نظم پنهان

» احمد شریف‌پور

داده‌هایی که ما ذخیره می‌کنیم؛ از آدرسی که پشت یک پاکت نوشته می‌شود تا اعداد و ارقامی که یک محقق از آزمایش‌هایش به دست می‌آورد و حتی فهرست تماس‌هایی که همه ما در تلفن‌های همراه‌مان داریم، به صورت کلی تنها برای یک منظور به کار می‌روند و آن نظم بخشیدن به دانسته‌هایمان است.

در این راستا، برقراری ارتباط میان دو یا چند چیز است که امکان رسیدن به هدفی خاص را برای ما فراهم می‌کند؛ دفترچه تلفنی که تعدادی عدد و رقم را به تعدادی انسان پیوند می‌زند و آدرسی که یک محل را به مجموعه‌ای از نام‌های خیابان‌ها متصل می‌کند.

از سوی دیگر یافتن الگوهای مشابه در میان این مجموعه داده‌های منظم، توان پیش‌بینی را برای ما به ارمغان خواهد آورد. هرچه داده‌ها بیشتر باشند و ارتباط و نظم آن‌ها دقیق‌تر و مفصل‌تر ثبت شده باشد، یافتن رابطه‌های جدید و نظم‌های پنهان بسیار ساده‌تر خواهد شد. خوانندگان قدیمی‌تر مجله به یاد دارند که در پرونده «از دانش ۲ تا پایانی بر دانش» در شماره ۹۳ ماهنامه گفتیم که در صورتی که شما بتوانید تمام یا حداقل بخش عمده داده‌های مربوط به یک موضوع را جمع‌آوری کنید، دیگر در آن حوزه به نظریه‌های علمی و آزمایش و تحقیق نیازی نخواهد بود چراکه داده‌های ثبت‌شده، خود به خود نظم حاکم و قوانین علمی را؛ که پیش‌تر باید با آزمایش و خطا و تحقیق جست‌وجو می‌شدند، به شما نشان خواهند داد.

در دنیای کنونی ما افزایش جمعیت، پیشرفت دانش، افزوده شدن مداوم تعداد و کیفیت حسگرها، ارزان شدن تجهیزات ذخیره‌سازی و علاقه روزافزون نوع بشر به یافتن نظام‌های پیچیده‌تر و جدیدتر، باعث شده است که حجم داده‌هایی که در حال ذخیره آن‌ها هستیم به شدت افزایش یابد.

یک مثال ساده از دنیای دیجیتال آشنای ما، سایت آمازون است. این سایت اکنون تقریبا هر روز به اندازه کل ظرفیت ذخیره‌سازی که در سال ۲۰۰۱ مورد استفاده قرار می‌داد، به ظرفیت مراکز داده خود می‌افزاید! اما در این میان شاید معضل کمبود فضا و ناکارآمدی فناوری‌های سنتی ذخیره‌سازی برای این حجم عظیم اطلاعات، چالش اصلی Big Data نباشد بلکه چالش اصلی را باید در نحوه کندوکاو داده‌ها و یافتن ارتباط‌ها و نظم‌های پنهان در لابه‌لای این حجم عظیم داده جست‌وجو کرد. نظمی که مثلاً ممکن است با تحلیل داده‌های خریدهای شما و دوستان‌تان در یک فروشگاه آنلاین، شما را به جامعه مصرف‌کنندگان یک محصول خاص و از آنجا به قلب یک کارخانه صنعتی پیوند بزند. یافتن چنین رابطه‌ای به یقین می‌تواند به افزایش تولید کارخانه و افزایش میزان فروش محصول منجر شود. در مثالی متفاوت و در صورت وجود چنین داده‌هایی، نظمی دیگر ممکن است ارتباط میان بیماران مبتلا به یک بیماری خاص با مسافرت به محلی معین را آشکار کرده و از این طریق محل شیوع یک عامل بیماری‌زا را یافته و از همه‌گیر شدن آن پیشگیری کند.

به هر حال، این افزایش انفجاری در میزان داده‌ها و نیاز روزافزون به استخراج نظام‌ها و ارتباطات پنهان در لابه‌لای این داده‌های موجود، کارایی سیستم‌ها و مدل‌های سنتی ذخیره‌سازی و کندوکاو داده را به چالش کشیده است. همین امر تبدیل به محرکی برای توسعه سیستم‌هایی شده است که از ابتدا برای کار با چنین حجمی از داده‌ها طراحی و بهینه‌سازی شده‌اند. سیستم‌هایی که می‌توانند توانایی‌های چندین هزار ماشین موازی را برای زیر و رو کردن داده‌ها بسیج کنند و روابطی گاه عجیب و غریب را از دل این توده داده‌ها بیرون بکشند.

دنیایی که در آن و در کمتر از سی سال پیش حجم یک فلاپی دیسک برای ذخیره تمام داده‌های یک کاربر کافی بود، دنیایی که در آن نیازی به تعبیه هارددیسک برای کامپیوترهای خانگی دیده نمی‌شد، امروز از سرورفارم، دیتاسنتر و انبار داده‌ها سخن می‌گوید و اعداد و ارقامش از پتابایت گذشته و به اگزابایت و زتابایت رسیده است. و این موضوع پرونده ویژه‌ای است که پیش‌رو دارید.

درباره ویژه‌نامه

این پرونده ویژه، خود به دو بخش قابل تفکیک است. بخش اول به بررسی پدیده Big Data می‌پردازد. در این بخش و در مقاله نخست یعنی «Big Data، مرزهای جدید نوآوری، رقابت و تولید» به طرح مسئله پرداخته و توضیح می‌دهیم که داده‌های عظیم یا Big Data اصولا چیست و چرا به چالشی بزرگ در دوره حاضر تبدیل شده است.

پس از آن در اینفوگرافی انفجار داده‌ها، به صورتی ساده و بصری با ابعاد این پدیده و بخش‌های درگیر با آن آشنا شده و دیدی کلی از عظمت داده‌هایی که از آن‌ها صحبت می‌کنیم به دست خواهید آورد. و درنهایت، مقاله «دیسک ذخیره‌سازی بزرگ در آسمان» به تفصیل به بررسی راهبرد شرکت‌های بزرگ و پیشرو اینترنتی نظیر گوگل و آمازون و مایکروسافت برای غلبه بر چالش داده‌های بزرگ پرداخته است و سیستم‌های فایلی و ابزارهایی را که برای این کار توسعه داده‌اند معرفی خواهد کرد.

بخش دوم به بررسی نسل جدید پایگاه‌های داده می‌پردازد که از اساس برای کار با حجم عظیمی از داده‌های بدون ساختار طراحی و پیاده‌سازی شده‌اند. در مقاله نخست این بخش، چیستی این پایگاه‌های داده مورد بحث قرار می‌گیرد. مقاله بعدی نخستین نمونه‌های چنین ابزارهایی را معرفی و مزایا و معایب هر یک را شرح می‌دهد.

در مقاله «فیل، بیرون از تاریکی» به معرفی تاریخچه و کاربردهای فریم‌ورک هادوپ می‌پردازیم که چگونه و از کجا شکل گرفت و چه شد که به یکی از محبوب‌ترین ابزارهای کار با داده در مقیاس عظیم تبدیل شد. شایستگی‌های این فریم‌ورک تا آن حد زیاد بود که حتی مایکروسافت را به روی آوردن به هادوپ و به قولی آشتی با دنیای اپن‌سورس وادار کرد و این موضوعی است که در مقاله بعدی مورد بررسی قرار گرفته است.

در ادامه، برای آشنایی بیشتر با پایگاه‌های داده NoSQL در سه مقاله کاربردی به معرفی نحوه پیاده‌سازی، پیکربندی و استفاده از سه نمونه مشهور یعنی Redis، کاساندرا و MongoDB پرداخته‌ایم. در انتها نیز شما با چند کتاب مشهور در حوزه NoSQL و یک فرهنگ لغت تخصصی ساده، می‌توانید به دانش خود در این زمینه افزوده و مسیر مناسب را برای ادامه مطالعات و پژوهش در این حوزه بیابید. مشکل ما با جنبه‌های مختلف موضوعاتی چنین مهم و فراگیر، دقیقاً همانند ابتدایی‌ترین مشکل غول‌های بزرگ اینترنتی با حجم عظیم داده‌هایشان، مسئله کمبود فضا است و به همین دلیل می‌دانیم که این پرونده هیچ‌گاه کامل و بی‌عیب نخواهد بود. تنها امید داریم این پرونده نیز توانسته باشد وظیفه خود را در مطرح کردن این چالش جدید و عظیم و برانگیختن حساسیت شما به خوبی انجام داده باشد. به هر حال شاید نظم و ارتباطی که شما در داده‌هایتان می‌یابید، دنیای فردا را به کل دگرگون کند...