بسمه تعالی

آشنایی با مفهوم Big Data

"داده های بزرگ" یا "عظیم داده" یا "کلان داده " ترجمه اصطلاح Big Data می باشد که معمولا به مجموعه از داده ها اطلاق می شود که اندازه آنها فراتر از حدی است که با نرم افزارهای معمول بتوان آنها را در یک زمان معقول اخذ، دقیق سازی، مدیریت و پردازش کرد. مفهوم «اندازه» در داده های بزرگ بطور مستمر در حال تغییر است و به مرور بزرگتر می شود.

داده های بزرگ (Big Data) مجموعه از تکنیک ها و تاکتیک هایی است که نیازمند شکل جدیدی از یکپارچگی هستند تا بتوانند ارزش های بزرگی را که در مجموعه های بزرگ، وسیع، پیچیده و متنوع داده پنهان شده اند، آشکار سازند.

از این رو با رشد روز افزون داده ها و نیاز به بهره برداری و تحلیل از این داده ها، بکارگیری زیرساخت های Big Data از اهمیت ویژه ای برخوردار شده است. عبارت Big Data مدت‌ها است که برای اشاره به حجم‌های عظیمی از داده‌ها که توسط سازمان‌های بزرگی مانند گوگل یا ناسا ذخیره و تحلیل می‌شوند مورد استفاده قرار می‌گیرد. اما به تازگی، این عبارت بیشتر برای اشاره به مجموعه‌های داده‌ای بزرگی استفاده می‌شود که به قدری بزرگ و حجیم هستند که با ابزارهای مدیریتی و پایگاه‌های داده سنتی و معمولی قابل مدیریت نیستند. مشکلات اصلی در کار با این نوع داده‌ها مربوط به برداشت و جمع‌آوری، ذخیره‌سازی، جست‌وجو، اشتراک‌گذاری، تحلیل و نمایش آن‌ها است. این مبحث، به این دلیل هر روز جذابیت و مقبولیت بیشتری پیدا می‌کند که با استفاده از تحلیل حجم‌های بیشتری از داده‌ها، می‌توان تحلیل‌های بهتر و پیشرفته‌تری را برای مقاصد مختلف، از جمله مقاصد تجاری، پزشکی و امنیتی، انجام داد و نتایج مناسب‌تری را دریافت‌کرد. بیشتر تحلیل‌های مورد نیاز در پردازش داده‌های عظیم، توسط دانشمندان در علومی مانند هواشناسی، ژنتیک، شبیه‌سازی‌های پیچیده فیزیک، تحقیقات زیست‌شناسی و محیطی، جست‌وجوی اینترنت، تحلیل‌های اقتصادی و مالی و تجاری مورد استفاده قرار می‌گیرد. حجم داده‌های ذخیره‌شده در مجموعه‌های داده‌ایBig Data ، عموماً به‌خاطر تولید و جمع‌آوری داده‌ها از مجموعه بزرگی از تجهیزات و ابزارهای مختلف مانند گوشی‌های موبایل، حسگرهای محیطی، لاگ نرم‌افزارهای مختلف، دوربین‌ها، میکروفون‌ها، دستگاه‌های تشخیصRFID ، شبکه‌های حسگر بی‌سیم وغیره با سرعت خیره‌کننده‌ای در حال افزایش است.

تلگرام
@dataacademy
09120637751
مرکز داده کاوی، متن کاوی و علوم مرتبط
آکادمی داده
www.dataacademy.ir

Big Data

برای ایجاد یک دید مناسب در خصوص Big Data و اهمیت آن، جامعه ای را تصور کنید که در آن جمعیت بطور نمایی در حال افزایش است، اما خدمات و زیرساخت های عمومی آن نتواند پاسخگوی رشد جمعیت باشد و از عهده مدیریت آن برآید. چنین شرایطی در حوزه داده در حال وقوع است. بنابراین نیازمند توسعه زیرساخت های فنی برای مدیریت داده و رشد آن در بخش هایی نظیر جمع آوری، ذخیره سازی، جستجو، به اشتراک گذاری و تحلیل می باشیم. دستیابی به این توانمندی معادل است با شرایطی که مثلا بتوانیم "هنگامی که با اطلاعات بیشتری در حوزه سلامت مواجه باشیم، با بازدهی بیشتری سلامت را ارتقا دهیم"، "در شرایطی که خطرات امنیتی افزایش پیدا میکند، سطح امنیت بیشتری را فراهم کنیم"، "وقتی که با رویدادهای بیشتری از نظر آب و هوایی مواجه باشیم، توان پیش بینی دقیقتر و بهتری بدست آوریم"، "در دنیایی با خودروهای بیشتر، آمار تصادفات و حوادث را کاهش دهیم"، "تعداد تراکنش های بانکی، بیمه و مالی افزایش پیدا کند، ولی تقلب کمتری را شاهد باشیم"، "با منابع طبیعی کمتر، به انرژی بیشتر و ارزانتری دسترسی داشته باشیم" و بسیاری موارد دیگر از این قبیل که اهمیت پنهان کلان داده را نشان می دهد.

چالش های حوزه کلان داده

در بحث Big Data ، ما نیاز داریم که داده ها را به منظور استخراج اطلاعات، کشف دانش و در نهایت تصمیم گیری در خصوص مسائل مختلف کاربردی به صورت صحیح مدیریت کنیم. مدیریت داده ها عموما شامل ۵ فعالیت اصلی میباشد.

۱. جمع آوری
۲. ذخیره سازی
۳. جستجو
۴. به اشتراک گذاری
۵. تحلیل

تا کنون چالشهای زیادی در حوزه کلان داده مطرح شده است که تا حدودی از جنبه تئوری ابعاد مختلفی از مشکلات این حوزه را بیان میکنند.

Big Data

- **حجم داده :**(Volume) حجم داده ها به صورت نمایی در حال رشد می باشد. منابع مختلفی نظیر شبکه های اجتماعی، لاگ سرورهای وب، جریان های ترافیک، تصاویر ماهواره ای، جریان های صوتی، تراکنش های بانکی، محتوای صفحات وب، اسناد دولتی و ... وجود دارد که حجم داده بسیار زیادی تولید می کنند.

- **نرخ تولید :**(Velocity) داده ها از طریق برنامه های کاربردی و سنسورهای بسیار زیادی که در محیط وجود دارند با سرعت بسیار زیاد و به صورت بلادرنگ تولید می شوند. بسیاری از کاربردها نیاز دارند به محض ورود داده به درخواست کاربر پاسخ دهند. ممکن است در برخی موارد نتوانیم به اندازه کافی صبر کنیم تا مثلا یک گزارش در سیستم برای مدت طولانی پردازش شود.

- **تنوع :**(Variety) انواع منابع داده و تنوع در نوع داده بسیار زیاد می باشد که در نتیجه ساختارهای داده ای بسیار زیادی وجود دارد. مثلا در وب، افراد از نرم افزارها و مرورگرهای مختلفی برای ارسال اطلاعات استفاده می کنند. بسیاری از اطلاعات مستقیما از انسان دریافت میشود و بنابراین وجود خطا اجتناب ناپذیر است. این تنوع سبب میشود جامعیت داده تحت تاثیر قرار بگیرد. زیرا هرچه تنوع بیشتری وجود داشته باشد، احتمال بروز خطای بیشتری نیز وجود خواهد داشت.

- **صحت :**(Veracity) با توجه به اینکه داده ها از منابع مختلف دریافت میشوند، ممکن است نتوان به همه آنها اعتماد کرد. مثلا در یک شبکه اجتماعی، ممکن است نظرهای زیادی در خصوص یک موضوع خاص ارائه شود. اما اینکه آیا همه آنها صحیح و قابل اطمینان هستند، موضوعی است که نمیتوان به سادگی از کنار آن در حجم بسیار زیادی از اطلاعات گذشت. البته بعضی از تحقیقات این چالش را به معنای حفظ همه مشخصه های داده اصلی بیان کرده اند که باید حفظ شود تا بتوان کیفیت و صحت داده را تضمین کرد. البته تعریف دوم در مولدهای کلان داده صدق میکند تا بتوان داده ای تولید کرد که نشان دهنده ویژگی های داده اصلی باشد.

- **اعتبار :**(Validity) با فرض اینکه دیتا صحیح باشد، ممکن است برای برخی کاربردها مناسب نباشد یا به عبارت دیگر از اعتبار کافی برای استفاده در برخی از کاربردها برخوردار نباشد.

- **نوسان :**(Volatility) سرعت تغییر ارزش داده های مختلف در طول زمان میتواند متفاوت باشد. در یک سیستم معمولی تجارت الکترونیک، سرعت نوسان داده ها زیاد نیست و ممکن است داده های موجود مثلا برای یک سال ارزش خود را حفظ کنند، اما در کاربردهایی نظیر تحلیل ارز و بورس، داده با نوسان زیادی مواجه هستند و داده ها به سرعت ارزش خود را از دست

میدهند و مقادیر جدیدی به خود می گیرند. اگرچه نگهداری اطلاعات در زمان طولانی به منظور تحلیل تغییرات و نوسان داده ها حائز اهمیت است. افزایش دوره نگهداری اطلاعات، مسلما هزینه های پیاده سازی زیادی را دربر خواهد داشت که باید در نظر گرفته شود.

- **نمایش :**(Visualization) یکی از کارهای مشکل در حوزه کلان داده، نمایش اطلاعات است. اینکه بخواهیم کاری کنیم که حجم عظیم اطلاعات با ارتباطات پیچیده، به خوبی قابل فهم و قابل مطالعه باشد از طریق روش های تحلیلی و بصری سازی مناسب اطلاعات امکان پذیری است.

- **ارزش :**(Value) این موضوع دلالت بر این دارد که از نظر اطلاعاتی برای تصمیم گیری چقدر داده حائز ارزش است. بعبارت دیگر آیا هزینه ای که برای نگهداری داده و پردازش آنها میشود، ارزش آن را از نظر تصمیم گیری دارد یا نه. معمولا داده ها میتوانند در لایه های مختلف جابجا شوند. لایه های بالاتر به معنای ارزش بیشتر داده می باشند. بنابراین برخی از سازمانها میتوانند هزینه بالای نگهداری مربوط به لایه های بالاتر را قبول کنند.

کاربردهای Big Data

Big Data نحوه کار سازمان ها و افراد را تحت تاثیر قرار می دهد. Big Data فرهنگی را در سازمان ها ایجاد می کند که از طریق آن کسب و کارها و مدیران فناوری اطلاعات را به سمت استفاده از تمامی ارزش های پنهان در داده ها سوق می دهد. ادراک این ارزش ها به همه کارکنان سازمان ها این امکان را می دهد که با بینش وسیع تری تصمیم گیری کنند، نزدیکی بیشتری با مشتریان داشته باشند، فعالیت های خود را بهینه کنند، با تهدیدات مقابله کنند و در نهایت سرمایه های خود را بر روی منبع جدیدی از سود سرشار پنهان در داده ها متمرکز سازند. سازمان ها برای رسیدن به این مرحله نیازمند معماری جدید، ابزارهای نو و فعالیت ها و تلاش های مستمری هستند تا بتوانند از مزیت های چهارچوب های مبتنی بر داده های بزرگ بهره مند گردند.