[۸] Python

[۹] Modis

[۱۰] Burch Works

نیز نادیده گرفته می‌شود و ارزشی برای آن نیست. منابع متفاوت و گوناگونی از داده‌ها با یکدیگر ترکیب می‌شود تا در پروژه‌های بزرگ بکار متخصصان بیاید، اطمینان از دقت و کیفیت این داده‌های اطلاعاتی برای دستیابی به موفقیت، امری حساس و بحرانی است.

کلان داده رشد نرم افزار را ایجاد می‌کند.

در آخرین گزارش ردیابی نیمه سالانه جهانی نرم‌افزار، شرکت داده‌های بین‌المللی (IDC) پیش‌بینی کرد، که تجارت جهانی نرم‌افزار سالانه ۵. ۹ درصد طبق پول رایج آمریکا رشد داشته است.

IDC معتقد است که میزان رشد سالانه نرم افزار طی سال‌های ۲۰۱۳تا ۲۰۱۸ نزدیک به ۶ درصد بوده است. میزان رشد نرم‌افزار در سال‌های ۲۰۱۳ تا ۲۰۱۸ در آسیا به جز ژاپن، آمریکای لاتین، و شرق مرکزی، آسیای میانه و آفریقا ۸. ۵ بوده است. در حالیکه در شمال آمریکا، اروپای غربی و ژاپن این میزان رشد ۵. ۹ درصد می‌باشد.

بصری ساختن داده‌ها مورد تقاضا است.

کار بصری ساختن بسیار مورد تقاضا است چون کار تحلیل داده‌ها را بسیار آسان می‌کند. طبق گزارش اداره اطلاعات هفتگی، تحلیل و ارزیابی مدیریتی، تقریبا نزدیک به ۴۵ درصد از ۴۱۴ پرسشنامه را شامل می‌شود. در پرسشنامه "چالش‌های کار آسان با نرم‌افزارهای پیچیده برای کاربران با مهارت‌های تکنیکی کمتر " دومین مانع برای پذیرش تولیدات تحلیلی، مورد استناد قرار گرفته است. سایت Match. com یک نرم‌افزار با چشم‌انداز بسیار زیبا طراحی کرده است و با این کار قابلیت‌های تحلیل را در دسترس استفاده کنندگان قرار می‌دهد که فقط مخصوص متخصصان خاص یا تحلیل گران نخبه نیست.