ششمین همایش بین‌المللی افق های نویـن در
مهندسی برق، کامپیوتـر و مکانیک
6th International Conference on the New Horizons in
Electrical Engineering, Computer and Mechanical
www.mhconf.ir

نرم افزارهای کار با کلان داده

محمدصالح احمدی[۱]، روزیتا جمیلی اسکوئی*[۲]

[۱]دانشکده مهندسی کامپیوتر و فناوری اطلاعات، دانشگاه آزاد اسلامی، واحد الکترونیکی، تهران، ایران، ahmadi.saleh۹٤@gmail.com

[۲] دانشکده مهندسی کامپیوتر و فناوری اطلاعات، دانشگاه آزاد اسلامی، واحد الکترونیکی، تهران، ایران، * rozita۲۰۲۰j@gmail.com

چکیده

امروزه کلان داده عبارات ناآشنایی در صنعت نیست از آنجایی که تعداد خیلی زیادی رکورد در صنعت ها وجود دارد اما میزان این زیاد بودن به راحتی قابل تصور نیست. کلان داده عصری از پردازش داده است. کار با کلان داده ها با سیستم های سنتی مدل رابطه داده بسیار سخت است و اغلب گرفتن خروجی مورد نظر در بازه زمانی مد نظر امکان پذیر نیست. این چالش ها نظیر ذخیره سازی داده، بررسی داده، تمیز کردن و اصلاح داده، اشتراک گذاری، انتقال، مخابره، تصویر سازی و مسائل امنیتی آن می باشد. برای مثالی از داده های زیاد، پتابایت ها که برابر ۱۰۲۴ ترابایت هست یا اگزابایت ها که برابر ۱۰۲۴ پتابایت است شامل میلیاردها میلیارد اطلاعات میلیون ها نفر در سرتاسر کره زمین است که این داده ها در شکل ها و فرمت های مختلفی وجود دارند. هدف این مقاله معرفی ابزارهای مختلف کار با کلان داده[۱] از جمله هدوپ[۲]، مقایسه کلی و بررسی میزان موفقیت، مزایا و معایب آن ها می باشد.

واژه‌های کلیدی

کلان داده، هدوپ، کاهش‌نگاشت، نواس‌کیوال، داده توزیع شده

۱. مقدمه

در دنیایی که اندازه داده امر مهمی است، عبارت کلان داده ارزشمند محسوب میشود. اطلاعات خصوصا در شرکت های بزرگ از ارزش مهم و حیاتی برخوردار است. نتایج بهینه سازی های انجام شده در این زمینه انقدر مهم و حیاتی بود که باعث شد شرکت های بزرگ نرم افزاری به فکر پیاده کردن این روش ها و تحقیق در رابطه با آنها بیوفتند.[۳] داده ها از منابع متنوعی تولید میشوند[۴] و سرعت پیشرفت تکنولوژی[۵] باعث افزایش مقدار داده شده است[۶]. این داده ها میتوانند به صورت ساختار یافته، نیمه ساختار یافته یا غیر ساختار یافته باشند. همچنین نیاز برای تبدیل این کلان داده به هوش تجاری که شرکت ها توانایی استفاده از آن را داشته باشند نیز وجود دارد. داده های بهتر منجر به تصمیم گیری بهتر و چینش استراتژی بهینه تر برای سازمان ها فارق از سایز آنها، جغرافیای محلی که در آن قرار دارند، سهم آن ها از بازار، میزان رضایت مشتری، طبقه بندی مشتری و دسته بندی های دیگر میشود. نتایج بررسی های مختلف آماری در این زمینه نشان میدهد مدیریت این حجم از داده با ابزارهای قبل از دوران کلان داده بسیار سخت و طاقت فرسا است.[۷] ابزارهای مختلفی برای برای آنالیز داده وجود دارد. هدوپ یکی از ابزارهایی است که برای کار با مقادیر بسیار زیاد داده مورد استفاده قرار میگیرد و الگوریتم کاهش‌نگاشت[۸] از پرکاربردترین الگوریتم ها در این زمینه است. کلان داده در شش زمینه طبقه بندی و مورد بررسی قرار میگیرد که هر کدام یکی از موارد جمع آوری داده، مانیتورینگ، ذخیره سازی، آنالیز و گزارش گیری را تحت تاثیر قرار میدهند. این شش دسته برابر با : حجم – شتاب – تنوع – صحت – اعتبار – ارزش میباشند. این مقاله به پنج بخش تقسیم شده است که شامل مقدمه، پیشینه تحقیق، کلان داده، محاسبه و آنالیز کلان داده و نتیجه گیری میشود.

شکل ۱. شش دسته کلان داده

۲. پیشینه تحقیق

Elgendy و همکاران[۹] عنوان کردند که در عصر اطلاعات، مقدار بسیار زیادی از داده های برای تصمیم گیری وجود دارد. آنها ارزش تحلیل آماری چنین داده هایی را در عصری که دانش در حال سرریز کردن می باشد را شدیدا بالا میدانستند. این بینش های اکتسابی از داده میتواند باعث ایجاد برتری برای سازندگانی که از این اطلاعات استفاده میکنند شود.

Inukollu و همکاران[۱۰] بیان کردند که نرم افزارهای کلان داده، سودمندی بالایی برای سازمان ها، کسب و کارها، شرکت ها و صنعت هایی با سایزهای مختلف دارد. آنها همچنین به بررسی راهکارهای امن ارتباطی بین رایانش ابری و هدوپ نیز پرداختند. بخش امنیت

رایانش ابری به سرعت در حال گسترش می باشد که شامل امنیت وسایل قابل حمل، امنیت جامعه ابری، امنیت رکوردها و محرمانگی فایل ها می باشد.

Suthaharan و همکاران[۱۱] مسائل امنیتی رایانش ابری، الگوریتم نگاشت‌کاهش و اتمسفر هدوپ را ذکر و بررسی کردند و پیشنهاد استفاده از تکنولوژی های ابری و همچنین استفاده از HDFS همراه با راهکارهای آن را ارائه کردند و به بررسی میزان بهره وری از اطلاعات دریافتی پرداختند.

Ali و همکاران[۱۲] به بیان الگوریتم ها و تکنیک های موثر برای آنالیز مقدار زیاد داده پرداختند. تضمین جایگزینی سرویس های درمانی و آماده سازی محصولات کشاورزی با استفاده از BD٤D برای توسعه ظرفیت استفاده از داده های زیاد را دادند.

Borodo و همکاران[۱۳] نتیجه گیری کردند که داده درحال رشد با نرخ بی سابقه ای است و که این امر باعث تولید این میزان از داده شده است. داده های عظیم چالش هایی نظر ذخیره سازی، محاسبه، امنیت و محرمانگی را به همراه دارد. تکنیک های آماری تضمین ارائه بینش عمیق به اطلاعات ورودی را میدهد که میتواند در زمین های مختلفی نظیر فروش زیورآلات و سرمایه گذاری های مالی استفاده شود.

۳. کلان داده:

برای تعریف کلی برای کلان داده تا به اینجای کار میتوان گفت: کلان داده معمولا شامل مجموعه داده هایی دارای سایزی فراتر از توانایی[۱۴] نرم افزارهای معمول برای ذخیره، نگه داری، مدیریت و پردازش داده در بازه زمانی قابل قبول می باشد.[۱۵]

۱.۳. ویژگی های کلان داده

۱.۱.۳. تنوع:

اطلاعات شما در فرمت های متنوعی موجود است که هر شامل پتانسیل میزان متنوعی از بررسی های آماری را دارا می باشند. برای مثال فروشگاه ها تنوع داده ای آنها تشکیل شده از ایمیل ها، اسناد پردازش سفارش، تصاویر کالا ها، ویدئو ها و همچنین رکوردهایی که در پایگاه داده های رابطه ای وجود دارد می باشد. تنوع در کلان داده اشاره به نوع های مختلف داده نظیر pdf, doc و mp۳ دارد. انواع مختلف داده بصورت ساختار یافته، نیمه ساختار یافته و ساختار نیافته[۱۶] می باشند.

۲.۱.۳. ارزش:

کلان داده ها در داده های زیاد شامل اطلاعاتی از سازمان مربوطه میشوند که بهینه سازی آن ها و استفاده از آنها در محاسبات و تبلیغات باعث فراهم آوری ارزش برای سازمان مربوطه میشوند. بنابراین با توجه به ارزش داده های موجود میتوان استراتژی های اطلاعاتی مناسبی انتخاب کرد.

۳.۱.۳. صحت:

هنگام سر و کار داشتن با حجم بالا، شتاب و تنوع داده های در میزان بالا، صحت به میزان درستی و حقیقی بودن داده ها اشاره دارد.

۴.۱.۳. اعتبار:

میزان ناهماهنگی های درون داده نیاز است که با تکنیک های شناخت ناهنجاری شناسایی شوند تا تاثیر شدیدی بر روی خروجی نتیجه گیری نداشته باشند. همچنین تعداد رکورد های بالا، تنوع آماری بالایی با توجه به نوع آنها دارند که باعث میشوند نتایج گسیخته آماری مختلفی بدست آید. این تنوع همچنین میتواند با سرعت ورود داده به سیستم نیز ادغام شود.

۵.۱.۳. شتاب:

داده ها در سرعت های بالایی به وجود می آیند. این بخش با مدیریت زمان در ارتباط مستقیم است. داده ها در سیستم های رایانه بزرگ، مدل های کاربر-سرور، اینترنت، موبایل و زیرساخت ابری مورد استفاده قرار میگیرند.[۱۷] در صورتی که داده های با نرخ سریعی تولید میشوند و در زمان مناسب پاسخ دریافت میکنید، بخش مربوط به شتاب کلان داده ها را بدست آورده اید.

در جدول زیر به بررسی این موارد در بخش های مختلف کلان داده می پردازیم:

جدول ۱. چالش های مختلف کلان داده در زمینه های مختلف

چالش ها	حجم	شتاب	تنوع	صحت	ارزش
محرمانگی داده	بله	بله	بله	خیر	خیر
امنیت داده	بله	بله	بله	خیر	خیر
ذخیره سازی داده	خیر	بله	بله	خیر	بله
پردازش داده	بله	بله	بله	خیر	بله
ذخیره داده و آنالیز	خیر	بله	بله	خیر	بله
کسب دانش و پیچیدگی محاسباتی	خیر	خیر	بله	بله	خیر
مقیاس پذیری و تصویر سازی داده	بله	بله	بله	خیر	بله
رشد داده	بله	بله	بله	خیر	خیر
تولید بینش در زمان مناسب	بله	خیر	خیر	بلخ	خیر
جامعیت منابع مختلف داده	خیر	خیر	بله	خیر	خیر
صحت داده	بله	بله	بله	بله	خیر
امنیت کلان داده	بله	بله	بله	خیر	خیر
مقاومت سازمانی	بله	بله	بله	خیر	خیر

۲.۳. کلان داده – دسته ها

کلان داده شامل داده هایی که توسط دستگاه ها و برنامه های مختلف تولید شده میشود. بعضی از این زمینه ها در ادامه بیان شده اند.

۱.۲.۳. داده های شبکه های اجتماعی:

شبکه های اجتماعی[۱۸] نظیر فیس بوک، توییتر و غیره می باشد[۱۹]. این شبکه های اطلاعات مختلف، پیشنهادات، دعوت نامه ها و غیره که توسط مردم زیاد در سرتاسر جهان تولید شده را شامل میشوند. کمپین ها، تبلیغات و غیره نیز از طریق این شبکه ها مخابره میشود.

۲.۲.۳. داده های موتور جستجو:

موتورهای جستجوی داده های زیادی را از پایگاه داده های مختلف واکشی میکند.

۳.۲.۳. داده های سوابق درمانی:

بیمارستان اطلاعات سوابق مختلف درمانی بیماران مختلف را میتوانند تولید کنند.

۴.۲.۳. داده های خرید های اینترنتی:

خرید محصولات مختلف آنلاین میتواند به شناخت ترجیحات و نظر کاربران برای کالای مختلف در بازه های زمانی مختلف منجر شود.

۵.۲.۳. داده های معامله های بورس:

اطلاعات معامله های بورسی شامل اطلاعاتی نظیر میزان سهم شرکت های مختلف میشود. این داده ها باعث ارائه بینش مناسب به خریداران سهم برای گرفتن تصمیمات مهم برای فعالیت در زمینه داد و ستد میشود.

۶.۲.۳. اطلاعات رزرو وسایل نقلیه:

رزرو وسایل نقلیه ای نظیر قطار، اتوبوس، هواپیما، تاکسی و غیره میتواند باعث تولید داده های مخصوص رزرو شود که بر مبنای مدل، سایز، فاصله و میزان در دسترس پذیری یک وسیله نقلیه است.

۷.۲.۳ داده های هوانوردی:

اطلاعات ضبط شده صوتی و تصویری، اطلاعات میزان عملکرد هوانوردها و غیره در زمره این اطلاعات قرار میگیرد.

۴. محاسبه و آنالیز کلان داده:

یکی از شناخته شده ترین متد ها برای تبدیل داده خام به اطلاعات قابل استفاده الگوریتم نگاشت‌کاهش نام دارد. این روش متدی است برای اینکه مقدار بزرگی از مجموعه داده ها را دریافت و عملیات محاسباتی بر روی آنها بین چند کامپیوتر بصورت موازی پردازش انجام دهد. معمولا به عنوان مدلی برای چگونگی برنامه استفاده میشود و به آن بعنوان پیاده سازی واقعی این مدل ارجاع داده میشود. انجام عمل پردازش به دلیل استفاده صحیح از کاربرد داده در این تکنیک سخت است. در اصل نگاشت‌کاهش از دو بخش تشکیل شده است. بخش تابع کاهش که خلاصه ای از داده ها بوسیله تجمیع آن ارائه میدهد. با توجه به تحقیق های وسیعی که اغلب توسط گوگل انجام شده است، عبارت نگاشت‌کاهش اکنون یک عبارت عمومی است که و بعنوان یک مدل عمومی توسط تکنولوژی های زیاد استفاده میشود.

سیستم با پردازش داده برای نمایش اطلاعات اصلی میتواند کار خود را آغاز کند. لایه محاسبه شاید متنوع ترین لایه سیستم باشد زیرا با توجه به خروجی مد نظر نیاز ها و به‌روش های رسیدن به آن میتواند تفاوت های زیادی داشته باشند. داده ها اغلب بصورت تکراری و متناوب پردازش میشوند. این تکرار توسط یک ابزار بصورت متناوب و یا توسط دسته ای از ابزارها برای دریافت خروجی های مختلف اطلاعاتی انجام میشود.

پردازش دسته ای یکی از متد های محاسبه بر روی مقدار زیادی مجموعه داده ای است. این فرآیند شامل شکستن کار به دسته های کوچک تر، تخصیص هر کدام از این دسته ها به یک ماشین مجزا، بهم آمیختن داده ها بر مبنای نتایج میانی و پس محاسبه و سرهم کردن نتایج نهایی است. این گام های معمولا بصورت مجزا با نام های جداسازی، نگاشت، بهم آمیختن، کاهش و سرهم کردن است. همچنین بصورت جامع هم به آنها به عنوان الگوریتم توزیع شده نگاشت‌کاهش هم نام مینهند. این استراتژیی است که توسط نگاشت‌کاهش در آپاچی هدوپ نیز استفاده شده است. پردازش دسته ای زمانی که با مجموعه داده های عظیم که نیازمند مقداری پردازش هستند بهترین کاربرد را دارد.

با اینکه پردازش دسته ای انتخاب مناسبی برای نوع های مختلف داده ای و محاسباتی است، حجم های دیگر کاری نیاز به پردازش لحظه ای بیشتری دارند. پردازش بلادرنگ نیازمند این است که اطلاعات به سرعت پردازش و آماده شوند و سیستم به محض دریافت اطلاعات جدید واکنش نشان دهد و در دسترس باشد. یکی از روش های دسترسی به چنین موردی پردازش جریان داده است که از آیتم های جداگانه تشکیل شده است. یکی دیگر از خصوصیت های پردازش بلادرنگ، محاسبه در حافظه است که با استفاده از داشتن اطلاعات در حافظه خوشه انجام میشود که از نوشتن آن بر روی دیسک جلوگیری کند.

آپاچی فلینک، آپاچی استورم و آپاچی اسپارک روش های مختلفی برای بدست آوردن پردازش بلادرنگ یا نزدیک به بلادرنگ دارند. هر کدام از این تکنولوژی ها، کاستی های خود را دارند که باعث میشود انتخاب بهترین روش برای حل مساله تحت تاثیر قرار بگیرد. بصورت کلی، پردازش بلادرنگ بیشتر برای آنالیز دسته های کوچکتر داده که به سرعت در حال تغییر یا افزوده شدن به سیستم هستند مناسب است.

مثال های بالا نشانگر چهارچوب های محاسباتی هستند. هرچند که روش های دیگری نیز برای آنالیز داده درون سیستم کلان داده نیز وجود دارند. این ابزارها معمولا به چهارچوب های مذکور افزوده میشوند تا خروجی های بیشتری برای کار با لایه های زیرین ارائه شود. برای مثال آپاچی هایو یک رابط انبار داده ای ارائه میدهد. آپاچی پیگ رابط پرسشگر سطح بالایی ارائه میدهد در حالی که ارتباطاتی شبیه به اس‌کیوال با داده میتواند توسط آپاچی دریل، آپاچی ایمپالا، آپاچی اسپارک اس‌کیوال و پِرِستُو انجام شود. برای یادگیری ماشین پروژه هایی نظیر آپاچی ماهاوت، آپاچی سیستم‌امآی و آپاچی اسپارک املیب میتوانند کاربردی باشند. برای برنامه نویسی آماری و آنالیز مستقیم نیز چهارچوب ها و برنامه های وسیعی در اکو سیستم کلان داده وجود دارند که پایتون و آر گزینه های مناسبی هستند.

۱.۴. ابزارهای استفاده شده در آنالیز کلان داده:

هدوپ به ۴ بخش اصلی تقسیم میشود:

- فایل سیستمی توزیع شده هدوپ (HDFS) که یک فایل سیستمی توزیع شده است که برای پهنای باند تجمیع شده مقدار بالا طراحی شده.

- یارن که بستری است برای مدیریت منابع هدوپ و برنامه ریزی زمانی برای برنامه هایی که بر روی زیرساخت هدوپ اجرا میشوند.

- کاهش‌نگاشت که همانگونه که قبلا توضیح داده شد مدلی است برای پردازش کلان داده.

- یک دسته کتابخانه اشتراکی برای استفاده ماژول های دیگر.

۲.۴. ابزار های دیگر کلان داده:

تعداد بیشماری ابزار متن باز برای کار با کلان داده وجود دارد که اکثر آنها برای ارائه توانایی های خاص پردازشی بهینه سازی مخصوصی برای دستگاه های سخت افزاری با مشخصات مشخص شده اند. بنیاد نرم افزار آپاچی (ASF) تعداد زیادی از این پروژه ها را پشتیبانی میکند. در زیر لیستی از این موارد را خواهیم داشت:

- آپاچی بیم، یک مدل یکپارچه برای تعریف خطلوله های پردازش موازی داده، هم دسته ای و هم جریانی می باشد. همچنین به توسعه دهندگان اجازه میدهند که یکبار کد بنویسند ولی بر روی موتورهای پردازش متفاوت کار کند.

- آپاچی هایو، یک انباره داده است بر روی هدوپ ساخته شده است. پروژه ای سطح بالا از آپاچی است که کار خواندن، نوشتند و مدیریت مجموعه داده های بزرگ را با استفاده از اس‌کیوال را بهبود میبخشد.

- آپاچی ایمپالا، یک موتور پرسشگر اس‌کیوال است که بر روی هدوپ اجرا میشود که در تعریف آن ذکر شده است که باعث بهبود سطح اجرای پرس و جو های اس‌کیوال همزمان با ارائه رابط کاربری آشنا میشود.

- آپاچی کافکا، به کاربران اجازه میدهد که خبررسان های بلادرنگ داده را تولید و اخبار آن را دریافت کنند. هدف آن افزودن اتکاپذیری سیستم های ارسال پیام برای داده های جریانی است.

- آپاچی لوسین، کتابخانه نرم افزاری جستجو و شاخص گذاری تمام متنی است که میتواند برای موتور های توصیه گر استفاده شود. همچنین پایه پروژه های جستجوی زیاد دیگری نظیر سول‌آر و الستیک‌سرچ می باشد.

- آپاچی پیگ، بستری برای آنالیز مجموعه داده های بزرگ می باشد که بر روی هدوپ اجرا میشود. یاهو آن را برای انجام وظایف کاهش‌نگاشت بر روی مجموعه داده های بزرگ در سال ۲۰۰۷ برای ASF توسعه داد.

- آپاچی سول‌آر، بستر جستجو سازمانی میباشد که بر مبنای لوسین که بالاتر ذکر شد، ساخته شده است.

- آپاچی جیراف، سیستم پردازش گراف های تکرار شونده می باشد.

- وبسایت Algorithms.io آنالیزهای پیش بینی کننده برای جریان داده تولید شده توسط ماشین ارائه میدهد.

- آپاچی زپلین، پروژه آنالیز تعاملی است که با استفاده از اس‌کیوال و دیگر زبان های برنامه نویسی اجرا میشود.

دسته بندی بسترهای مختلف کلان داده را در زیر مشاهده میکنید:

جدول ۲. بسترهای کلان داده متن باز

بسترها	توضیحات	اهمیت
آپاچی هدوپ	فایل سیستم	هدوپ بر روی ویندوز اجرا میشود
کاهش‌نگاشت	فارق از وابستگی به بستر اجرایی	مقدار زیاد داده را میتواند پردازش کند
گرید گِین	کاهش‌نگاشت	با HDFS هدوپ مطابقت دارد
سیستم های HPCC	توسط لکسیس نکسیس توسعه داده شده است	از هدوپ بسیار سریع تر است
آپاچی استورم	موتور پردازش موازی بلادرنگ	تقریبا تمام زبان های برنامه نویسی را پشتیبانی میکند
آپاچی اسپارک	ذخیره سازی خافظه	سریع تر از کاهش‌نگاشت
سامُوا	سیستم استخراج اطلاعات	به عنوان افزونه برای چهارچوب های الحاقی سیستم های توزیع شده

پروژه های دیگر ابزار کلان داده نیز به اختصار در زیر ذکر شده است:

- الستیک‌سرچ، که موتور جستجوی در سطح سازمانی است که بر مبنای لوسین ساخته شده است. بخشی از پشته الستیک است که قبلا با نام ای‌ال‌کی شناخته میشد که علت نام گذاری آن، مولفه های تشکیل دهنده آن بودند که شامل کیبانا، الستیک‌سرچ و لاگ‌اِستَش بودند. این ابزار از داده های ساخت یافته و نیافته استفاده میکند.

- کروز کنترل، که توسط لینکداین توسعه داده شد تا خوشه آپاچی کافکا را در مقیاس بزرگ اجرا کند.

- تنسرفلو، یک کتابخانه نرم افزاری برای یادگیری ماشین است که از زمانی که گوگل در اواخر سال ۲۰۱۵ آن را متن باز کرد به سرعت گسترش پیدا کرد. یکی از علل ستایش آن این بود که یادگیری ماشین را به دلیل سادگی در استفاده آن برای همه آسان کرد.

در زیر طبقه بندی تکنولوژی های کلان داده را میبینیم:

جدول ۳. تکنولوژی های مختلف کلان داده

تکنولوژی/دسته بندی	هدوپ	اسپراک	فلینک
پردازش نوع دسته	دسته کوچک	دسته جریانی	دسته
ذخیره سازی عمومی	HDFS	بدون ذخیره اصلی	بدون ذخیره اصلی
آنالیز	SQL,ML,ETL,GRAPH	ML,CEP,SQL,GRAPH	گراف پرس و جو گر اِکس

همانگونه که اهمیت و سایز کلان داده افزایش پیدا میکند، لیست ابزار های متن باز کار با آن نیز مسلما افزایش می یابد.

۳.۴. تصویر سازی نتایج:

با توجه به نوع اطلاعاتی که در سیستم های کلان داده پردازش میشود، شناخت روندها و تغییرات در آنها با گذر زمان معمولا مهم تر از خود مقدار داده ها است. تصویر سازی داده ها معمولا کاربردی ترین روش شناخت روندها و شناختن معانی در حجم های زیاد داده است.

پردازش بلادرنگ اغلب برای تصویرسازی برنامه های وضعیت های سرور استفاده میشود. داده ها اغلب تغییر میکند و موج های بزرگ در وضعیت ها معمولا نشانگر تاثیر عمده بر سلامت سیستم ها یا سازمان ها دارد. در این موارد، پروژه هایی نظیر پرومتئوس میتوانند برای پردازش جریان های داده بعنوان سری های زمانی پایگاه داده و تصویر سازی آن، اطلاعات سودمند باشد.

یکی از روش های معروف تصویر سازی داده با استفاده از پشته الستیک می باشد که پیشتر با عنوان پشته ای‌ال‌کی شناخته میشد. این پشته شامل الستیک‌سرچ برای شاخص گذاری داده، کیبانا برای تصویر سازی و لاگ‌اِستَش میشود که میتواند برای سیستم های کلان داده رابط تصویری ایجاد کند تا نتایج محاسبات یا نتایج خام را به تصویر کشد. پشته ای شبیه به این را میتوان با آپاچی سول‌آر برای شاخص گذاری داده و کپی کیبانا با نام بانانا برای تصویر سازی بدست آورد که این پشته سیلک نام دارد.

تکنولوژی تصویرسازی دیگری که معمولا برای علوم داده های تعاملی مورد استفاده قرار میگیرد نوت بوک داده نام دارد. این پروژه ها اجازه شناسایی و تصویر سازی تعاملی برای داده ها که مناسب به اشتراک گذاری، نمایش یا همکاری باشد را میدهد. از مثال های معروف در این زمینه میتوان به ژوپیتر نوت‌بوک و آپاچی زپلین اشاره کرد.

۴.۴. واژه نامه کلان داده:

کلان داده: یک عبارت چتری می باشد که برای مجموعه داده هایی که امکان مدیریت معقول آن ها به دلیل حجم، شتاب و تنوع آنها توسط کامپیوترها یا ابزارها سنتی وجود ندارد بکار میرود. این عبارت همچنین به تکنولوژی ها و استراتژی های کار با این نوع داده نیز اطلاق میشود.

پردازش دسته ای: پردازش دسته ای یک استراتژی محاسباتی است که شامل پردازش داده در دسته های بزرگ است. این امر معمولا برای کارهایی که حساسیت زمانی ندارند و بر روی ست های داده ای عظیم کار میکنند، ایده آل است . پردازش شروع میشود و نتایج بعدا توسط سیستم ارائه میشوند.

رایانش خوشه ای: رایانش خوشه ای شامل واکشی منابع از ماشین های متعدد و استفاده از توانایی های تجمیعی آنها برای انجام کارها میشود. کامپیوترهای خوشه ای نیازمند لایه مدیریت خوشه هستند که ارتباطات بین نودها مجزا را مدیریت و تخصیص کار بین آنها را هماهنگ میکند.

دریاچه داده: دریاچه داده عبارتی است که برای مخازن بزرگ داده های تجمیع شده به صورت وضعیت تقریبا خام بکار برده میشود. این عبارت معمولا اشاره به داده های جمع آوری شده در سیستم های کلان داده دارد که ممکن است بصورت ساختارنیافته و مکررا متغیر باشد. این امر با انبارهای داده متفاوت است که در ادامه به آن میپردازیم.

داده کاوی: داده کاوی عبارت گستره ای برای تلاش برای یافتن الگوها در میزان مجموعه های بزرگ داده می باشد. پردازشی که برای پالودن حجم زیادی از داده به اطلاعات قابل فهم و یکپارچه اطلاق میشود.

انبارهای داده: انبارهای داده مخازن بزرگ و مرتب شده داده ای هستند که میتوان از آنها برای آنالیز و گزارش گیری استفاده کرد. برخلاف دریاچه داده، انبار داده از داده های تمیز شده، تجمیع شده با دیگر منابع که همچنین عموما مرتب سازی مناسبی شده اند تشکیل شده است. انبارهای داده معمولا در رابطه با ارتباط با کلان داده صحبت میشوند اما اغلب جزو مولفه های سیستم های مرسوم هستند.

ETL: این عبارت مخفف استخراج، تبدیل و بارگزاری است و به عمل دریافت داده خام و تبدیل آن برای استفاده سیستم اطلاق میشود. این عمل بصورت سنتی با پردازش های در رابطه با انبارهای داده مرتبط بود اما ویژگی های امروزه آن در پردازش های خطلوله های سیستم های کلان داده نیز دیده میشود.

هدوپ: یک پروژه آپاچی است که از اولین موفقیت های متن باز در زمینه کلان داده بود. از فایل های سیستمی توزیع شده با نام HDFS تشکیل شده است که همراه با مدیریت خوشه ای و زمانبندی منابع با نام YARN (که به اختصار به معنای یک مذاکره گر منابع دیگر) می باشد. توانایی پردازش دسته ای توسط موتور محاسباتی کاهش‌نگاشت ارائه میشود. سیستم های دیگر آنالیز و محاسبه نیز میتوانند همراه با هدوپ و در کنار کاهش‌نگاشت عمل کنند.

رایانش درون-حافظه ای: رایانش درون حافظه ای استراتژیی است که شامل جابجایی کامل مجموعه داده های عملیاتی بصورت کامل درون حافظه تجمیعی خوشه میشود. محاسبات میانی بر روی دیسک نوشته نمیشوند و در حافظه نگهداری میشوند. این امر مزیت شدیدا بالایی برای سیستم هایی نظیر آپاچی اسپارک در زمینه سرعت ورودی/خروجی سیستم در مقایسه با کاهش‌نگاشت از هدوپ فراهم میکند.

یادگیری ماشین: یادگیری ماشین مطالعه و تمرین طراحی سیستم هایی است که میتوانند بر مبنای داده های ورودی یاد بگیرند، خود را تطبیق دهند و بهبود ببخشند. این سیستم اغلب با پیاده سازی الگوریتم های آماری و پیش بینی کننده ایجاد میشود که این موارد میتوانند رفتار صحیح و بینش مناسب را با توجه به داده های جدید دریافتی انجام دهند.

الگوریتم نگاشت‌کاهش: این الگوریتم کلان داده (و نه موتور محاسباتی هدوپ) الگوریتمی است برای برنامه ریزی کارها بر روی خوشه های محاسباتی. این روند شامل تقسیم مساله و نگاشت آن بر روی نودهای دیگر و انجام محاسبه بر روی آنها برای ایجاد نتایج میانی و دسته بندی نتایج شبیه مجموعه ها و سپس کاهش نتایج به یک نتیجه به ازای هر مجموعه میشود[۲۰].

نواس‌کیوال: عبارت گسترده ای است برای اشاره به پایگاه داده هایی که خارج از سیستم سنتی مدل رابطه ای هستند. این پایگاه های داده ضعف های دیگری نسبت به مدل های رابطه ای دارند اما با توجه به معماری متمرکز بر توزیع و انعطاف پذیری آنها اغلب مناسب کلان داده ها هستند.

پردازش جریان: تمرین انجام محاسبه بر روی آیتم های داده مجزا زمانی که در حال عبور از یک سیستم می باشند. این مورد اجازه آنالیز لحظه ای داده هایی که به سیستم داده شده اند را میدهد و برای عملیات های حساس زمانی که شتاب بالایی دارند مناسب می باشد.

۵. نتیجه گیری:

کلان داده عبارت وسیعی است که به سرعت در حالت گسترش است. با اینکه این امر برای همه نوع محاسبات مناسب نیست اما سازمان ها توجه خود را برای بعضی از کارهای خود متوجه کلان داده کرده اند و از آن برای آنالیز داده های موجود و کسب و کار خود استفاده میکنند. سیستم های کلان داده منحصرا مناسب پیدا کردن الگوهایی که شناسایی آن ها در داده های سخت است استفاده میشوند و بینش مناسبی را که به دست آوردن آن با روش های سنتی غیرقابل اکتساب است را ارائه میدهند. با پیاده سازی صحیح کلان داده سازمان ها میتوانند ارزش بالایی از داده هایی که برایشان فراهم است کسب کنند.

منابع

[۱] Lynch CJN. How do your data grow? ۲۰۰۸;٤٥٥(۷۲۰۹):.۹-۲۸
[۲] Kakhani MK, Kakhani S, Biradar SJIJoAoIiE, Management. Research issues in big data analytics. ۲۰۱٥;۲(۸):.۳۲-۲۲۸
[۳] Oussous A, Benjelloun F-Z, Lahcen AA, Belfkih SJJoKSU-C, Sciences I. Big Data technologies: A survey. ۲۰۱۸;۳۰(٤):.٤۸-٤۳۱
[٤] Jin X, Wah BW, Cheng X, Wang YJBDR. Significance and challenges of big data research. ۲۰۱٥;۲(۲):.٦٤-٥۹
[٥] Acharjya DP, Ahmed KJIJoACS, Applications. A survey on big data analytics: challenges, open research issues and tools. ۲۰۱٦;۷(۲):.۸-٥۱۱
[٦] Chen CP, Zhang C-YJIs. Data-intensive applications, challenges, techniques and technologies: A survey on Big Data. ۲۰۱٤;.٤۷-۲۷٥:۳۱٤
[۷] Singh H, Bhargava S, Ganeshan S, Kaur R, Sethi T, Sharma M, et al. Big data analysis of traditional knowledge-based Ayurveda medicine. ۲۰۱۸;۳(٥):e.۰۰۲۰
[۸] Gandomi A, Haider MJIjoim. Beyond the hype: Big data concepts, methods, and analytics. ۲۰۱٥;۳٥(۲):.٤٤-۱۳۷
[۹] Elgendy N, Elragal A, editors. Big data analytics: a literature review paper. Industrial conference on data mining; ۲۰۱٤: Springer.
[۱۰] Inukollu VN, Arsi S, Ravuri SRJIJoNS, Applications I. Security issues associated with big data in cloud computing. ۲۰۱٤;٦(۳):.٤٥
[۱۱] Suthaharan SJER. Big data classification. ACM SIGMETRICS Perform. ۲۰۱٤;.۳-٤۱:۷۰
[۱۲] Ali A, Qadir J, ur Rasool R, Sathiaseelan A, Zwitter A, Crowcroft JJBDA. Big data for development: applications and techniques. ۲۰۱٦;۱(۱):.۲٤-۱
[۱۳] Borodo SM, Shamsuddin SM, Hasan SJIJoEE, Science C. Big data platforms and techniques. ۲۰۱٦;۱(۱):.۲۰۰-۱۹۱
[۱٤] Ikhlaq S, Keswani BJIJoAER. A comparative study of big data computational approaches. ۲۰۱۷;۱۲(۱۹):.٦-۸۱۳۱
[۱٥] Katal A, Wazid M, Goudar RH, editors. Big data: issues, challenges, tools and good practices. ۲۰۱۳ Sixth international conference on contemporary computing (IC۳); ۲۰۱۳: IEEE.
[۱٦] Prasad BR, Agarwal SJIJoDT, Application. Comparative study of big data computing and storage tools: a review. ۲۰۱٦;۹(۱):.٦٦-٤٥
[۱۷] Paranjpe MSJdm. BIG DATA TOOLS: A SOLUTION TO CYBER SECURITY. ۲۰۲۱;.(۲)۱۰
[۱۸] Kitchin RJBd, society. Big Data, new epistemologies and paradigm shifts. ۲۰۱٤;۱(۱):.۲۰٥۳۹٥۱۷۱٤٥۲۸٤۸۱
[۱۹] Kambatla K, Kollias G, Kumar V, Grama AJJop, computing d. Trends in big data analytics. ۲۰۱٤;۷٤(۷):.۷۳-۲٥٦۱
[۲۰] Del Río S, López V, Benítez JM, Herrera FJIS. On the use of mapreduce for imbalanced big data using random forest. ۲۰۱٤;.۳۷-۲۸٥:۱۱۲