انجمن علمی تجارت الکترونیکی ایران

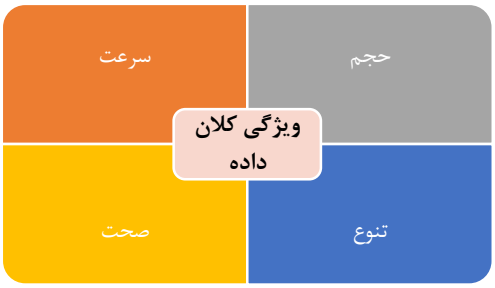
فصل‌نامه علمی-آموزشی تجارت الکترونیکی

شماره ششم، تابستان ۱۳۹۵

کلان داده

۱–کلان‌داده[1] چیست؟

کلان‌داده، گونه‌ای از تجمیع مجموعه‌های داده‌ای بسیار بزرگ و پیچیده است به طوری که پردازش کردن آن‌ها از طریق ابزارهای مدیریت پایگاه داده معمول یا ابزارهای پردازش داده سنتی، دشوار باشد. چالشی که در مورد کلان‌داده‌ها وجود دارد، پیرامون محورهای کلیدی هشت‌گانه‌ای مطرح می‌شود: نحوه‌ی به‌دست آوردن، نحوه‌ی گزینش، نحوه‌ی ذخیره‌سازی، نحوه‌ی جست‌وجو، نحوه‌ی به‌اشتراک‌گذاری، نحوه‌ی انتقال، نحوه‌ی تحلیل و نحوه‌ی نمایش. بر اساس گزارشی که دانشمندان IBM داده‌اند، کلان‌داده می‌تواند دارای چهار بعد اصلی حجم، تنوع، سرعت، و صحت به شرح زیر باشد.



۱–۱– حجم

عوامل بسیاری وجود دارند که به افزایش حجم داده‌ها کمک می‌کنند. برای مثال، داده‌های مبتنی بر تراکنش‌ها طی سال‌ها می‌توانند ذخیره‌سازی شود یا داده‌های غیرساخت‌یافته‌ای که در رسانه‌های اجتماعی جریان دارند از این دسته هستند. در گذشته، افزایش حجم داده‌ها، یک مشکل از بُعد ذخیره‌سازی به حساب می‌آمد اما با کاهش هزینه‌های ذخیره‌سازی، مشکلات دیگری در این زمینه پدیدار شدند. از جمله‌ی این مشکلات می‌توان به تشخیص میزان ارتباط بین داده‌ها و چگونگی تحلیل آن‌ها به منظور ایجاد ارزش اشاره کرد.

۱–۲ تنوع

امروزه داده‌ها در انواع و فرمت‌های مختلفی می‌توانند قرار بگیرند. داده‌های ساختاری-عددی، اطلاعات ایجاد شده از برنامه‌های کاربردی برخط، اسناد مبتنی بدون ساختار، ایمیل، ویدئو، صوت، داده‌های سهام و معاملات مالی از جمله انواع مختلف داده‌هایی است که در پیرامون ما وجود دارند و می‌توانند به انواع مختلفی ذخیره‌سازی شوند.

۱–۳ سرعت

داده‌ها امروزه در سرعت بی‌سابقه‌ای جریان دارند و می‌بایست در به صورت به موقع به آن‌ها رسیدگی شود و تحلیل‌های مرتبط از آن‌ها استخراج گردد. برای مثال، داده‌های RFID، سنسورها یا اندازه‌گیری‌های هوشمند

---

[1] Big data

نیاز دارند تا به صورت بلادرنگ مورد رسیدگی قرار گیرند. پاسخ‌دهی به سرعت داده‌ها یکی از چالش‌های اساسی سازمان‌ها می‌باشد.

۴-۱ صحت

صحت کلان‌داده‌ها به جهت‌گیری‌ها، نویزها و ناهنجاری‌های داده‌ای اشاره دارد. پرسش اصلی در این حوزه آن است که آیا داده‌هایی که در حال ذخیره شدن هستند، ارتباط معناداری با مسئله‌ی مورد تحلیل دارند؟ به بیان دیگر، اعتماد در تحلیل کلان‌داده‌ها یکی از بزرگترین چالش‌ها در مقایسه با مواردی دیگر نظیر حجم و سرعت می‌باشد. برای این منظور، در حوزه‌ی مربوط به راهبردهای کلان داده، می‌بایست تیم کاری برای حفظ اطلاعات به صورت تمیز و پیش‌پردازش آن‌ها برای حذف اطلاعات کثیف تشکیل شود و به طور مداوم این موضوع را مد نظر داشته باشد.

۲- کلان داده را چه کسی ایجاد می‌کند؟

ظهور رسانه‌های اجتماعی و حضور عامه‌ی مردم در این رسانه‌ها و استفاده‌ی افراد از دستگاه‌های هوشمند، سبب شده است تا منابع تولید کلان داده متنوع شده و محدود به طیف خاصی از کاربردها نباشد. بنابراین، کلان داده را می‌توان از طریق رسانه‌ها یا شبکه‌های اجتماعی، زیرساخت‌های مربوط به دانش، دستگاه‌های هوشمند و شبکه‌ها و فناوری‌های مربوط به سنسور تولید نمود. «زیرساخت‌های دانش» نیز از تجهیزات آزمایشگاهی تا ابزارهای رصد و ماهواره‌های فضایی را شامل می‌شود. باید توجه نمود که امروزه توانمندی در جمع‌آوری داده، مانعی برای پیشرفت و نوآوری به حساب نمی‌آید و توانمندی در مدیریت، تجزیه و تحلیل، خلاصه کردن، تصویرسازی و کشف دانش از بین داده‌ها از جمله عوامل کلیدی برای نوآوری و پیشرفت به حساب می‌آیند.

۳- دلیل اهمیت کلان‌داده‌ها

در رابطه با موضوع کلان‌داده‌ها مسئله‌ی اصلی پیرامون دریافت مقدار زیاد داده شکل نمی‌گیرد و همانطور که گفته شد، سؤالاتی در رابطه با نحوه‌ی مدیریت، تجزیه و تحلیل و نمایش این داده‌ها به وجود می‌آید. اهمیت کلان‌داده‌ها آنجایی شکل می‌گیرد که کاربرد آن‌ها و نحوه‌ی استفاده از آن‌ها مشخص شود. چشم‌انداز امیدوارکننده در این زمینه آن است که سازمان‌ها می‌توانن کلان‌داده را از هر منبعی تهیه کنند و آن را برای یافتن پاسخ‌های خود مورد تجزیه و تحلیل قرار دهند. این پاسخ‌ها در نهایت به سازمان کمک می‌کند تا هزینه‌های مالی و زمانی خود را کاهش دهد، محصولات جدید را توسعه دهد و فرایند تصمیم‌گیری کسب‌وکار

را هوشمندانه پیگیری نماید. به بیان دیگر، با ترکیب کلان‌داده‌ها با ابزارهای تجزیه و تحلیل امکانات زیر برای سازمان‌ها و افراد به وجود خواهد آمد:

- شناسایی علل اصلی شکست، مسائل و نقص‌ها به صورت نزدیک به بلادرنگ و صرفه‌جویی میلیون‌ها دلار به طور سالانه
- بهینه‌سازی مسیرها برای هزاران وسیله‌ی نقلیه‌ای که در مسیر هستند.
- تجزیه و تحلیل میلیون‌ها SKU [۲] برای تعیین قیمت و به حداکثر رساندن سود.
- تولید کوپن‌های خرده‌فروشی در نقطه‌ی فروش [۳] بر اساس خریدهای گذشته‌ی مشتری.
- ارائه‌ی پیشنهادهای شخصی‌سازی شده به مشتریان در حالی که آن‌ها در منطقه‌ی مشخصی قرار دارند.
- محاسبه مجدد ریسک کلی سبد سهام در عرض چند دقیقه.
- شناسایی سریع مشتریانی که جزء خریداران و مشتریان مهم به حساب می‌آیند.
- استفاده از تحلیل جریان کلیک داده‌ها برای شناسایی رفتارهای جعلی

۴– کلان‌داده چقدر بزرگ است؟

خلق و ایجاد داده‌ها در حال تجربه‌ی یک نرخ رکورد است. بر اساس مطالعه‌ی IDC [۴] که یکی از شرکت‌های برجسته در حوزه‌ی داده می‌باشد، پیش‌بینی شده است که بین سال‌های ۲۰۰۹ تا ۲۰۲۰ داده‌های دیجیتالی رشدی ۴۴ برابری تا ۳۵ زتابایت [۵] در سال را تجربه کنند. مهم‌تر از این حجم بزرگ، شناسایی این موضوع اهمیت دارد که این رشد انفجاری داده نتیجه‌ی رشد انفجاری دستگاه‌هایی است که در محدوده‌ی شبکه‌ها قرار دارند که از جمله‌ی آن‌ها می‌توان به سنسورهای یکپارچه شده، گوشی‌های هوشمند و تبلت‌ها اشاره کرد. همه‌ی این داده‌ها موجب ایجاد ارزش‌ها و فرصت‌های جدیدی برای تحلیل داده‌ها خواهد شد. پرسشی که همواره در این زمینه مطرح می‌شود پیرامون حجم فعلی کلان‌داده‌ها و حجم آن‌ها در پنج سال بعدی است. پاسخ دادن این موضوع، امر پیچیده‌ای است که همچنان تلاش‌ها برای تخمین دقیق ادامه دارد. باید توجه کرد که کلان‌داده مربوط به یک فناوری واحد، معماری مشخص یا یک مورد استفاده‌ی تنها نیست. اما بر اساس برخی مشاهدات IDC در بازار محاسبات سطح بالا [۶] (که مبدأ تولد کلان‌داده به حساب می‌آید)، می‌توان برخی شاخص‌ها برای آن را ذکر کرد:

---

[۲] مخفف Stock Keeping Unit (SKU) می‌باشد که یک کد معرف محصولات و خدمات را ارائه می‌دهد و به طور معمول توسط ماشین‌ها قابل خواندن است.

[۳] Point of sale

[۴] International Data Corporation

[۵] زتابایت $۱۰^{۲۱}$ بایت (یک میلیارد ترابایت) می‌باشد.

[۶] High-performance

- در سال ۲۰۱۰ حدود ۱۲,۴ میلیارد دلار برای تجهیزات ذخیره‌سازی مربوط به محاسبات سطح بالا خرج شده است و این رقم در سال ۲۰۱۴ به حدود ۱۷ میلیارد دلار افزایش یافته است. با در نظر گرفتن آنکه تنها سی درصد از این تجهیزات به موضوع کلان‌داده اختصاص داشته باشد، می‌توان گفت که در سال ۲۰۱۰ چیزی حدود ۳,۷ میلیارد دلار برای تجهیزات ذخیره‌سازی این حوزه خرج شده است. از سوی دیگر فضای ذخیره‌سازی ابری (مانند فیس‌بوک، یوتیوب، فلیکر و آی‌تونز) حدود ۲۸,۸ درصد از کل ظرفیت ذخیره‌سازی در سال ۲۰۱۰ و حدود ۴۸,۶ از ظرفیت در سال ۲۰۱۴ را شامل می‌شود. می‌توان گفت که اگر تنها ۵٪ از این ظرفیت را برای کلان‌داده در نظر بگیریم، این ظرفیت در سال ۲۰۱۰ بیش از ۲۵۰ پتابایت در نظر گرفته خواهد شد.
- در سال ۲۰۱۱ نیز IDC تخمین زده است که حدود ۱۴,۷ میلیارد دلار برای تجهیزات ذخیره‌سازی و سرور جهت امور تصمیم‌گیری شامل انباره‌های داده و امور تحلیل داده اختصاص یافته است. این عدد، بیش از ۱۷٪ از کل هزینه‌های خرج شده برای سرورها و منابع ذخیره‌ی داده در کل بازار می‌باشد.
- در سال ۲۰۱۰ توزیع‌کنندگان سرور بیش از ۵۱ میلیون هسته‌ی پردازشگر را جابه‌جا کرده‌اند که رقمی بیش از سه برابر مقدار پنج سال گذشته‌ی خود می‌باشد.

بر اساس موارد گفته شده، IDC تخمین زده است که در سال ۲۰۱۵ بیش از ۱,۹ میلیون سرور مورد جابه‌جایی و توزیع قرار بگیرد که این رقم، بیش از ۲۱٪ کل سرورها و ۱۳٪ بیش از مقدار سال ۲۰۱۰ می‌باشد.

بنابراین، برای پیش‌بینی حجم کلان‌داده‌ها می‌بایست به شاخص‌هایی از قبیل ظرفیت‌های ذخیره‌سازی در کل جهان، حجم مخاطبان شبکه‌ها و رسانه‌های اجتماعی و میزان استفاده‌ی عموم از ابزارهای هوشمند و روزمره استناد نمود و همانطور که IDC پیش‌بینی کرده است در سال ۲۰۲۰ چیزی در حدود ۳۵ زتابایت کلان‌داده وجود داشته باشد که می‌تواند انواع مختلفی اعم از داده‌های عددی تا اسناد غیرساخت‌یافته‌ی پیچیده داشته باشد.

شماره‌های تماس: ۸۸۹۹۱۵۶۰ – ۸۸۹۹۱۵۴۰

وب‌سایت: www.ieca.ir

ایمیل: info@ieca.ir