موضوع تحقیق:

کلان داده

آرمان قانعی

استاد یغمایی

تعریف کلان داده (Big Data) چیست؟

کلان‌داده، مه‌داده یا بزرگ‌داده به انگلیسی big data :معمولاً به مجموعه داده‌هایی گفته می‌شود که بیش از حد بزرگ یا پیچیده هستند که نمی‌توان با نرم‌افزارهای کاربردی پردازش داده سنتی آنها را پردازش کرد.

داده‌های با ورودی‌های زیاد (ردیف‌ها) توان آماری بیشتری را ارائه می‌دهند، در حالی که داده‌های با پیچیدگی بالاتر (ویژگی‌ها یا ستون‌های بیشتر) ممکن است به نرخ کشف نادرست بالاتری منجر شود. در حقیقت می‌توان گفت، مه‌داده حجم وسیعی از اطلاعات است که اگر حجم آن کم باشد قابل تفسیر و برداشت نیست.

چالش‌های تجزیه و تحلیل مه‌داده‌ها شامل جمع‌آوری داده‌ها، ذخیره‌سازی داده‌ها، تجزیه و تحلیل دادها، جستجوی آنها، اشتراک گذاری، انتقال، مصورسازی داده، پرسمان، به روز رسانی، حریم خصوصی اطلاعات و تعیین منبع داده است. کلان داده در ابتدا با سه مفهوم کلیدی مرتبط بود: حجم، تنوع و سرعت.

بدون سرمایه‌گذاری کافی در تخصص برای صحت کلان داده‌ها، حجم و تنوع داده‌ها می‌تواند هزینه‌ها و خطراتی را ایجاد کند که بیش از ظرفیت سازمان برای ایجاد و گرفتن ارزش از داده‌های بزرگ است.

استفاده کنونی از واژه مه‌داده به استفاده از تحلیل پیشگویانه، تحلیل رفتار کاربر یا برخی دیگر از روش‌های پیشرفته تجزیه و تحلیل داده اشاره دارد که ارزش را از داده‌های بزرگ استخراج می‌کند و به ندرت به اندازهٔ خاصی از مجموعه داده‌ها می‌پردازد. «شکی وجود ندارد که حجم داده‌های موجود در حال حاضر واقعاً زیاد است، اما این مهم‌ترین ویژگی این اکوسیستم داده جدید نیست.» تجزیه و تحلیل مجموعه داده‌ها می‌تواند همبستگی‌های جدیدی را برای «پیدا کردن روندهای تجاری، پیشگیری از بیماری‌ها، مبارزه با جرم و جنایت و غیره» پیدا کند.

انواع کلان داده کدام است؟

۱ -داده‌های بدون ساختار

همه داده‌ها به روش ساختاریافته مرتب نمی‌شوند. تمام داده‌های سازمان‌نیافته شما، «داده‌های بدون ساختار» هستند. تقریباً ۸۰٪ از داده‌های سراسر جهان ساختار ندارند. هیچ‌کس متن مکالمه‌های تلفنی خود را یادداشت نمی‌کند یا هر توییتی که می‌فرستد را با یک نشانه معنادار مشخص نمی‌کند. تقریباً هر کاری که شما با کامپیوتر انجام می‌دهید، داده‌های بدون ساختار تولید می‌کند. این داده‌ها، ماهیت پیچیده‌ای دارند، فضای بیشتری را اشغال می‌کنند و بی‌نظمی و به هم ریختگی‌شان، مدیریت و درک آنها را دشوار می‌کند.

اما برای به‌دست‌آوردن اطلاعات مفید، داده‌ها باید قابل تفسیر باشند. باوجوداینکه تفسیر داده‌ها زمان و انرژی بیشتری می‌خواهد، اما نتیجه این تفسیر ارزشمندتر از جمع‌آوری ساده داده‌های بدون ساختار است.

سخت‌ترین قسمت تحلیل داده‌های بدون ساختار این است که به یک برنامه یاد بدهیم که اطلاعات به‌دست‌آمده را درک کند. برای این کار باید اطلاعات را برای برنامه، به فرم‌هایی از داده‌های ساختاریافته ترجمه کرد. این کار آسانی نیست و از قالبی به قالب دیگر متفاوت است. معمولاً برای این کار از روش‌هایی مانند تجزیه متن، پردازش طبیعی زبان و... استفاده می‌کنند.

مانند پست های رسانه های اجتماعی، فایل های صوتی، تصاویر و نظرات مشتریان باشند. این نوع داده‌ها را نمی‌توان به‌راحتی در پایگاه‌های داده رابطه‌ایِ سطری‌ستونیِ استاندارد نمایش داد. در گذشته شرکت هایی که می خواستند حجم زیادی از داده¬های بدون ساختار را جستجو، مدیریت و یا تحلیل کنند، مجبور بودند از فرایندهای دستی پرزحمتی استفاده کنند. درمورد ارزش بالقوه‌ی تحلیل و درک چنین داده¬هایی شکی وجود نداشت، اما هزینه ی انجام این کار اغلب آن‌قدر گزاف بود که ارزش آن را نداشت و با توجه به زمان صرف‌شده، نتایج اغلب حتی قبل از ارائه، منسوخ شده بودند. به‌جای صفحات گسترده یا پایگاه‌های داده‌ی رابطه ای، داده های بدون ساختار معمولاً در دریاچه های داده(data lake)، انبارهای داده (data warehouses) و پایگاه¬های داده ی NoSQL ذخیره می‌شوند.

۲ -داده‌های نیمه‌ساختاریافته

داده‌های نیمه‌ساختاریافته، ترکیبی از داده‌های ساختاریافته و بدون ساختار هستند. برای این داده‌ها الگوهای معنادار و جدول‌های مخصوص طراحی نشده است. اما برای داده‌ها برچسب‌ها و نشانه‌هایی برای معناکردن وجود دارند که داده‌ها را برای ضبط و طبقه‌بندی و ساخت پرونده در مجموعه‌های داده آماده می‌کند که ذخیره‌سازی آن را نسبت به داده‌های ساختار نیافته آسان‌تر می‌کند.

داده های نیمه‌ساختاریافته، ترکیبی از داده های ساختاریافته و بدون ساختار هستند. ایمیل‌ها مثال خوبی از این نوع داده هستند، زیرا شامل داده‌های بدون ساختار در متن پیام و همچنین ویژگی‌های سازمان یافته‌ی بیشتری مانند فرستنده، گیرنده، موضوع و تاریخ هستند.

دستگاه‌هایی که از برچسب‌گذاری جغرافیایی(geo-tagging)، برچسب‌های زمانی (time stamp) یا تگ‌های معنایی (semantic tag) استفاده می‌کنند نیز می‌توانند داده‌های ساختاریافته را در کنار محتوای بدون ساختار ارائه دهند. به عنوان مثال، یک تصویر ناشناس از گوشی هوشمند هنوز می‌تواند زمان و مکانی که عکس در آن گرفته شده است را به شما بگوید. یک پایگاه داده‌ی مدرن که از فناوری هوش مصنوعی بهره می‌برد، نه‌تنها می‌تواند فوراً انواع مختلف داده‌ها را شناسایی کند، بلکه می‌تواند درلحظه الگوریتم‌هایی را برای مدیریت و تحلیل مؤثر مجموعه‌داده‌های مختلف تولید کند.

۳ -داده‌های ساختاریافته

این نوع داده‌ها بسیار سازمان‌یافته هستند. تصور کنید صدها صفحه داده، در ستون‌ها و ردیف‌هایی مرتب شده‌اند. برای تمام عنوان‌ها توضیح وجود دارد، متغیرها را به‌راحتی می‌شود تشخیص داد و اعداد قابل‌درک و ملموس هستند. معلوم است که کار با این داده‌ها آسان است و برای برنامه‌ریزی به‌راحتی می‌توان داده‌ها را مرتب و جمع‌آوری کرد.

داده‌های ساختاریافته از الگوها پیروی می‌کنند. این الگوها مشخص می‌کنند که هر داده کجاست و معنی آن چیست.

برای مثال، یک پایگاه‌داده حقوق و دستمزد کارمندان، داده‌هایی مثل اطلاعات شناسایی کارکنان، نرخ دستمزدشان، ساعت کار، نحوه پرداخت اضافه‌کاری و غیره را به ما می‌دهد. الگوی داده‌های ساختاریافته، برای هر برنامه‌ای که از آن استفاده می‌کنیم، هر یک از این موارد را تعریف خواهد کرد. پس این برنامه برای کشف معنای واقعی هرکدام از این موارد، مجبور نیست به جستجوی داده بپردازد، بلکه می‌تواند مستقیماً به جمع‌آوری و پردازش آن بپردازد.

داده‌های ساختاریافته در حجم بالا، لزوماً کلان‌داده محسوب نمی‌شوند؛ زیرا مدیریت داده‌های ساختاریافته به‌خودی‌خود نسبتاً ساده هستند و بنابراین معیارهای تعریف‌کننده‌ی کلان‌داده را برآورده نمی‌کنند.

نحوه کار بیگ دیتا؛ کلان داده چگونه کار می‌کند؟

اطلاعات بیگ دیتا را به‌کمک تکنیک‌ها و فناوری‌های مختلفی می‌توانید پردازش و تجزیه‌وتحلیل کنید. ازجمله تکنیک‌های پردازش اطلاعات می‌توان به انجام آزمون A/B یا ماشین لرنینگ یا پردازش به‌کمک هوش مصنوعی اشاره کرد. پیش از اینکه کسب‌وکارها از کلان داده‌ها استفاده کنند، باید منابع و مکان‌ها و رفتارهای کاربران را بررسی کنند. پنج مرحله مهم برای استفاده از بیگ دیتا وجود دارد که عبارت‌اند از:

۱. تنظیم استراتژی کلان داده

در سطوح پیشرفته، استراتژی کلان داده طرحی است که به بازیابی، ذخیره، جمع‌آوری و مدیریت دیتا کمک می‌کند و آن را بهبود می‌بخشد. بدین‌ترتیب، احتمال موفقیت کسب‌وکار افزایش پیدا می‌کند. استراتژی بیگ دیتا براساس اهداف تجاری تعیین می‌شود؛ به‌همین‌دلیل، نیاز است که استراتژی را به چشم دارایی ارزشمند تجاری نگاه کرد.

۲. شناسایی منابع کلان داده

داده‌های حاصل از اینترنت اشیاء و هوش مصنوعی تکنولوژی پیشرفته‌ای در زمینه‌های مختلف به‌وجود آورد؛ مانند وسایل پوشیدنی در تجهیزات پزشکی یا ابزارهای پیشرفته تجهیزات صنعتی. با بررسی اطلاعات به‌دست‌آمده ازطریق هوش مصنوعی و اینترنت اشیاء، اطلاعات رسانه‌های اجتماعی و داده‌های عمومی یا هر منبع دیگر، نیاز است تعیین کنید که کدام منبع به تحلیل و بررسی نیاز دارد.

۳. بازیابی و مدیریت و ذخیره اطلاعات

سیستم‌های محاسباتی پیشرفته سرعت و قدرت لازم برای دسترسی به کلان داده‌ها را فراهم می‌کنند. برخی از اطلاعات به ذخیره در فضای سنتی نیاز دارند و برخی دیگر ازطریق ذخیره در فضاهای ابری، هزینه و فضای کمتری اشغال می‌کنند.

۴. تجزیه‌وتحلیل داده‌ها

به‌کمک فناوری‌های جدید در تجزیه‌وتحلیل، سازمان‌ها می‌توانند هر داده‌ای را بررسی و تجزیه‌وتحلیل کنند. هوش مصنوعی و ماشین لرنینگ بهترین راه‌های تحلیل داده‌های بزرگ هستند.

۵. گرفتن تصمیمات هوشمندانه و مبتنی‌بر داده

داده‌هایی با نتایج مطمئن و مدیریت‌شده به تصمیم‌گیری درست منجر خواهند شد. هر کسب‌وکاری برای برنده‌شدن دربرابر رقبایش، به داده‌های باارزش و مطمئن نیاز دارد تا از این طریق تصمیم درستی برای پیشرفت کسب‌وکارش بگیرد. درنتیجه، سازمان‌های داده‌محور (مجموعه‌هایی که براساس داده‌های دقیق تصمیم‌گیری می‌کنند) عملکرد بهتری خواهند داشت.

ابزارهای بیگ دیتا؛ نرم‌افزارهای تخصصی برای تحلیل دیتا

همان‌طورکه پیش‌ازاین گفتیم، برای استفاده از بیگ‌دیتا به ابزارهای خاصی نیاز دارید. این ابزارها با مشخص‌کردن نیاز و هدفتان کار تحلیل را راحت‌تر می‌کنند. البته اگر قصد دارید خودتان داده‌ها را مستقیماً تحلیل کنید، شاید چندان به‌کارتان نیاید. برخی از ابزارهای تخصصی برای تحلیل دیتا ازاین‌قرارند:

- SAS
- IBM SPSS Modeller
- knim
- orange
- Xplenty
- Improvado
- Analytics

پنج «V» تعریف‌کننده‌ی بیگ دیتا چیست؟

بزرگ‌بودن یک مجموعه‌داده، لزوماً به‌معنی کلان‌داده بودن آن نیست. برای واجد شرایط بودن به‌عنوان کلان‌داده، داده ها باید حداقل دارای پنج ویژگی زیر باشند:

بیگ دیتا چیست؟ کلان داده چیست؟ big data چیست؟

حجم (Volume): در حالی که حجم به‌هیچ‌وجه تنها مؤلفه‌ای نیست که مجموعه‌ای از داده‌ها را به‌عنوان کلان‌داده‌ها را تعریف می‌کند، اما مطمئناً یک ویژگی اصلی است. برای مدیریت و استفاده‌ی کامل از بیگ‌دیتا، الگوریتم‌های پیشرفته و تحلیل‌های مبتنی بر هوش مصنوعی موردنیاز است. اما قبل از هرچیز، باید روشی امن و قابل‌اعتماد برای ذخیره‌سازی، سازماندهی و بازیابی ترابایت‌ها داده که توسط شرکت‌های بزرگ نگهداری می‌شوند، وجود داشته باشد.

سرعت (Velocity): در گذشته، هر داده-ای که تولید می-شد باید قبل از تحلیل یا بازیابی، در یک پایگاه داده-ی سنتی به‌صورت دستی وارد می-شد. امروزه، فناوری کلان‌داده به پایگاه‌های داده اجازه می‌دهد تا داده‌ها را در حین تولید، پردازش، تحلیل و پیکربندی کنند؛ گاهی فقط در چند میلی‌ثانیه. این بدان معناست که کسب‌وکارها می‌توانند داده‌های بلادرنگ را برای دریافت فرصت‌های مالی، پاسخ‌گویی به نیازهای مشتریان، خنثی کردن کلاهبرداری‌ها و رسیدگی به هر فعالیت دیگری که فاکتور سرعت در آن حیاتی است، استفاده کرد.

تنوع (Variety): مجموعه‌های داده‌ای که از داده‌های ساختاریافته تشکیل شده‌اند، صرف‌نظر از حجمشان، لزوماً کلان‌داده نیستند. کلان‌داده معمولاً از ترکیبی از داده-های ساختاریافته، بدون ساختار و نیمه‌ساختاریافته تشکیل شده است. پایگاه‌های داده سنتی و راه‌حل‌های مدیریت داده، فاقد انعطاف‌پذیری و گستردگی- برای مدیریت مجموعه‌داده‌های پیچیده و مختلفی که کلان‌داده را تشکیل می‌دهند، هستند.

صحت (Veracity): در حالی که فناوری پایگاه داده-ی مدرن این امکان را برای شرکت‌ها فراهم می‌کند که مقادیر و انواع کلان‌داده را جمع‌آوری و تفسیر کنند، تنها زمانی این کار ارزشمند است که دقیق و به‌موقع انجام شود. در پایگاه‌های داده سنتی که فقط با داده‌های ساختاریافته پر می‌شدند، خطاهای نحوی و اشتباهات تایپی عامل معمول درمورد نقص در دقت داده‌ها بودند. اما در داده-های بدون ساختار، مجموعه-ی جدیدی از چالش-های جدیدی درمورد صحت داده‌ها وجود دارد. جهت‌گیری‌های انسانی و مسائل مربوط به منشأ داده-ها و... همگی می-توانند بر کیفیت و صحت داده-ها تأثیر بگذارند.

ارزش (Value): بدون شک، نتایجی که از تحلیل کلان‌داده-ها به دست می‌آیند اغلب جذاب و غیرمنتظره هستند. اما برای کسب‌وکارها، تحلیل کلان‌داده--ها باید اینسایت‌هایی ارائه دهد که به آن‌ها کمک کند تا رقابتی‌تر و انعطاف‌پذیرتر شده و به مشتریان خود خدمات بهتری ارائه کنند. فناوری‌های مدرن کلان داده، ظرفیت جمع‌آوری و بازیابی داده‌ها را فعال کرده و مزایایی قابل‌اندازه‌گیری برای سودآوری و انعطاف‌پذیری عملیاتی فراهم می‌کنند.

مزایای استفاده از کلان داده در کسب‌وکار

وجود کلان‌داده‌ها و جمع‌آوری آن‌ها می‌تواند برای هر کسب‌وکاری مفید باشد. تجزیه‌وتحلیل این داده‌ها با سیستم‌های نرم‌افزاری پیشرفته، به سازمان‌ها در تصمیم‌گیری‌ها و تعیین استراتژی کمک می‌کند و این باعث کاهش هزینه‌ها و افزایش درآمد در هر کسب‌وکاری می‌شود. آنچه در ادامه می‌آید مثال‌هایی از کاربرد تحلیل کلان داده در بهبود وضعیت کسب‌وکارهاست.

جذب و نگهداری مشتری

کسب‌وکارها برای بقاء باید رویکرد درستی در بازاریابی محصولاتشان داشته باشند. شرکت‌ها با استفاده از کلان‌داده‌ها، می‌توانند بفهمند که مشتریان دقیقاً به دنبال چه چیزی هستند. با بهره‌گیری از شیوه‌های نوین، الگوهای مصرف مشتریان را زیر نظر بگیرند و با شناسایی این الگوها و یافتن راه‌هایی که مشتریان را خوشحال‌تر می‌کند، وفاداری آنها به محصول را بالاتر ببرند.

عملکرد هدفمند و متمرکز

کسب‌وکارها می‌توانند با استفاده از کلان داده‌ها، بدون صرف هزینه‌های زیاد برای تبلیغات، محصولات و خدمات خود را به شکلی هدفمند در بازارهایی که برای هر محصول مناسب است، بفروشند. کلان داده‌ها به تحلیل پیشرفته عادات مشتریان کمک می‌کنند. به طور مثال می‌توان رفتار خرید آنلاین مشتریان را با دقت خوبی شناخت. این آمار در نهایت به شرکت‌ها این امکان را می‌دهد که کمپین‌های هدفمند، متمرکز و موفق طراحی کنند و با برآورده کردن انتظارات مشتریان، وفاداری به برند را افزایش دهد.

شناسایی ریسک‌های بالقوه

این روزها،کسب‌وکارها با انواع و اقسام ریسک‌ها روبرو هستند. عوامل مختلفی می‌تواند ادامه حیات کسب‌وکار را تهدید کند. در این فضای ناامن، سازمان‌ها به مدیریت ریسک بیش از هر زمانی نیازمندند. کلان‌داده‌ها در ایجاد راه‌حل‌های جدید برای مدیریت ریسک نقشی اساسی دارند. آنها می‌توانند مدل‌های مدیریت ریسک را اثربخش‌تر کرده و استراتژی‌های هوشمندانه‌تری برای جلوگیری از ضرر و زیان در سازمان‌ها ایجاد کنند.

نوآوری در تولید محصولات

برای رقابت در بازار امروز دیگر نمی‌توان به غریزه اعتماد کرد. اکنون شرکت‌ها می‌توانند با استفاده از کلان‌داده‌ها، فرایندهایی را برای ردیابی بازخورد مشتری، موفقیت محصول و مقایسه با رقبا پیاده‌سازی کنند. کلان‌داده‌ها می‌توانند به شرکت‌ها کمک کنند تا در کنار ارتقاء محصولات موجود، به خلق و نوآوری محصولات جدید بپردازند. با جمع‌آوری حجم زیادی از داده‌ها، شرکت‌ها قادر به تشخیص نیازهای به‌روز مشتریان هستند.

پنج سوال از متن تحقیق در آخر فایل

1. بیگ دیتا یا کلان داده به طور خلاصه یعنی چه؟ : به مجموعه داده‌هایی گفته می‌شود که بیش از حد بزرگ یا پیچیده هستند که نمی‌توان با نرم‌افزارهای کاربردی پردازش داده سنتی آنها را پردازش کرد.

2. انواع کلان داده را نام ببرید. کدام یک تقریبا کلان داده حساب نمی شود و چرا؟
ساختار نیافته - نیمه ساختار یافته - ساختار یافته
ساختار یافته | زیرا مدیریت داده‌های ساختاریافته به‌خودی‌خود نسبتاً ساده هستند و بنابراین معیارهای تعریف‌کننده‌ی کلان‌داده را برآورده نمی‌کنند.

3. پنج «V» تعریف‌کننده‌ی بیگ دیتا را بنویسید.
حجم (Volume) - سرعت (Velocity) - تنوع (Variety) - صحت (Veracity) - ارزش (Value)

4. مزایای استفاده از کلان داده در کسب کار را نام ببرید.
جذب و نگهداری مشتری - عملکرد هدفمند و متمرکز - شناسایی ریسک‌های بالقوه - نوآوری در تولید محصولات

5. پنج مرحله مهم برای استفاده از بیگ دیتا چی هستند؟
تنظیم استراتژی کلان داده - شناسایی منابع کلان داده - بازیابی و مدیریت و ذخیره اطلاعات
تجزیه‌وتحلیل داده‌ها - گرفتن تصمیمات هوشمندانه و مبتنی‌بر داده