بسمه تعالی

مقدمه‌ای بر کلان داده و فناوری‌های مرتبط

۱. مقدمه

بیگ دیتا یا کلان داده چند سالیست که در ادبیات فناوری اطلاعات به یک اصطلاح فراگیر تبدیل شده است و در این مقاله سعی شده است، این حوزه نوین، به صورت اجمالی معرفی گردد.

اگر بخواهیم تعریفی از کلان داده ارائه کنیم می‌توانیم آنرا مجموعه داده‌هایی بدانیم که اندازه آنها فراتر از حدی است که با نرم افزارها و روشهای معمول بتوان آنها را در یک زمان قابل قبول، دریافت، ذخیره، مدیریت و پردازش کرد. دراین تعریف، حجم داده مشخص نشده است چون میزان کلان بودن داده همزمان با پیشرفت فناوری‌های ذخیره‌سازی و پردازش اطلاعات روز به روز و عموماً به‌خاطر تولید داده توسط تجهیزات و ابزارهای مختلف دیجیتال در حال افزایش است. گوشی‌های موبایل، حسگرهای محیطی، لاگ نرم‌افزارهای مختلف، دوربین‌ها، میکروفون‌ها، دستگاه‌های تشخیص RFID ، شبکه‌های حسگر بی‌سیم، ایستگاه‌های هواشناسی، سامانه‌های امواج رادیویی، مبادلات شبکه‌های اجتماعی آنلاین، متون و اسناد اینترنتی، داده‌های نجوم، اطلاعات پزشکی و سلامت بیماران، اطلاعات سامانه‌های خرید از فروشگاه‌ها، پژوهش‌های زمین‌شناسی و غیره نمونه‌هایی از داده‌ها در مقیاس کلان هستند. مقیاسی که امروزه از گیگابایت و ترابایت به پتا بایت و اگزا بایت و زتا بایت در حال حرکت است

برای ایجاد یک دید مناسب در خصوص کلان داده 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 بی شتری را فراهم کنیم"، "وقتی که با رویدادهای بی شتری 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.

۲. چالش‌ها و خصوصیات کلان داده

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. این چالش ها در ابتدا سه بعد اصلی حجم داده، نرخ تولید و تنوع به عنوان ۳ V'sمطرح شدند ولی در ادامه چالش های بیشتری در ادبیات موضوع توسط محققان مطرح شده است:

- حجم داده :(Volume) حجم داده های درون سازمان و خارج آن به مدد پدیده اینترنت، دستگاه های الکترونیکی و موبایل ها، زیر ساخت های شبکه و سایر منابع هر ساله رشد نمایی دارد و پیش بینی شده است که تا سال ۲۰۲۰ ما ده زتابایت داده در جهان خواهیم داشت.

- 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 که اغلب باید در لحظه پردازش و ذخیره شوند.
- تنوع (Variety): انواع منابع داده و تنوع در نوع داده بسیار زیاد می باشد که در نتیجه ساختارهای داده ای بسیار زیادی وجود دارد و بیشتر حجم داده دنیا هم بی-ساختار و بسیار متنوع است. بخشی از داده-ها امروزه در بانکهای اطلاعاتی، بخشی در صفحات وب، بخشی به صورت XML و JSON و بقیه نیز در فایلها با قالب های متفاوت ذخیره شده اند که عمل پردازش آنها را پیچیده می-کند.
- 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
- 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 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): آیا هزینه ای که برای نگهداری داده و پردازش آنها میشود، ارزش آن را از نظر تصمیم گیری دارد یا نه و ارزش و فایده موردنظر را برای یک سازمان خواهند داشت؟

به طور کلی، تفاوتهای اصلی کلان داده و داده های سنتی در جدول زیر بیان شده است.

معیار	***داده های سنتی***	***کلان داده***
اندازه	گیگا بایت تا ترابایت	پتابایت تا اگزابایت
معماری	متمرکز	توزیع شده
ساختار	دارای ساختار	بی-ساختار یا نیم-ساختار
مدل داده	مدل داده ثابت	بدون شِمای مشخص
ارتباط داخلی	ارتباطات پیچیده بین رکوردها	فاقد ارتباطات داخلی پیچیده

جدول ۱: مقایسه داده های کلاسیک با کلان داده

۳. ابزار های ذخیره و پردازش در حوزه کلان داده

رهیافتهایی که امروزه در بخش پردازش کلان داده مطرح هستند، دارای چندین خاصیت مشترک هستند:

- اجرا بر روی سخت افزار موجود که باعث می‌شود بتوان با هزینه کم امکان پردازش موازی و ارتقای سخت افزاری را فراهم کرد.
- استفاده از ابزارهای تحلیل و مصورسازی پیشرفته برای سهولت کاربر نهایی.
- استفاده همزمان از ابزارها و کتابخانه های مختلف که معماری داده یک سازمان را شکل می دهند.
- استفاده از بانک های اطلاعاتی غیر رابطه ای (NoSql) به عنوان جزئی از معماری و بسترداده سازمان

دو رهیافت اصلی که امروزه در پردازش و تحلیل کلان داده بیشترین رواج را دارند عبارتند از :

هدوپ و بانکهای اطلاعاتی NoSQL

۳-۱ هدوپ

هدوپ یک چهارچوب متن-باز برای پردازش، ذخیره و تحلیل حجم عظیم داده های توزیع شده و بدون ساختار است۶. منشاء اصلی پیدایش این چهارچوب پردازشی به شرکتهای جستجوی اینترنتی یاهو و گوگل باز می-گردد که برای ایندکس کردن صفحات وب و جستجوی آنها نیاز به ابزار و مدل های جدید پردازشی داشتند. این چهارچوب برای پردازش موازی داده ها در سطح پتابایت و اگزابایت که بر روی رایانه-های معمولی توزیع شده اند، به گونه-ای طراحی شده ا ست که کلا ستر ت شکیل دهنده آن به راحتی و بسته به نیاز، قابل گسترش است.

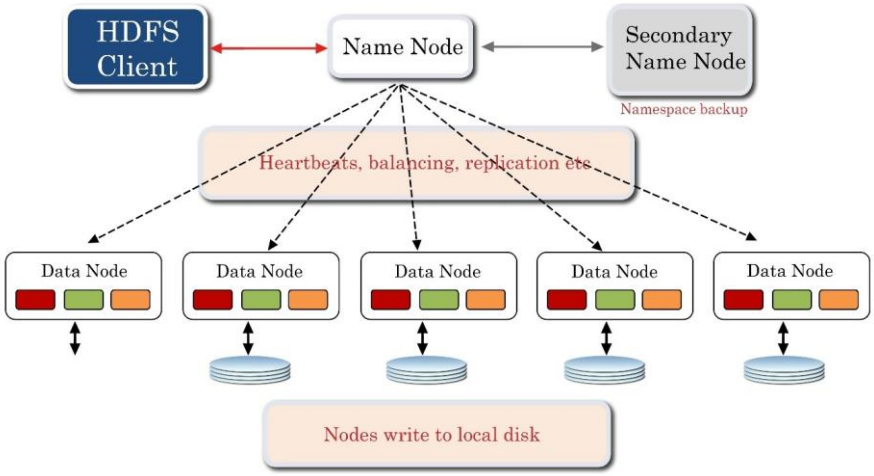
هدوپ چگونه کار می کند

در این سامانه فایلهای داده ای با حجم بالا مانند فایلهای ثبت تراکنش، خوراک خوان شبکه های اجتماعی و سایر منابع داده ای ابتدا بخش بندی شده و در شبکه توزیع می شوند.

وظیفه تقسیم، ذخیره و بازیابی فایلهای حجیم بر روی یک کلاستر هدوپ را سیستم فایل توزیع شده آن به نام HDFS بر عهده دارد. برای بالابردن ضریب اطمینان سیستم ، هر بخش از فایل در چندین رایانه توزیع می شود تا در صورت از کارافتادن یک سیستم، آن فایل باز هم قابل بازیابی باشد.

در هدوپ سه نوع گره محاسباتی یا رایانه داریم . مدیر نام ، وظیفه تقسیم فایلها و ذخیره آدرس هر بخش از آن را برعهده دارد. برر سی دوره-ای گره ها و تعیین از رده خارج شدن آنها هم جزء وظایف این مولفه از سیستم مدیریت فایل هدوپ است.

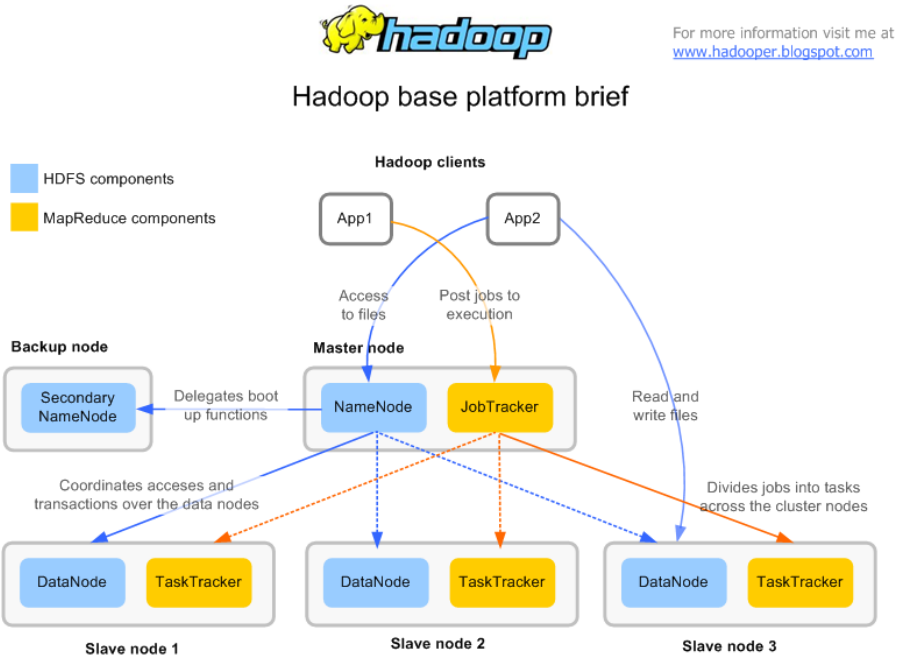
گره داده که تک تک رایانه های عضو هدوپ را در بر می گیرد، بلاک های فایل را در بردارد که برای مدیریت بهتر آنها، به ازای مجموعه ای از این گره-های داده ، یک گره مدیریت نام در سامانه هدوپ وجود دارد. نوع سوم ، گره نام ثانویه ا ست که یک رونوشت از اطلاعات گره مدیریت نام بر روی آن قرار دارد تا در صورت از کار افتادن آن گره ، اطلاعات آن از بین نرود. شکل ۱ شمایی کلی از مولفه مدیریت فایل هدوپ را نشان می دهد.



شکل ۱ : ساختار سیستم فایل HDFS

بعد از توزیع داده ها در سامانه هدوپ ، تحلیل و پردازش آنها بر عهده بخش نگا شت و تجمیع آن ا ست. شکل ۲ این فرایند را به صورت بصری نمایش می دهد. در مرحله اول،کاربر درخوا ست خود را که معمولاً یک پرس و جو به زبان جاوا ست را به گرهی که وظیفه اجرای درخواست ها را بر عهده دارد (مدیر درخواست Job tracker) – ارسال می-کند. در این مرحله مدیر درخواست بررسی می کند که به چه فایل هایی برای پاسخ به پرس و جوی کاربر نیاز دارد و به کمک گره مدیریت نام ، گره های داده حاوی آن بخش ها را در کلاستر می یابد (عمل نگاشت).

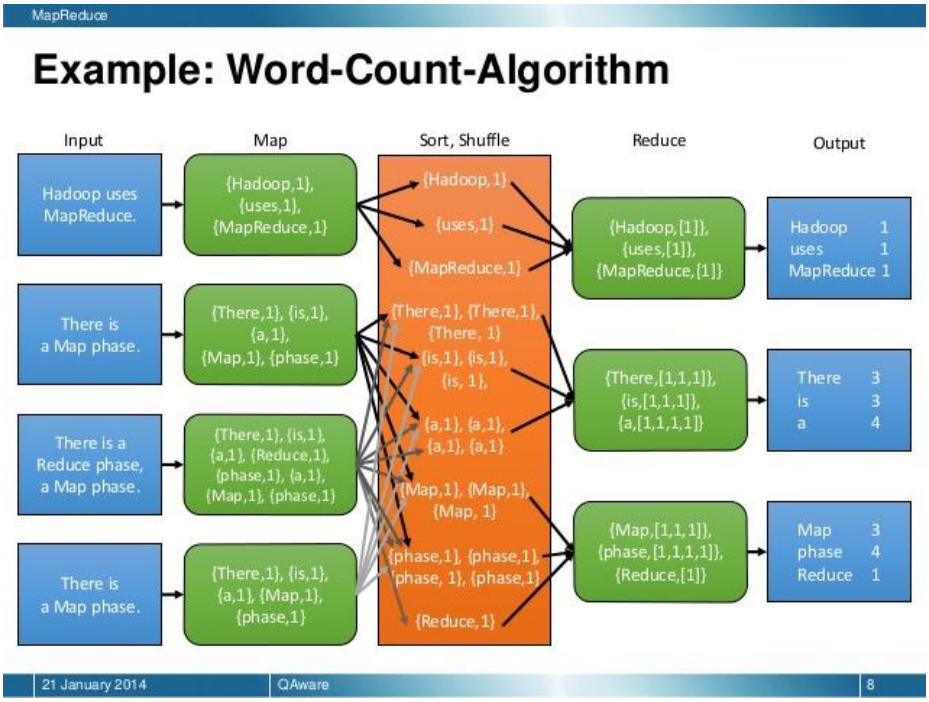
سپس این درخواست به تک تک آن گره ها ارسال میگردد. این گره ها که هنگام پردازش به آنها مدیر وظیفه می گوئیم مستقلاً و به صورت موازی کار پردازش داده های خود را (اجرای تابع نگاشت) انجام می دهند.



شکل ۲ : ساختار عملیاتی هدوپ و فرآیند نگاشت و تجمیع

پس از اتمام کار هر مدیر وظیفه، نتایج در همان گره ذخیره می گردد. پس از آماده شدن نتایج میانی که طبیعتا چون وابسته به داده های موجود در روی یک گره ا ست ، محلی و ناقص خواهد بود ، مدیر درخوا ست، فرمان تجمیع را به این گره ها ارسال می کند تا پردازش نهایی را بر روی نتایج انجام داده و نتیجه درخوا ست کاربر در یک گره محا سباتی نهایی ذخیره گردد. در این مرحله ، نگاشت و تجمیع به اتمام رسیده است و پردازش بعدی بر روی نتایج حاصل بر عهده تحلیل گران حوزه کلان داده است . این پردازش می تواند به صورت مستقیم بر روی نتایج انجام شود و یا با انتقال داده های حاصله به بانک های اطلاعاتی رابطه ای و یا انباره های داده، از روشهای کلاسیک تحلیل داده استفاده شود.

مثالی از نحوه شمارش کلمات در یک کلا ستر هدوپ با روش نگا شت و تجمیع در شکل زیر نمایش داده شده ا ست . فایلهای ورودی در HDFS ذخیره شده اند و عملیات نگا شت در هر گره محا سباتی بدین صورت انجام میگیرد که به ازای هر کلمه که از فایل خوانده میشود، یک زوج (کلمه ، تعداد) ایجاد می کند که تعداد اولیه آن یک خواهد بود. در مرحله بعدی این زوجهای ایجاد شده مرتب سازی می شوند و در مرحله تجمیع، کلمات کنار هم که یکسان هستند با هم ادغام شده و اعداد آنها با هم جمع میشود و سرانجام فایل نهایی که شمارش تعداد هر کلمه در آن آمده است، ایجاد می گردد.



مزایا و معایب هدوپ

مهم ترین مزیت هدوپ توانایی پردازش و تحلیل حجم عظیم داده-های بدون ساختار یا شبه-ساختار که تاکنون امکان پردازش آنها به صورت بهینه (هزینه و زمان) مقدور نبوده است.

مزیت بعدی هدوپ به امکان گسترش ساده و مقیاس پذیری افقی (سهولت افزودن سیستم به کلا ستر هدوپ بدون نیاز به ارتقاء سخت افزاری یک سیستم) آن بر می گردد که به راحتی می توان تا سطح اگزا بایت داده ها را مورد تحلیل قرار داد

و دیگر لازم نیست شرکتها بر روی داده های نمونه و زیرمجموعه ای از داده های اصلی کار کنند و به کمک هدوپ امکان بررسی تمام داده ها فراهم شده است.

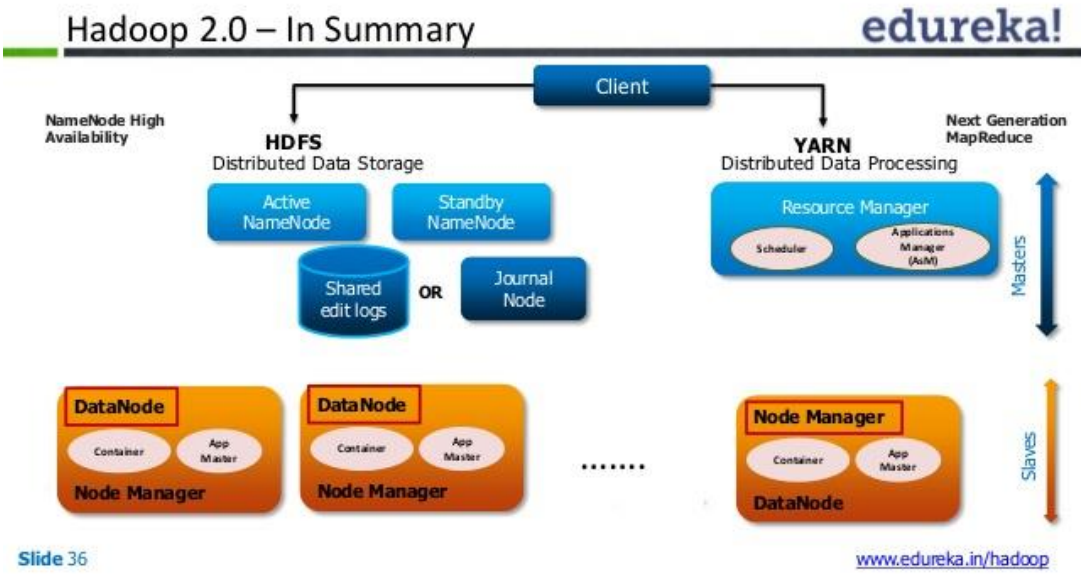
مزیت دیگر هدوپ هم هزینه راه اندازی اندک آن است که دلیل اصلی آنهم رایگان بودن آن است و نیز عدم نیاز به سخت افزار حرفه ای و گران . بخصوص با رواج رایانش ابری و قیمتهای مناسب آن برای پردازشهای موردی و نیز ابرهای خصوصی، راه اندازی یک سامانه هدوپ به فرآیندی چند ساعته تبدیل شده است.

از طرف دیگر هدوپ و زیر مجموعه های آن همگی در مراحل اولیه توسعه هستند و غیر بالغ و نوپا هستند. این امر خود باعث تغییر و اصلاح مداوم این چهارچوب می شود که هزینه آموزش مداوم را به سازمانها تحمیل می کند.

از سوی دیگر نوپا بودن این مدل نرم افزاری باعث می شود افراد کمی مهارت لازم برای ایجاد و کار با سامانه های مبتنی بر هدوپ را دارند و برای بسیاری از شرکتها کمبود نیروی انسانی متخصص مهمترین چالش آنها در استفاده از این سامانه خواهد بود.

مشکل دیگر هدوپ که ماهیت ذاتی دارد، عدم توانایی پردازش بلادرنگ داده هاست. چون مدیر درخواست باید منتظر تکمیل کار تک تک گره های محاسباتی سامانه بماند تا بتواند جواب نهایی را به کاربر تحویل دهد . هر چند با رشد سریع فناوریهای بانکهای اطلاعاتی NoSQL و تلفیق آن با هدوپ ، این مشکل نیز تا حدی رفع خواهد شد.

امروزه نسخه دوم هدوپ با بهبود فرآیند مدیریت منابع، لایه ای جدید به سامانه هدوپ اضافه کرده است با نام YARNکه وظیفه مدیریت منابع سیستم مانند حافظه، دیسک ، شبکه و غیره را بر عهده دارد که با این توصیف، در لایه پایین هدوپ ما سیستم HDFS را برای ذخیره داده ها داریم و در لایه میانی ، YARNوظیفه مدیریت منابع سیستمی را برعهده دارد و در لایه بالا هم عملیات پردازش داده با مکانیزم نگاشت و تجمیع انجام می پذیرد.



می توان به جای لایه فوقانی یعنی روش کلاسیک و سنتی نگاشت و تجمیع (Map/Reduce) در دنیای کلان داده از روش‌های نوینی مانند آپاچی تز (TEZ) و یا اسپارک استفاده کرد که بسته به کاربرد، اسپارک سرعتی ده تا صد برابری نسبت به روش معمول نگاشت و تجمیع دارد.

۲-۳ بانک‌های اطلاعاتی NoSQL

هدوپ به طور خاص برای پردازش کلان داده شکل گرفته است و نیازهای ذخیره و بازیابی کلان داده در آن دیده نشده است. شرکتهایی مانند گوگل ، فیس بوک ، آمازون، توئیتر و مانند آن که روزانه نیاز به ذخیره چندین گیگابایت تا چندین ترابایت بایت داده را دارند و نیز بازیابی سریع و موثر اطلاعات برایشان امری حیاتی است، دست به ابداع نوع جدیدی از بانکهای اطلاعاتی زده اند که به طور خاص برای ذخیره و بازیابی خودکار و حرفه ای کلان داده طراحی شده اند.

این نوع از بانکهای اطلاعاتی که دیگر مفاهیم کلاسیک پایگاه داده مانند جدول و رکورد در آنها معنای خود را از دست داده است، به بانکهای اطلاعاتی NoSQL یا Not Only SQL معروف شده اند و به عنوان یکی از روشهای اصلی ذخیره کلان داده در دنیای فناوری اطلاعات مطرح هستند. امروزه بیش از ۱۲۰ بانک اطلاعاتی در زمره این گروه قرار گرفته اند.

وجود این بانکهای اطلاعاتی به تحلیلگران سازمانی این امکان را می دهد تا بدون درگیر شدن در جزئیاتی مانند مدل نگاشت و تجمیع، داده ها را ذخیره کرده و با امکاناتی که خود بانکهای اطلاعاتی در اختیار آنها می گذارند به تحلیل آنها بپردازند. بعضی از این بانکهای نوین مانند کاساندرا و HBase می توانند همراه به هدوپ به کار گرفته شوند.

مشکل اصلی در استفاده از این بانکهای اطلاعاتی علاوه بر نوپا بودن و توسعه سریع آنها، عدم پشتیبانی آنها از مفاهیم تراکنش، جامعیت و سازگاری داده و استقلال عملیات است که به خاطر افزایس بهره وری و سرعت انجام گرفته است.

۴. کاربردهای کلان داده

- کشف خطا و یا کشف نفوذ به شبکه با ذخیره و آنالیز لاگ شبکه در یک سازمان یا وب سایت.
- تنظیم قیمت صحیح محصول در جهت فروش بیش تر، طراحی محل قرارگیری محصولات در فروشگاه با توجه به اطلاعات آماری حرکت خریداران، کشف راه کارهای ترغیب مشتری در خرید مجدد از فروشگاه، مدیریت زنجیره عرضه، تقسیم بندی مشتریان، پیشنهاد دقیق کالا در زمان مناسب از جمله موارد استفاده از کلان داده با تجزیه و تحلیل اطلاعات مربوط به سبد خرید مشتریان خواهد بود.
- پیش بینی میزان ریسک مرتبط با یک طرح اقتصادی و تشخیص الگوی شک برانگیز در استفاده از کارت اعتباری در حوزه بانکداری. کشف نفوذ و یا تقلب، کلاهبرداری و یا پولشویی با استفاده از تجزیه و تحلیل تراکنشهای مالی مشتریان با دیگر منابع اطلاعاتی نیز، امروزه بسیار کاربردی شده است.
- شخصی سازی خدمات از دیگر حوزه های فعال کاربرد کلان داده است و بسته به رفتار قبلی کاربر و داده هایی که از او داریم ، پیشنهاد خود را به او کاملا اختصاصی ارائه دهیم مثلا برای پیشنهاد وام به یک مشتری ، نمایش تبلیغات، پیشنهاد خودرو ، نمایش نوع خروجی جستجوهای کاربر و مثالهایی از این دست، می توان از کلان داده استفاده کرد.

منابع:

عزیزی وامرزانی، حا مد، و مریم خادمی، ۱۳۹۳، کلان داده، کاربرد ها و چالش های آن، ه مایش ملی الکترونیکی دستاوردهای نوین در علوم مهندسی و پایه، تهران، مرکز پژوهشهای زمین کاو

بنائی، ســید مجتبی و ســید هادی موســوی، ۱۳۹۱، رهیافت های نوین در هوش تجاری، اولین کارگاه ملی رایانش ابری، تهران، دانشگاه صنعتی امیرکبیر،

http://itresearches.ir/