اکنون کلان داده

مؤلف:

اوریلی

مترجمین:

دکتر حامد رجبی قمی

مهندس مریم کاردگر

فن‌آوری نوین

عنوان و نام پدیدآور	:	اکنون کلان‌داده/ مولف اوریلی؛ مترجمین حامد رجبی‌قمی، مریم کاردگر.
مشخصات نشر	:	بابل: فناوری نوین، ۱۴۰۰.
مشخصات ظاهری	:	۱۲۸ ص.
شابک	:	۵۰۰۰۰۰ ریال :۰-۳۳ -۷۳۹۳-۶۲۲-۹۷۸
وضعیت فهرست نویسی	:	فیپا
یادداشت	:	عنوان اصلی: .Big data now, ۲۰۱۴
یادداشت	:	کتاب حاضر در سالهای مختلف توسط ناشران و مترجمان متفاوت منتشر شده است.
موضوع	:	داده‌های کلان
موضوع	:	Big data
موضوع	:	داده‌کاوی
موضوع	:	Data mining
موضوع	:	هوش مصنوعی
موضوع	:	Artificial intelligence
شناسه افزوده	:	رجبی قمی، حامد، ۱۳۶۰-، مترجم
شناسه افزوده	:	کاردگر، مریم، ۱۳۶۶ -، مترجم
شناسه افزوده	:	شرکت اورایلی مدیا
شناسه افزوده	:	O'Reilly Media, Inc
رده بندی کنگره	:	QA۹/۷۶
رده بندی دیویی	:	۰۰۵/۷۴
شماره کتابشناسی ملی	:	۷۵۷۱۰۲۵
وضعیت رکورد	:	فیپا

فن‌آوری نوین بابل، کد پستی ۴۷۱۶۷-۷۳۴۴۸ تلفن: ۳۲۲۵۶۶۸۷-۰۱۱

اکنون کلان داده

ترجمه: حامد رجبی قمی، مریم کاردگر.

نوبت چاپ: چاپ اول

سال چاپ: بهار ۱۴۰۰

شمارگان: ۲۰۰

قیمت: ۵۰۰۰۰ تومان

نام چاپخانه و صحافی: دفتر فنی سورنا

شابک: ۰-۳۳ -۷۳۹۳-۶۲۲-۹۷۸

نشانی ناشر: بابل، چهارراه نواب، کاظم‌بیگی، جنب مسجد منصور کاظم‌بیگی، طبقه اول

طراح جلد: کانون آگهی و تبلیغات آبان (احمد فرجی)

فروشگاه و پخش کتاب چاپی: تهران، تلفن ۶۶۴۰۰۲۲۰– ۶۶۴۰۰۱۴۴

تهران، خ اردیبهشت، نبش وحید نظری، پلاک ۱۴۲ تلفکس: ۶۶۴۰۰۲۲۰-۶۶۴۰۰۱۴۴

فهرست مطالب

فصل ۱ مقدمه

در اولین نسخه "Big Data Now"، تیم O'Reilly تولد و توسعه زودهنگام ابزارهای داده و علم داده را دنبال کردند. حال در این نسخه دوم، می‌خواهیم ببینم زمانی که کلان داده رشد می‌کنند چه اتفاقی رخ می‌دهد: چه چیزی در حال اعمال است، کجا نقشی ایفا می‌کند و پیامدهای- به‌طور یکسان خوب و بد - رشد داده.

نسخه ۲۰۱۰۲، Big Data Now را در پنج بخش سازمان‌دهی کردیم:

افزایش سرعت با کلان داده- اطلاعات ضروری در مورد ساختارها و تعاریف کلان داده

ابزارهای کلان داده، تکنیک‌ها و استراتژی‌ها- راهنمایی کارشناسانه برای تبدیل نظریه‌های کلان داده به محصولات کلان داده.

کاربرد کلان داده- نمونه‌هایی از کلان داده‌ها در عمل، شامل نگاهی به ضعف داده‌ها

آنچه در کلان داده دنبال می‌شود - افکار در خصوص نحوه تکامل کلان داده و نقش آن در صنایع و حوزه‌ها.

کلان داده و مراقبت سلامتی- یک بخش ویژه برای بررسی احتمالات در زمانی که داده و مراقبت سلامتی با یکدیگر همراه می‌شوند.

علاوه بر Big Data Now، با تحلیلاتی که بر روی O'Reilly Radar انجام می‌دهیم و از طریق Strata coverage و events series، می‌توانید به توسعه‌های اخیر داده دسترسی داشته باشید.

فصل ۲ افزایش سرعت با کلان داده

۱–۲. کلان داده چیست؟

کلان داده، داده‌ای با ظرفیت پردازشی بیش‌تر از ظرفیت پردازشی سیستم‌های پایگاه داده متعارف است. داده بیش‌ازحد بزرگ است، خیلی سریع رشد می‌کند، یا در محدودیت‌های معماری پایگاه داده شما نمی‌گنجد. برای دستیابی به منفعتی از این داده، باید روش دیگری برای پردازش آن انتخاب کنید.

شعار فناوری اطلاعات داغ ۲۰۱۲، کلان داده به‌عنوان روش‌های مقرون‌به‌صرفه برای کاهش حجم، سرعت و تنوع داده‌های انبوه در حال ظهور هستند.

در این میان، داده حاوی الگوها و اطلاعات ارزشمندی است که قبلاً به دلیل حجم کار موردنیاز برای استخراج‌شان پنهان بودند. با مشارکت شرکت‌هایی مثل Walmart و Google این توانایی تا حدودی میسر گشته است اما با هزینه بسیار بالا. تجهیزات سخت‌افزاری امروزی، معماری ابر و نرم‌افزارهای منبع باز پردازش داده کلان داده را با حداقل منابع امکان‌پذیر ساخته است. پردازش کلان داده حتی برای گاراژهای راه‌انداز کوچکی که می‌توانند زمان سرویس‌دهنده در ابر را با هزینه پایین اجاره کنند نیز میسر شده است.

ارزش کلان داده برای یک سازمان دو نوع است: استفاده تحلیلی و ارائه محصولات جدید. تحلیل داده‌های بزرگ، بینش‌های پنهان موجود در داده (ازجمله تأثیر نظیر به نظیر بر مشتریان، آشکارشده با تحلیل تراکنش‌های خریداران، و داده‌های اجتماعی و جغرافیایی) که پردازش آن‌ها بسیار هزینه‌بر بود را آشکار می‌کند. علی‌رغم ماهیت نسبتاً ایستای گزارش‌های از پیش تعیین‌شده، این‌که بتوانیم هر قلم داده را در یک زمان معقول پردازش کنیم، نیاز مبرم به نمونه‌برداری را حذف و یک رویکرد تحقیقاتی برای داده‌ها باز می‌کند.

راه‌اندازی‌های موفق وب در ده گذشته، نمونه‌های بزرگی از کلان داده به‌کاررفته به‌عنوان یک فعال‌ساز برای محصولات و دستگاه‌های جدید است. به‌عنوان‌مثال، فیس‌بوک با ترکیب سیگنال‌های بسیاری از واکنش‌های کاربران و دوستان‌شان، توانست تجربیات بسیار شخصی کاربر را کشف و نوع جدیدی از تبلیغات را ایجاد کند. این‌که بخش زیادی از ایده‌ها و ابزارهای پایه کلان داده از گوگل، یاهو، آمازون و فیس‌بوک پدید آمده است، تصادفی نیست.

ظهور کلان داده در شرکت‌ها یک همتای ضروری را برای آن ارمغان آورده است: چابکی. بهره‌برداری موفق از مقادیر در داده‌های بزرگ نیازمند آزمایش و اکتشاف است. چه محصولات جدیدی ایجاد کنیم یا به دنبال روش-هایی برای دستیابی به مزیت رقابتی باشیم، شغل نیازمند کنجکاوی و کارآفرینی است.

۱–۱–۲. کلان داده چه شکلی هستند؟

به‌عنوان یک اصطلاح کلی، همانند اصطلاح **"ابر"** که فناوری‌های متنوعی را پوشش می‌دهد، اصطلاح **"کلان دادهی "** نیز می‌تواند بسیار مبهم باشد. داده‌های ورودی در سیستم‌های کلان داده می‌توانند برگرفته از شبکه‌های

اجتماعی، لاگ‌های سروری وب، حسگرهای جریان ترافیکی، تصاویر ماهواره، جریان‌های پخش صوتی، تعاملات بانکی، MP۳های موسیقی راک، محتوای صفحات وب، اسکن اسناد دولتی، مسیرهای GPS، سنجش از راه دو خودروها، اطلاعات بازار مالی، و غیره باشد. آیا همه این‌ها یک چیز یکسان هستند؟

برای روشن کردن مباحث، معمولاً از سه V (برای **حجم**[۱]، **سرعت**[۲]، و **تنوع**[۳]) برای مشخص کردن جنبه‌های مختلف کلان داده استفاده می‌شود که لنز مفیدی برای مشاهده و درک ماهیت داده‌ها و پلت‌فرم‌های نرم‌افزاری موجود برای بهره‌برداری از آن‌ها هستند. احتمالاً تا حدودی با یکی از این Vها سروکار خواهید داشت.

حجم

مزیت به دست آمده از توانایی پردازش حجم بزرگی از اطلاعات، جذابیت اصلی تجزیه‌وتحلیل کلان داده است. داشتن داده‌های بیش‌تر منجر به داشتن مدل‌های بهتر می‌شود: بیت‌های ساده ریاضی می‌توانند بدون دلیل برای حجم-های زیاد داده مؤثر باشند. آیا اگر شما بتوانید با در نظر گرفتن حدوداً ۳۰۰ فاکتور به‌جای ۶ فاکتور، پیش‌بینی انجام دهید، در این صورت می‌توانید تقاضا را بهتر پیش‌بینی کنید؟ این مسئله، مهم‌ترین چالش در ساختارهای IT متعارف است که نیازمند فضای ذخیره‌سازی مقیاس‌پذیر و یک روش توزیعی برای پرس‌وجوها است. بسیاری از شرکت‌ها دارای مقدار زیادی داده بایگانی‌شده، احتمالاً در قالب لاگ‌ها هستند اما توانایی پردازش آن‌ها را ندارند.

با فرض این‌که حجم داده‌ها بیش‌تر از حدی است که زیرساخت‌های پایگاه داده رابطه‌ای بتواند با آن مقابله کند، گزینه‌های پردازشی به‌طور گسترده به انتخاب بین معماری‌های پردازش موازی انبوه- انبار داده‌ها یا پایگاه داده‌هایی مثل Greenplum و روش‌های مبتنی بر Apache Hadoop تقسیم می‌شوند. این انتخاب اغلب با درجه‌ای که یکی از Vهای دیگر- تنوع- در آن نقش خواهد داشت، اعلان می‌شود. معمولاً، روش‌های انبار داده‌ها شامل طرح‌های از پیش تعیین‌شده، مجموعه داده‌های منظم و به‌آرامی تکامل‌یافته هستند. از طرفی دیگر، Apache Hadoop هیچ جایی در ساختار داده‌هایی که پردازش می‌کند، ندارد.

در حقیقت، هادوپ پلت‌فرمی برای توزیع مسائل محاسباتی بر روی تعدادی سرویس‌دهنده است و به‌عنوان اولین نرم‌افزار منبع باز منتشرشده، از روش نگاشت کاهش[۴] که گوگل در استفاده از آن پیش‌گام است، برای جمع‌آوری شاخص‌های جست‌وجویش استفاده می‌کند. MapReduce در هادوپ شامل توزیع یک مجموعه داده بر روی

---

۱ volume
۲ velocity
۳ variety
۴ MapReduce

چندین سرویس‌دهنده و انجام عملیاتی بر روی داده است: گام "نگاشت[۵]". سپس نتایج جزئی مجدداً باهم ترکیب می‌شوند: گام "کاهش[۶]".

هادوپ برای ذخیره داده از سیستم فایل توزیعی خودش، HDFS که دسترسی به داده را از چندین گره محاسباتی میسر می‌سازد، استفاده می‌کند. یک الگوی استفاده معمول از هادوپ شامل سه گام است:

- بارگذاری داده در HDFS
- عملیات نگاشت‌کاهش و
- بازیابی نتایج از HDFS

این فرآیند که ماهیتاً یک عمل دسته‌ای است، برای وظایف محاسباتی تحلیلی یا غیرتعاملی مناسب است. به همین دلیل هادوپ خودش یک روش پایگاه داده‌ای یا انبار داده نیست بلکه می‌توان به‌عنوان یک‌یک مکمل تحلیلی[۷] عمل کند.

یکی از معروف‌ترین کاربران هادوپ، فیس‌بوک است که مدلش از الگوی پیروی می‌کند. پایگاه داده MySQL داده‌های اصلی را ذخیره می‌کند و سپس این داده‌ها در هادوپ منعکس می‌شود که در آن محاسباتی ازجمله ایجاد توصیه‌هایی بر اساس علایق دوستان شما انجام می‌شود. سپس فیس بوک نتایج را به‌منظور استفاده در صفحات کاربران به MySQL برمی‌گرداند.

سرعت

اهمیت سرعت داده‌ها- نرخ افزایشی جاری شدن داده در یک سازمان- مطابق الگویی مشابه الگوی حجم است. مسائلی که قبلاً محدود به بخش‌های صنعتی بودند اکنون در محیط‌های وسیع‌تری وجود دارند. شرکت‌های تخصصی مثل معامله‌گران مالی از سیستم‌هایی که با انتقال سریع داده مقابله می‌کردند، به نفع خود استفاده کردند. حال نوبت ماست.

چرا این‌طور است؟ عصر اینترنت و تلفن همراه بدین معنی است که شیوه تحویل محصول و سرویس به مشتری به‌طور فزاینده‌ای ابزاری شده است و جریان داده‌ای را به ارائه‌دهنده برمی‌گرداند. خرده‌فروشان آنلاین قادرند با هر کلیک و تعامل، تاریخچه بزرگی از مشتریان را کامپایل کنند: نه فقط فروش‌های نهایی. کسانی که می‌توانند با پیشنهاد خریدهای اضافی، برای رسیدن به مزیت رقابتی، به‌سرعت از اطلاعات استفاده کنند. در عصر گوشی‌های

---

[۵] map
[۶] reduce
[۷] analytical adjunct

هوشمند، ازآنجایی‌که مشتریان به همراه خود منبعی از تصاویر جغرافیایی و داده‌های صوتی را حمل می‌کنند، مجدداً نرخ جریان داده افزایش کرده است.

مسئله فقط سرعت داده‌های ورودی نیست: مثلاً انتقال سریع داده‌ها به فضای ذخیره‌سازی عمده برای پردازش دسته‌ای آتی نیز امکان‌پذیر است. اهمیت در سرعت چرخه بازخورد، گرفتن داده از داده به‌واسطه یک تصمیم نهفته است. یک IBM تجاری این نکته را بیان می‌کند که اگر تصویر لحظه‌ای مکان ترافیک را نداشتید ، نمی‌توانستید از جاده عبور کنید. زمان‌هایی وجود دارد که نمی‌توانید منتظر گزارشی برای اجرا یا تکمیل کر هادوپ باشید. اصطلاح صنعتی این‌گونه انتقال سریع داده **"جریان داده‌ها"** یا **"پردازش رویداد پیچیده"** نام دارد. اصطلاح دوم که در دسته محصولات قبل از پردازش جریانی داده‌ها قرار دارد به ارتباطات گسترده‌تری دست‌یافته است و به نظر می‌رسد به نفع جریان داده‌ها، کاهش‌یافته است. دو دلیل اصلی برای در نظر گرفتن پردازش جریانی وجود دارد: اولی زمانی است که داده‌های ورودی خیلی سریع و کامل ذخیره می‌شوند: برای حفظ الزامات ذخیره‌سازی عملی، همانند جریان‌های داده، باید چندین سطح تحلیلی رخ دهد. در پایان مقیاس، Large Hadron Collider در CERN داده‌های خیلی زیادی تولید کرده است که دانشمندان باید اکثریت آن را حذف کنند- با امید این‌که هیچ‌چیز مفیدی را از بین نبرند. دلیل دوم در نظر گرفتن پردازش جریانی، جایی است که برنامه به داده‌ها پاسخ سریع می‌دهد. به دلیل ظهور برنامه‌های موبایل و بازی‌های آنلاین، این وضعیت به میزان زیادی مرسوم شده است.

دسته‌بندی محصولات برای مدیریت جریان داده‌ها به محصولات انحصاری ازجمله جریان‌های InfoSphere در IBM و lesspolished تقسیم می‌شود و همچنان چارچوب‌های منبع باز نشات گرفته از صنعت وب نیز وجود دارند: طوفان تویتر و یاهو S۴.

همان‌طور که در بالا ذکر گردید، این مسئله فقط در مورد داده‌های ورودی نیست. سرعت خروجی سیستم نیز ممکن است بسیار مهم باشد. هرچه حلقه بازخورد محدودتر باشد، مزیت رقابتی بیش‌تر خواهد بود. نتایج ممکن است به‌طور مستقیم به یک محصول مثل توصیه‌های فیس‌بوک یا داشبوردهای به‌کاررفته برای تصمیم‌گیری بروند که برای افزایش سرعت به‌ویژه در وب نیاز است و توسعه ذخیره‌سازی مقادیر کلیدی و پایگاه داده‌های ستونی را برای بازیابی سریع اطلاعات از قبل محاسبه‌شده، بهینه‌سازی کرده است. این پایگاه داده‌ها بخشی از یک **دسته چتری**[۸] شناخته‌شده به نام NoSQL را تشکیل می‌دهند و زمانی به‌کاررفته می‌روند که مدل‌های رابطه‌ای مناسب نباشند.

[۸] umbrella category

تنوع

داده‌ها به‌ندرت در فرمی بی‌نقص و آماده برای پردازش هستند. یک موضوع رایج در سیستم کلان داده این است که داده‌های منبع متنوع هستند و دارای ساختارهای رابطه‌ای منظم نیستند. این داده‌ها می‌توانند متنی از شبکه‌های اجتماعی، تصویر یا داده خامی از یک منبع حسگر باشند. هیچ‌کدام برای ادغام شدن در یک برنامه آماده نیستند.

حتی در وب، که ارتباطات کامپیوتر به کامپیوتر باید تضمین‌هایی به همراه داشته باشد، ماهیت داده‌ها کثیف است. مرورگرهای مختلف داده‌های مختلفی ارسال می‌کنند، کاربران داده‌ها را دریغ می‌کنند، آن‌ها ممکن است از نسخه-های نرم‌افزاری مختلف یا از فروشندگان مختلف برای برقراری ارتباط با شما استفاده کنند و می‌توانید شرط ببندید که اگر بخشی از فرآیند توسط انسان انجام شود، خطا و ناسازگاری وجود خواهد داشت.

یک استفاده رایج پردازش داده‌های بزرگ، گرفتن داده‌های غیر ساخت‌یافته و استخراج معانی مرتب، برای مصرف توسط انسان یا به‌عنوان داده‌های ورودی برای یک برنامه است. این چنین مثالی یک قطعنامه موجودیت است، فرآیند تعیین دقیق چیزی که نام به آن اشاره دارد. آیا این‌جا شهر لندن، انگلستان یا لندن تگزاس است؟ تا زمانی که منطق تجاری شما دخیل است، نمی‌خواهید حدس زده شوید.

فرآیند انتقال از داده‌های منبع به داده‌های کاربردی پردازش‌شده شامل از دست رفتن اطلاعات است. هنگامی‌که مرتب می‌شوید، پرتاب وسایل را ترک خواهید کرد. این اصل کلان داده است: زمانی که می‌توانید همه‌چیز را حفظ کنید. ممکن است در بیت‌هایی که دور می‌ریزید، سیگنال‌های مفیدی وجود داشته باشد. اگر داده‌های منبع را از دست بدهید، هیچ بازگشتی وجود ندارد.

علی‌رغم محبوبیت و شناخت ماهیت پایگاه داده‌های رابطه‌ای، این‌طور نیست که داده‌ها همواره باید در مقصد باشند، حتی اگر مرتب‌شده باشند. انواع خاصی از داده‌ها برای کلاس‌های خاصی از پایگاه داده بهتر هستند. به‌عنوان‌مثال، اسناد کدگذاری شده در XML، زمانی که در یک XML اختصاصی مثل MarkLogic ذخیره شوند، متنوع‌ترین اسناد هستند. روابط شبکه‌های اجتماعی دارای ماهیت گروهی هستند، و پایگاه داده‌های گرافی مثل Neo۴J عملیات را ساده‌تر و مؤثرتر بر روی آن‌ها انجام می‌دهند.

حتی در مواردی که هیچ عدم سازگاری در نوع داده رادیکالی وجود ندارد، عیب پایگاه داده رابطه‌ای مربوط به ماهیت ایستای طرح‌هایش است. در یک محیط اکتشافی چابک، نتایج محاسبات با کشف و استخراج سیگنال‌های بیش‌تر تکامل خواهند یافت. پایگاه داده‌های NoSQL نیمه ساخت‌یافته به‌منظور ایجاد انعطاف‌پذیری، این نیازمندی

را برآورده می‌سازند: آن‌ها ساختار کافی برای سازمان‌دهی داده فراهم می‌کنند، اما نیازی به استخراج طرح داده قبل از ذخیره‌سازی آن ندارند.

۲-۱-۲. در عمل

ماهیت کلان داده را کشف و چشم‌انداز کلان داده را از یک سطح بالا بررسی کردیم. به‌طورمعمول، هنگام استقرار، ابعادی هستند که باید فرای انتخاب ابزار در نظر گرفته شوند.

ابر یا در خانه؟

در حال حاضر اکثریت راه‌حل‌های کلان داده در سه فرم ارائه می‌شوند: تنها نرم‌افزار، به‌عنوان یک دستگاه[۹] یا مبتنی بر ابر. تصمیمات در خصوص این‌که کدام مسیر باید انتخاب شود، به فاکتورهای دیگری مثل محلیت داده، حریم خصوصی و تنظیمات، منابع انسانی و الزامات پروژه بستگی دارد. بسیاری از سازمان‌ها یک راه‌حل ترکیبی را انتخاب می‌کنند: استفاده از منابع ابری مبتنی بر تقاضا برای تکمیل کارها در خانه.

کلان داده بزرگ است

این یک حقیقت محض است که داده‌ای که برای پردازش عادی بسیار بزرگ باشد، برای انتقال به‌جای دیگر نیز بزرگ است. اولویت‌های IT در حال تغییر هستند: این برنامه است که باید انتقال داده شود، نه داده. اگر بخواهید داده‌های سرشماری ایالت متحده را تحلیل کنید، اجرای کد آن در پلت‌فرم سرویس‌های تحت وب آمازون که میزبان این‌گونه داده‌های محلی است بسیار ساده‌تر است و دیگر نیازی به‌صرف زمان یا هزینه برای انتقال آن ندارید.

حتی اگر داده برای انتقال خیلی بزرگ نباشد، محلیت می‌تواند مسئله‌ساز باشد، به‌خصوص با این به‌روزرسانی سریع داده‌ها. سیستم‌های معاملات مالی برای به دست آوردن سریع‌ترین اتصال به داده‌های منبع، در مراکز داده تجمع کردند، بنابراین دلیل، تفاوت میلی‌ثانیه‌ای در زمان پردازشی نماد یک مزیت رقابتی است.

کلان داده کثیف است

همه‌چیز به زیرساخت مربوط نیست. کاربران کلان داده به‌طور پیوسته گزارش می‌دهند که ۸۰ درصد تلاش‌های انجام‌شده برای مقابله با داده‌ها به پاک کردن آن در ابتدا مربوط می‌شود، همان‌طور که Pete Warden در واژه‌نامه کلان داده خود مشاهده نمود: "من احتمالاً زمان بیش‌تری را برای تبدیل داده‌های منبع کثیف به چیزی قابل‌استفاده صرف می‌کنم، نسبت به این‌که فرآیند تحلیل داده را انجام دهم".

---

[۹] appliance

به دلیل هزینه بالای جمع‌آوری و پاک‌سازی داده، این‌که آنچه را که واقعاً نیاز دارید، به خود بسپارید. بازار داده-ها ابزارهایی برای دستیابی به داده‌های مشترک هستند و اغلب می‌توانید به بهبود آن‌ها کمک کنید. کیفیت نیز ممکن است متغیر باشد اما به‌طور فزاینده‌ای یک معیار رقابتی در بازار داده‌ها خواهد بود.

فرهنگ

پدیده کلان داده به ظهور علم داده‌ها بسیار وابسته است، رشته‌ای که ریاضیات، برنامه‌نویسی، و غریزه علمی را باهم ترکیب می‌کند. بهره‌مندی از ابزارهای داده به معنی سرمایه‌گذاری در تیم‌هایی با این مهارت و اتصال آن‌ها به تمایلات سازمانی به‌منظور درک و استفاده از داده‌ها است.

در این گزارش، "ساختن تیم‌های علوم داده"، D.J. Patil دانشمندان داده‌ای را با معیارهای کیفی زیر مشخص کرده است:

- **تخصص فنی:** بهترین دانشمندان داده‌ای معمولاً دارای تخصص‌های عمیق در برخی رشته‌های علمی هستند.
- **کنجکاوی:** تمایل به بررسی عمیق و کشف و تبدیل یک مسئله به یک مجموعه فرضیات واضحی که قابل‌آزمایش باشد.
- **داستان‌سرایی:** توانایی استفاده از داده‌ها برای گفتن یک داستان و قادر به برقراری ارتباط مؤثر با آن.
- **باهوشی:** توانایی برخورد با مسئله به روش‌های خلاق مختلف.

ماهیت گسترده پروژه‌های تجزیه‌وتحلیل کلان داده می‌تواند دارای جنبه‌های ناراحت‌کننده‌ای باشد: داده باید به چندین سیلو شکسته شود تا قابل‌استخراج باشد و سازمان باید نحوه برقراری ارتباط و تفسیر نتایج تحلیل‌ها را یاد بگیرد.

مهارت‌های داستان‌سرایی و باهوشی فاکتورهای بسیار مهمی هستند که درنهایت تعیین می‌کنند که آیا مزایای کارهای تحلیلی در یک سازمان جذب می‌شوند یا خیر. هنر و تجربه بصری سازی داده در ایجاد پل بر روی شکاف بین انسان و کامپیوتر به‌منظور ایجاد هدفمند بینش تحلیلی بیش‌ازپیش مهم‌تر شده است.

بدانید کجا می‌خواهید بروید

درنهایت، به خاطر داشته باشید که کلان داده، نوشدارو نیست. می‌توانید الگوها و سرنخ‌هایی در داده‌های خود پیدا کنید، اما بعد چی؟ Christer Johnson، سرگروه IBM در تحلیل پیشرفته در آمریکای شمالی توصیه زیر را برای شروع کار تجارت با کلان داده گفته است: اول، تصمیم بگیرید که چه مسئله‌ای را می‌خواهید حل کنید. انتخاب یک مسئله تجاری واقعی، مثلاً چگونه می‌توانید برای افزایش هزینه‌های مشتری، استراتژی تبلیغ خود را تغییر دهید،

در پیاده‌سازی به ما کمک می‌کند. علاوه بر روحیه شرکتی، داشتن یک هدف مشخص نیز به‌شدت برای کلان داده مفید است.

۲–۲. Apache Hadoop چیست؟

Apache Hadoop نیروی محرک رشد صنعت کلان داده بوده است که به همراه فناوری‌های مربوطه مثل Hive و Pig اغلب خواهید شنید. اما چه‌کاری انجام می‌دهد و چرا به همه استراتژی‌هایش مثل Oozie و Zookeeper و Flume نیاز دارید؟

Hadoop توانایی پردازش ارزان حجم زیادی از داده را بدون در نظر گرفتن ساختارش به ارمغان می‌آورد. منظور ما از حجم زیاد، ۱۰ تا ۱۰۰ گیگابایت و بالاتر است. تفاوت این توانایی پردازش نسبت به گذشته چیست؟

انبار داده‌های سازمانی موجود و پایگاه داده‌های رابطه‌ای در پردازش داده‌های ساخت‌یافته پیشرفته هستند و می‌توانند حجم عظیمی از اطلاعات را ذخیره کنند، اما هزینه‌ای نیز دارند: این الزام برای ساختار، انواع داده‌هایی که می‌توانند پردازش شوند را محدود می‌کند و منجر به یک حالت سکون می‌شود که باعث می‌شود انبار داده‌ها برای اکتشاف چابک داده‌های ناهمگن کلان مناسب نباشد. میزان تلاش موردنیاز برای انبار داده اغلب بدین معنی است که منابع داده‌ای باارزش درون سازمان‌ها هرگز استخراج نمی‌شوند. این‌جا است که Hadoop می‌تواند تفاوت بزرگی ایجاد کند.

این تحقیق مؤلفه‌های اکوسیستم هادوپ را بررسی و توابع هرکدام را شرح می‌دهد.

۱–۲–۲. هسته هادوپ: نگاشت‌کاهش

در پاسخ به مسئله ایجاد شاخص‌های جست‌وجوی تحت وب، چارچوب نگاشت‌کاهش ایجادشده در گوگل یک نیروگاه قوی برای اکثر سیستم‌های پردازش کلان داده امروزی است. علاوه بر هادوپ، نگاشت‌کاهش نیز در پایگاه داده‌های MPP و NoSQL ازجمله Vertica یا MongoDB مشاهده می‌شود.

نوآوری مهم نگاشت‌کاهش، قابلیت گرفتن یک پرس‌وجو در یک مجموعه داده، تقسیم آن، و سپس اجرای موازی آن در چندین گره است. توزیع محاسبات مسئله بزرگ بودن داده به‌منظور قرار گرفتن بر روی یک ماشین را حل می‌کند. با ترکیب این تکنیک با سرویس‌دهنده‌های لینوکس یک روش مقرون‌به‌صرفه برای آرایه‌های محاسباتی عظیم به دست می‌آید.

در هسته، هادوپ یک پیاده‌سازی نگاشت‌کاهش منبع باز است. مطابق گزارش یاهو، هادوپ در سال ۲۰۰۶ پدیدار گشت، و به گفته سازنده آن، Doug Cutting در سال ۲۰۰۸ به قابلیت **"مقیاس وبی[۱۰]"** دست یافت.

با تکامل پروژه هادوپ، مؤلفه‌های بیش‌تری برای بهبود قابلیت استفاده و کارکرد به آن اضافه گردید. نام **"هادوپ"** برای نمایش کل این اکوسیستم به وجود آمده است که به‌موازات ظهور لینوکس بوده است: این نام شدیداً به هسته لینوکس اشاره دارد، اما به‌عنوان نام سیستم‌عامل کامل نیز پذیرفته‌شده است.

۲-۲-۲. لایه‌های پایین‌تر هادوپ: HDFS و نگاشت‌کاهش

در بالا در مورد توانایی نگاشت‌کاهش در توزیع محاسبات بر روی چند سرویس‌دهنده صحبت کردیم. برای انجام محاسبات، هر سرویس‌دهنده باید به داده‌ها دسترسی داشته باشد. این نقش HDFS، به‌عنوان سیستم فایل توزیع‌شده هادوپ است.

HDFS و نگاشت‌کاهش قوی هستند. سرویس‌دهنده‌ها در خوشه هادوپ ممکن است شکست بخورند و فرآیند محاسبات را قطع نکنند. HDFS تضمین می‌کند که داده‌ها به‌واسطه افزونگی در خوشه، تکرار می‌شوند. با تکمیل محاسبات، گره نتایج خود را در HDFS می‌نویسد.

هیچ محدودیتی برای داده‌هایی که توسط HDFS ذخیره می‌شوند، وجود ندارد. داده‌ها می‌توانند بدون ساختار و بدون طرح باشند. در عوض، پایگاه داده‌های رابطه‌ای نیاز دارند تا داده‌ها ساخت‌یافته باشد و طرح‌ها قبل از ذخیره-سازی داده تعریف‌شده باشند. در HDFS، ارائه‌دهنده کد مسئول معنی‌دار بودن داده‌ها است.

برنامه‌نویسی هادوپ در سطح نگاشت‌کاهش یک نمونه کار با APIهای جاوا است که در آن فایل داده‌ها به‌صورت دستی در HDFS بارگذاری می‌شود.

۳-۲-۲. بهبود قابلیت برنامه‌نویسی: Pig و Hive

کار کردن مستقیم با APIهای جاوا می‌توانند خسته‌کننده و مستعد خطا باشد و همچنین استفاده از هادوپ را به برنامه‌نویسان جاوا محدود می‌کند. هادوپ دو راهکار برای برنامه‌نویسی راحت‌تر هادوپ ارائه کرده است.

- Pig یک زبان برنامه‌نویسی است که وظایف رایج در کار کردن با هادوپ را ساده می‌کند: بارگذاری داده، نمایش تبدیلات انجام‌شده بر روی داده‌ها و ذخیره‌سازی نتایج نهایی. عملیات داخلی Pig می‌توانند با داده‌های نیمه ساخت‌یافته مثل فایل‌های لاگ سروکار داشته باشند و این زبان با استفاده از جاوا قابل بسط است تا بتواند انواع داده‌های سفارشی و تبدیلات را پشتیبانی کند.

---

[۱۰] web scale

❖ Hive به هادوپ قابلیت کار کردن به‌عنوان یک انبار داده را می‌دهد و ساختار را به‌سوی داده‌ها در HDFS سوق می‌دهد و سپس اجرای پرس‌وجو بر روی داده را با استفاده از نحوی همانند نحو SQL میسر می‌سازد. همانند Pig، قابلیت‌های هسته Hive نیز قابل گسترش هستند.

انتخاب بین Hive و Pig ممکن است گیج‌کننده باشد. Hive برای امور ذخیره‌سازی داده‌های با ساختار ایستا و نیازمند تجزیه‌وتحلیل مکرر مناسب‌تر است. نزدیک بودن Hive به SQL باعث شده است تا گزینه ایده آلی برای ادغام هادوپ با دیگر ابزارهای هوشمند تجاری باشد.

Pig به توسعه‌دهنده چابکی بیش‌تری برای کشف مجموعه داده‌های بزرگ، توسعه اسکریپت‌های موجز برای انتقال جریان داده به‌منظور پیوستن به برنامه‌های بزرگ را می‌دهد. Pig یک لایه نازک‌تر نسبت به Hive در هادوپ است و مزیت اصلی آن کاهش چشم‌گیر حجم کد موردنیاز در مقایسه با استفاده مستقیم از APIهای جاوا در هادوپ است. بدین ترتیب، مخاطب موردنظر Pig در درجه اول توسعه‌دهنده نرم‌افزار است.

۴-۲-۲. بهبود دسترسی به داده: HBase، Sqoop و Flume

هادوپ اساساً یک سیستم دسته گرا است. داده‌ها در HDFS بارگذاری، پردازش و سپس بازیابی می‌شوند. این نوعی محاسبات بازگشتی و اغلب به دسترسی تعاملی و تصادفی به داده‌ها نیاز است.

HBase یک پایگاه داده ستون گرایی است که در بالای HDFS اجرا می‌شود. هدف این پروژه که پس از BigTable گوگل مدل‌سازی شده است، میزبانی میلیاردها سطر داده برای دسترسی سریع است. نگاشت‌کاهش می‌تواند از HBase هم به‌عنوان یک منبع و هم به‌عنوان یک مقصد برای محاسباتش استفاده کند، و از Hive و Pig نیز می‌توان به‌صورت ترکیبی با HBase استفاده نمود.

HBase برای دسترسی تصادفی به داده محدودیت‌هایی را در نظر گرفته است: کارایی Hive با HBase، ۴ تا ۵ مرتبه کندتر از HDFS است و حداکثر حجم داده‌ای که می‌توان در HBase ذخیره نمود، تقریباً در حد پتابایت است درحالی‌که حداکثر حجم ذخیره‌سازی HDFS بیش از ۳۰PB است.

HBase برای تجزیه‌وتحلیل داده‌ها مناسب نیست و بیش‌تر برای ادغام کلان داده به‌عنوان بخشی از یک برنامه بزرگ مناسب هستند. موارد استفاده شامل ورود، شمارش، و ذخیره‌سازی داده‌های سری زمانی هستند.

تاریخچه هادوپ

Ambari	توسعه، پیکربندی و نظارت
Flume	جمع‌آوری و ورود لاگ و داده‌های رویدادی
HBase	تغییر مقیاس پایگاه داده ستون گرا به میلیاردها سطر
HCatalog	اشتراک طرح و نوع داده در Pig، Hive و نگاشت‌کاهش
HDFS	فایل سیستم افزونه توزیع‌شده برای هادوپ
Hive	دسترسی به انبار داده‌ها همانند SQL
Mahout	کتابخانه‌ای از الگوریتم‌های یادگیری ماشین و داده‌کاوی
MapReduce	محاسبات موازی در خوشه‌های سروری
Pig	زبان برنامه‌نویسی سطح بالا برای محاسبات هادوپ
Oozie	تنظیم (ارکستراسیون) و مدیریت گردش کار
Sqoop	ورود داده از پایگاه داده‌های رابطه‌ای
Whirr	استقرار ابری-آگنوستیک خوشه‌ها
Zookeeper	مدیریت پیکربندی و هماهنگ‌سازی

ورود و خروج داده

بهبود قابلیت همکاری با مابقی جهان داده‌ها با Sqoop و Flume میسر گردید. Sqoop یک ابزار طراحی‌شده برای ورود داده از پایگاه داده‌های رابطه‌ای به هادوپ با ورود مستقیم به HDFS یا Hive است. Flume برای ورود مستقیم جریانی از لاگ‌ها به HDFS طراحی‌شده است.

دوستانه بودن SQL در Hive بدین معنی است که می‌توان از آن به‌عنوان یک نقطه ادغام با جهان وسیعی از ابزارهای پایگاه داده که قابلیت برقراری ارتباط از طریق درایورهای پایگاه داده JBDC یا ODBC را دارند، استفاده نمود.

۲-۲-۵. هماهنگی و گردش کار: Zookeeper و Oozie

با رشد مجموعه سرویس‌های در حال اجرا به‌عنوان بخشی از یک خوشه هادوپ، به سرویس‌های هماهنگ‌سازی و نام‌گذاری نیاز است. با توجه به این‌که آمدورفت گره‌های محاسباتی امکان‌پذیر است، اعضای خوشه باید با یکدیگر همگام باشند، باید بدانند کجا به سرویس‌ها دسترسی دارند و نحوه پیکربندی خود را نیز باید بدانند که اهداف Zookeeper است.

سیستم‌های تولیدی که از هادوپ استفاده می‌کنند اغلب دارای خط لوله‌های پیچیده‌ای از تبدیلات وابسته به یکدیگر هستند. به‌عنوان‌مثال، ورود یک دسته جدید از داده‌ها منجر به ورودی‌هایی می‌شوند که سپس منجر به محاسبات جدیدی در مجموعه داده‌های وابسته می‌شوند. مؤلفه Oozie ویژگی‌هایی برای مدیریت جریان کاری و وابستگی‌ها فراهم و نیاز به توسعه‌دهندگان برای کد کردن راه‌حل‌های سفارشی را از بین می‌برد.

۲-۲-۶. مدیریت و استقرار: Ambari و Whirr

یکی از ویژگی‌های رایج افزوده‌شده به هادوپ از طریق توزیع‌کنندگانی مثل IBM و مایکروسافت، نظارت و مدیریت است. در مراحل اولیه، هدف Ambari افزودن این ویژگی‌ها به پروژه اصلی هادوپ است. Ambari برای کمک به مدیران سیستم به‌منظور راه‌اندازی و پیکربندی هادوپ، ارتقاء خوشه‌ها و نظارت بر خدمات در نظر گرفته‌شده است و از طریق API می‌تواند با دیگر ابزارهای مدیریت سیستم ادغام شود.

Whirr به‌عنوان یک بخش نه‌چندان محدود، یک مؤلفه شدیداً مکمل است که روشی برای اجرای سرویس‌ها شامل هادوپ بر روی پلت‌فرم‌های ابری را ارائه می‌نماید. Whirr یک ابر خنثی است و در حال حاضر از سرویس-های Amazon EC۲ و Rackspace پشتیبانی می‌کند.

۲-۲-۷. یادگیری ماشین: Mahout

داده‌های هر سازمانی مختلف و متناسب با نیازهای‌شان است. اما تنوع تحلیل‌های انجام‌شده بر روی این داده‌ها بسیار کم است. پروژه Mahout کتابخانه‌ای از پیاده‌سازی‌های محاسبات تحلیلی رایج توسط هادوپ است و موارد استفاده آن شامل فیلتر مشارکتی کاربر، توصیه‌های کاربری، خوشه‌بندی و دسته‌بندی است.

۲-۲-۸. کاربرد هادوپ

به‌طورمعمول، از هادوپ به شکل توزیع‌شده استفاده می‌شود. همانند لینوکس که قبل از هادوپ بود، فروشندگان مؤلفه‌های اکوسیستم Apache Hadoop را باهم ادغام و ابزارها و ویژگی‌های مدیریتی را در آن گنجاندند.

اگرچه نصب ابری و مدیریت‌شده نگاشت‌کاهش هادوپ به‌خودی‌خود توزیع‌شده نیست، اما در سرویس نگاشت-کاهش کشسان آمازون در دسترس است.

چرا کلان داده، بزرگ است: سیستم عصبی دیجیتال

کل داده‌های موجود در "کلان داده" از کجا آمده‌اند؟ و چرا کلان داده فقط نگرانی شرکت‌هایی مثل فیس‌بوک و گوگل نیست؟ پاسخ این است که شرکت‌های تحت وب پیش‌گام این عرصه هستند. به‌وسیله شبکه‌های اجتماعی،

تلفن همراه و فناوری ابری، یک انتقال مهم رخ‌داده است که همه ما را به سمت جهان داده‌ها سوق داده و این شرکت‌ها امروزه ساکن هستند.

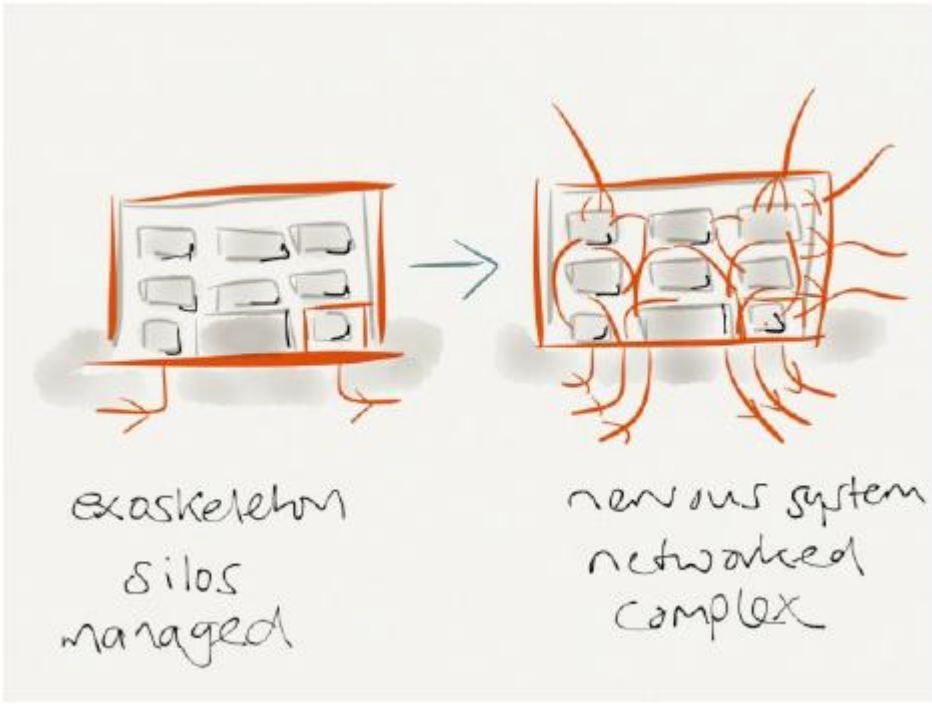
از اسکلت خارجی به یک سیستم عصبی

تا چند سال پیش، کاربرد اصلی سیستم‌های کامپیوتری در جامعه و به‌خصوص تجارت، به‌عنوان یک سیستم پشتیبانی دیجیتال بوده است. برنامه‌های دیجیتالی شده که در دنیای واقعی وجود دارند، شامل برنامه واژه پردازش، برنامه حقوق و دستمزد و برنامه موجودی هستند. این سیستم‌ها از طریق، فروشگاه‌ها، مردم، تلفن، حمل‌ونقل و غیره رابط‌هایی با دنیای واقعی برقرار کردند. عبارت جدید "اداره بدون کاغذ" به انتقال پردازش‌های کاغذی پیشین به کامپیوتر اشاره دارد. این سیستم‌های کامپیوتری تشکیل یک اسکلت دیجیتال دادند و از تجارت در جهان واقعی پشتیبانی می‌کنند.

ورود اینترنت و وب، منجر به افزودن ابعاد جدیدی شده و عصر تجارت کاملاً دیجیتال را به ارمغان آورده است. تعامل با کاربران، پرادخت‌ها و اغلب تحویل محصول توسط سیستم‌های کامپیوتری به‌طور کامل قابل انجام است. داده فقط درون اسکلت باقی نمی‌ماند، بلکه یکی از عناصر کلیدی در عملیات است. در دورانی هستیم که تجارت و جامعه به یک سیستم عصبی دیجیتال دستیابی دارند.

همان‌طور که طرح زیر نشان می‌دهد، یک سازمان دارای سیستم عصبی دیجیتال با تعداد زیادی از جریان‌های ورودی و خروجی داده و یک شبکه سطح بالا مشخص می‌شود که هردو به‌صورت داخلی و خارجی منجر به افزایش جریان داده و درنتیجه افزایش پیچیدگی می‌شوند.

این انتقال نماد مهم بودن کلان داده است. تکنیک‌های ارائه‌شده برای مقابله با اطلاعات به‌هم‌پیوسته و ناهمگن به‌دست‌آمده از شرکت‌های تحت وب گسترده، ابزارهای اصلی ما برای انتقال به عملیات دیجیتال بومی خواهند بود. نمونه‌های اولیه این فناوری شامل کشف تقلب در معاملات مالی و اشکال‌زدایی و بهبود روند استخدام در HR است: و در حال حاضر تقریباً هر کسی به جریان عظیم اطلاعات در شبکه‌های اجتماعی توجه دارد.



نمودار انتقال

با گسترش فناوری در تجارت، هر قدم برداشته‌شده منجر به جهشی در حجم داده‌شده است. سؤال منطقی افراد درگیر با کلان داده این است که بپرسند چرا زمانی که تجارت آن‌ها گوگل یا فیس‌بوک نیست، کلان داده به آن‌ها اعمال می‌شود؟

پاسخ این سؤال به توانایی تجارت‌های اینترنتی در انجام ۱۰۰ درصد آنلاین فعالیت‌های‌شان بستگی دارد. سیستم عصبی دیجیتال آن‌ها به‌راحتی از ابتدا تا انتهای عملیات گسترش می‌یابد. اگر کارخانه، فروشگاه یا بخش‌های دیگری از جهان واقعی را در تجارت خود دارید، بیش‌تر درگیر ترکیب آن‌ها با سیستم عصبی دیجیتال هستید.

اما "بیش‌تر" بدین معنی نیست که اتفاق نخواهد افتاد. موتور وب، رسانه‌های اجتماعی، تلفن همراه و ابر تجارت-های بیش‌تری را به جهان مبتنی بر داده هدایت می‌کند. در انگلستان، سرویس دیجیتال دولتی تحویل سرویس به شهروندان را یکپارچه‌سازی می‌کند. نتایج شامل بهبود تجربیات شهروندان است و برای اولین بار بسیاری از ادارات توانستند تصویر واقعی نحوه انجام آن‌ها را مشاهده کنند. در هر خرده‌فروشی، شرکت‌هایی نظیر Square، American