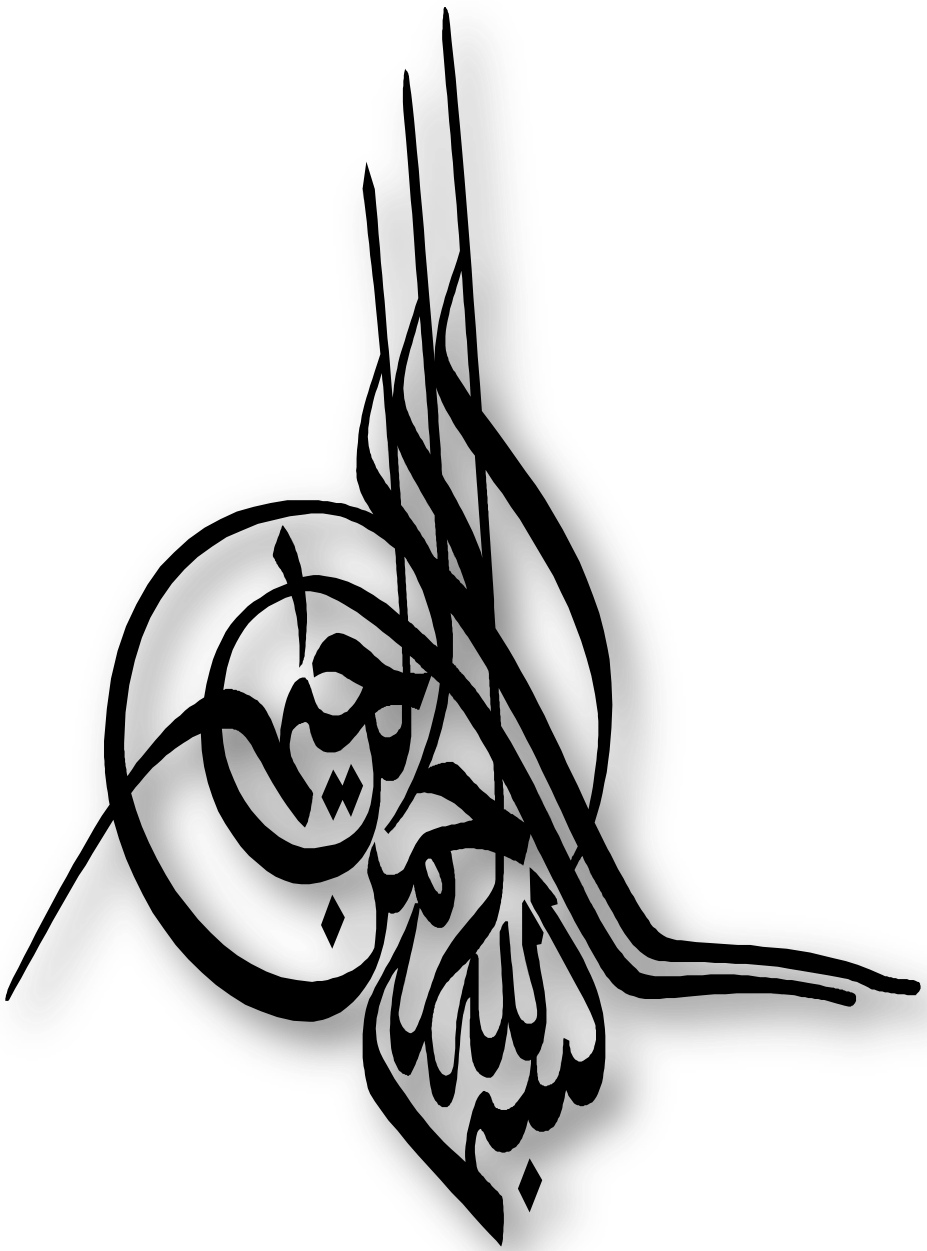

بسم الله الرحمن الرحيم

دانشکده فنی و مهندسی واحدری

گزارش درس سمیناردر مقطع کارشناسی ارشد

گرایش مهندسی نرم افزار

عنوان رساله:

تجزیه و تحلیل کاربرد کلان داده ها

استاد راهنما:

دکتر علی رضوی

نگارنده:

ویدا سپاسی

بهمن ۱۴۰۰

فهرست مطالب

عنوان..... صفحه

فصـل سـوم

مروری بر کارهای انجام شده

فصـــل چهارم

نحوه ذخیره سازی و کاربردهای کلان داده

فصـل پنجم

جمع‌بندی و پیشنهادها

فهرست تصاویر

عنوان..... صفحه

فهرست جداول

عنوان ..... صفحه

فصـــل اول

کلیات پژوهش

مقدمه

در دنیای پیچیده امروزی که مدام در حال تغییر و تحول حوزه های گوناگون فناوری و تکنولوژی می باشد ایجاد ظرفیت ها وتوانایی های لازم جهت حفظ فضای کسب وکار و رقابت با سایر رقبا به امری کاملا ضروری برای بقا تبدیل گردیده است. چرا که قابلیت پاسخ گویی به نیازهای متنوع و گوناگون مشتریان؛ نیازمند درک صحیح و شناخت این نیازها و انجام سریع اقدامات لازم با هدف عقب نماندن از فرصت های بازار می باشد.

با در نظر گرفتن شرایط حاکم بر بازارهای امروزی کاملا می توان متوجه شد که دریافت و شناخت نیازهای دائما در حال تغییر مشتریان و مردم مهم ترین فعالیت و اقدامات شرکت های فعال بازار را شکل می دهد چرا که هر کدام از این شرکت ها تنها در صورتی می توانند به حیات خود ادامه دهند که خود را با این تقاضا ها هماهنگ ساخته و گوی رقابت را از سایر رقبا بربایند.

سازمانها و شرکتها با به کارگیری و پیاده سازی فناوری IT توانایی انجام بهتر و ساده تر وظایف خود را افزایش میدهند و از این طریق قادرند روش کار خویش را دگرگون سازند. مزایایی که IT در سازمانها ایجاد میکند از جمله در صرفه جویی هزینه ها، جلوگیری از خطاهای انسانی، افزایش بهبود کارآیی و اثر بخشی سازمانی بسیار قابل تامل میباشد. به همین خاطر، امروزه سرانه هزینه IT به ازای یک نیروی انسانی به عنوان یکی از شاخص های توسعه ملی کشورها مطرح میشود(کمپ،۲۰۱٤).

BigDataاصطلاحی است برای مجموعه داده های حجیم که بزرگ، متنوع، با ساختار پیچیده و با دشواری هایی برای ذخیره سازی، تحلیل و تصویرسازی (نمایش)، پردازش های بیشتر یا نتایج می باشد. پروسه تحقیق بر روی داده های حجیم جهت آشکارسازی الگوهای مخفی و راز همبستگی ها ، تجزیه و تحلیل BigData نامیده میشود. این مطالعات مفید برای سازمان ها وشرکت ها در جهت کسب بینش غنی تر و عمیق تر و موفقیت در رقابت کمک میکند. به همین دلیل اجراهای BigData نیاز دارند تا در صورت امکان ، تحلیل شوند و به طور دقیق اجرا شوند. این پژوهش به بررسی کاربرد BigData در تجزیه و تحلیل های مرتبط به کسب و کار شرکت ها در محیط رقابتی می پردازد (ایناو و لوین۲۰۱۳)

۱-۱ تعریف مسئله و بیان سؤال‌های اصلی تحقیق

با سپری شدن سال های متمادی در صنعت IT ورشد تجارب گوناگون در حوزه جمع آوری، ذخیره سازی و بازیابی اطلاعات، موضوع معنا بخشیدن به داده ها و آسان کردن فرآیند تصمیم سازی، مرکز توجه کارشناسان فناوری اطلاعات و متخصصان علم مدیریت و کسب و کار قرار گرفته است. می دانیم که تصمیم سازی به معنای ساختن و پیشنهاد انواع تصمیماتی است که در شرایط خاص می توان اتخاذ کرد. بنابراین داده های عظیم یا همان BigData قابلیت آن را دارند که با استفاده از تکنیک های پیشرفته ارزیابی و تحلیل داده ها، اطلاعاتی بسیار با ارزش و گرانبهایی را در اختیارشرکت ها قرار دهند. این تکنیک های پیشرفته که شامل طیف وسیعی از نرم افزار های کاربردی، مدل های کسب وکار و الگوهای مختلف می باشد به سازمان ها کمک می کند تا داده های پراکنده موجود را به اطلاعاتی کارگشا، قابل تبادل (به اشتراک گذاشتن میان مدیران) و قابل ذخیره سازی برای تصمیمات بعدی تبدیل کنند(مانیکا و همکاران، ۲۰۱۱).

امروز هر سازمانی که خواستارتوسعه بازاریابی و فروش محصولات وخدمات خود در هر محدوده ای باشد، باید شناخت بازار را که مستلزم درک جامعی از نیازمندی های مشتریان و فضای عمومی عرضه و تقاضای تجارت در منطقه فعالیتش است را سرلوحه توسعه کسب وکار خود قرار دهد وبه این ترتیب، فروش محصولات و خدمات شرکت مربوطه افزایش یافته و در نهایت اعتبار سازمان ارتقا می یابد. درواقع بازاریابی فرزند اقتصاد رقابتی است. در دورانی که سرعت اتصال به اینترنت و حجم دستگاه های متصل به آن کم بود تحلیل سازمانی بیش تر بر مبنای ادراک مدیران و به صورت شهودی انجام می شد. اما به مرور بارشد انتظارات سازمان ها و مشتریان نسبت به دریافت خدمات سریع، تحلیل بلا درنگ به عنوان ابزاری ارزشمند معرفی گردید که قادر است به کسب و کار های گوناگون جهت اتخاذ تصمیمات هوشمندانه و همسو با نیازهای مشتریان به منظور صرفه جویی در وقت و هزینه، کمک شایانی نماید. تحلیل بلا درنگ در کنار رشد و نوآوری های ایجاد شده در صنعت تراشه سازی و سایر فناوری ها قادر به مرتب نمودن اطلاعات حجیم سازمان ها در عرض یک میلینیو ثانیه بودند، ابداع گردید و دارای محبوبیتی فراگیر شد. در مقابل کارمندان سازمان ها می توانند براساس نوع انتخاب خود، ازاین فناوری به صورت لجستیک یا استراتژیک استفاده نمایند (کاسترو و همکاران، ۲۰۰۷)

عبارت BigData مدت ها است که برای اشاره به حجم های عظیمی از داده ها که توسط سازمان های بزرگی مانند گوگل یا ناسا ذخیره و تحلیل می شوند مورد استفاده قرار گیرد. اما به تازگی، این عبارت بیشتر برای اشاره به مجموعه های داده ای بزرگی استفاده می شود که به قدری بزرگ و حجیم هستند که با ابزارهای مدیریتی و پایگاه های داده سنتی و معمولی قابل مدیریت نیستند. مشکلات اصلی در کار با این نوع داده ها مربوط به برداشت و جمع آوری، ذخیره سازی، جست و جو، اشتراک گذاری، تحلیل و نمایش آن ها است. این مبحث، به این دلیل هر روز جذابیت و مقبولیت بیشتری پیدا می کند که با استفاده از تحلیل حجم های بیشتری از داده ها، می توان تحلیل های بهترو پیشرفته تری را برای مقاصد مختلف، از جمله مقاصد تجاری، پزشکی و امنیتی، انجام

داد و نتایج مناسب تری را دریافت کرد. بیشتر تحلیل های مورد نیاز در پردازش داده های عظیم، توسط دانشمندان در علومی مانند هواشناسی، ژنتیک، کانکنومیک(علوم مرتبط با نگاشت سیستم عصبی)، شبیه سازی های پیچیده فیزیک، تحقیقات زیست شناسی و محیطی، جست وجوی اینترنت، تحلیل های اقتصادی و مالی و تجاری مورد استفاده قرار می گیرد(لوی و ویلنسکای، ۲۰۱۱).

حجم داده های ذخیره شده در مجموعه های داده ای BigData، شبکه های حساس بی سیم و غیره با سرعت خیره کننده ای در حال افزایش است، به طوری که در هر روز، ۲/۵ کوادریلیارد بایت (هر کوادریلیارد برابر ۱۰به توان ۲۷ است)داده در حال تولید است. نکته جالب توجه دراین زمینه آن است که ۹۰ درصد داده هایی که اکنون در اختیار ما است، تنها در۲ سال اخیر تولید شده است.

یکی از مهم ترین مسائل مرتبط با داده های عظیم، مشکل بودن کار با آن ها به وسیله پایگاه های داده ای رابطه ای و بسته های نرم افزاری تصویر نگاری داده ها و نرم افزارهای آماری رومیزی است. چرا که این داده ها، برای پردازش شدن در یک زمان معقول به نرم افزارهای به شدت موازی شده با قابلیت اجرا روی ده ها، صد ها یا هزاران سرور نیاز دارند. البته مفهوم BigData برای شرکت ها و سازمان های مختلف تعابیر متفاوتی دارد و هر کدام، بسته به کاربرد و نیازمندی هایی که دارند، در حجمی خاص و با شرایطی خاص به روش های جدیدی برای آسان کردن کار با این نوع داده ها روی می آورند. از این رو است که BigData برای بعضی سازمان ها، تنها صد ها گیگابایت حجم دارد در حالی که برای برخی، ده ها و صد ها ترابایت یا حتی مضاربی از اگزابایت وزتابایت از انواع داده های عظیم محسوب می شوند

یکی از بهترین تعابیری که در زمینه توصیف و تبیین BigData و چالش های پیش روی آن به کار رفته است، ایده دوگ لنی(Laney Doug)در گزارش سال۲۰۰۱ موسسه META group (اکنون گارتنر) بود که در آن عنوان شده است داده ها در سه بعد مختلف در حال رشد هستند (بک و موستو، ۲۰۰۸.)

این سه بعد عبارتند از حجم، سرعت و تنوع داده ها که روز به روز نرخ رشد آن ها با سرعتی باور نکردنی افزایش می یابد. به همین دلیل، توصیف BigData تنها با حجم های عظیم و نحوه مدیریت آن ها کار درستی نیست و باید دیگر جنبه های این مفهوم مهم وکلیدی را نیزدر نظر گرفت.

براین اساس، با توجه به افزایش روز افزون استفاده از تجهیزات تولید با جمع آوری داده ها و همچنین روی آوردن تعداد بیشتری از شرکت ها و افراد به شکل های جدیدی از زندگی دیجیتالی، اهمیت مفهوم BigData و نحوه برنامه ریزی و تعیین راهبرد های مناسب برای بهره برداری صحیح از آن، دو چندان شده و نیاز به توسعه ابزارها و امکانات مناسب برای مدیریت بهتر آن ها بیش از پیش مشخص می شود(بیکر، ۲۰۱۲).

برنامه های بازاریابی، به تحلیل اطلاعات کلان داده ها درباره ی چگونگی جذب بیشتر مشتریان و ایجاد حس وفاداری و برقراری ارتباط پایدار با آنها، نیاز دارند. ارتباطی که بتواند با پایداری در بلندمدت، نیازهای مشتریان را در مقابل رقبا، مرتفع کند. یافتن روش هایی برای عدم رغبت مشتریان به امتحان کردن یا انتخاب محصول و

خدمات رقیب، استراتژی ای بسیار مهمتر از تبدیل مشتریان رقیب به مصرف کنندگان کالاها یا خدمات خود است. پاسخ سوالات اساسی و یافتن راه حلهای مطلوب، نیازمند استفاده ی متفاوت از کلان داده ها است. الزامی است به جای یافتن راه حل هایی برای هدف گیری مشتریان بالقوه، به دنبال استفاده از کلان داده ها جهت کسب صفات ارزشمند در محصولات و خدمات بود.

سؤالاتی که در این تحقیق سعی به پاسخ به آنها داریم:

۱. مکانیزم فعالیت و چارچوب های کاربردی کلان داده را بیان نمایید .
۲. از چه روشهایی برای تجزیه و تحلیل کلان داده استفاده می شود؟
۳. کاربردهای کلان داده را عنوان نمایید ؟

۲-۱ سابقه و ضرورت انجام تحقیق

اهمیت و ضرورت تحقیق از دو منظر ایجابی و سلبی مورد بررسی قرار خواهد گرفت.

از منظر ایجابی، تعریف پژوهش حاضر می تواند زمینه ساز شفافیت بیشتر در داده های تراکنش بانکی یا معاملات دو طرفه (سازمان- مشتری) شود و از این طریق منجر به افزایش سودآوری سازمان متبوع شود. همچنین تعریف این پروژه و ارزیابی میزان دستیابی به نتایج مورد انتظار، می تواند زمینه ساز جهت دهی بهتر پژوهش های دیگر در حوزه تشخیص تخلف و مسائل عمده مربوط به آن شود.

از منظر سلبی نیز، نبود این پژوهش هزینه های پیدا و پنهان زیادی را متوجه سازمان ها خواهد کرد. در حال حاضر بسیاری از سازمان ها چه در ایران و چه در خارج از ایران از نبود داده های برچسب خورده کافی جهت مدل سازی مسئله برای خود شکایت دارند. لذا عدم تعریف پژوهش هایی نظیر پژوهش حاضر، مسئله موجود را همچنان باز نگه خواهد داد حال آنکه ارائه الگویی جهت باز ساخت داده های تخلف از میان تخلف های اندک کشف شده، می تواند برای حالت های جدید و نادیده، عملکرد بالایی ارائه دهد.

۳-۱چه کاربردهایی از انجام این تحقیق متصور است؟

پروژه ه ی حاضر از منظر کاربرد برای دو گروه قابل استفاده خواهد بود.

الف) سازمان ها و کسب و کارها

سازمان ها امروزه سرشار از داده هستند. این داده ها مخصوصا در حوزه مالی با دقت بالا در حال ذخیره سازی است. همچنین وقوع تخلف های صورت گرفته از صنعت بیمه گرفته تا برق دزدی و نظام بانکداری، به وفور در حال صورت گرفتن است و همین طور با توجه به عدم کشف داده های تخلف کافی در عمده موارد، کمبود برچسب داده های تخلف در سازمان ها مشهود است بنابراین، شرکت و سازمان ها با استفاده از این مدل پیشنهادی، می توانند با دقت بالاتری موارد مشکوک به تخلف خود را شناسایی کنند.

ب) پژوهشگران

با توجه به رویکرد نوآورانه این پژوهش، پژوهشگران دیگر می توانند با استفاده از نتایج به دست آمده از اجرای این مدل، در جهت تعریف پروژه های جدید استفاده نمایند.

۱-٤ روش انجام تحقیق

پژوهش حاضر با بیان ویژگی ها و خصوصیات منحصر به فرد ابزار داده های عظیم یا همان BigData ، ازنظر هدف، کاربردی و از نظر جمع آوری داده، توصیفی از نوع پیمایشی تلقی گردیده و با در نظر گرفتن این امر که نتایج حاصل از آن درتمامی صنعت ها و فعالیت های اقتصادی در هر زمینه ای قابل بهره برداری و استفاده می باشد کاربردی قلمداد می گردد.

۱-۵ مراحل انجام تحقیق

در ابتدا به مطالعه در زمینه کلان داده و مفـاهیـم عمـومـی کلان داده و شاخصه های آن پرداخته شده است. در ادامه به بیان کارهای انجام شده با استفاده از فناوریهای مرتبط و تولید و اکتساب کلان داده پرداخته‌ایم و و همچنین کاربرد کلان داده ها و نحوه تجزیه و تحلیل و معماری کلان داده ها را مورد بررسی قرار داده ایم. معماری کلان داده ها نیز مورد مطالعه قرار گرفته است.

۱-٦ ساختار گزارش سمینار

در این فصل به بیان اهداف کلی و مباحث مطرح شده می‌پردازیم.

در فصل دوم به مفـاهیـم عمـومـی کلان داده و شاخصه های آن پرداخته شده است.

در فصل سوم به بیان کارهای انجام شده با استفاده از فناوریهای مرتبط و تولید و اکتساب کلان داده پرداخته‌ایم.

در فصل چهارم ابتدا به بیان نحوه ذخیره ساری کلان داده ها و همچنین کاربرد کلان داده ها و نحوه تجزیه و تحلیل و معماری کلان داده ها پرداخته شده است.

در فصل پنجم به نتایج گرفته شده از تحقیق پرداخته شده و پیشنهاداتی برای کارهای آینده بیان شده است.

فصـــل دوم

کلان داده و شاخصه های آن

مقدمه

کلان داده، به مجموعه داده هایی اطلاق می گردد که حجم و سرعت تولید آنها بیش از ظرفیت و امکانات بانک های اطلاعاتی مرسوم برای ضبط، ذخیره سازی ، مدیریت و تحلیل داده است و نتوان آنها را با یک پردازشگر معمولی پردازش کرد. در واقع، عبارت BigData مدتها است که برای اشاره به حجمهای عظیمی از داده ها 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 مجموعه های داده ای بزرگی استفاده میشود که به قدری بزرگ و حجیم هستند که با ابزارهای مدیریتی و پایگاه های داده سنتی و معمولی قابل مدیریت نیستند. مشکلات اصلی در کار با این نوع داده ها مربوط به برداشت و جمع آوری، ذخیره سازی، جست وجو ، اشتراک گذاری ، 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 داده ها، می توان تحلیلهای بهتر و پیشرفته تری را برای مقاصد مختلف، از جمله مقاصد تجاری، پزشکی و امنیتی، انجام داد و نتایج مناسبتری را دریافت کرد.

بیشتر تحلیل های مورد نیاز در پردازش داده های عظیم، توسط دانشمندان در علومی مانند هواشناسی، ژنتیک، شبیه سازی های پیچیده فیزیک، تحقیقات زیست شناسی و محیطی، جست وجوی اینترنت، تحلیلهای اقتصادی و مالی وتجاری مورد استفاده قرار می گیرد. حجم داده های ذخیره شده در مجموعه های داده های BigData ، عموماً به خاطر تولید وجمع آوری داده ها از مجموعه بزرگی از تجهیزات و ابزارهای مختلف مانند گوشی های موبایل، حسگرهای محیطی، لاگ نرم افزارهای مختلف، دوربین ها، میکروفون ها، دستگاه های تشخیص RFID، شبکه های حسگر بی سیم وغیره با سرعت خیره کننده ای در حال افزایش است.

امروزه به واسطه ی افزایش استفاده از اینترنت و شبکه های اجتماعی حجم داده های موجود در فضای اینترنت بسیارافزایش داشته است به نحوی که هرگونه پردازش به روی آن از عهده ی کامپیوترهای معمولی خارج است داده های تولیدشده از انواع منابع با حجم بسیار زیاد سرعت بالا و ساختمان داده متفاوت را کلان داده می نامند. امروزه کلان داده به عنوان یک منبع غنی و ارزشمند برای رنج گسترده ای از سازمانها به رسمیت شناخته شده است.

کلان داده [۱]ها اصطلاحی است که به داده های یکپارچه با رشد سریع و ناگهانی اطلاق می شود و بیشتر برای توصیف مجموعه داده های بسیار زیاد استفاده می شود . در این فصل کلان داده ها را تعریف کرده و سیر تکامل آن را در بیست سال گذشته مورد بررسی قرار گرفته وهمچنین چهار ویژگی معرف کلان داده ها معروف به چهار Vرا بیان نموده که عبارت اند ا : حجم داده ها [۲]، تنوع داده ها [۳]، سرعت تولید داده ها [۴]و ارزش آن ها .[۵]در این فصل چالش های موجود در زمینه ی کلان داده ها بررسی می شود.

امروزه چالش اصلی همه حوزه های شبکه و پایگاه داده، موضوع داده های عظیم(کلان داده ها)است .داده های عظیم مفهومی است که به تازگی مطرح شده و بطور کلی به افزایش حجم اطلاعات غیرساختارمند ویکپارچه درکنارذخیره سازی و پردازش آنها می پردازد. آنها بر روی پایگاه داده ها که به شکل حجیم رشد می کنند، ذخیره می شوند وضبط ، شکل دهی، ذخیره سازی، مدیریت، به اشتراک گذاری، تحلیل و نمایش آنها از طریق ابزارهای نوعی نرم افزار پایگاه داده ها ، دشوار می شود.

از این رو تحلیل های دقیق بر روی این داده های عظیم، منجر به تصمیم گیری های با اطمینان بیشتری شده وتصمیمات بهتر، می تواند معنای کارایی بیشتر عملیات، کاهش هزینه ها و کاهش ریسک ها باشد.

داده های بزرگ تقاضا برای متخصصان در این حوزه را به شدت بالا برده است و شرکت هاییچون Oracle، IBM ، Microsoft، SAP و.... بیش از ۱۵ میلیارد دلار برای توسعه نرم افزارهای مدیریت و تحلیل داده سرمایه گذاری کرده اند. داده های بزرگ نحوه کار سازمان ها و افراد را تحت تاثیر قرار می دهد .داده های بزرگ فرهنگی را در

---

[۱] BigData
[۲] Volume
[۳] Variety
[۴] Velocity
[۵] Value

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.

۱-۲ طلوع عصر کلان داده

داده در بیست سال گذشته در مقیاس بزرگ در حوزه های مختلف افزایش یافته بود . مطابق با گزارش شرکت بین المللی داده (IDC) که در سال ۲۰۱۱ منتشر گردید ، حجم داده کپی شده و ایجاد شده کلی در جهان معادل ۱٫۸ ZB بوده است که در عرض پنج سال تقریبا تا ۹ برابر افزایش یافته است. این ویژگی حداقل هر دوسال دیگر در آینده نزدیک دوبرابر خواهد شد .

اصطلاح کلان داده تحت افزایش انفجاری داده جهانی به طور عمده برای توضیح مجموعه داده های بیشمار استفاده می گردد. معمولا کلان داده در مقایسه با مجموعه های داده سنتی شامل حجم های زیاد داده ساخت نیافته می باشد که به تحلیل انی بیشتر نیاز دارد. بعلاوه، همچنین کلان داده سبب فرصت های جدید برای کشف مقادیر تازه می گردد، در رسیدن به درک عمیق از مقادیر مخفی شده به ما کمک می کند و همچنین چالش های جدید را ایجاد می کند یعنی چگونه چنین مجموعه داده هایی را سازماندهی و مدیریت نمود.

صنایع در سال های اخیر به پتانسیل بالا کلان داده علاقمند شده اند و بعضی آژانس های دولتی یک سری طرح های مهم را برای تسریع جستجو و کاربرد های کلان داده اعلان کرده اند. بعلاوه موضوعات در مورد کلان داده اغلب در رسانه های عمومی نظیر اکونومیست، نیویورک تایمز و رادیو عمومی ملی پوشش داده می شوند. همچنین دو مجله علمی مهم یعنی Nature و Science یک سری ستون های خاص را برای بحث در مورد چالش ها و تاثیرات کلان داده باز کرده اند. عصر کلان داده فراتر از همه شک ها وارد عرصه شده است.

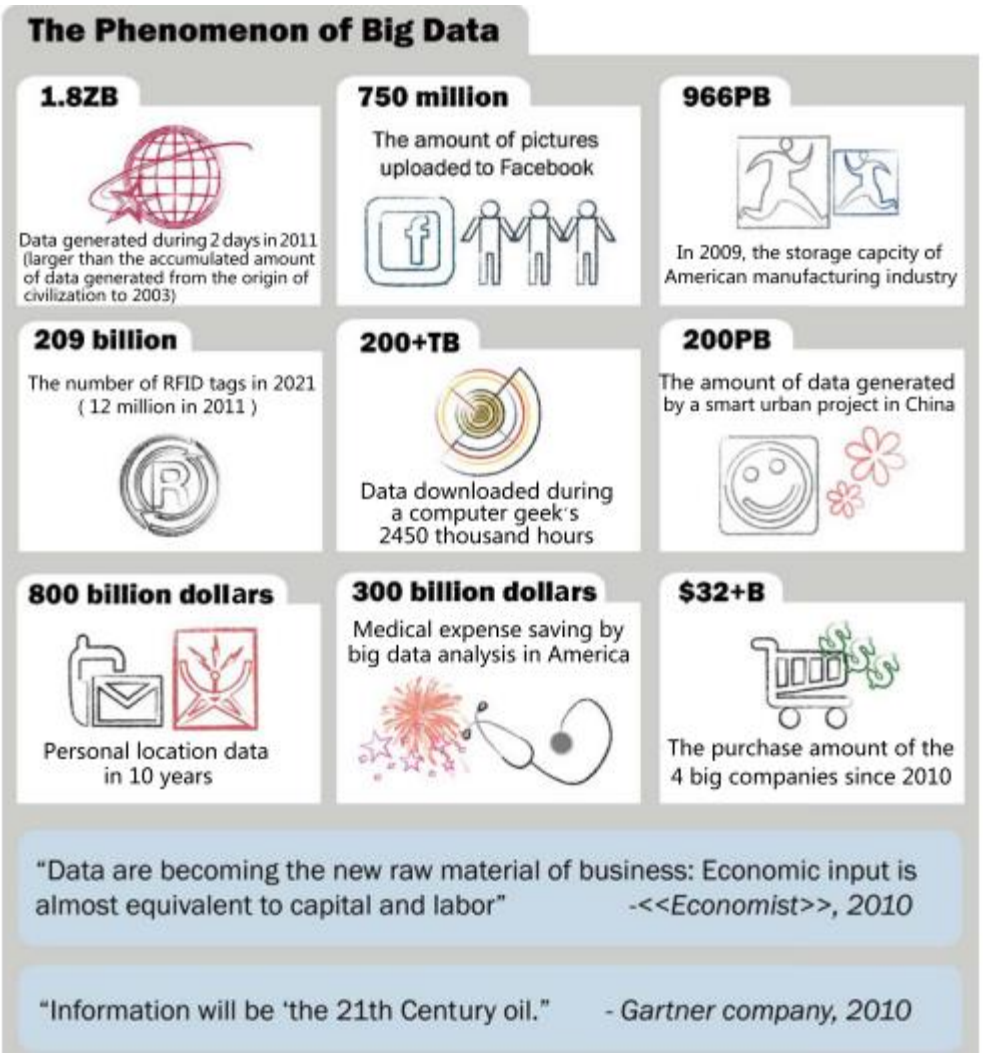
امروزه کلان داده مرتبط با سرویس دهی شرکت های اینترنتی به سرعت رشد می کند. برای مثال، داده های صد ها پتابایتی (PB) در گوگل پردازش می گردند، فیسبوک هر ماه بیش از ۱۰ پتابایت داده ثبتی را تولید می کند، شرکت چینی بایدو در حدود ده ها پتابایت داده را پردزاش می نماید و شرکت تائوبائو که زیرمجموعه Alibaba می باشد، داده ده ها ترابایتی (TB) را برای تجارت انلاین در هر روز تولید می کند. جهش حجم داده جهانی در تصویر۱ نشان داده می شود. در حالی که مقدار مجموعه داده بزرگ به طور خیره کننده ای در حال ترقی می باشد، همچنین سبب یک سری مشکلات چالش برانگیز می گردد که به راه حل های فوری نیاز دارد:

جدید ترین پیشرفت ها فناوری اطلاعات باعث ساده تر شدن تولید داده ها می شوند. برای مثال، بطور میانگین معادل ۷۲ ساعت ویدیو در هر دقیقه در یوتیوب آپلود می شود. از اینرو، ما با چالش اصلی جمع اوری و یکی کردن داده انبوه از منابع بی نهایت پراکنده (توزیعی) مواجه هستیم.

رشد سریع رایانش ابری و اینترنت اشیاء (IoT) باعث بهبود بیشتر رشد تند داده ها می شود. نگهداری، سایت های دسترسی و کانال ها برای دارایی داده از طریق رایانش ابری مهیاء می گردند. حسگر ها در الگوی IoT در سراسر جهان در حال جمع اوری هستند و داده هایی را انتقال می دهند که قرار است ذخیره شوند و در ابر پردازش گردند. چنین داده هایی در کمیت و روابط متقابل بسیار برتر از ظرفیت های معماری های فناوری اطلاعات و زیرساحت ویژگی های موجود خواهند بود و همچنین الزام زمان واقعی اشان تا حد زیادی بر ظرفیت رایانش در دسترس فشار خواهند اورد. داده های در حال رشد سبب ایجاد مشکل در نحوه ذخیره سازی و مدیریت این مجموعه داده های ناهمگن عظیم با الزامات معتدل در زیرساخت نرم افزاری و سخت افزاری می شوند. ما در بررسی ناهمگنی، مقیاس پذیری، زمان واقعی، پیچیدگی و محرمانگی کلان داده بایستی به طور موثر در سطوح مختلف در طول تجزیه و تحلیل، مدل سازی، بصری سازی و پیش بینی برای آشکار سازی ویژگی ذاتی اش و بهبود تصمیم گیری درمجموعه های داده کاوش نماییم.

۲-۲ تعریف و ویژگی های کلان داده

کلان داده یک مفهوم انتزاعی می باشد. همچنین کلان داده جدا از حجم های داده از یک سری خصیصه های دیگر برخوردار بوده است که تفاوت بین خودش و "داده کلان" یا " داده بسیار بزرگ" را مشخص می سازند.



تصویر۱: زیاد شدن دایمی کلان داده

فعلا هر چند اهمیت کلان داده به طور کلی شناخته شده است، هنوز مردم ایده هایی متفاوتی در مورد تعریف کلان داده دارند. بطور کلی، کلان داده بایستی به معنی مجموعه های داده باشد که از طریق فناوری اطلاعات سنتی و ابزار نرم افزاری/سخت افزاری بر طبق یک زمان قابل تحمل قابل درک، کسب، مدیریت و پردازش نمی باشد. موسسات علمی و فناورانه، محققان تحقیقاتی، تحلیل گران داده و اصحاب فنی به دلیل نگرانی های مختلف از تعاریف متفاوت برای کلان داده استفاده می کنند. تعاریف زیر ممکن است به ما کمک نمایند تا به درک بهتر از نامگذاری های اجتماعی، اقتصادی و فناورانه برجسته کلان داده دست یابیم.

کلان داده در سال ۲۰۱۰ توسط Apache Hadoop به صورت زیر تعریف می گردد: کلان داده را نمی توان گرفت، مدیریت نمود و از طریق کامپیوتر ها در یک گستره قابل قبول پردازش نمود. آژانس جهانی مشاوره McKinsey&Company در ماه می ۲۰۱۱ بر اساس این تعریف به توصیف کلان داده پرداخته است و اعلان

کرده است کلان داده می تواند مرز آینده برای نواوری، رقابت و بهره وری باشد. کلان داده بایستی همانند مجموعه های داده ای تعریف گردد که نمی توان کسب نمود، ذخیره کرد و از طریق نرم افزار پایگاه داده کلاسیک مدیریت نمود. این تعریف دو معنای ضمنی را در بر می گیرد که عبارتند از: اولا، حجم هایی مجموعه های داده که با استاندارد کلان داده تطبیق می یابند، در حال تغییر می باشند و ممکن است با گذشت زمان یا با پیشرفت های فناورانه رشد نمایند؛ ثانیا حجم های مجموعه های داده که با استاندارد کلان داده در کاربرد های مختلف تطبیق می یابند با همدیگر فرق دارند. اکنون، کلان داده به طور معمول در دامنه چند ترا بایت تا تاچندپتابایت قرار می گیرد. برحسب تعریف McKinsey&Company می توان مشاهده نمود که حجم مجموعه داده نمی تواند تنها معیار برای کلان داده باشد. مقیاس داده در حال رشد زیاد و مدیریت اش که از طریق فناوری های پایگاه داده سنتی قابل هندل کردن نبود از جمله دو خصیصه کلیدی بعدی محسوب می شوند.

حقیقت امر این است کلان داده حتی در ابتدای سال ۲۰۰۱ تعریف شده بود. دوگ لانی تحلیل گیر META یک سری چالش ها و فرصت هایی را تعریف کرده است که سبب داده زیاد در مدل ۳V شده اند یعنی حجم، سرعت و تنوع در گزارش تحقیق افزایش می یابد. هر چند اساسا چنین مدلی برای تعریف کلان داده استفاده نشده است، گارتنر و چندین شرکت دیگر ار جمله IBM و تعدادی از دپارتمان های تحقیقاتی مایکروسافت هنوز از مدل "۳Vs " برای توصیف کلان داده در طی ده سال جاری استفاده کرده اند. حجم در مدل "۳Vs " بدان معنی است که مقیاس داده با تولید و جمع اوری بخش های زیاد داده به طور فزاینده بزرگ می شود؛ سرعت به معنی بجا بودن کلان داده می باشد بویژه جمع اوری و تحلیل داده و اقدامات دیگر بایستی با سرعت و بموقع انجام بگیرند تا استفاده حداکثری از ارزش تجاری کلان داده صورت گیرد؛ تنوع در واقع انواع مختلف داده را مشخص می نماید که شامل داده های نیمه ساخت یافته و ساخت یافته نظیر صوت، ویدیو، صفحه وب و متن و همچنین داده های ساخت یافته سنتی می باشد.

از اینرو ، شرکت های دیگر از جمله IDC به عنوان یکی از تاثیر گذارترین رهبران در حوزه کلان داده و زمینه های تحقیقاتی ان یک سری ایده های متفاوت دارند. کلان داده در گزارش سال ۲۰۱۱ IDC این گونه تعریف شده است: فناوری های کلان داده یک نسل جدید از فناوری ها و معماری ها را توصیف می کنند که برای مقدار استخراج تجاری از حجم های بسیار بزرگ با انواع گسترده داده تا توانمند سازی برای گرفتن داده، کشف و یا تحلیل سرعت بالا طراحی شده اند. رفرنس با این تعریف در واقع مشخصه های کلان داده را به صورت چهار Vخلاصه می کند که عبارتند از حجم [۱](حجم زیاد)، تنوع [۲](جنبه های مختلف)، سرعت [۳](تولید سریع) و مقدار [۴](مقدار عظیم اما با دانسیته بسیار پایین) که درتصویر ۲ نشان داده شده اند. این تعریف برای چهار V ها با

---

[۱] Volume
[۲] Variety
[۳] Velocity
[۴] Value

شناخت همه جانبه ای همراه شده بود چون معنی و ضرورت کلان داده را برجسته می سازد یعنی مقادیر عظیم مخفی را کشف می کند. بحرانی ترین مشکل در کلان داده با این تعریف مشخص می گردد و این مشکل به نوعی این گونه بیان می گردد که چگونه مقادیر از مجموعه های داده در مقیاس بزرگ ، تنوع زیاد و تولید سریع کشف می گردند. همانطور که جای پاریخ جانشین مهندسی فیسبوک بیان کرده است : اگرشما از داده های جمع اوری شده بهره برداری ننماید، تنها می توانستید یک دسته از داده را بجای کلان داده در اختیار داشته باشید.

بعلاوه، کلان داده در NIST این گونه تعریف می گردد: کلان داده بایستی به معنی داده ای باشد که حجم داده، سرعت اکتساب یا نمایش داده از ان داده باعث محدودیت ظرفیت استفاده از روش های رابطه ای سنتی برای انجام تحلیل موثر می گردد یا داده ای که ممکن است به طور موثر با فناوری های مهم زوم افقی پردازش شوند که بر جنبه فناورانه کلان داده تمرکز می کند. NIST مشخص می کند که روش ها یا فناوری های کارامد بایستی توسعه یافته و برای تجزیه و تحلیل و پردازش کلان داده استفاده گردند.

بحث های زیادی در مورد تعریف کلان داده از حوزه های دانشگاهی و صنعتی وجود داشته اند. همچنین تحقیق کلان داده علاوه بر پدیدآوری یک تعریف مناسب بایستی بر چگونگی استخراج مقدار داده، چگونگی استفاده ان و چگونگی تغییر شکل یک دسته داده به کلان داده تمرکز نماید.

۲-۳چالش های کلان داده

سیل داده های در حال افزایش شدید در عصر کلان داده سبب چالش های عظیمی در حوزه های اکتساب، ذخیره سازی، مدیریت و تحلیل داده می شود. سیستم های تحلیل و مدیریت داده سنتی بر اساس سیستم مدیریت پایگاه داده رابطه ای (RDBMS[۱]) هستند. از اینرو، چنین سیستم هایی تنها برای داده ساخت یافته بکار می روند و برای داده های نیمه ساخت یافته یا غیر ساخت یافته کارایی ندارند. علاوه بر این، RDBMS ها به طور فزاینده ای از سخت افزار گران و گران تر استفاده می کنند. مشهود است که RDBMS های سنتی قادر نبودند تا به حجم عظیم و ناهمگن کلان داده رسیدگی نمایند. جامعه تحقیق یک سری راه حل ها را از چشم انداز های متفاوت پیشنهاد داده است. برای مثال، رایانش ابری مورد استفاده قرار می گیرد تا الزاماتی نظیر بهره وری هزینه، الاستیسیته و به روز رسانی هموار/تنزیل رتبه در زمینه زیرساخت کلان داده را برطرف نمایند. پایگاه های داده NoSQL و سیستم های فایل پراکنده برای راه حل های ذخیره سازی دایمی و مدیریت مجموعه های داده آشفته مقیاس بزرگ از جمله گزینه های خوب محسوب می شوند. این قبیل چارچوب های برنامه ریزی در پردازش وظایف خوشه ای بویژه برای رتبه بندی صفحه وب به موفقیت شایانی دست یافته بودند. کاربرد های متنوع کلان داده را می توان بر اساس این فناوری ها یا پلت فرم های نواورانه توسعه داد. علاوه بر این، کلان داده برای بکار گیری سیستم های تحلیل کلان داده غیر بدیهی می باشد.

[۱] Relational Database Management System

موانع توسعه کاربرد های کلان داده در بعضی مطالب قبلی مورد بحث قرار می گیرند. چالش های کلیدی وجود دارند که در ذیل فهرست می شوند:

۱-۳-۲نمایش داده

بعضی مجموعه های داده دارای سطوح معین ناهمگنی در نوع، ساختار، معناشناسی، سازماندهی، دانه ای بودن و دسترسی پذیری می باشند. هدف نمایش داده این است تا داده ها را برای تحلیل کامپیوتری و تفسیر کاربر قابل فهم تر سازد. معهذا، نمایش نامناسب داده به کاهش ارزش داده اصلی خواهد انجامید و حتی می تواند تحلیل موثر داده را با اشکال مواجه نماید. نمایش داده کارامد بایستی ساختار ، دسته و نوع داده و همچنین فناوری های یکپارچه را برای میسر سازی فعالیت های کافی در مجموعه های داده متفاوت منعکس سازد.

۲-۳-۲کاهش افزونگی (حشو) و فشردگی داده ها

معمولا، سطح بالا افزونگی در مجموعه های داده وجود دارد. کاهش افزونگی و فشردگی داده برای کاهش هزینه غیر مستقیم کل سیستم در یک فرضیه اثبات شده قبلی موثر است که مقادیر بالقوه داده تحت تاثیر قرار نمی گیرند. برای مثال، بخش عمده داده های تولید شده از طریق شبکه های حسگر تا حد زیادی افزونه هستند که می توان این داده ها را فیلتر نمود و بر اساس مرتبه بزرگی فشرده نمود.

۳-۳-۲مدیریت چرخه عمر داده

رایانش و حسگری نافذ یک سری داده ها را در نرخ ها و مقیاس های غیر منتظره در مقایسه با پیشرفت های نسبتا کند سیستم های ذخیره سازی تولید می کنند. ما با چالش های زیاد فشرده سازی مواجه هستیم و یکی از ان چالش ها را می توان این گونه عنوان نمود که سیستم ذخیره سازی فعلی قادر به پشتیبانی از چنین داده گسترده ای نمی باشد. بطور کلی، مقادیر پنهان شده در کلان داده به تازگی داده بستگی دارند. از اینرو، اصل اهمیت داده مرتبط با مقدار تحلیلی بایستی توسعه یابد تا تصمیم گرفته شود کدام داده بایستی ذخیره گردد و کدام داده بایستی کنار گذاشته شود.

٤-۳-۲مکانیزم تحلیلی

سیستم تحلیلی کلان داده بایستی حجم های زیاد داده ناهمگن را در یک زمان محدود پردازش نماید. از اینرو، RDBMS های سنتی دقیقا با نبود مقیاس پذیری و توسعه پذیری طراحی می شوند که نمی توانستند الزامات عملکرد را براورده سازند. پایگاه های داده غیر رابطه ای مزیت های منحصر به فرد اشان را در پردازش داده ساخت نیافته نشان داده اند و کار تبدیل شدن به جریان اصلی در تحلیل کلان داده را شروع کرده اند. در هر صورت، هنوز یک سری مشکلات برای پایگاه های داده غیر رابطه ای در عملکرد و کاربرد های خاص اشان

وجود دارند. ما بایستی یک راه حل سازشی را بین RDBMS ها و پایگاه های داده غیر رابطه ای بیابیم. برای مثال، بعضی شرکت ها از معماری پایگاه داده ترکیبی استفاده کرده اند که مزیت های هر دو نوع پایگاه داده را یکی می سازد (برای مثال فیسبوک و Taobao). تحقیق بیشتر در مورد پایگاه داده دارای حافظه و داده نمونه مبتنی بر تحلیل تقریبی الزامی است.

محرمانگی داده اکثر تهیه کننده ها یا مالکان سرویس کلان داده در زمان حال نمی توانستند به طور موثر این قبیل پایگاه های داده عظیم را به دلیل ظرفیت محدود اشان حفظ نمایند و تحلیل کنند. این افراد بایستی به حرفه ای ها یا ابزاری برای تحلیل این قبیل داده ها که ریسک های ایمنی بالقوه را افزایش می دهند، تکیه نمایند. برای مثال، معمولا مجموعه داده تراکنشی از یک مجموعه داده عملیاتی کامل تشکیل می گردد تا فرایند های تجاری کلیدی را هدایت نمایند. چنین داده ای حاوی جزییات با پایین ترین اطلاعات دانه دانه ای و بعضی اطلاعات حساس نظیر شماره های کارت اعتباری می باشند. بنابراین، تحلیل کلان داده ممکن است یک طرف ثالث را برای پردازش در زمانی تحویل دهد که اقدامات بازدارنده مناسب برای محافظت از این قبیل داده های حساس برای تضمین ایمنی اش پذیرفته می شوند.

۲-۳-۵مدیریت انرژی

مصرف انرژی سیستم های رایانش کامپیوتر اصلی از چشم انداز های اقتصادی و زیست محیطی تا حد زیادی مورد توجه قرار گرفته است. پردازش، ذخیره سازی و انتقال کلان داده با افزایش حجم داده و تقاضا های تحلیلی به طور طور اجتناب ناپذیری با مصرف بیشتر و بیشتر انرژی الکتریکی همراه خواهند شد. ازاینرو، کنترل مصرف توان سطح سیستم و مکانیزم مدیریت بایستی برای کلان داده ایجاد گردد در حالی که بسط پذیری و دسترس پذیری تضمین می گردند.

۲-۳-٦بسط پذیری و مقیاس پذیری

سیستم تحلیلی کلان داده بایستی از مجموعه داده های فعلی و آینده پشتیبانی نماید. الگوریتم تحلیلی بایستی قادر به پردازش مجموعه های داده پیچیده تر و در حال توسعه فزاینده گردد.

۲-۳-۷ همکاری

تحلیل کلان داده یک تحقیق بین رشته ای محسوب می گردد که به کارشناسانی در رشته های مختلف نیاز دارد تا برای برداشت بالقوه از کلان داده همکاری نمایند. معماری شبکه کلان داده جامع بایستی ایجاد گردد تا به دانشمندان و مهندسان در رشته های مختلف کمک نماید و این افراد به انواع مختلف داده دست یابند و از مهارت اشان برای همکاری برای تکمیل اهداف تحلیلی استفاده نمایند.

۲-٤ اهداف مطالعه

با توجه به ویژگی‌های گفته شده در مورد کلان داده و کاربرد آن بر آن شدیم که در این تحقیق به بررسی روش‌های داده‌کاوی در زمینه سلامت و پزشکی بپردازیم. تمرکز اصلی بر روی روش‌های دسته‌بندی جهت بالا بردن دقت و همچنین ارائه روش‌های بهینه و کاربردی جهت ارتقای تشخیص و پیش‌بینی در این حوزه می‌باشد.

۲-۵ جمع بندی

در این فصل بیان کردیم که داده‌کاوی به‌عنوان یک گام در فرآیند آشکار کردن دانش مطرح شده است و این گام یک گام اساسی می‌باشد، همچنین روشهای کشف دانش در داده‌کاوی را بیان کردیم که می‌توان از روش‌های داده‌کاوی در زمینه اطلاعات سلامت و پزشکی استفاده کرد.

فصـــل ســـوم

مروری بر کارهای انجام شده

مقدمه

در این فصل به بررسی کارهای انجام شده در زمینه کلان داده و **مرکز داده** خواهیم پرداخت.

۳-۱-فناوری های مرتبط

این بخش با هدف رسیدن به درک عمیق از کلان داده به معرفی چندین فناوری اساسی خواهد پرداخت که دارای رابطه بسیارنزدیک با کلان داده از جمله رایانش ابری، اینترنت اشیاء، مرکز داده و Hadoop هستند .

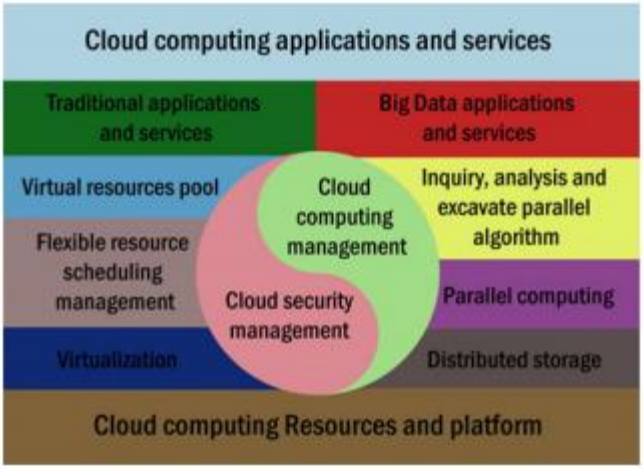
۱-۱-۳رابطه بین رایانش ابری و کلان داده

رایانش ابری دارای رابطه نزدیکی با کلان داده می باشد. مولفه های کلیدی رایانش ابری در شکل سه نشان داده می شوند. کلان داده را می توان هدف فعالیت متمرکز بر محاسبه محسوب نمود و بر ظرفیت ذخیره سازی سیستمن ابر تاکید می کند. هدف اصلی رایانش ابری این است تا از رایانش عظیم و منابع ذخیره سازی تحت مدیریت متمرکز استفاده نماید تا کاربرد های کلان داده را با ظرفیت رایانش ریزدانه فراهم نماید. توسعه رایانش ابری یک سری راه حل ها را برای ذخیره سازی و پردازش کلان داده فراهم می نماید. از طرف دیگر، ظهور کلان داده باعث تسریع توسعه رایانش ابری می گردد. فناوری ذخیره سازی پراکنده بر اساس رایانش ابری می تواند به طور موثری کلان داده را مدیریت نماید؛ ظرفیت رایانش موازی بر اساس رایانش ابری می تواند کارامدی اکتساب و تحلیل کلان داده را بهبود بخشد .

ولو این که چندین فناوری در در رایانش ابری و کلان داده وجود دارند که دارای هم پوشانی هستند، این فناوری ها در دو جنبه زیر فرق دارند. اولا، مفاهیم تا یک میزان معین متفاوت هستند. معماری فناوری اطلاعات در رایانش ابری تغییر شکل می دهد در حالی کلان داده بر تصمیم گیری تجاری تاثیر می گذارد. از اینرو، کلان داده به رایانش ابری همانند یک زیرساخت اساسی برای فعالیت هموار وابسته است .

ثانیا، کلان داده و رایانش ابری دارای مشتریان هدف متفاوت می باشند. رایانش ابری یک فناوری و محصول می باشد که کارکنان اطلاع رسانی اصلی (CIO) را به عنوان راه حل فناوری اطلاعات پیشرفته هدف قرار می دهد. کلان داده یک محصول می باشد که کارکنان اجرایی ارشد (CEO) را که بر عملیات های کسب و کار تمرکز می کنند، هدف قرار می دهد. نظر به این که تصمیم گیرندگان ممکن است بواسطه رقابت بازار به طور مستقیم تحت فشار قرار گیرند، آنها بایستی مخالفان را در روش های رقابتی تر شکست دهند. این دو فناوری همراه با پیشرفت های کلان داده و رایانش ابری به طور قطع و به طور زیاد شونده به همدیگر گره می خورند. رایانش ابری با کارکرد هایی مشابه با کارکرد های کامپیوتر ها و سیستم های عامل یک سری منابع سطح سیستم را فراهم می کند؛ کلان داده در سطح بالاتر وبا حمایت رایانش ابری کار می کند و عملکرد هایی مشابه با پایگاه های داده و ظرفیت پردازش داده کارامد را فراهم می کند. کسینجر رئیس EMC نشان داده است که کاربرد کلان داده بایستی بر اساس رایانش ابری باشد.

تکامل کلان داده بواسطه رشد سریع تقاضا های کاربرد و رایانش ابری توسعه یافته از فناوری های مجازی می باشد. از اینرو، رایانش ابری نه تنها محاسبه و پردازش کلان داده را فراهم می کند بلکه همچنین خودش یک شیوه سرویس دهی می باشد. پیشرفت های رایانش ابری تا یک میزان معین به بهبود توسعه کلان داده می انجامد که هر دو انها مکمل هم هستند.



تصویر ۲ مولفه های کلیدی محاسبات ابری

۲-۱-۳ رابطه بین اینترنت اشیاء و کلان داده

مقدار بیشمار حسگر های شبکه بندی در الگو اینترنت اشیاء درون انواع دستگاه ها و ماشین ها در جهان واقعی تعبیه می شوند. چنین حسگر هایی که در میدان های مختلف بکار گرفته شده اند می توانند انواع مختلف داده ها نظیر داده های زیست میحطی، داده های جغرافیایی، داده های نجومی و داده های پشتیبانی را راجمع اوری نمایند. تجهیزات موبایل، تاسیسات حمل و نقل، تاسیسات عمومی و ابزار الات خانگی می توانستند همگی از جمله تجهیزات اکتساب داده در اینترنت اشیاء باشند که در تصویر ۳ نمایش داده می شود.

کلان داده ای که از طریق اینترنت اشیاء تولید شده است در مقایسه با کلان داده عمومی و به خاطر انواع مختلف داده های جمع اوری شده با مشخصه های متفاوتی ظاهر شده است که کلاسیک ترین مشخصه ها از بین انها عبارتند از ناهمگنی، تنوع، خصیصه غیر ساخت یافته، نویز و افزونگی بالا. هر چند داده جاری اینترنت اشیاء در کلان داده یک بخش برتر نمی باشد، حسگر های کمیت تا سال ۲۰۱۳ به یک تریلیون خواهند رسید و سپس داده های اینترنت اشیاء مهمترین بخش کلان داده خواهند بود. در گزارش شرکت اینتل اشاره شده است که کلان داده در اینترنت اشیاء از سه مشخصه برخوردار بوده است که با الگو کلان داده مطابقت دارد:

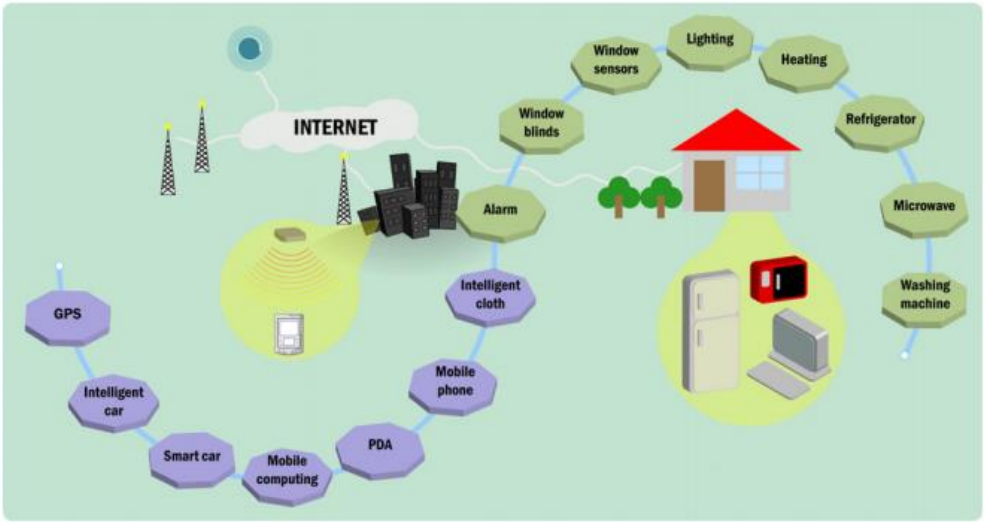
(۱) ترمینال های فراوان تولید کننده حجم های زیاد داده؛

(۲) داده تولید شده از اینترنت اشیاء به طور معمول نیمه ساخت یافته یا غیر ساخت یافته می باشد؛ (۳) داده اینترنت اشیاء تنها زمانی که تحلیل می گردد، سودمند است.

اکنون ظرفیت پردازش داده اینترنت اشیاء کمتر از داده های جمع اوری شده می باشد و تسریع در معرفی فناوری های کلان داده برای بهبود توسعه اینترنت اشیاء بی نهایت اضطراری می باشد. تعدادی از اپراتور های اینترنت اشیاء از اهمیت کلان داده اگاهند چون موفقیت اینترنت اشیاء به یکپارچه سازی موثر کلان داده و

رایانش ابری وابسته می باشد. بکار گیری گسترده اینترنت اشیاء یک سری شهر ها را وارد عصر کلان داده خواهد نمود.

پذیرش کلان داده برای کاربرد های اینترنت اشیاء یک نیاز اجباری می باشد در حالی که توسعه کلان داده از قبل با تاخیر می باشد. شناخت در مورد این موضوع گسترش یافته است که این دو فناوری وابسته به هم هستند و بایستی به طور مشترک توسعه یابند: از یک طرف، بکار گیری گسترده اینترنت اشیاء رشد بالا داده را در مقدار و دسته هدایت می کند از اینرو فرصتی برای کاربرد و توسعه کلان داده فراهم می گردد ؛ از طرف دیگر، کاربرد فناوری کلان داده با اینترنت اشیاء می تواند پیشرفت های جستجو و مدل های کسب و کار اینترنت اشیاء را تسریع نماید.



تصویر۳ : نمایش تجهیزات اکتساب داده در اینترنت اشیاء

۳-۱-۳ مرکز داده

مرکز داده در الگو کلان داده نه تنها یک پلت فرم برای ذخیره سازی متمرکز داده می باشد بلکه همچنین مسئولیت های بیشتر نظیر کسب داده، اداره کردن داده و استفاده ابزاری از مقادیر و کارکرد های داده رابرعهده می گیرد. مراکز داده عمدتا نگران "داده" و نه "مرکز " هستند. مرکز دارای حجم های زیاد داده می باشد و داده ها را مطابق با هدف هسته اش و مسیر توسعه سازماندهی و مدیریت می نماید که نسبت به داشتن سایت و منبع خوب ارزشمند تر است. ظهور کلان داده سبب فرصت های توسعه عالی و چلش زید برای مراکز داده می گردد. کلان داده یک الگوی در حال ظهورمی باشد که رشد انفجاری زیر ساخت و نرم افزار مرکز داده مرتبط را بهبود خواهد بخشید. شبکه مرکز داده فیزیکی یک هسته برای حمایت از کلان داده می باشد اما اکنون یک زیرساخت کلیدی می باشد که به فوریت مورد نیاز است.

کلان داده به مرکز داده ای نیاز دارد که پشتیبانی پشت صحنه قدرتمند را مهیاء می نماید. الگوی کلان داده دارای الزامات سخت تر در ظرفیت ذخیره سازی و ظرفیت پردازش و همچنین ظرفیت انتقال شبکه بوده است. شرکت ها بایستی توسعه مراکز داده را برای بهبود ظرفیت پردازش سریع و موثر کلان داده تحت نسبت قیمت / عملکرد محدود به حساب اورند. مرکز داده بایستی زیرساختی را با تعداد زیاد گره ها فراهم نماید، شبکه داخلی سرعت بالا بسازد، گرما را به طور موثر پخش نماید و دارای داده بک اپ (پشتیبان) موثر باشد. تنها زمانی که یک مرکز داده با بهره وری انرژی بالا، باثبات، ایمن، بسط پذیر و افزونه ساخته می شود، فعالیت عادی اپلیکیشن های کلان داده ممکن است تضمین شود.

رشد اپلیکیشن های کلان داده باعث تسریع در تحول و نواوری مراکز داده می گردد. چندین اپلیکیشن کلان داده توانستند معماری های منحصربه فرد اشان را توسعه دهند و به طور مستقیم توسعه فناوری های ذخیره سازی، شبکه و رایانش مرتبط با مرکز داده را بهبود می بخشند. ظرفیت های رایانش و پردازش داده مرکز داده با رشد دایمی حجم های داده ساخت یافته و غیر ساخت یافته بایستی تا حد قابل توجه ای ارتقاء یابند. بعلاوه، نظر به این که مقیاس مرکز داده به طور فزاینده ای در حال توسعه می باشد، همچنین آن یک موضوع مهم است که چگونه هزینه عملیاتی برای توسعه مراکز داده کاهش یابد.

کلان داده می تواند کارهای بیشتری را به مرکز داده اعطاء نماید. مرکز داده در الگوی کلان داده نه تنها بایستی به امکانات سخت افزاری مرتبط باشد بلکه همچنین ظرفیت های نرم یعنی ظرفیت های اکتساب، پردازش، سازماندهی، تحلیل و کاربرد کلان داده را تقویت نماید. مرکز داده می تواند در تجزیه و تحلیل داده موجود، کشف مشکلات در عملیات کسب و کار و توسعه راه حل ها ازکلان داده به پرسنل کسب و کار کمک کند.

٤-١-٣ رابطه بین هادوپ (Hadoop) و کلان داده

هادوپ در اپلیکیشن های کلان داده در بخش صنعت یعنی فیلترینگ اسپم ایمیل، جستجو شبکه، تجزیه و تحلیل جریان کلیک و پیشنهاد اجتماعی به طور گسترده ای استفاده می شود. علاوه بر این، اکنون تحقیق دانشگاهی قابل توجهی بر اساس هادوپ صورت می گیرد. بعضی موارد نمونه در زیر نشان داده می شوند. همانطور که در ژوئن ۲۰۱۲ بیان شده است، یاهو در ٤۲۰۰۰ سرور در چهار مرکز داده به راه اندازی هادوپ مبادرت ورزید تا از محصولات و خدمات اش یعنی جستجو و فیلترینگ اسپم ایمیل و موارد دیگر پشتیبانی نماید. فعلا، بزرگ ترین خوشه هادوپ دارای ٤۰۰۰ گره بوده است اما تعداد گره ها با انتشار هادوپ ۲ تا ۱۰۰۰۰ عدد افزایش خواهد یافت. فیسبوک در همان ماه اعلان کرده بود که خوشه هادوپ انها می تواند ۱۰۰ پتابایت داده را پردازش نماید که تا ۰,۵ پتابایت در هر روز تا نوامبر ۲۰۱۲ رشد کرده بود . بعلاوه، تعدادی از شرکت ها نوع اجراء تجاری و یا پشتیبانی هادوپ از جمله Cloudera ، IBM ، MapR ، EMC و Oracle را تهیه می کنند.

حسگرها در میان سیستم ها و ماشین آلات صنعتی مدرن به طور گسترده ای بکار رفته اند تا اطلاعات را برای پایش زیست محیطی و پیش بینی شکست و موارد دیگر جمع اوری نمایند . باهاگا و دیگران یک چارچوب را برای سازماندهی داده و زیر ساخت رایانش ابری پیشنهاد کرده اند که CloudView لقب گرفته است . CloudView از معماری های ترکیبی، گره های محلی و خوشه های راه دور بر اساس هادوپ استفاده می کند تا داده های تولید شده ماشینی را تجزیه و تحلیل نماید. گره های محلی برای پیش بینی شکست های انی استفاده می شوند؛ خوشه ها برای تحلیل آفلاین پیچیده یعنی تحلیل داده مورد محور بر اساس هادوپ استفاده می شوند.

رشد نمایی داده های ژنوم و افت شدید هزینه ترتیب دهی باعث تغییر شکل علوم زیستی و بیوپزشکی به علوم داده محور می گردد. گوناراتنه و همکارانش از زیرساخت های رایانش ابری، آمازون AWS، مایکروسافت Azune و چارچوب پردازش داده مبتنی بر MapReduce، هادوپ و مایکروسافت DryadLINQ استفاده کرده اند تا دو اپلیکیشن بیو پزشکی موازی را ران نمایند:

(۱) مونتاژ قطعات ژنوم؛

(۲) کاهش ابعاد در تحلیل ساختار شیمیایی.

مجموعه های داده ۱٦٦ بعدی در اپلیکیشن بعدی استفاده شده اندکه شامل ۲٦۰۰۰۰۰۰ نقطه داده می باشند. محققان عملکرد کل چارچوب ها را بر حسب کارامدی، هزینه و دسترس پذیری مقایسه کرده اند. محققان بر اساس بررسی به نتیجه دست یافته بودند که کوپلینگ ضعیف برای جستجو در ابر الکترون به طور فزاینده ای بکار گرفته خواهد شد و چارچوب فناوری برنامه ریزی موازی (MapReduce) ممکن است یک واسط با خدمات راحت تر و کاهش هزینه های غیر ضروری را برای کاربر فراهم نماید.

۵-۱-۳ تولید و اکتساب کلان داده

ما چندین فناوری کلیدی نظیر رایانش ابری، اینترنت اشیاء، مرکز داده و هادوپ را معرفی کرده ایم که به کلان داده مرتبط هستند. ما در ادامه کار بر زنجیره ارزش کلان داده تمرکز خواهیم کرد که به طور معمول می توان به چهار فاز تقسیم نمود: تولید داده، اکتساب داده، ذخیره سازی داده و تجزیه و تحلیل داده. اگر ما داده را همانند یک ماده خام فرض کنیم، تولید داده و اکتساب داده فرایند بهره برداری هستند، ذخیره سازی داده یک فرایند ذخیره سازی می باشد و تحلیل داده یک فرایند تولید می باشد که از ماده خام برای خلق ارزش جدید بهره می برد.

۳-۲ تولید داده

اولین گام در کلان داده با تولید داده شروع می شود. مقدار عظیم داده بر حسب وردی های جستجو، پست های اجتماع هعای اینترنتی، ثبت های اتاق های گفتکو و پیام های میکروبلاگ با در نظر گرفتن داده اینترنتی به عنوان مثال تولید می شوند. آن داده ها دارای رابطه نزدیک با زندگی روزانه مردم می باشند ودارای خصیصه های مشابه دانسیته کم و ارزش بالا می باشند. چنین داده اینترنتی ممکن است به تنهایی بی ارزش باشد اما اطلاعات سودمند نظیر عادات و سرگرمی های کاربران را از طریق بهره بردرای از کلان داده متراکم می توان شناسایی نمود و حتی این امکان وجود دارد تا رفتاری های کاربران و حالت های عاطفی را پیش بینی نمود.

علاوه بر این، مجموعه داده هایی که از طریق منابع داده طولی و یا پراکنده تولید شده اند دارای مقیاس بزرگ تر، بی نهایت متنوع و پیچیده هستند. چنین منابع داده ای عبارتند از حسگر ها، ویدیو ها، جریان های کلیک و یا کل منابع داده در دسترس دیگر. اکنون، منابع مهم کلان داده عبارتند از اطلاعات تجاری و عملیات در شرکت ها، اطلاعات حسگری و لجستیک در اینترنت اشیاء، اطلاعات فعل و انفعال انسانی و اطلاعات موقعیت در دنیای اینترنتی و داده های تولید شده در تحقیق علمی و غیره. اطلاعات بسیاربهتر از ظرفیت های معماری های فناوری اطلاعات و زیر ساخت های شرکت های موجود هستند در حالی که همچنین نیازمندی زمان واقعی اش تا حد زیادی بر ظرفیت رایانش موجود تاکید می کند.

۳-۲-۱ داده های شرکت

شرکت IBM در سال ۲۰۱۳ یک تحلیلی را صادر کرده است : کاربرد های کلان داده برای جهان واقعی که مشخص می سازئ که داده داخلی شرکت ها از جمله منابع مهم کلان داده می باشند. عمدتا داده های داخلی شرکت ها شامل داده انلاین تجاری و داده تحلیل انلاین می باشند و بخش عمده این داده ها یک سری داده های استاتیک تاریخی هستند و از طریق RDBMS ها در وضعیت ساخت یافته مدیریت می شوند. بعلاوه، داده تولید، داده فهرست موجودی، داده فروش ها و داده مالی و موارد دیگر داده های داخلی شرکت را تشکیل می دهند که در نظر دارد تا فعالیت های داده محور و اطلاعاتی را در شرکت ها را بگیرد تا کل فعالیت های شرکت ها را در شکل داده داخلی ثبت نماید.

فناوری اطلاعات و داده دیجیتالی در چند دهه گذشته تا حد بسیار زیادی در بهبود سوداوری سازمان های تجاری سهیم بوده اند. براورد می گردد که حجم داده های کسب و کار کل شرکت ها در جهان ممکن است در هر ۱,۲ سال به دوبرابر برسد که گردش تجاری در این دوره از طریق اینترنت، شرکت به شرکت و شرکت ها با مصرف کننده ها در هر روز به ٤٥٠ میلیارد دلار خواهد رسید. افزایش دایمی حجم داده تجاری به تحلیل انی موثر تر نیاز دارد تا پتانسیل اش به طور کامل برداشت گردد. برای مثال، میلیون ها عملیات ترمینال و بیش از ۵۰۰۰۰۰ جستجو از فروشنده های شخص ثالث در هر روز در سایت امازون پردازش می گردد. یک میلیون

تجارت مشتری در هر ساعت در سایت شرکت والمارت پردازش می گردد و چنین داده های تجاری وارد پایگاه داده با ظرفیت بیش از ۲,۵ پتابایت می شوند. اکامی تعداد ۷۵ میلیون رویداد را در هر روز برای تبلیغات هدف اش تجزیه و تحلیل می کند.

۲-۲-۳ داده اینترنت اشیاء

همانطور که در بخش های قبلی بحث شده است، اینترنت اشیاء یک منبع مهم برای کلان داده می باشد. کلان داده در میان شهر های هوشمند که بر اساس اینترنت اشیاء ساخته شده اند، می تواند از بخش صنعتی، مشاورزی، ترافیک، حمل و نقل، مراقبت پزشکی، بخش های عمومی و خانواده ها و موارد دیگر ناشی گردد.
معماری شبکه ان مطابق با فرایند های اکتساب داده و انتقال در اینترنت اشیاء می تواند به سه لایه تقسیم گردد : لایه حسگری، لایه شبکه و لایه کاربرد. لایه حسگری مسئول اکتساب داده می باشد و عمدتا از شبکه های حسگر تشکیل می گردد. لایه شبکه مسئولیت انتقال و پردازش داده را برعهده دارد که در انجا انتقال نزدیک ممکن است به شبکه های حسگر تکیه نماید و انتقال راه دور بایستی به اینترنت وابسته باشد. در نهایت ، لایه کاربرد از کاربرد های خاص اینترنت اشیاء پشتیبانی می کند.
داده تولید شده از اینترنت اشیاء مطابق با مشخصه های اینترنت اشیاء از خصیصه های زیر برخوردار بوده است:
داده مقیاس بزرگ : حجم های انبوه تجهیزات اکتساب داده در اینترنت اشیاء به طور پراکنده استفاده می شوند که ممکن است به داده عددی ساده نظیر مکان یا داده چند رسانه ای پیچیده یعنی فیلم ویدیویی جستجو دست یابند. نه تنها داده های جدیدا کسب شده بلکه همچنین داده های تاریخی بر طبق قالب زمانی معین بایستی ذخیره شوند تا تقاضا های تحلیل و پردازش براورده شوند. بنابراین، داده ای که از طریق اینترنت اشیاء تولید شده است از طریق مقیاس های بزرگ شناخته می شوند.
ناهمنگی: داده کسب شده به دلیل دستگاه های اکتساب داده متنوع متفاوت می باشد و چنین داده ای یک ناهمگنی را بطور برجسته نشان داده می شود .
زمان قوی و همبستگی فضا : هر یک از وسیله های اکتساب داده در اینترنت اشیاء در یک مکان جغرافیایی ویژه قرار می گیرند و هر تکه داده دارای مُهر زمانی بوده است. همبستگی زمانی و فضایی از جمله ویژگی های مهم داده از اینترنت اشیاء می باشند. همچنین زمان و فضا در طول تحلیل و پردازش داده از جمله ابعاد مهم برای تحلیل اماری می باشند.
گزارش های داده موثر برای یک بخش کوچک از کلان داده : مقدار زیاد نویز ممکن است در طول اکتساب و انتقال داده در اینترنت اشیاء رخ می دهند. تنها مقدار کمی داده غیر عادی در میان مجموعه های داده کسب شده از طریق دستگاه های اکستاب ارزشمند می باشد. برای مثال، تعداد کمی از فریم های ویدیویی در

طول اکتساب ویدیو ترافیکی که نقض مقررات ترافیک را کشف می کنند و تصادفات ترافیک ارزشمند تر از مواردی هستند که تنها جریان عادی ترافیک را پیدا می کنند.

۳-۲-۳ داده بیو پزشکی

وقتی یک سری فناوری های سنجش زیستی با بازده بالا به طور نواورانه در شروع قرن بیست و یکم توسعه می یابند، تحقیق خط مقدم در مورد رشته بیوپزشکی وارد عصر کلان داده می شود. مکانیزم حکمفرما در پشت پدیده بیولوژیکی پیچیده از طریق ساخت سیستم های تئوریک و مدل های تحلیلی هوشمند، کارامد و دقیق ممکن است آشکار گردند. نه تنها توسعه آینده بیوپزشکی را می توان مشخص نمود بلکه همچنین نقش های مهم را می توان در توسعه سری هایی از صنایع استراتژیک مهم مرتبط با اقتصاد ملی، معیشت مردم و امنیت ملی همراه با کاربرد های مهم نظیر مراقب پزشکی، تحقیق و توسعه دارو جدید و تولید غلات فرض نمود.

تکمیل HGP (پروژه ژنوم انسانی) و توسعه ادامه دار فناوری ترتیب گذاری نیز به کاربرد های گسترده کلان داده در میدان منجر می گردند. توده های داده تولید شده از طریق ترتیب گذاری ژن کار تحلیل تخصصی رل مطابق با تقاضا های متفاوت کاربرد انجام می دهند تا ان را با تشخیص ژن کلینیکی ترکیب کنند و اطلاعات ارزشمند را برای تشخیص اولیه و درمان شخصی بیماری فراهم می کنند. یک ترتیب گذاری ژن انسانی ممکن است ۱۰۰ داده خام ٦۰۰ گیگا بایتی را تولید نماید. در بانک ژن ملی چین در شن ژن در حدود ۱,۳ میلیون نمونه از جمله ۱,۱۵ میلیون نمونه انسانی و ۱۵۰۰۰۰ نمونه حیوانی ، گیاهی و میکروارگانیسم وجود دارند. تعداد ۱۰ میلیون نمونه بیولوژیک قابل ردیابی تا آخر سال ۲۰۱۳ ذخیره خواهند شود و این ویژگی تا آخر سال ۲۰۱۵ به ۳۰ میلیون خواهد رسید. قابل پیش بینی است که ترتیب دهی ژن با توسعه فناوری های بیوپزشکی سریع تر و راحت تر خواهد شد و از اینرو کلان داده بیوپزشکی را به طور دایمی فراتر از همه ابهامات رشد می دهد.

بعلاوه، داده تولید شده از مراقبت بهداشتی و تحقیق و توسعه پزشکی به سرعت ترقی می کند. برای مثال، دانشگاه مرکز پزشکی پیتزربوگ (UPMC) در حدود ۲ ترابایت داده را ذخیره کرده است. شرکت امریکایی Explorys یک پلت فرم را برای جمع اوری داده کلینیکی، عملیات و داده تعمیر و نگهداری و داده مالی فراهم می کند. اکنون اطلاعات ۱۳ میلیون نفر با ٤٤ مقاله داده در مقیاس حدود ٦۰ ترابایتی جمع اوری شده اند که به ۷۰ ترابایت در سال ۲۰۱۳ خواهد رسید. دیگر شرکت امریکایی یعنی Practice Fusion در زمینه مدیریت ثبت های پزشکی الکترونیکی حدود ۲۰۰۰۰۰ بیمار فعالیت می کند.

دیگر شرکت های شناخته شده حوزه فناوری اطلاعات نظیر گوگل، مایکروسافت و IBM صرف نظر از چنین شرکت های اندازه کوچک و متوسط به طور گسترده ای در زمینه تحلیل محاسباتی و تحقیق روش های مرتبط با کلان داده بیولوژیکی بازده بالا برای سهام در بازار عظیم سرمایه گذاری کرده اند که معروف به Next Internet می باشد. IBM در کنفرانس استراتژی سال ۲۰۱۳ پیش بینی می کند که حرفه ای های پزشکی با افزایش زیاد

تصاویر پزشکی و ثبت های پزشکی الکترونیکی ممکن است از کلان داده برای استخراج اطلاعات کلینیکی سودمند از حجم های داده استفاده می کنند تا به سابقه پزشکی و دست یابند و اثرات درمان را پیش بینی نمایند از اینرو مراقبت بیمار بهبود یافته و هزینه کاهش می یابد. پیش بینی می گردد که میانگین حجم داده هر بیمارستان تا سال ۲۰۱۵ از ۱٦۷ ترابایت به ٦٦٥ ترابایت افزایش خواهد یافت.

٤-۲-۳ تولید داده از دیگر فیلد ها

وقتی کاربرد های علمی در حال افزایش هستند ، مقیاس مجموعه های داده به تدریج توسعه می یابد و توسعه بعضی اصول تا حد زیادی به تحلیل حجم های داده وابسته می باشد. در اینجا، ما چندین کاربرد را بررسی می کنیم. کاربرد ها هر چند در فیلد های علمی مختلف وجود دارند، دارای تقاضا مشابه و زیاد شونده در تحلیل داده می باشند. اولین مثال به بیولوژی محاسباتی ربط دارد. بانک ژن یک پایگاه داده متوالی نوکلوئتید می باشد که توسط مرکز نواوری بیوتکنولوژی ملی ایالات متحده نگهداری شده است. داده در این پایگاه داده ممکن است هر ده ماه دوبرابر شود. بانک ژن تا آگوست ۲۰۰۹ دارای بیش از ۲۵۰ میلیارد پایگاعه از ۱۵۰۰۰۰ موجودات زنده متفاوت بوده است. دومین مثال به ستتاره شناسی ربط دارد . SDSS[1] که بزرگ ترین پروژه بررسی آسمان در ستاره شناسی می باشد در حدود ۲۵ ترابایت داده را از سال ۱۹۹۸ تا ۲۰۰۸ ثبت کرده است. وقتی وضوح تلسکوپ تا سال ۲۰۰٤ بهبود یافته است، حجم داده تولید شده در شب بهتر از ۲۰ترابایت خواهد بود. اخرین کاربرد به فیزیک پرانرژی ربط دارد. آزمایش اطلس از LHC سازمان اروپایی برای تحقیق هسته ای در شروع سال ۲۰۰۸ به تولید داده خام در دو پتابایت بر ثانیه منجر می گردد و در حدود ۱۰ ترابایت داده پردازش شده در هر سال ذخیره می گردد.

بعلاوه، حسگری و رایانش غیرفعال در سراسر طبیعت، تجارت، اینترنت، دولت و محیط های اجتماعی در حال تولید داده ناهمگن با پیچیدگی بی سابقه هستند. این مجموعه های داده دارای مشخصه های داده منحصر به فرد در مقیاس، ابعاد زمانی و دسته داده می باشند. برای مثال، داده های موبایل با توجه به موقعیت ها، حرکت، درجات نزدیکی، ارتباطات، چند رسانه ای، استفاده از اپلیکیشن ها و محیط صورتی ثبت گردیدند. این قبیل مجموعه های داده درون دسته های متفاوت برای انتخاب راه حل های مناسب و امکان پذیر برای کلان داده مطابق با محیط کاربرد و الزامات می باشند.

٥-۲-۳ اکتساب کلان داده

اکتساب کلان داده به عنوان دومین فاز سیستم کلان داده بزرگ شامل جمع اوری داده، انتقال داده و پیش پردازش داده می باشد. یک زمانی که ما در طول اکتساب کلان داده یک سری داده خام را جمع می کردیم می

[1] Sloan Digital Sky Survey

بایستی از مکانیزم انتقال کارامد برای ارسال ان به سیستم مدیریت ذخیره سازی مناسب استفاده نماییم تا از کاربرد های تحلیل متفاوت حمایت کنید. مجموعه های داده جمع اوری شده ممکن است در بعضی اوقات شامل داده افزونه بسیار زیاد یا داده بی استفاده باشند که فضای ذخیره سازی را به صورت غیر ضروری افزایش می دهند. برای مثال، افزونگی بالا در میان مجموعه های داده که از طریق حسگر ها برای پایش محیطی جمع آوری شده اند بسیار متداول می باشد. فناوری فشرده سازی داده را می توان برای کاهش افزونگی اعمال نمود. بنابراین، عملیات های پیش پردازش داده برای تضمین ذخیره سازی و بهره برداری کافی داده الزامی هستند.

٦-٢-٣جمع اوری داده

جمع اوری داده عبارتست از بهره برداری از تکنیک های ویژه جمع اوری داده برای کسب داده خام از محیط تولید داده خاص. چهار روش جمع اوری داده در زیر نشان داده می شوند که عبارتند از:

پرونده های ثبت (log file): پرونده های ثبت که به عنوان یکی از روش های جمع اوری داده استفاده شده اند یک سری فایل ها را به صورت خودکار ثبت می کنند که از طریق سیستم منبع داده برای ثبت فعالیت ها در فرمت های فایل تعیین شده برای تحلیل بعدی ثبت می شوند. برای مثال، تعداد کلیک ها، نرخ های کلیک کردن، بازدید ها و دیگر ثبت های ویژگی کاربران وب در وب سرو های ثبت می شوند. وب سرور ها برای ضبط و ثبت فعالیت های کاربران در سایت های وب عمدتا سه فرمت پرونده ثبت را منظور می کنند که عبارتند از: فرمت پرونده ثبت عمومی (NSCA)، فرمت ثبت توسعه یافته (W٣C) و فرمت ثبت IIS (مایکرو سافت). همه این سه نوع پرونده های ثبت ممکن است در بعضی مواقع برای ذخیره سازی اطلاعات ثبت برای بهبود کارامدی جستجو حجم زیاد ذخیره ثبت استفاده شوند. همچنین تعدادی از دیگر پرونده های ثبت بر اساس جمع اوری داده از جمله شاخص های سهام در کاربرد های مالی و تعیین وضعیت های عملیاتی در پایش شبکه و مدیریت ترافیک وجود دارند .

حسگری: حسگر ها در زندگی روزانه برای اندازه گیری کمیت های فیزیکی تبدیل پردازش بعدی (ذخیره سازی) متداول هستند. داده حسگری ممکن است به صورت موج صوت، صدا، ارتعاش، خودرو، شیمیایی، جریان، آب و هوا، فشار، دما و موارد دیگر دسته بندی شوند. ثانیا، اطلاعات از طریق شبکه های سیم دار یا بی سیم انتقال می یابد. برای کاربرد هایی نظیر سیستم تجسس ویدیویی که ممکن است به راحتی استفاده و مدیریت گردند، شبکه حسگر سیمی یک راه حل راحت برای کسب اطلاعات مربوطه می باشد. بعضی مواقع موقعیت دقیق پدیده خاص ناشناخته است و بعضی مواقع محیط ذکر شده دارای انرژی یا زیر ساخت های ارتباط نیست. ارتباط بی سیم بایستی برای میسر سازی انتقال داده در میان گره های حسگر تحت قابلیت ارتباط و انرژی محدود استفاده شوند. شبکه های حسگر بی سیم در سال های اخیر به طور چشمگیری علاقمند پیدا کرده اند و برای کاربرد های مختلف نظیر جستجو زیست میحطی، پایش کیفیت آب، مهندسی شهری و پایش

عادات حیوانات وحشی اعمال شده اند. معمولا شبکه حسگر بی سیم شامل تعداد زیاد گره های حسگر پراکنده جغرافیایی می باشد که هر گره یک دستگاه میکرو با توان باتری می باشد. چنین حسگر هایی در موقعیت های تعیین شده بواسطه کاربرد و بر حسب نیاز بکار گرفته می شوند تا داده های حسگری از راه دور را جمع اوری نمایند. هر زمان حسگر ها بکار گرفته می شوند، ایستگاه پایه یک سری اطلاعات کنترل برای پیکربندی/ مدیریت شبکه یا گره های حسگر یا جمع اوری داده ارسال خواهد کرد. داده حسگری بر اساس این قبیل اطلاعات کنترلی در گره های حسگر متفاوت اسمبل می گردد و به عقب برای ایستگاه پایه برای پردازش بیشتر ارسال می شود. علاقمندان برای مطالعه بیشتر به رفرنس برای بحث های دقیق تر رجوع نمایند.

روش های دستیابی به داده شبکه: اکنون انتساب داده شبکه با استفاده از ترکیبی از یک خزنده وب، سیستم قطعه بندی کلمه، سیستم وظیفه و سیستم شاخص و موارد دیگر می باشد. خزنده وب یک برنامه می باشد که توسط موتور های جستجو برای دانلود کردن و ذخیره سازی صفحات وب استفاده شده اند. خزنده وب از یک منبع یاب یکنواخت (URL) از یک صفحه وب اولیه برای دسترسی به دیگر صفحات وب لینک شده شروع می شود که در طول این مدت می تواند URL ها ذخیره ساخته و کل موارد بازیابی شده را مرتب می سازد. خزنده وب به یک URL با ترتیب قبلی از طریق صف URL دست می یابد و سپس صفحات وب را دانلود می نماید و کل URL ها را در صفحات وب دانلود شده شناسایی می نماید و URL های جدید را استخراج می نماید که قرار است در صف قرار داده شوند. این فرایند تا زمانی تکرار می شود که خزنده وب متوقف می شود. اکتساب داده از طریق خزنده وب به طور گسترده ای در کاربرد های مبتنی بر صفحات نظیر موتور های جستجو یا ذخیره سازی موقت داده استفاده می گردد. فناوری های استخراج صفحه وب سنتی یک سری راه حل های کارامد چند تایی را به طور برجسته نشان می دهند و تحقیق شایان توجهی در این حوزه صروت گرفته است. وقتی کاربرد های صفحه وب پیشرفته تر ظاهر می شوند، بعضی راهبرد های استخراج در پیشنهاد می شوند تا از عهده کاربرد های اینترنت غنی برامد.

فناوری های اکتساب داده شبکه فعلی عمدتا شامل فناوری ثبت بسته مبتنی بر Libpcap سنتی، فناوری گرفتن بسته بدون کپی و همچنین تعدادی از نرم افزاری های پایش شبکه تخصصی نظیر Wireshark ، SmartSniff و WinNetCap می باشند .

فناوری ثبت بسته منبتنی بر Libpcap : Libpcap (کتابخانه ثبت بسته) یک نو کتابخانه عملکرد ثبت بسته داده شبکه می باشد که در حد زیادی استفاده می گردد. Libpcap یک ابزار معمولی می باشد که به هر نوع سیستم خاص وابسته نمی باشد و عمد تا برای ثبت داده در لایه لینک دیتا استفاده می شود. این ابزار بخاطر سادگی، استفاده راحت و قابل حمل بودن مورد توجه است اما دارای کارامدی نسبتا پایینی می باشد. از اینرو، تلفات بسته قابل ملاحظه ای در محیط شبکه سرعت بالا ممکن است در زمان استفاده از Libpcap رخ دهد.

فناوری ثبت داده بدون کپی : عبارت به اصطلاح بدن کپی (ZC) بدان معنی است که هیچ نوع کپی برداری در طول دریافت و ارسال در گره رخ نمی دهد. بسته های داده در زمان حسگری به طور مستقیم از بافر کاربر اپلیکیشن ها شروع می شوند ، واسط های شبکه را مرور می نمایند و به شبکه بیرونی می رسند. واسط های شبکه در زمان دریافت به طور مستقیم بسته های داده را به بافر کاربر می فرستند. ایده اصلی بدون کپی این است تا دفعات کپی داده، تماس های سیستم و بار گذاری سی پی یو کاهش یابند در حالی که دیتا گرام ها از تجهیزات شبکه به فضای برنامه کاربر عبور داده می شود. فناوری بدون کپی در ابتدا ازفناوری دسترسی حافظه مستقیم (DMA) استفاده می کندتا دیتاگرام های شبکه را به طور مستقیم به فضای ادرس از قبل تخصیص یافته توسط کرنل(هسته) سیستم ارسال نماید تا ازمشارکت سی پی یو جلوگیری بعمل اورد. در ضمن فناوری می تواند حافظه داخلی دیتاگرام ها را در کرنل سیستم با ان برنامه کشف ترسیم نماید یا منطقه ذخیره در فضای کاربر بسازد و ان را یا فضای کرنل ترسیم نماید. سپس برنامه کشف به طور مستقیم به حافظه داخلی دسترسی دارد تا کپی بردای از حافظه داخلی از کرنل سیستم به فضای کاربر را کاهش دهد و مقدار تماس های سیستم به حداقل برسد.

تجهیزات موبایل: اکنون تجهیزات موبایل به طور گسترده ای استفاده می شوند. وقتی عملکرد های دستگاه موبایل به طور فزاینده ای قدرتمند تر می شود، این وسایل یک سری ابزار پیچیده تر و چند تایی اکتساب داده و همچنین تنوع داده را نشان می دهند. دستگاه های موبایل از طریق سیستم موقعیت یابی به اطلاعات مکان جغرافیایی دست می یابند؛ اطلاعات صوتی را از طریق میکروفن ها می گیرند ؛ به تصاویر، ویدیو ها، اطلاعات خیابان ها، بارکد های دو بعدی و دیگر اطلاعات چند رسانه ای از طریق دوربین ها بدست می اورند؛ اطلاعات ژست ها و دیگر داده های زبان بدن کاربر از طریق صفحه نمایش های لمسی و حسگر های گرانش دست می یابد. اپراتور های بی سیم در عرض چند سال در بهبود سطح سرویس اینترنت موبایل از طریق کسب و تحلیل این قبیل اطلاعات موفق بوده اند. برای مثال، خود آیفون یک جاسوس موبایل می باشد. آیفون می تواندداده های بی سیم و اطلاعات مکان جغرافیایی را شناسایی نماید و سپس این داده ها را برای پردازش به شرکت اپل ارسال می نماید و کاربر از این موضوع مطلع نیست. سیستم های عامل تلفن هوشمند نظیر اندروید گوگل و ویندوز فون مایکروسافت صرف نظر از شرکت اپل می توانند اطلاعات را در وضعیت مشابه جمع اوری نمایند.
چندین روش دیگر جمع اوری داده به همراه سه روش اکتساب داده تشریح شدده وجود دارند. برای مثال، چندین ابزار ویژه را در آزمایشات علمی می توان برای جمع اوری داده آزمایشی نظیر اسپکترومتر های مغناطیسی و رادیو تلسکوپ ها استفاده نمود. ما می توانیم روش های جمع اوری داده را از چشم انداز های متفاوت دسته بندی نماییم. روش های جمع اوری داده را از چشم انداز منابع داده ای می توان به دو دسته تقسیم نمود: روش های جمع اوری ثبت کننده از طریق منابع داده و روش های جمع اوری ثبت کننده از طریق دیگر ابزار کمکی.

۷-۲-۳ حمل داده

داده ها به محض تکمیل جمع اوری داده خام به زیرساخت ذخیره سازی داده برای پردازش و تحلیل انتقال خواهند یافت. همانطور که در بخش ۲-۳ بحث گردید، کلان داده عمدتا در مرکز داده ذخیره می گردد . طرح کلی داده بایستی برای بهبود کارامدی رایانش یا تسهیل نگهداری سخت افزاری تنظیم گردد. از جهات دیگر، انتقال داده داخلی ممکن است در مرکز داده رخ دهد. از اینرو، انتقال داده از دو فاز تشکیل می گردد: انتقال های Inter-DCN و انتقال های Intra-DCN .

انتقال های Inter-DCN: این نوع انتقال ها از منبع داده به مرکز داده صورت می گیرند که معمولا به زیرساخت شبکه فیزیکی موجود دست می یابد. به دلیل رشد سریع تقاضا های ترافیکی، زیرساخت شبکه فیزیکی در اکثر مناطق جهان از سیستم های انتقال فیبر نوری حجم بالا، نرخ بالا و مقرون به صرف تشکیل می شوند. تجهیزات و فناوری های مدیریت پیشرفته نظیر معماری شبکه تسهیم سازی بر اساس تقسیم طول موج (WDM) مبتنی بر IP برای کنترل هوشمند و مدیریت شبکه های فیبر نوری در بیست سال گذشته توسعه یافته اند. WDMیک فناوری می باشد که سیگنال های حامل نوری چند تایی را با طول موج های مختلف تسهیم می کند و سپس انها را به فیبر نوری یکسان لینک اپتیکی جفت می نماید. لیزر های با طول های موج متفاوت در چنین فناوری هایی یک سری سیگنال های متفاوت را حمل می نمایند. شبکه های ستون فقراتی به مراتب با سیستم های انتقال اپتیکی WDM با نرخ کانال تکی ٤٠ گیگابیت در ثانیه استفاده شده اند. واسط های عددی ۱۰۰ گیگابیت در ثانیه در زمان فعلی در آینده نزدیک در دسترس خواهند بود. از اینرو، فناوری های انتقال اپتیکی سنتی بواسطه پهنای باند گلوگاه الکترونیکی محدود می شوند. اخیرا، مدولاسیون تقسیم فرکانس عمود برهم (OFMD) که در ابتدا برای سیستم های بی سیم طراحی شده بود، به عنوان یکی از فناوری های کاندید مهم برای انتقال اپتیکی سرعت بالا اینده مورد توجه می باشد. OFMD یک فناوری موازی چند حامل می باشد. این فناوری جرزیان داده سرعت بالا را برای تبدیل ان به جریان خای داده فرعی سرعت پایین دسته بندی می کند که قرار است بر روی حامل های جانبی قائم چند تایی انتقال یابند. OFDM که با فاصله بندی کانال ثابت WDM مقایسه شده است به یک سری طیف های فرکانس کانال فرعی اجازه می دهد تا با همدیگر هم پوشانی داشته باشند. از اینرو، این فناوری انعطاف پذیر و از نوع شبکه بندی اپتیکی کارامد می باشد.

Intra-DC: انتقال های Intra-DCN از نوع جریان های ارتباط داده درون مراکز داده می باشند. انتقال های Intra-DCN به مکانیزم ارتباط درون مرکز داده وابسته هستند (یعنی در پلیت های اتصال فیزیکی، تراشه ها، حافظه های داخلی سرور های داده، معماری های شبکه مراکز داده و پروتکل های ارتباط). مرکز داده شامل قفسه های سرور یکپارچه چند تایی می باشد که بع شبکه های داخلی اش به همدیگر متصل هستند. این روز ها، شبکه های اتصال داخلی اکثر مراکز داده ها دارای ساختار های Fat Tree (شبکه جهانی برای ارتباط کارامد)،

دو لایه یا سه لایه و مبتنی بر جریان های شبکه چند محصولی می باشند. قفسه ها در ساختار جانمایی دو لایه از طریق سوئیچ های قفسه بالا (TOR) یک گیگا بین بر ثانیه وصل می شوند و سپس این سوئیچ های قفسه بالا با سوئیچ های متمرکز ۱۰ گیگا بایت بر ثانیه در ساختار های مکان نگر متصل می شوند. ساختار مکان نگر سه لایه ای از نوعی است که با یک لایه در راس ساختار مکان نگر دو لایه تقویت می گردد و چنین لایه ای از سوئیچ های هسته ۱۰ گیگابایت بر ثانیه تا ۱۰۰ گیگابایت تشکیل می گردد تا سوئیچ های متمرکز را در ساختار مکان نگر متصل نماید. همچنین دیگر ساختار های مکان نگر وجود دارند که بهبود شبکه های مرکز داده را می توان هدف انها محسوب نمود. به خاظر کفایت سوئیچ های بسته الکترونیکی، افزایش پهنا های باند ارتباط دشوار می باشد در حالی که مصرف انرژی را در حد کم نگه می دارد. اتصال داخلی اپتیکی در بین شبکه ها در عرض سال های گذشته به دلیل موفقیت شایانی که از طریق فناوری های اپتیکی بدست امده بود باعث جلب توجه زیادی شده است. اتصال داخلی اپتیکی یک راه حل مصرف انرژی بازده بالا، تاخیر کم و کم انرژی می باشد. اکنون فناوری های اپتیکی تنها برای لینک های نقطه به نقطه در مراکز داده استفاده می شوند. این قبیل لینک های اپتیکی یک نوع اتصال را برای سوئیچ هایی که از فیبر چند حالتی کم هزینه (MMF) با نرخ داده ۱۰ گیگابایت بر ثانیه استفاده می کنند، فراهم می نمایند. اتصال داخلی اپتیکی (سوئیچینگ در حوزه اپتیکی) شبکه ها در مراکز داده یک راه حل امکان پذی می باشد که می تواند پهنای باند انتقال سطح Tbs (ترابایت بر ثانیه) را با مصرف انرژی کم فراهم نماید. اخیرا، تعدادی از طرح های اتصال داخلی برای شبکه های مرکز داده پیشنهاد می شوند. تعدادی از طرح ها یک سری مسیر های اپتیکی را برای به روز رسانی شبکه های موجود پیشنهاد می دهند و طرح های دیگر به طور کامل سوئیچ های فعلی را تعویض مینمایند. ژو و همکارانش با لینک های بی سیم در باند فرکانس ٦٠ گیگاد هرتز برای تقویت لیک های سیمی به عنوان یک فناوری تقویت کننده موافقت می کنند. همچنین مجازی سازی شبکه بایستی برای بهبود کارامدی و بهره برداری از شبکه های مرکز داده مورد توجه قرار گیرد.

۸-۲-۳ پیش پردازش داده

مجموعه های داده جمع اوری شده به دلیل تنوع زیاد منابع داده با توجه به نویز، افزونگی و پایداری و موارد دیگر فرق دارند و ذخیره سازی داده های بی معنی بدون شک یک کار بیهوده است. بعلاوه، بعضی روش های تحلیلی دارای الزامات جدید در مورد کیفیت داده می باشند . ازاینرو ، ما با هدف انجام تحلیل داده موثر بایستی داده را تحت بعضی شرایط پیش پرازش نماییم تا داده های منابع مختلف را یکپارچه نماییم که نه تنها می توان هزینه ذخیره سازی را کاهش داد بلکه همچنین می توان دقت تحلیل را بهبود بخشید . بعضی تکنیک های پیش پردازش داده رابطه ای در زیر مورد بحث قرار می گیرند :

یکپارچه سازی: یکپارچه سازی داده را می توان سنگ بنای اطلاع رسانی تجاری مدرن محسوب نمود که شامل ترکیبی از داده های منابع مختلف می باشد و بازنگری متحد الاشکلی را از داده برای کاربران تهیه می کند. این یک حوزه تحقیق بالغ برای پایگاه داده سنتی می باشد. دو روش از نظر تاریخی به طور گسترده ای شناخته شده هستند: انبار داده و فدارسیون داده. انبار کردن داده شامل فرآیندی می باشد که ETL (استخراج ، تبدیل و بارگذاری) نامیده شده است. استخراج داده شامل وصل شدن به سیستم های منبع، انتخاب، جمع اوری، تحلیل و پردازش داده های ضروری می باشد. تبدیل عبارتست از اجرای سری هایی از قواعد برای تبدیل داده های استخراج شده به فرمت های استاندارد. بارگذاری به معنی وارد کردن دادهای استخراج شده و تبدیل شده به زیرساخت ذخیره سازی هدف می باشد. بارگذاری پیچیده ترین رویه در میان سه روش فوق است که شامل عملیات هایی نظیر تبدیل، کپی برداری، تعویض داده، استاندارد سازی، غربالگری و سازماندهی داده می باشد . پایگاه داده مجازی را می توان برای صف بندی و تجمیع داده از منابع داده متفاوت ساخت اما چنین پایگاه داده ای حاوی داده نیستند. برعکس، پایگاه داده مجازی شامل اطلاعات یا ابر داده مرتبط با داده واقعی و موقعیت هایش می باشد. این دو رویکرد برای خواندن انبار داده نمی توانند الزامات عملکرد بالا جریان های داده یا برنامه های جستجو و کاربرد ها را براورده سازند. داده در این دو رویکرد در مقایسه با صف ها به صورت دینامیک می باشد و بایستی در طول انتقال داده پردازش گردد. معمولا، روش های یکپارچه سازی داده با موتور های پردازش جریان و موتور های جستجو همراه می شوند.

تعویض داده: تعویض داده یک فرایندی است که داده های غیر دقیق ، ناقص یا غیر منطقی را شناسایی می کند و سپس این داده ها را برای بهبود کیفیت داده اصلاح یا پاک می کند. معمولا، پاکسازی داده شامل پنج تکنیک مکمل می باشد که عبارتند از: تعریف و تعیین انواع خطا، جستجو و شناسایی خطا ها، اصلاح خطا ها، مستند سازی مثال های خطا و انواع خطا و اصلاح تکنیک های ورودی داده برای کاهش خطا های آینده. فرمت داده ها، کامل بودن، عقلانیت و محدودیت بایستی مورد بررسی قرار گیرند. پاکسازی داده برای حفظ پایداری داده بسیار مهم است که به طور گسترده ای در حوزه های مختلف نظیر بانکداری، بیمه، صنعت خرده فروشی، ارتباطات راه دور و کنترل ترافیک استفاده می شود.

بخش عمده در تجارت الکترونیکی به طور الکترونیکی جمع اوری می گردد که ممکن است دارای مشکلات کیفیت داده جدی باشند . عمدتا مشکلات کیفیت داده کلاسیک از کمبود های نرم افزاری، خطا های بومی شده یا پیکر بندی اشتباه سیستم ناشی می گردند. محققان در مورد پاکسازی داده در تجارت الکترونیک از طریق خزنده ها و کپی کردن مجدد منظم اطلاعات حساب و مشتری بحث کرده اند.

مشکل پاکسازی داده RFID مورد بررسی قرار گرفته بود. RFID در چندین کاربرد مورد استفاده قرار می گیرد که برای مثال می توان از مدیریت فهرست موجودی و ردیابی هدف نام برد. از اینرو، RFID اصلی دارای مشخصه هایی نظیر کیفیت پایین می باشد که شامل مقادیر زیاد داده غیر عادی محدود شده از طریق طراحی

فیزیکی می باشد و تحت تاثیر نویز های زیست محیطی می باشد . مدل احتمالی توسعه یافته بود تا اتلاف داده در محیط های موبایل برطرف گردد. خواسنویوا و همکارانش یک سیستم را برای اصلاح خودکار خطا های داده ورودی از طریق تعریف محدودیت های یکپارچه سازی داده پیشنهاد داده اند.

هربرت و همکارانش یک چارچوبی را بنام BIO-AJAX پیشنهاد داده اند تا داده های بیولوژیک را برای انجام محاسبه بیشتر و بهبود کیفیت جستجو استاندارد سازند. بعضی خطا ها و تکرار ها با BIO-AJAX ممکن است حذف گردند و فناوری های داده کاوی مشترک را می توان به طور موثر تری اجراء نمود.

حذف افزونگی: افزونگی داده به تکرارها یا مازاد داده گفته می شود که معمولا در چندین پایگاه داده رخ می دهد. افزونگی داده می تواند هزینه انتقال داده غیر ضروری را افزایش دهد و سبب کمبود هایی در سیستم های ذخیره سازی یعنی هرز فضای ذخیره، ناسازگاری داده، کاهش قابلیت اطمینان داده و آسیب به داده می گردد. از اینرو، انواع روش های کاهش افزونگی نظیر کشف افزونگی، فیلترینگ داده و فشرده سازی داده پیشنهاد شده اند. این قبیل روش ها ممکن است برای مجموعه های داده مختلف یا محیط های کاربردی اعمال گردند. از اینرو، کاهش افزونگی ممکن است سبب اثرات منفی معین شود. برای مثال، فشرده سازی و غیر فشرده سازی داده سبب زحمت محاسباتی اضافی می شوند. بنابراین، مزیت های کاهش افزونگی و هزینه بایستی به دقت متوازن گردد. داده های جمع اوری شده از حوزه های مختلف به طور فزاینده ای در فرمت های ویدیویی یا تصویری ظاهر خواهند شد. معروف است که تصاویر و ویدیو ها حاوی افزونگی قابل ملاحظه ای نظیر افزونگی موقتی، افزونگی سریالی، افزونگی اماری و افزونگی حسگری هستند. فشرده سازی ویدیویی به طور گسترده ای برای کاهش افزونگی در داده ویدیویی استفاده می گردد که در چندین استاندازد کد گذاری ویدیویی (MPEG-۲,MPEG-٤,H.٢٦٣, H.٢٦٤/AVC) مشخص شده اند محققان در رفرنس [٧٤] به بررسی مشکل فشرده سازی ویدیویی در سیستم تجسس ویدیویی با شبکه حسگر ویدیویی پرداخته اند. محققان یک روش جدید مبتنی بر MPEG-٤ را از طریق بررسی افزونگی بافتی مرتبط با پس زمینه و پیش زمینه در صحنه پیشنهاد می دهند. نسبت پیچیدگی و فشرده سازی پایین رویکرد پیشنهادی از طریق نتایج ارزیابی ثابت شده بودند .

حذف داده تکراری در انتقال یا ذخیره سازی داده تعمیم یافته یک فناوری فشرده سازی داده خاص می باشد که در نظر دارد تا کپی های داده تکراری را حذف نماید. بلوک های داده انفرادی یا قطعات داده با حذف داده های تکراری به شناسایی کننده (برای مثال استفاده از الگوریتم درهم سازی) تخصیص می یابند. اگر بلوک داده جدید دارای یک شناسایی کننده ای بوده است که همانند شناسایی کننده فهرست شده در لیست شناسایی می باشد، وقتی تحلیل حذف داده تکراری ادامه می باشد یک بلوک داده جدید به عنوان افزونه فرض خواهد شد و با بلوک داده ذخیره شده مشابه تعویض خواهد گردید. حذف داده تکراری می توان تا حد زیادی الزام ذخیره سازی را کاهش دهدکه برای سیستم ذخیره سازی کلان داده به طور خاص اهمیت دارد. اهداف داده خاص صرف نظر از روش های پیش پردازش داده فوق الذکر بایستی یک سری عملیات های دیگر نظیر استخراج

خصیصه را انجام دهند. این قبیل عملیات ها نقش مهمی در جستجو چند رسانه ای و تحلیل DNA ایفاء می کنند. معمولا، بردار های خصیصه چند بعدی (یا نقاط خصیصه ابعادی بالا) برای توصیف این قبیل اهداف داده استفاده می شوند و بردار های خصیصه ابعادی برای بازیابی توسط سیستم ذخیره می شوند. معمولا انتقال داده برای پردازش منابع داده ناهمگن پراکنده بویژه مجموعه های داده تجاری استفاده می شود. در واقع، ایجاد یک تکنیک پیش پردازش داده متحد الاشکل که برای کل انواع مجموعه های داده اجراء شدنی می باشد در بررسی مجموعه های داده مختلف غیر ممکن یا غیر بدیهی می باشد. مشکل، الزامات عملکردی و دیگر فاکتور های مجموعه های داده در خصیصه خاص بایستی مورد توجه قرار گیرد تا استراتژی پیش پردازش داده مناسب انتخاب گردد.

۳-۳ جمع بندی

در این فصل به مروری بر کارهای انجام شده پرداختیم. همان‌طور که گفته شد فناوریهای مختلف و تولید داده های مختلف را مورد بررسی قرار دادیم

فصـــل چهارم

نحوه ذخیره سازی و کاربردهای کلان داده

مقدمه

۱-٤ ذخیره سازی کلان داده

رشد انفجاری داده یک سری الزامات شدید تر در زمینه ذخیره سازی و مدیریت را را به همراه داشته است. ما دربخش حاضر بر ذخیره سازی کلان داده تمرکز می کنیم. ذخیره سازی کلان داده به ذخیره سازی و مدیریت مجموعه های داده مقیاس بزرگ گفته می شود در حالی که به قابلیت اطمینان و دسترس پذیری دسترسی داده دست می یابیم. ما موضوعات مهم از جمله سیستم های ذخیره سازی حجیم، سیستم های ذخیره سازی پراکنده و مکانیزم های ذخیره سازی کلان داده را بازنگری خواهیم کرد .از یک طرف، زیرساخت ذخیره سازی به تهیه سرویس ذخیزه سازی اطلاعات با فضای ذخیره سازی مطئمن نیاز دارد؛ از طرف دیگر بایستی واسط دسترسی قدرتمند را برای صف بندی و تحلیل مقدار زیاد داده فراهم نماید.

وسیله ذخیره سازی داده به طور سنتی به عنوان تجهیزات کمکی سرور استفاده می شود تا داده با RDBMS های ساخت یافته را ذخیره، مدیریت، جستجو و تحزیه و تحلیل نماید. از اینرو، نیاز اجباری برای تحقیق در مورد ذخیزه سازی داده وجود دارد.

۱-۱-٤ سیستم ذخیره سازی برای داده های حجیم

انواع سیستم های ذخیره سازی برای براورده سازی تقاضا های داده حجیم معرفی می شوند. فناوری های ذخیره سازی داده حجیم فعلی را می توان به صورت ذخیره سازی پیوست شده مستقیم (DAS) و ذخیره سازی شبکه دسته بندی نمود در حالی که ذخیره سازی شبکه را می توان به ذخیره سازی پیوست شده شبکه (NAS) و شبکه ذخیره سازی (SAN) دسته بندی نمود.

انواع هارددیسک ها در DAS به طور مستقیم به سرور ها وصل می شوند و مدیریت داده متمرکز بر سرور می باشد به نحوی که دستگاه های ذخیره سازی ی سری تجهیزات پیرامونی هستند که هر یک از انها یک مقدار معین از منبع I/O را می گیرند و از طریق طریق نرم افزار اپلیکیشن انفرادی مدیریت می گردد. DAS به همین خاطر تنها برای وصل شدن داخلی به سرور های با مقیاس کوچک مناسب می باشد. از اینرو، DAS به دلیل مقیاس پذیری اش یک نوع راندمان غیر مطلوب را در زمانی که ظرفیت ذخیره سازی افزایش می یابد نمایش خواهد داد و برای مثال قابلیت به روز رسانی و بسط پذیری تا حد زیادی محدود می شوند. از اینرو، عمدتا DAS در کامپیوتر های شخصی و سرور های اندازه کوچک استفاده می گردد.

ذخیره سازی شبکه از شبکه استفاده می کند تا یک واسط یکسان رابرای دسترسی و اشتراک داده برای کاربران تهیه نماید. تجهیزات ذخیره سازی داده شبکه شامل تجهیزات مبادله داده، آرایه دیسک، آرایه نوار و دیگر رسانه های ذخیره سازی و همچنین نرم افزار ذخیره سازی خاص می باشد.

NAS به طور واقعی یک وسیله ذخیره سازی کمکی شبکه می باشد. NAS از طریق یک هاب یا سوئیچ بوسیله پروتکل های TCP/IP به طور مستقیم به شبکه وصل می گردد. داده در NAS در شکل فایل ها ارسال می گردد. دشواری I/O در سرور NAS در مقایسه با DAS تا حد زیادی کاهش می یابد چون سرور به طور غیر مستقیم از طریق شبکه به وسیله ذخیره سازی دسترسی می یابد.

در حالی که NAS شبکه محور می باشد، SAN به طور ویژه برای ذخیره سازی داده با شبکه متمرکز بر پهنای باند و مقیاس پذیر یعنی شبکه سرعت بالا با اتصالات فیبر نوری طراحی می گردد. مدیریت ذخیره سازی داده در SAN نسبتا مستقل از شبکه منطقه محلی ذخیره سازی می باشد که در انجا سوئیچینگ داده چند مسیری در میان هر نوع گره داخلی مورد استفاده قرار می گیرد تا به حداکثر مقیاس تسهیم داده و مدیریت داده دست یافت.

DAS، NAS و SAN از سازماندهی سیستم ذخیره سازی داده می توان به سه بخش تقسیم نمود:

(۱) آرایه دیسک : آرایه دیسک را می توان اساس و زیرساخت سیستم ذخیره سازی و تضمین اساسی برای ذخیره سازی داده به حساب اورد.

(۲) اتصال و سیستم های فرعی شبکه که اتصال بین یک یاچند آرایه دیسک و سرور را فراهم می کنند.

(۳) نرم افزار مدیریت ذخیره سازی که به تسهیم داده، ریکاوری در زمان خطر و دیگر وظایف مدیریت ذخیره سازی سرور های چند تایی رسیدگی می کند.

۲-۱-٤ سیستم ذخیره سازی پراکنده

اولین چالش که بواسطه کلان داده ایجاد شده بود این است که چگونه سیستم ذخیره سازی پراکنده مقیاس بزرگ را برای تحلیل و پردازش داده کارامد توسعه داد. فاکتور های زیر برای استفاده از سیستم پراکنده برای ذخیره سازی داده حجیم بایستی به حساب اورده شوند:

۱-۲-۱-٤ پایداری:

سیستم ذخیره سازی پراکنده به سرور های چند تایی نیاز دارد تا به طور همکارانه داده را ذخیره نماید. وقتی سرو های بیشتر وجود داشته باشند، احتمال شکست سرور ها بیشتر خواهد بود. معمولا داده به قطعات چند تایی تقسیم می گردد تا در سرور های متفاوت ذخیره شوند تا دسترس پذیری در مورد شکست سرور را تضمین نمایند. از اینرو، شکست های سرور و ذخیره سازی موازی ممکن است سبب ناپایداری در میان کپی های متفاوت داده یکسان شوند. پایداری به تضمینی گفته می شود که کپی های چند تایی داده یکسان به یک شکل هستند.

۲-۲-۱-٤ دسترس پذیری:

سیستم ذخیره سازی پراکنده در مجموعه های سرور چند تایی کار می کند. وقتی سرورهای بیشتری استفاده می شوند، شکست های سرورها اجتناب ناپذیر می گردند. اگر کل سیستم به طور جدی برای براورده سازی درخواست های مشتری بر حسب خواندن و نوشتن تحت تاثیر قرار نگیرد این وضعیت مطلوب خواهد بود. این ویژگی را دسترس پذیری می نامند.

۳-۲-۱-٤ تلرانس تقسیم داده :

سرور های چند تایی در سیستم ذخیره سازی پراکنده از طریق شبکه وصل می شوند. شبکه می توانست دارای شکست های گره/لینک یا هضم موقتی باشد .سیستم پراکنده بایستی داری سطح معین تحمل برای مشکلات ناشی از شکست های شبکه باشد. این وضعیت مطلوب خواهد بود که ذخیره سازی پراکنده هنوز به خوبی در زمانی که شبکه دسته بندی می گردد، کار می کند.

فرضیه CAP در سال ۲۰۰۰ توسط اریک بروئر پیشنهاد شده است که سیستم پراکنده نمی توانست الزامات را در پایداری ، دسترس پذیری و تحمل دسته بندی داده به طور همزمان براورده سازد؛ دو الزام از سه مورد را در اخرین حد می توان به طور همزمان بدست اورد. ساس گیلبرت و نانسی لینچ از MIT صحت فرضیه CAP را در سال ۲۰۰۲ اثبات کرده اند. نظر به این که پایداری، دسترس پذیری و تحمل دسته بندی داده را نمی توان به طور همزمان بدست اورد ، ما با نادیده گرفتن سیستم CA می توانیم تحمل دسته بندی داده را داشته باشیم، با نادیده گرفتن دسترس پذیری می توانیم دارای CP باشیم و سیستم AP را می توانیم در صورتی در اختیار داشته باشیم که برای پایداری اهمیتی قایل نشویم.

سه سیستم در ادامه مورد بحث قرار می گیرند.

سیستم های CA دارای تحمل دسته بندی داده نیستند یعنی این سیستم ها نمی توانستند به شکست های شبکه رسیدگی نمایند. از اینرو، معمولا سیستم های CA همانند سیستم های ذخیره سازی با یک سرور تکی نظیر پایگاه های داده رابطه ای مقیاس کوچک به نظر می رسند. این قبیل سیستم ها یک کپی تکی از داده را نمایش می دهند به نحوی که پایداری به سادگی تضمین می گردد. دسترس پذیری از طریق طراحی عالی پایگاه های داده رابطه ای تضمین می گردد. از اینرو، چون سیستم های CA نمی توانستند به شکست های شبکه رسیدگی کنند در نتیجه نمی توانستند برای استفاده چندین سرور توسعه یابند. از اینرو، اکثر سیستم های ذخیره سازی مقیاس بزرگ از نوع سیستم های CP و AP هستند .

سیستم های CP در مقایسه با سیستم های CA می توانند تحمل دسته بندی داده را تضمین نمایند. از اینرو، سیستم های CP را می توان توسعه داد تا به سیستم های پراکنده تبدیل شوند. معمولا سیستم های CP چندین کپی از داده یکسان را حفظ می کنند تا سطح تحمل خطا را تضمین نمایند. همچنین سیستم های CP می توانند پایداری داده را تضمین نمایند یعنی کپی های چند تایی داده یکسان تضمین می گردند که به طور کامل یکسان باشند. از اینرو، CP نمی توانست دسترس پذیری عالی را به دلیل هزینه بالا تضمین پایداری ضمانت نماید. بنابراین، سیستم های CP برای سناریو های با بار متوسط اما الزامات سخت در دقت داده سودمند هستند. BigTable و Hbase از نمونه های دو سیستم CP معروف هستند.

همچنین سیستم های AP می توانند تحمل دسته بندی داده را تضمین نمایند. از اینرو، سیستم های AP با سیستم های CP در این مورد فرق دارند که سیستم های AP می توانند دسترس پذیری را تضمین نمایند. از اینرو، سیستم های AP تنها پایدرای نهایی را بجای پایداری قوی در دو سیستم قبلی تضمین می کنند. از اینرو، سیستم های AP تنها برای سناریو های با درخواست های متوالی اعمال می گردند اما الزامات بسیار بالا در مورد دقت وجود ندارد. برای مثال، در چندین بازدید همزمان برای داده در سیستم های سرویس های شبکه شبکه سازی اجتماعی انلاین (SNS) یک سری بازید های همزمان برای داده وجود دارند اما مقدار معین خطا های داده تحمل پذیر می باشند. علاوه براین، نظر به این که سیستم های AP یک نوع پایدرای نهایی را تضمین می کنند، داده دقیق را می توان هنوز بعد از مقدار تاخیر معین کسب نمود. بنابراین، سیستم های AP می توانند تحت شرایطی و بدون الزامات انی سخت استفاده گردند. Dynamo و Cassandra از جمله دو سیستم AP شناخته شده می باشند.

۳-۱-٤ مکانیزم ذخیره سازی کلان داده

تحقیق و جستجو چشمگیر در مورد کلان داده باعث بهبود توسعه مکانیزم های ذخیره سازی در این حوزه گردید. مکانیزم های ذخیره سازی موجود کلان داده را می توان به سه سطح از پایین به بالا دسته بندی نمود : (۱) سیستم های فایل، (۲) پایگاه های داده و (۳) مدل های برنامه ریزی.

سیستم های فایل را می توان اساس و زیرساخت کاربرد ها در سطوح بالاتر محسوب نمود. GFS گوگل یک سیستم فایل پراکنده توسعه پذیر برای حمایت از کاربرد های متمرکز بر داده پراکنده مقیاس بزرگ می باشد. GFS از سرور های معمول ارزان برای دستیابی به تلرانس خطا استفاده می کند و سرویس های عملکرد بالا را برای مشتریان فراهم می نماید. GFS از کاربرد های فایل مقیاس بزرگ با خواندن تکراری تز نسبت به نوشتن حمایت می کند. از اینرو، همچنین GFS دارای محدودیت هایی نظیر نقطه شکست تکی و عملکرد های ضعیف برای فایل های کوچک بوده است. چنین محدودیت هایی از طریق Colossus که جانشین GFS گردید، برطرف شده اند.

بعلاوه، دیگر شرکت ها و محققان دارای راه حل های خودشان برای براورده کردن تقاضا های متفاوت برای ذخیره سازی کلان داده می باشند. برای مثال، HDFS و KOSMOSFS از مشتقات کد های منبع باز GFS هستند. فیسبوک از Haystack برای ذخیره سازی مقادیر زیاد عکس های اندازه کوچک استفاده می کند. همچنین Taobao یک TFS و FastDSF را توسعه داده است. در نتیجه، سیستم های فایل پراکنده بعد از چند سال توسعه و فعالیت کسب و کار دارای بلوغ نسبی می باشند. از اینرو ما بر دو سطح دیگر در مابقی این بخش تمرکز خواهیم کرد.

۱-۳-۱-٤ فناوری پایگاه داده

فناوری پایگاه داده به مدت بیش از سی سال در حال رشد و پیشرفت بوده است. انواع سیستم های پایگاه داده برای رسیدگی به مجموعه های داده در مقیاس های متفاوت و کاربرد های مختلف پشتیبانی توسعه می یابند. مجموعه های داده رابطه ای سنتی از عهده چالش ها در زمینه دسته ها و مقیاس هایی که از طریق کلان داده ایجاد شده اند بر نمی ایند. مجموعه های داده NoSQL (برای مثال مجموعه های داده رابطه ای غیر سنتی) برای ذخیره سازی کلان داده معروف تر می شوند .پایگاه های داده NoSQL می توانند شیوه های انعطاف پذیر، پشتیبانی از کپی برداری ساده و راحت، AIP ساده، پایداری نهایی و حمایت از حجم زاد داده را نشان می دهند. ما سه پایگاه داده اصلی NoSQL زیر را در این بخش بررسی خواهیم کرد: پایگاه های داده مقدار-کلید، پایگاه های داده ستون محور و پایگاه های داده سند محور که هر یک بر اساس مدل های داده معین می باشند.

پایگاه های داده مقدار-کلید: این پایگاه های داده از مدل داده ساده تشکیل می گردند و داده مشابه با مقادیر کلید ذخیره می شود. هر کلید منحصر به فرد است ومشتریان می توانند مقادیر صف بندی شده را مطابق با کلید ها وارد کنند. چنین پایگاه های داده ای می توانند ساختار ساده را نشان دهند و پایگاه های داده مقدار–کلید مدرن با توسعه پذیری بالا و زمان پاسخ صف کوتاه تر از پایگاه های داده رابطه ای شناخته می شوند. تعدادی از پایگاه های داده مقدار– کلیدی در پنج سال گذشته ظاهر شده اند و همینطور توسط سیستم داینامو امازون برای استفاده تهییج شده اند. ما داینامو و چندین پایگاه داده مقدار – کلید نمونه دیگر را معرفی خواهیم کرد:

داینامو (Dynamo): داینامو را می توان سیستم ذخیره سازی داده مقدار– کلیدی پراکنده توسعه پذیر و با دسترسی بالا میباشد. این سیستم برای ذخیره سازی و مدیریت وضعیت بعضی سرویس های هسته استفاده می گردد که می توان با دسترسی به کلید در پلت فرم تجارت الکترونیکی امازون محقق نمود. حالت عمومی پایگاه های داده رابطه ای ممکن است داده غیر معتبر و مقیاس داده و دسترس پذیری محدود ایجاد کنند در حالی که داینامو می تواند با واسط هدف– کلید ساده به راحتی بر این مشکلات غلبه نماید که از عملیات خواندن و نوشتن ساده تشکیل می گردد. داینامو از طریق دسته بندی داده، کپی برداری داده و مکانیزم های ویرایش هدف به الاستیسیته و دسترس پذیری دست می یابد. طرح دسته بندی داده داینامو به درهم سازی یکنواخت تکیه می کند که دارای مزیت های مهم بوده است که گره در حال عبور به طور مستقیم بر گره های مجاور تاثیر می گذارد و بر گره های دیگر تاثیر نمی گذارد تا بار را برای ماشین های ذخیره سازی اصلی چند تایی تقسیم نماید. داینامو می تواند داده را با مجموعه های N سرور ها کپی نماید که N در ان سرور ها یک پارامتر قابل پیکربندی می باشد تا به دسترس پذیری و دوام بالا دست یابند. همچنین سیستم داینامو می تواند یکنواختی نهایی را برای انجام به روز رسانی غیر همزمان در کل کپی ها مهیاء نماید.

Voldemort: همچنین Voldemort یک سیستم ذخیره سازی مقدار– کلید می باشد که در ابتدا برای LinkedIn توسعه یافته بود و هنوز توسط ان استفاده می شود. کلید واژه ها و مقادیر در Voldemort یک سری اشیای ترکیبی هستند که از جداول و تصاویر تشکیل شده اند. واسط Voldemort شامل سه فعالیت ساده می باشد: خواندن، نوشتن و حذف که همه این کار ها از طریق کلید واژه ها تایید می شوند. Voldemort میتواند کنترل همزمان به روزرسانی همزمان از ویرایش های چندتایی را فراهم نماید اما پایداری داده را تضمین نمی کند. از اینرو، Voldemort از قفل گذاری خوش بینانه برای به روز رسانی چند ثبتی یکنواخت پشتیبانی می کند. وقتی اختلاف بین به روز رسانی و دیگر عملیات ها اتفاق می افتد، عملیات به روز رسانی متوقف خواهد شد. مکانیزم کپی برداری داده سیستم Voldemort همانند مکانیزم داینامو می باشد. Voldemort نه تنها داده را در RAM ذخیره می سازد بلکه همچنین اجازه می دهد تا داده درون موتور ذخیره سازی جاگذاری گردد. Voldemort به طور ویژه از دو موتور ذخیره سازی پشتیبانی می کند که عبارتند از Berkeley DB و Random Access Files.

پایگاه داده مقدار–کلید در واقع تنها چند سال قبل ظاهر شده است. دیگر سیستم های ذخیره سازی مقدار– کلیدی که عمیقا تحت تاثیر Amazon Dynamo DB بوده است عبارتند از Redis، Tokyo Cabinet و Tokyo Tyrant، Memacached و Memcahe DB ، Riak و Scalaris که همه انها توسعه پذیری را ا طریق توزیع کلمات کلید درون گره ها فراهم می کنند . Voldemort، Riak، Tokyo Cabinet و Memecached را می توان در دستگاه های ذخیره سازی وصل شده به سیستم استفاده نمود تا داده را در RAM یا دیسک ها ذخیره نمود. دیگر سیستم های ذخیره سازی می توانند داده را درRAM ذخیره نمایند و بک آپ دیسک را مهیاء نمایند یا به کپی برداری و ریکاوری برای اجتناب از بک اپ گیری تکیه نمایند.

پایگاه داده ستون محور: داده ها در پایگاه داده ستون محورمطابق با ستون ها و نه ردیف ها ذخیره و پردازش می شوند. ستون ها و ردیف ها در گره های چند تایی قطعه بندی می شوند تا قابلیت توسعه تحقق یابد. پایگاه های داده ستون محور عمدتا از BigTable گوگل الهام گرفته شده اند. ما در بخش حاضر در ابتدا در مورد BigTable بحث می کنیم و سپس چندین ابزار اشتقاقی را معرفی می کنیم.

BigTable: BigTable یک سیستم ذخیره سازی داده ساخت یافته پراکنده می باشد که برای پردازش داده مقیاس بزرگ (دسته PB) در میان هزاران سرور تجاری طراحی می گردد. ساختار اساسی BigTable یک نگاشت مرحله به مرحله چند ابعادی با ذخیره سازی کم پشت، پراکنده و پایدار می باشد. شاخص های نگاشت عبارتند از کلید ردیف، کلید ستون و مُهر های زمان و هر مقدار در نگاشت یک آرایه بیتی تجزیه و تحلیل نشده می باشد. هر کلید ردیف در BigTable یک رشته کاراکتری ٦٤ کیلو بیتی می باشد. ردیف ها از طریق ترتیب لغت نویسی ذخیره می شوند و به طور دایمی به صورت Tablet ها برای توازت بار قطعه بندی می شوند (برای مثال واحد های توزیع). از اینرو، خواندن ردیف کوتاه داده می تواند بی نهایت موثر باشد چون تنها شامل ارتباط با بخش کوچک ماشین ها می باشد. ستون ها مطابق با پیشوند های کلید ها گره بندی می شوندو از اینرو خانواده های ستون شکل می گیرند. این خانواده های ستون یک سری واحد های اسایی برای کنترل دسترسی می باشند. مهر های زمانی یک سری عدد های صحیح ٦٤ بیتی برای تشخیص ویرایش های متفاوت مقادیر سلول هستند. مشتریان ممکن است به طور انعطاف پذیر یک سری اعداد ویرایش های سلول ذخیره شده را مشخص نمایند. این ویرایش ها در ترتیب نزولی مُهرهای زمانی مرحله به مرحله هستند از اینرو جدید ترین ویرایش همیشه خوانده خواهد شد.

خلق و حذف Tablet ها و خانواده های ستون و همچنین اصلاح ابر داده خوشه ها، جداول و خانواده های ستون در BigTable API به طور برجسته نشان داده می شوند. اپلیکیشن های کلاینت ممکن است مقادیر BigTable، مقادیر پرس و جو ستون ها یا مجموعه های داده فرعی مرروگر در جدول را جاگذاری نمایند یا حذف کنند. همچنین BigTable از بعضی مشخصه های دیگر نظیر پردازش تراکنش در ردیف تکی پشتیبانی می کند. کاربران می توانند از چنین خصیصه هایی برای انجام پردازش داده پیچیده تر استفاده کنند.

هر تکنیک که از طریق BigTable اجراء شده است شامل سه مولفه اصلی می باشد : سرور Master، سرور Tabletو کتابخانه مشتری. تنها Bigtable به یک مجموعه از سرور Master اجازه می دهد تا پراکنده باشد و مسئول جداول توزیع برای سرور Tablet می باشد ، سرور های Tablet اضافه شده یا برداشته شده کشف می شوند و توازن بار انجام می گیرد. بعلاوه، همچنین می تواند الگو BigTable را اصلاح نماید یعنی خلق جداول و خانوده های ستون و جمع اوری داده ناخواسته ذخیره شده در GFS و همچنین فایل های حذف شده یا از رده خارج شده و استفاده از انها در مثال های خاص BigTable. هر سرور جدول یک مجموعه Tablet را مدیریت می کند و مسئول خواندن و نوشتن Tablet بارگذاری شده می باشد . وقتی Tablet ها بسیار بزرگ

هستند ، این Tablet ها از طریق سرور قطعه بندی خواهند شد. کتابخانه مشتری اپلیکیشن استفاده می گردد تا با مثال های BigTable ارتباط ایجاد نمود.

BigTable مبتنی بر چندین مولفه گوگل از جمله GFS، سیستم مدیریت خوشه، فرمت فایل SSTable و chubby می باشد. GFSاستفاده می گردد تا داده و پرونده های ثبت را ذخیره شوند. سیستم مدیریت خوشه مسئولیت زمان بندی وظیفه، تسهیم منابع، پردازش شکست های ماشین و پایش وضعیت های ماشین را برعهده دارد. فرمت فایل SSTable استفاده می گردد تا داده BigTable به طور داخلی ذخیره گردد و نگاشت بین کلید ها و مقادیر دایمی، مرحله به مرحله و غیر قابل تغییر را همانند هر نوع رشته های بایت فراهم می کند. BigTable از Chubby برای وظایف زیر در سرور استفاده می کند:

(۱) تضمینی که در انجا در نهایت یک کپی مستر فعال در هر زمان وجود دارد ؛

(۲) مکان خود راه انداز داده BigTable ذخیره می گردد ؛

(۳) به سرور Tablet رجوع می گردد ؛

(٤) ریکاوری خطا در مورد شکست های سرور Table انجام می گیرد ؛

(۵) اطلاعات الگو Bigtable ذخیره می شود ؛

(٦) جدول کنترل دسترسی ذخیره می گردد.

Cassandra: سیستم Cassandra از نوع ذخیره پراکنده می باشد تا مقدار عظیم داده ساخت یافته توزیع شده در میان سرور های تجاری چند تایی را مدیریت نمایند. سیستم توسط فیسبوک توسعه یافته بود و به ابزار منبع باز در سال ۲۰۰۸ تبدیل شده بود. این سیستم یک سری ایده ها و مفاهیم آمازون داینامو و BigTable گوگل را بویژه در یکی کردن فناوری سیستم پراکنده داینامو با مدل داده BigTable می پذیرد. جداول در Cassandra در شکل نگاشت ساخت یافته چهار بعدی پراکنده هستند که در انجا چهار ابعاد از جمله ردیف، ستون، خانواده ستون و ستون برتر وجود دارند. ردیف از طریق کلید رشته با طول اختیاری تشخیص داده می شود. مقدار ستون هایی که قرار است خوانده یا نوشته شوند مهم نیست، عملیات در ردیف ها به صورت خودکار است. ستون ها می توانند خوشه ها را تشکیل دهند که خانوده های ستون نامیده می شود و مشابه با مدل داده Bigtable هستند. Cassandra دو نوع خانوداده ستون را فراهم می کند: خانوده های ستون و ستون های برتر. ستون برتر شامل تعداد ستون های اختیاری مرتبط با اسامی یکسان می باشد. خانوده ستون شامل ستون ها و ستون های برتر می باشد که ممکن است به طور دایمی در طول زمان اجراء در خانواده ستون جاگذاری می گردد. مکانیزم های کپی و دسته بندی Cassandra بسیار مشابه با مکانیزم های داینامو هستند تا به پایداری برسند.

ابزار اشتقاقی Bigtable: چون کد BigTable را نمی توان از طریق مجوز منبع باز بدست اورد ، تعدادی از پروژه های منبع باز برای اجرای مفهوم منبع باز برای توسعه سیستم های نشابه نظیر HBase و Hypertable رقابت می کنند .

Hbase نسخه کپی شده BigTable می باشد با جاوا برنامه ریزی شده است و بخشی از چارچوب MapReduce سیستم Hadoop Apache می باشد. Hbase با HDFS جایگزین GFS می گردد. Hbase محتوی به روز شده را بر طبق RAM می نویسد و به طور منظم انها را درون فایل ها در ردیسک ها می نویسد. عملیات های ردیف از نوع اتمی هستند که به قفل گذاری سطح ردیف و پردازش تراکنش مجهز شده اند و برای مقیاس بزرگ اختیاری می باشد. دسته بندی و توزیع به طور شفاف فعالیت می کنند و دارای فضا برای هش کلاینت یا کلید ثابت می باشند.

HyperTable مشابه با BigTable توسعه یافته بود تا به مجموعه ای از سیستم های عملکرد بالا، توسعه پذیر، ذخیره سازی پراکنده و پردازش برای داده ساخت یافته و ساخت نیافته دست یابد. HyperTable به سیستم های فایل پراکنده یعنی HDSF و مدیر قفل توزیع شده تکیه می کند. نمایش داده، پردازش و مکانیزم دسته بندی با موارد BigTable مشابه هستند. HyperTable دارای زبان پرس و جو خودش بوده است که زبان پرس و جو HyperTable (HQL) نامیده شده است و خلق، اصلاح و جداول اساسی پرس و جو را برای کاربران مجاز می سازد.

نظر به این که عمدتا پایگاه های داده ذخیره سازی ستون محور از BihTable تقلید می کنند، طراحی های این پایگاه های داده بجزء برای مکانیزم همزمانی و چندین خصیصه دیگر مشابه هستند. برای مثال، Cassandra بر پایداری ضعیف کنترل همزمان ویرایش های چند تایی تاکید می کند در حالی که HBase و HyperTable بر پایداری قوی از طریق قفل ها یا پرونده ها تمرکز می کنند.

پایگاه داده سند: ذخیره سازی سند در مقایسه با ذخیره سازی مقدار– کلید می تواند از شکل های داده پیچیده تر پشتیبانی نماید. نظر به این که اسناد از حالت های سخت تبعیت نمی کنند در نتیجه نیازی نیست تا نقل مکان حالت انجام گیرد. بعلاوه، جفت های مقدار کلید را می توان هنوز ذخیره نمود. ما سه نمونه مهم از سیستم های ذخیره سازی سند را بررسی خواهیم کرد که عبارتند از MongoDB، SimpleDB و CouchDB.

MongoDB : MongoDB یک پایگاه داده سند محور و منبع باز می باشد. این پایگاه داده سند ها را همانند Binary JSON(BSON) ذخیره می کند که با شی شباهت دارد. هر سند دارای یک میدان ID همانند کلید مقدماتی بوده است. پرس و جو در MongoDB با ترکیب مشابه با JSON بیان می گردد. درایور پایگاه داده پرس و جو را همانند شی BSON با MongoDB ارسال می گردد. کار پرس و جو در کل اسناد توسط سیستم مجاز می گردد. شاخص ها را برای انجام پرس و جو سریع می توان در حوزه های قابل جستجو اسناد ایجاد نمود. عملیات کپی در MongoDB را می توان با پرونده های ثبت در گره های اصلی اجراء نمود که از کل عملیات های سطح بالا انجام گرفته در پایگاه داده پشتیبانی می کنند. برده دار ها در طول کپی برداری به دلیل آخرین همزمان سازی در مورد کل عملیات های نوشتن پرس و جو می کنند و عملیات ها را در پرونده های

ثبت در پایگاه های داده محلی اجرا می نماید. MongoDB ازتوسعه افقی با تسهیم خودکار برای توزیع داده در میان هزاران گره از طریق متوازن سازی خودکار بار و افزونگی پشتیبانی می کند.

SimpleDB : SimpleDB یک پایگاه داده پراکنده می باشد و وب سرویس امازون می باشد. داده در SimpleDB بر طبق دامین های مختلف سازماندهی می گردد که داده ها در این دامین ها ذخیره، کسب و پرس و جو می گردند. دامین ها عبارتند از ویژگی ها و مجموعه های جفت نام/مقدار پروژه ها. تاریخ با ماشین های متفاوت در مراکز داده متفاوت کپی می گردد تا ایمنی داده و بهبود عملکرد را تضمین نماید. این سیستم از دسته بندی خودکار داده پشتیبانی نمی کند و از اینرو نمی توانست با تغییر حجم داده توسعه یابد. SimpleDB به کاربران اجازه می دهد تا با SQL پرس و جو کنند. حایز اهمیت است که SimpleDB میتواند پایداری نهایی را تضمین نماید اما از کنترل همزمانی نسخه Muti (MVCC) پشتیبانی نمی کند. از اینرو، اختلاف ها را نمی توان از سمت کلاینت کشف نمود.

CouchDB: Apache CouchDB یک پایگاه داده سند محور می باشد که با Erlang نوشته شده است. داده در CouchDB بر طبق اسنادی سازماندهی می گردد که از فیلد های نام گذاری شد با کلیدها/اسامی و مقادیر نام گذاری شده اند که که همانند اشیای JSON ذخیره می شوند و در دسترس قرار می گیرند. هر سند با شناسایی کننده منحصر به فرد مهیاء می گردد. CouchDB اجازه دسترسی به اسناد پایگاه داده را از طریق RESTful HTTP API می دهد. اگر لازم است تا سندی اصلاح گردد، کلاینت بایستی سند کل را برای اصلاح ان دانلود نماید و سپس به عقب به پایگاه داده برگرداند. بعد از این که سند یک بار دیگر بازنویسی می گردد، شناسایی کننده به روز خواهد شد. CouchDB از کپی برداری بهینه برای رسیدن به مقیاس پذیری بدون مکانیزم مشترک استفاده می کند. چون انواع CouchDB ها ممکن است در امتداد با دیگر تراکنش ها به طور همزمان اجراء شوند، هر نوع Replication Topology را می توان ساخت. پایداری CouchDB بر مکانیزم کپی سازی تکیه می کند. CouchDB از MVCC با ثبت های هش تاریخی پشتیبانی می کند.

معمولا کلان داده ها در صد ها و حتی هزاران سرور تجاری ذخیره می شوند. از اینرو، مدل های موازی سنتی نظیر واسط عبور پیام (MPI) و پردازش چند تایی باز (OpenMP) ممکن نیست برای پشتیبانی از این قبیل برنامه های موازی مقیاس بزرگ کافی باشند. اخیرا، تعدادی از مدل های برنامه ریزی موازی پیشنهادی به طور موثری باعث بهبود عملکرد NoSQL و کاهش شکاف با پایگاه های داده رابطه ای شده اند. بنابراین، این مدل ها به اساس تجزیه و تحلیل داده های حجیم تبدیل شده اند.

MapReduce : MapReduce یک مدل برنامه ریزی ساده اما قدرتمند برای رایانش مقیاس بزرگ با استفاده از تعداد خوشه های زیای کامپیوتر های شخصی تجاری می باشد تا به پردازش و توزیع موازی خودکار دست یابد. مدل رایانش در MapReduce تنها دارای دو کارکرد یعنی Map و Reduce می باشد که هر دو آنها توسط کاربران برنامه ریزی می شوند. عملکرد Map این است که جفت های مقدار–کلیدی ورودی را پردازش می کند

و جفت های مقدار–کلیدی واسط را تولید می کند. سپس، کل مقادیر واسط مرتبط با کلید یکسان در MapReduce ترکیب می شوند و انها را به عملکرد Reduce ارسال می کند که بیشتر مجموعه مقدار را درون مجموعه کوچک تر فشرده می سازد. MapReduce از مزیتی برخوردار بوده است که از گام های پیچیده برای پدیدآوری اپلیکیشن های موازی یعنی زمان بندی داده، تلرانس خطا و ارتباطات گره داخلی اجتناب می ورزد. تنها کاربر ملزم است تا دو کارکرد برای توسعه کاربرد موازی برنامه ریزی نماید. چارچوب اولیه MapReduce از مجموعه های داده چند تایی در وظیفه پشتیبانی نکرده است که از طریق بعضی پیشرفت های جدید کاهش یافته است.

برنامه ریزها در چند دهه گذشته با زبان تشریحی پیشرفته SQL که اغلب در پایگاه داده رابطه ای برای توصیف وظیفه و تحلیل مجموعه داده استفاده گردیده اند، آشنا می شوند. از اینرو، چارچوب فشرده MapReduce تنها دو کارکرد غیر شفاف را مهیاء می نماید که نمی توانند کل عملیات های متداول را پوشش دهند. از اینرو، برنامه ریز ها بایستی برای برنامه ریزی کارکرد های اصلی زمان صرف نمایند که به ور نوعی به سختی حفظ می شوند و از نو استفاده می گردند. تعدادی از سیستم های زبان پیشرفته با هدف بهبود راندمان برنامه ریزی پیشنهاد شده اند که عبارتند از sawzall از گوگل، Pig Latin از یاهو ،Hive از فیسبوک و Scope از مایکروسافت.

Dyrad: Dyrad یک موتور اجرای توزیعی با مصرف عمومی برای پردازش اپلیکیشن های موازی داده های دانه ای–درشت می باشد. ساختار عملیاتی Dyrad یک نگاره سازی ناچرخه ای جهت دار می باشد که راس ها در این گراف ها یک سری برنامه ها را نشان می دهند و لبه ها معرف کانال های داده می باشند. عملیات ها در Dyrad در راس های در خوشه ها اجراء می گردد و داده ها از طریق کانال ها از جمله سند ها، اتصالات TCP و حافظه مشترک FIFO انتقال می یابد. منابع در گراف عملیات منطقی در طول عملیات به طور خودکار با منابع فیزیکی نگاشته می شوند.

ساختار عملیات Dyrad از طریق یک برنامه مرکزی هماهنگ می گردد که مدیر وظیفه نامیده شده است و می تواند در خوشه ها یا ایستگاه های کار از طریق شبکه اجراء گردد. مدیر وظیفه شامل دو بخش می باشد :

(۱) کد های کاربرد که برای ساخت گراف ارتباط وظیفه استفاده می شوند.

(۲) کد های کتابخانه برنامه که برای هماهنگی منابع در دسترس استفاده میگردند. کل انواع داده ها به طور مستقیم بین راس ها انتقال می یابند. از اینرو، مدیر وظیفه تنها مسئول تصمیم گیری می باشد که هر نوع انتقال داده را با مشکل مواجه نمی نماید.

توسعه دهنده های کاربرد در Dyrad می توانند به طور انعطاف پذیری هر نوع نگاره سازی ناچرخه ای جهت دار را انتخاب کنند تا شیوه های ارتباط کاربرد را توضیح دهند و مکانیزم های انتقال داده را بیان نمایند. بعلاوه، Dyrad به راس ها اجازه می دهد تا از هر مقدار داده ورودی و خروجی استفاده کنند در حالی که mapReduce تنها از یک مجموعه ورودی و خروجی پشتیبانی می نماید.

DyradLINQ یک زبان پیشرفته از Dyrad می باشد و برای یکی کردن محیط اجرای زبان مشابه با SQL استفاده می گردد.

All-Psirs: این سیستم به طور خاص برای کاربرد های زیست سنجی ها، اطلاع رسانی زیستی و داده کاوی طراحی می گردد. سیستم بر مقایسه جفت های عنصر در دو مجموعه داده در کاربرد معین تمرکز می کند. سیستم All-Pairs را می توان همانند سه چند تایی بیان نمود (مجموعه A ، مجموعه B و تابع F) که تابع F در این مجموعه ها برای مقایسه کل اجزاء در مجموعه A و B استفاده می گردد. نتیجه مقایسه یک ماتریس خروجی M می باشد که همچنین محصول دکارتی یا پیوستن عرضی A و B نامیده می شود.

سیستم All-Piars در چهار فاز اجراء می گردد: مدل سازی سیستم، توزیع داده ورودی، مدیریت وظیفه دسته و جمع اوری نتیجه. مدل تقریبی عملکرد سیستم در فاز یک ساخته خواهد شد تا ارزیابی نماید چه مقدار منبع سی پی یو مورد نیاز است و چگونه دسته بندی وظیفه را انجام داد. درخت پوشا در فاز دوم برای انتقال های داده ساخته می شود که حجم کار هر دسته بندی را برای بازیابی داده ورودی به طور موثر انجام می دهد. در فاز سوم بعد از این که جریان داده به گره های مناسب تحویل داده می شود، موتور All-Pairs یک ارایه پردازش دسته را برای وظایف در دسته بندی ها خواهد ساخت در حالی که انها را در سیستم پردازش دسته ای مرحله به مرحله قرار می دهد و دستور اجرای گره را برای کسب داده تدوین می نماید. در فاز آخر بعد از این تکمیل وظیفه سیستم پردازش دسته ای ، موتور استخراج نتایج را جمع اوری خواهد کرد و انها را در ساختار مناسب ترکیب می نماید که معمولا لیست فایل تکی می باشد که کل نتایج در این لیست به طور ترتیبی قرار می گیرند.

Pregel: سیستم Pregel گوگل می تواند کار پردازش گراف های اندازه بزرگ یعنی تحلیل گراف های شبکه و سرویس های شبکه سازی اجتنماعی را تسهیل نماید. وظیفه محاسباتی از طریق گراف جهت دار بیان می گردد که از طریق راس ها و لبه های جهت دار تشکیل می گردد. هر راس به مقدار تعریف شده کاربر و قابل اصلاح ربط دارد و هر لبه جهت دار مرتبط با راس منبع از مقدار تعیین شده کاربر و شناسایی کننده راس هدف تشکیل می گردد. وقتی گراف ساخته می شود، برنامه یک سری محاسبات تکراری را انجام می دهد که گام های عالی نامیده شده است و نقاط همزمان سازی جهانی در میان انها تا زمان تکمیل الگوریتک و تکمیل خروجی تنظیم می شوند. محاسبات راس در هر گام عالی موازی هستند و کار تعیین شده کاربر برای بیان منطق الگوریتم معین توسط هر راس اجرا می گردد. هر راس ممکن است خودش و لبه های خروجی اش را اصلاح نماید، پیام فرستاده شده به دیگر راس ها را دریافت نماید و حتی ساختار مکان نگر کل گراف را اصلاح کند. لبه ها با محاسبات مشابه تهیه نمی گردند. کار های هر راس ممکن است از طریق تعلیق کنار گذاشته شوند. وقتی کل راس ها در وضعیت غیر فعال بدون هر نوع پیام برای ارسال می باشند، اجرای برنامه کل تکمیل می گردد.

خروجی برنامه Pregel یک مجموعه ای مشتمل بر خروجی مقادیر از کل راس ها می باشد. ورودی و خروجی برنامه Pregel یک سری گراف های جهت دار هم ریخت هستند.

دیگر محققان با الهام از مدل های برنامه ریزی فوق بر شیوه های برنامه ریزی برای وظایف محاسباتی پیچیده تر یعنی محاسبات تکراری، محاسبات حافظه تلرانس عیب، محاسبات افزایشی و تصمیم گیری کنترل جریان مرتبط با داده تمرکز کرده اند.

٤-٢ تجریه و تحلیل کلان داده

تجزیه و تحلیل کلان داده عمدتا شامل روش های تحلیل برای داده سنتی و کلان داده، معماری تحلیلی برای کلان داده و نرم افزار استفاده شده برای کاوش و تحلیل کلان داده می باشد. تجزیه و تحلیل داده را می توان فاز نهایی و بسیار مهم در زنجیره ارزش کلان داده با هدف استخراج مقادیر سودمند، تهیه پیشنهادات یا تصمیمات نام گذاشت. سطوح متفاوت مقادیر بالقوه را می توان از طریق تحلیل مجموعه های داد در فیلد های متفاوت تولید نمود. از اینرو، تحلیل داده یک حوزه گسترده می باشد که به کرات تغییر می نماید و بی نهایت پیچیده است. ما روش ها، معماری ها و ابزار تجریه و تحلیل کلان داده را در این بخش معرفی می کنیم.

١-٢-٤ تجزیه و تحلیل داده سنتی

تجزیه و تحلیل داده سنتی با این هدف انجام می گیرد تا از روش های آماری مناسب برای تحلیل داده حجیم، تمرکز، استخراج و اصلاح داده سودند مخفی در دسته مجموعه های داده آشفته و شناسایی قانون لاینفک موضوع اصلی برای به حداکثر رساندن ارزش داده استفاده نماید. تجزیه و تحلیل داده نقش راهنمایی کننده بزرگی در ایجاد طرح های توسعه ای برای یک کشور، درک تقاض های مشتری برای تجارت و پیش بینی روند بازار برای شرکت ها ایفاء می کند. تجزیه و تحلیل کلان داده را می توان همانند تکنیک تحلیل برای نوع خاص داده فرض نمود. از اینرو، روش های تحلیل داده مختلف هنوز برای تجزیه و تحلیل کلان داده استفاده می شوند .چندین روش سنتی تجزیه و تحلیل داده در ادامه مورد بررسی قرار می گیرند و تعدادی از انها از علم کامپیوتر و امار هستند.

٢-٢-٤ تجزیه و تحلیل خوشه

این یک روش اماری برای گروه بندی اهداف و بویژه دسته بندی اهداف مطابق بابعضی خصیصه ها می باشد. تجزیه و تحلیل خوشه مورد استفاده قرار می گیرد تا اهداف با خصیصه های ویژه را متمایز سازد و انها را مطابق با این خصیصه ها به چندین دسته (خوشه) تقسیم نمود به نحوی که اهداف در دسته یکسان دارای همگنی بالا خواهند بود در حالی که دسته های متفاوت دارای ناهمگنی بالا خواهند بود. تجزیه و تحلیل خوشه یک روش بررسی نظارت نشده بدون اموزش داده می باشد.

۳-۲-٤ تجزیه و تحلیل فاکتور

این روش اساسا توصیف رابطه میان اجزای مختلف را با تنها چندین فاکتور هدف قرار می دهد یعنی گروه بندی چندین متغیر دارای رابطه نزدیک درون فاکتور و تعداد کمی از فاکتور ها استفاده می شوند تا بخش زیادی اطلاعات داده اصلی را آشکار سازند.

تحلیل همبستگی: این یک روش تحلیلی برای تعیین قانون روابط نظیر همبستگی، وابستگی همبسته و محدودیت متقابل در میان پدیده های مشاهده شده و متناوبا انجام پیش بینی و کنترل می باشد. چنین روابطی ممکن است به دو نوع دسته بندی شوند:

(۱) تابع، انعکاس رابطه وابستگی شدید میان پدیده ها که همچنین رابطه وابستگی قطعی نامیده شده است؛

(۲) همبستگی، تعدادی از روابط وابستگی تعیین نشده یا غیر دقیق و مقدار عددی یک متغیر ممکن است با چندین مقدار عددی متیغر دیگر شباهت دارند و چنین مقادیر عددی یک نوسان منظم را پیرامون مقادیر میانه اشان نشان می دهند.

تجریه و تحلیل رگرسیون: یک ابزار ریاضی برای آشکار سازی همبستگی ها بین یک متغیر و چندین متغیر دیگر می باشد. تجزیه و تحلیل رگرسیون بر اساس گروهی از آزمایشات یا داده های مشاهده شده یک سری روابط وابستگی را در میان متغیر های مخفی از طریق تصادفی بودن شناسایی می کند. تجزیه و تحلیل رگرسیون ممکن است همبستگی ها میان متغیر هایی را که قرار است ساده و منظم باشند را پیچیده تر و غیر معین می سازد.

A/B Testing: همچنین تست منحنی پیوند نامیده می شود. این روش یک فناوری برای تعیین نحوه بهبود متغیر های هدف از طریق مقایسه گروه تست شده می باشد. کلان داده به تعداد زیادی از تست ها نیاز خواهد داشت که قرار بود اجراء و تجزیه و تحلیل گردند.

تحلیل آماری: تحلیل اماری بر اساس فرضیه اماری یا یک شاخه از ریاضیات کاربردی می باشد. تصادفی بودن و عدم قطعیت در فرضیه آماری با فرضیه احتمال مدل سازی می شوند. تجزیه و تحلیل اماری می تواند توصیف و استنباط کلان داده را فراهم می کند. تجزیه و تحلیل اماری توصیفی می تواند مجموعه های داده را خلاصه می سازد و توصیف می نماید در حالی که تجزیه و تحلیل آماری استنباطی می تواند نتیجه گیری ها را از موضوع داده با تغییرات تصادفی استخراج نماید. تجزیه و تحلیل اماری به طور گسترده ای در رشته های مراقبت بهداشتی و اقتصادی اعمال می گردد.

الگوریتم های داده کاوی: داده کاوی یک فرایند برای استخراج اطلاعات و دانش مخفی، ناشناخته اما به طور بالقوه سودمند از یک داده حجیم، ناقص، نویزی، فازی و تصادفی می باشد. کنفرانس بین المللی IEEE در سال ۲۰۰٦ در سری های داده کاوی (ICDM) می تواند ده الگوریتم داده کاوی بی نهایت تاثیر گذار را از طریق تکنیک انتخاب سخت از جمله C٤,٥، میانه های K، ماشین بردار مجازی، Apriori، EM، Naïve Bayesو cart

و موارد دیگر شناسایی نماید. این ده الگوریتم می توانند دسته بندی، خوشه بندی، رگرسیون، یادگیری آماری، تجزیه و تحلیل مجموعه و کاوش لینک را پوشش دهند و همه انها از جمله مشکلات بسیار مهم در تحقیق داده کاوی هستند.

۳-٤ روش های تحلیل کلان داده

مردم با طلوع عصر کلان داده در مورد این موضوع نگران هستند که چگونه اطلاعات کلیدی را از داده حجیم با سرعت بالا استخراج نماید تا مقادیر را برای شرکت ها و اهداف بیاورند. اکنون روش های پردازش مهم کلان داده در زیر نشان داده می شوند.

Bloom Filter: Bloom Filter شامل سری هایی از کار های هش می باشد. قاعده Bloom Filter این است تا مقادیر درهم داده را از طریق استفاده از آرایه رقم دو دویی بجای خود داده ذخیره نمایند که در اصل یک شاخص طرح بیتی می باشد که از توابع هش برای ذخیره سازی فشرده سازی پراتلاف داده استفاده می کند . این روش دارای بعضی مزیت ها نظیر راندمان بالای فضا و سرعت پرس و جو بالا بوده است بلکه همچنین دارای معایبی در تشخیص نادرست و حذف می باشند.

درهم سازی: درهم سازی روشی است که اساسا داده را به مقادیر عددی طول ثابت کوتاه تر یا مقادیر شاخص تبدیل می کند. بعضی مزیت های درهم سازی عبارتند از خواندن و نوشتن سرعت بالا پرس و جو اما تابع هش عالی را به سختی می یابد.

ایندکس(فهرست سازی): ایندکس همیشه یک روش موثر برای کاهش هزینه خواندن و نوشتن می باشد و جاگذاری، حذف، اصلاح و سرعت های پرس و جو در پایگاه های داده رابطه ای سنتی را بهبود می بخشند که داده ساخت یافته و فناوری های دیگر را مدیریت می کنند که داده های نیم ساخت سافته و ساخت نیافته را اداره می کنند. از اینرو، ایندکس دارای معایبی بوده است که دارای هزینه اضافی برای ذخیره سازی فایل های ایندکی بوده است که بایستی به طور دینامیکی در زمانی که داده به روز می گردد، حفظ می گردد.

Triel: همچنین درخت trie نامیده شده است و یک گونه از درخت هش می باشد. عمدتا برای بازیابی سریع و امار های فراوانی کلمه اعمال می گردد. ایده اصلی Triel این است تا از پیشوند های متداول رشته های کاراکتر استفاده نماید تا مقایسه در رشته های کاراکتر را با بیشترین میزان کاهش دهد تا کارامدی پرس و جو را بهبود بخشد.

رایانش موازی: رایانش موازی که با رایانش سریالی سنتی مقایسه شده است به استفاده همزمان از چندین منبع رایانس برای تکمیل وظیفه محاسبه گفته می شود. ایده اصلی رایانش موازی این است تا مشکل را تجزیه نمود و آنها را به چندین فرایند مجزاء اختصاص داد که قرار است به طور مستقل کامل شوند تا به پردازش مشترک

دست یافت . احتمالا یک سری مدل های رایانش موازی کلاسیک شامل MPI (واسط عبور پیام)، MapReduce و Dryad هستند.

هر چند سیستم ها یا ابزار رایانش موازی نظیر MapReduce و Dryad برای تجزیه و تحلیل کلان داده سودمند هستند، انها ابزار سطوح پایین هستند که به سختی استفاده گردیده و یادگرفته می شوند. از اینرو، تعدادی از ابزار یا زبان های برنامه ریزی موازی سطح بالا بر اساس این سیستم ها توسعه می یابند. این قیل زبان های سطح بالا عبارتند از sawzall، Pigو Hive که برای MapReduce استفاده شده اند و همچنین scope و DryadLINQ که برای Dryad استفاده شده اند.

٤-٤ معماری تجزیه و تحلیل کلان داده

معماری های تحلیل متفاوت به دلیل ٤V های کلان داده بایستی برای الزامات کاربردی متفاوت مورد توجه قرار گیرند .

جدول ١ : مقایسه MPI ، MapReduced و Dryad

	MPI	MapReduced	Dryad
بکار گیری	گره رایانش و ذخیره سازی داده که به طور مجزاء مرتب شده اند (داده بایستی گره رایانش را حرکت دهد)	رایانش و ذخیره سازی داده که در گره یکسان مرتب شده است (رایانش بایستی به داده نزدیک شود)	رایانش و و ذخیره سازی داده در گره یکسان هماهنگ شده بود (رایانش بایستی به داده نزدیک باشد)
مدیریت منبع /زمان بندی	-	صف کاری(گوگل)	شفاف نیست
برنامه ریزی سطح پایین	MPI API	MapReduce API	Dryad API
برنامه ریزی سطح بالا	-	Pig,Hive.Jaql..	Scope,DryadLNQ
ذخیره سازی داده	سیستم فایل محلی، NFS	GFS (گوگل)	NTFS
دسته بندی وظیفه	دسته بندی دستی کاربر	خودکار سازی	خودکار سازی
ارتباط	پیام رسانی ، دسترسی حافظه راه دور	فایل ها (FSمحلی، DFS)	فایل ها، TCP Pipes، FIFO های حافظه مشترک
تحمل عیب	نقطه بررسی	اجرای مجدد وظیفه	اجرای مجدد وظیفه

۱-٤-٤ تجزیه و تحلیل انی در برابر آنلاین

تجزیه و تحلیل کلان داده را مطابق با الزامات مناسبت و شایستگی می توان به صورت تجزیه و تحلیل انی و آفلاین دسته بندی نمود.

تجزیه و تحلیل انی: این روش عمدتا در تجارت الکترونیک و امورمالی استفاده می گردد . نظر به این که داده به طور دایمی تغییر می کند ، تجزیه و تحلیل سریع داده مورد نیاز می باشد و نتایج تحلیلی بایستی با تاخیر بسیار کوتاه برگشت داده شوند. معماری های موجود اصلی تجزیه و تحلیل انی عبارتند از:

(۱) خوشه های پردازش موازی با استفاده از پایگاه های داده رابطه ای سنتی

(۲) پلت فرم های رایانش حافظه محور. برای مثال، Greenplum از EMC و HANA از SAP دارای معماری های تجزیه تحلیل انی هستند.

تجزیه و تحلیل آفلاین: معمولا برای کاربرد های بودن الزامات در زمان پاسخ یعنی فراگیری ماشین، تحلیل اماری و الگوریتم های پیشنهاد استفاده می گردد. معمولا تجزیه و تحلیل آفلاین این کار را از طریق وارد کردن ثبت ها به درون یک پلت فرم خاص از طریق ابزاری اکتساب داده انجام می دهد. تعدادی از شرکت های اینترنتی از معماری تجزیه و تحلیل افلاین مبتنی بر هادوپ تحت محیط کلان داده استفاده می کنند تا هزینه تبدیل فرمت داده را کاهش دهند و کارامدی اکتساب داده را بهبود می بخشند. مثال ها عبارتند از Scribe ابزار منبع باز فیسبوک، Kafka ابزار منبع باز LinkdIn، Timetunnel ابزار منبع باز Tabao و Chukwa هادوپ و موارد دیگر. این ابزار می توانند تقاضا های اکتساب و تبدیل داده را با صد ها مگابایت در ثانیه براورده سازند.

۲-٤-٤ تجزیه و تحلیل در سطوح مختلف

تجزیه و تحلیل کلان داده را می توان در سطح حافظه، سطح هوش تجاری (BI) و سطح حجیم دسته بندی نمود که در زیر بررسی می شوند.

تجزیه و تحلیل سطح حافظه: این تجزیه و تحلیل برای موردی می باشد که در انجا کل حجم داده کمتر از حداکثر حافظه خوشه می باشد. این روز ها، حافظه خوشه سرور بهتر از صد ها گیگابایت می باشد در حالی که حتی سطح ترابایت متداول است. بنابراین، فناوری پایگاه داده داخلی ممکن است استفاده گردد و هات دیتا بایستی در حافظه باقی بماند تا کارامدی تحلیلی را بهبود بخشد. تجزیه و تحلیل سطح حافظه برای تحلیل انی بی نهایت مناسب است. MongoDB یک معماری نمونه تحلیلی سطح حافظه می باشد. ظرفیت و عملکرد تحلیل داده سطح حافظه با توسعه SSD با بهبود بیشتری مواجه شده است و به طور گسترده ای اعمال گردید.

تجزیه و تحلیل هوش تجاری (BI): این تحلیل در زمانی برای موردی استفاده می گردد که مقیاس داده برتر از سطح حافظه باشد اما ممکن است به درون محیط تحلیل هوش تجاری وارد می شود. اخیرا محصولات هوش تجاری جریان اصلی با طرح های تحلیل داده مهیاء می گردند تا از سطح بر روی TB پشتیبانی نمایند.

تجزیه و تحلیل انبوه: این روش برای یک مورد در زمانی استفاده می شود که مقیاس داده به طور کامل برتر از ظرفیت های محصولات BI و پایگاه های داده رابطه ای سنتی بوده است. اکنون، بخش عمده تحلیل انبوه از HDFS هادوپ استفاده می کند تا داده را ذخیره نماید و از MapReduce برای تحلیل داده استفاده نمایند. بخش عمده تحلیل انبوه به دسته تحلیل آفلاین نعلق دارد.

۳-٤-٤ تجزیه و تحلیل با پیچیدگی متفاوت

پیچیدگی زمان و فضای الگوریتم های تجزیه و تحلیل داده تا حد زیادی با همدیگر مطابق با انواع مختلف داده و تقاض های کاربرد فرق دارند. برای مثال، برای کاربرد هایی که با پردازش موازی قابل جوابگویی می باشد، الگوریتم پراکنده ممکن است طراحی گردد و مدل پردازش موازی ممکن است برای تحلیل داده استفاده شوند .

٥-٤ ابزار کاوش و تحلیل کلان داده

بعضی ابزار ها نظیر نرم افزار آماتور و حرفه ای، نرم افزار تجاری گران و نرم افزار منبع باز برای کاوش و تحلیل کلان داده در دسترس هستند. ما در بخش حاضر به طور مختصر در مورد پنج نرم افزار عالی با کاربرد گسترده بازنگری انجام می دهیم و این کا را مطابق با این این سئوال می پرسیم که علم تجزیه و تحلیل، داده کاوی، نرم افزار کلان داده که شما در ۱۲ ماه قبلی برای پروژه واقعی استفاده کرده اید چیست؟

R (۳۰,۷ ٪) : R زبان برنامه ریزی منبع باز و محیط نرم افزاری می باشد که برای کاوش/ تحلیل و بصری سازی داده طراحی می گردد. در حالی که وظایف متمرکز بر رایانش اجراء می گردند، برنامه ریزی کد با C، C++ و Fortran ممکن است در محیط R فراخوانده شوند. بعلاوه، کاربران ماهر می توانند به طور متسقیم در C تماس گرفته شوند. در واقع، R عبارتست از تحقیق زبان S که یک زبان تفسیر شده توسعه یافته توسط لابراتوار های AT&T Bell می باشد و برای کشف داده، تحلیل اماری و طرح های ترسیم استفاده شده اند. R در مقایسه با S رایج تر است چون منبع باز می باشد. R در رتبه درجه یک در بررسی ۲۰۱۲ KDNuggets قرار دارد. علاوه بر این، در بررسی زبان های طراحی که شما برای داده کاوی/تحلیل در سال گذشته در سال ۲۰۱۲ استفاده کرده اید، R در رتبه یک بوده است و بر SQL و جاوا غلبه کرده است. تولید کننده های پایگاه داده نظیر Teradata و Oracle به دلیل محبوبیت R یک سری محصولات پشتیبان R را منتشر کرده اند.

اکسل (۲۹,۸ ٪) : اکسل که مولفه هسته در مایکروسافت آفیس می باشد یک سری قابلیت های تجزیه و تخلیل اماری و پردازش داده قدرتمند را تهیه می کند. وقتی اکسل نصب می گردد، تعدادی از برنامه های افزودنی پیشرفته نظیر Analysis ToolPak و Solver Add-in با عملکرد های قدرتمند برای تحلیل داده در ابتدا یکپارچه می شوند اما چنین برنامه های افزودنی را می توان تنها در صورتی استفاده نمود که کاربران انها را میسر می سازند. همچنین اکسل تنها نرم افزار تجاری در میان پنج مورد عالی می باشد.

Rapid-I rapidmine(۲۶,۷ ٪): Rapidminderیک نرم افزار منبع باز می باشد که برای داده کاوی، ماشین فراگیر و تحلیل پیش گویانه می باشد. Rapidminder در بررسی Kdnuggets در سال ۲۰۱۱ با تکرار بیشتری نسبت به R استفاده شده است (در رتبه یک قرار گرفته است). برنامه های فراگیری ماشین و داده کاوی که از طریق RapidMiner تهیه شده اند عبارتند از استخراج، تبدیل و بارگذاری(ETL)، پیش پردازش داده و مدل سازی بصری سازی، ارزیابی و استقرار.

جریان داده کاوی در XML توصیف می گردد و از طریق واسط کاربر گرافیکی (GUI) نمایش داده شده است. RapidMiner در جاوا نوسته می شود. این برنامه می تواند یادگیرنده و روش ارزیابی Weka را یکپارچه می سازد و با R کار می کند. کارکرد های RpidMiner با اتصال فرآیند ها از جمله انواع اپراتور ها اجراء می شوند. جریان کلی را می توان همانند خط تولید کارخانه با ورودی داده اصلی و خروجی نتایج مدل فرض کرد. اپراتور ها را می توان همانند بعضی کار های خاص با مشخصه های ورودی و خروجی متفاوت مورد توجه قرار داد.

KNMINE (۲۱,۸٪) : KNIME (Konstanz Information Miner) یک پلت فرم یکپارچه سازی داده غنی، پردازش داده، تحلیل داده و داده کاوی منبع باز، هوشمند و دوستار کاربر می باشد. این پلت فرم به کاربران اجازه می دهد تا جریان های داده یا کانال های داده را در وضعیت بصری شده خلق نمایند تا بعضی یا کل تکنیک های تحلیلی را به طور انتخابی اجراء کنند و نتایج تحلیلی، مدل ها و دیدگاه های متقابل را فراهم کند. KNIME با زبان جاوا نوشته شده بود و کار های بیشتر همانند برنامه های افزودنی را بیشتر فراهم می کند. کاربران از طریق فایل های برنامه افزودنی می توانند ماژول های پردازش را برای فایل ها، تصاویر و سری های زمانی جاگذاری نمایند و آنها را درون انواع پروژه های منبع باز یعنی R و Weka یکپارچه می نمایند. یکپارچه سازی داده، پاکسازی، تبدیل، فیلترینگ، امار ها، کاوش و در نهایت بصری سازی داده توسط KNIME کنترل می شوند. فرایند توسعه کامل تحت محیط بصری شده انجام می گیرد. KNIME همانند چارچوب توسعه پذیر و مبتنی بر ماژول طراحی می گردد. وابستگی بین واحد های پردازش و محفظه های داده وجود ندارد و ان را با محیط پراکنده و توسعه مستقل منطبق می سازد. بعلاوه، توسعه KNIME ساده است. توسعه دهنده ها می توانند به راحتی انواع گره ها و دیدگاه های KNIME را توسعه دهند.

Weka/Pentaho (۱۴,۸ ٪): wekaکه مخفف Waikato Environment for Knoweldge Analysisمی باشد یک فراگیری ماشین منبع باز و رایگان و نرم افزار داده کاوی نوشته شده به زبان جاوا می باشد. Weka می تواند کار هایی نظیر پردازش داده، انتخاب خصیصه، دسته بندی، رگرسیون، خوشه بندی، قاعده اجتماع و بصری سازی را فراهم نماید. Penato یکی از محبوب ترین نرم افزارهای هوش تجاری منبع باز می باشد. این نرم افزار شامل پلت فرم وب سرور و چندین ابزار پشتیبانی گزارش دهی، تحلیل، نمودار سازی، یکپارچه سازی داده و داده کاوی می باشد که همگی از جنبه های هوش تجاری هستند. الگوریتم های پردازش داده weka در Penato یکپارچه می شوند و می توان به طور مستقیم تماس گرفت.

٤-٦ کاربرد های کلان داده

ما در بخش قبلی کار تجزیه و تحلیل کلان داده را بررسی کرده ایم که فاز نهایی و مهم زنجیره ارزش کلان داده می باشد. تجزیه وتحلیل کلان داده می تواند مقادیر سودمندی را برای داوری ها، پیشنهادات، حمایت ها یا تصمیمات مهیاء نماید.

از اینرو، تحلیل داده شامل دامنه وسیعی از کاربرد ها می باشد که به کرات تغییر می کنند و بی نهایت پیچیده می باشند. ما در بخش حاضر در ابتدا تکامل منابع داده را بازنگری می نماییم. سپس ما شش فیلد تحلیل داده بسیار مهم از جمله تحلیل داده ساخت یافته، تحلیل تست، تحلیل وب سایت، تحلیل چند رسانه ای، تحلیل شبکه و تحلیل موبایل را بررسی می نماییم. در نهایت، ما چندین فیلد کاربرد کلیدی کلان داده را معرفی می نماییم.

٤-٧ تکامل های کاربرد

اخیرا، کلان داده به عنوان یک فناوری تحلیلی پیشرفته پیشنهاد شده است که معمولا شامل برنامه های پیچیده و مقیاس بزرگ تحت روش های تحلیلی خاص می باشد. درواقع، کاربرد های داده محور در دهه های قبلی ظاهر شده اند. برای مثال، هوش تجاری در ابتداری دهه ۱۹۹۰ به فناوری شایع برای کاربرد های تجاری تبدیل شده است و موتور های جستجو شبکه مبتنی بر پردازش داده کاوی انبوه در ابتدای قرن بیست و یکم ظاهر گردیده است. تعدادی از کاربرد های بالقوه و تاثیر گذار از فیلد های مختلف و داده و مشخصه های تحلیل اشان در زیر مورد بحث قرار می گیرند.

تکامل کاربرد های تجاری: جدی ترین داده تجاری به طور معمول از نوع داده ساخت یافته می باشد که توسط شرکت هایی از سیستم های به ارث مانده جمع اوری شده بود و سپس در RDBMS ها ذخیره شده اند. الزامات تحلیلی که در چنین سیستم هایی استفاده شده اند در دهه ۱۹۹۰ شایع بودند و درک مستقیم و ساده بوده است و برای مثال در شکل های گزارش ها، داشبورد، پرس و جو ها از وضعیت، هوش تجاری مبتنی بر جستجو، پردازش تراکنش آنلاین، بصری سازی متقابل، کارت های امتیاز، مدل سازی پیش گویانه و داده کاوی بودند. شبکه ها و صفحه وب جهانی از شروع قرن بیست ویکم یک فرصت منحصر به فرد را برای سازمان با هدف نمایش آنلاین و فعل و انفعال مستقیم با مشتریان فراهم کرده اند. محصولات فراوان و اطلاعات مشتری نظیر ثبت های داده جریان کلیک و رفتار کاربر را می توان از صفحه جهانی وب [1]بدست اورد. تحلیل متن و تکنیک های کاوش وب می توانند کارهایی نظیر بهینه سازی طرح کلی محصول، تحلیل تجارت مشتری، پیشنهادات محصول و تحلیل ساختار بازار را انجام دهند. همانطور که گزارش شده است، کمیت تلفن های موبایل و تبلت ها در ابتدا برتر از لپ تاپ ها و کامپوتر عای شخصی در سال ۲۰۱۱ بود. تلفن های موبایل و

[1] WWW

اینترنت اشیاء مبتنی بر حسگر ها یک نسل جدید از کاربرد های نواورانه را باز می کنند و به ظرفیت بسیار زیاد برای پشتیبانی از عملیات حسگری مکان، مردم محور و اگاه از مفهوم نیاز دارند.

تکامل اپلیکیشن های شبکه: نسل اولیه اینترنت عمدتا در زمینه خدمات ایمیل و WWW فعال بوده است . تحلیل متن، داده کاوی و تحلیل صفحه وب برای کاوش محتویات ایمیل و ایجاد موتور های جستجو اعمال شده اند. این روز ها صرف نظر از رشته و اهداف طراحی اشان بخش عمده کاربرد ها مبتنی بر وب هستند . داده شبکه یک درصد مهمی از حجم داده جهانی را پیش بینی می کند. وب به یک پلت فرم متداول برای صفحه وصل شده داخلی تبدیل شده اند که پر از انواع مختلف داده نظیر متن، تصاویر، صوت، ویدیو و محتویات فعل و انفعالی می باشد. از اینرو، فناوری های پیشرفته زیادی برای داده های نیمه ساخت یافته یا ساخت نیافته استفاده شده اند که بموقع ظاهر شده اند. برای مثال، تحلیل تصویر می تواند اطلاعات سودمند از تصاویر (برای مثال تشخیص چهره) استخراج نماید. فناوری های تجزیه و تحلیل چند رسانه ای را می توان برای خودکار سازی سیستم های تجسس ویدیویی برای کسب و کار، اجرای قانون و کاربرد های نظامی اعمال کرد. رسانه اجتماعی آنلاین نظیر نشست های اینترنتی، جوامع انلاین، بلاگ ها، سرویس های شبکه بندی اجتماعی و وب سایت های چند رسانه ای اجتماعی از سال ۲۰۰٤ یک سری فرصت های زیاد را برای خلق، آپلود کردن و تسهیم محتویات فراهم می نماند.

تکامل کاربرد های علمی: تحقیق علمی در رشته های متعدد در واقع کسب داده انبوه با حسگر های بازده بالاو ابزاری نظیر فیزیک نجوم، اقیانوس شناسی، ژن شناسی و تحقیق زیست محیطی می باشد. اخیرا بنیاد علمی ملی ایالات متحده یک برنامه BigData را برای بهبود تلاش ها در راستای استخراج دانش و بینش ها از مجموعه های بزرگ و پچیده داده دیجیتالی اعلان کرده است. تعدادی از قواعد تحقیق علمی باعث توسعه پلت فرم های کلان داده شده اند و به نتایج سودمندی دست یافته اند. برای مثال، iPlant در بیولوژی از زیرساخت شبکه، منابع رایانش فیزیکی، محیط هماهنگی، منابع ماشین مجازی، نرم افزار تحلیل همکارانه داخلی و سرویس داده برای کمک به محققان، مربیان و دانشجویان در غنی سازی علوم گیاهی استفاده می کند. مجموعه داده های iPlant از نظر شکل وازجمله در مشخصه یا داده مرجع، داده آزمایشی، داده مدل یا انالوگ، داده مشاهده و دیگر داده استخراج شده از تنوع بالایی برخوردارند.

۸-٤ فیلد های تجزیه و تحلیل کلان داده

همانگونه که بحث تبادل نظر شده است، ما می توانیم جستجو تحلیل داده را به شش فیلد فنی کلیدی یعنی تحلیل داده ساخت یافته، تحلیل داده متنی، تحلیل داده وب، تحلیل داده چند رسانه ای، تحلیل داده شبکه و تحلیل داده موبایل تقسیم نماییم. چنین دسته بندی در نظر دارد تا بر مشخصه های داده تاکید نماید اما تعدادی از فیلد ها ممکن است از فناوری های اساسی مشابه بهره ببرند. چون تحلیل داده دارای گسترده وسیعی بوده است

و پوشش جماع به راحتی حاصل نمی گردد، ما بر مشکلات کلیدی و فناوری ها در تحلیل داده در بحث های ذیل تمرکز خواهیم کرد.

۱-۸-٤تجزیه و تحلیل داده ساخت یافته

اپلیکیشن های تجاری و تحقیق علمی ممکن است داده ساخت یافته انبوه را تولید نمایند که مدیریت و تحلیل در انها به فناوری های تجاری شده بالغ نظیر RDBMS، انبار داده، OLAP و BPM (مدیریت فرایند کسب و کار) تکیه می کنند. تحلیل داده عمدتا مبتنی بر داده کاوی و تحلیل اماری می باشد که هر دو انها در سی سال گذشته مورد بررسی قرار گرفته اند.

از اینرو، تحلیل داده هنوز یک حوزه تحقیق بسیار فعال می باشد و تقاضا های کاربرد جدید باعث هدایت توسعه روش های جدید می شوند. برای مثال، فراگیری ماشین اماری بر اساس مدل های ریاضی دقیق و الگوریتم های قدرتمند برای کشف وضعیت غیر عادی و کنترل انرژی استفاده شده اند. بهره برداری از مشخصه های داده، زمان و کاوش فضا می تواند ساختار های دانش پنهان در جریان های داده سرعت بالاو حسگر ها را استخراج نماید. داده کاوی محافظت از اسراز شخصی که در حوزه محافظت از اطلاعات محرمانه در تجارت الکترونیک، دولت الکترونیک و کاربرد های مراقبت بهداشتی مورد توجه قرار گرفته است یک حوزه تحقیق نوظهور می باشد. کاوش فرایند در دهه گذشته به حوزه تحقیق جدید بویژه در تحلیل فرایند با داده رویداد تبدیل می گردد.

۲-۸-٤ تحلیل داده متن

رایج ترین فرمت ذخیره سازی اطلاعات به صورت متنی یعنی ایمیل ها، اسناد تجاری، صفحات وب و رسانه اجتماعی می باشد. از اینرو، تحلیل متن به نظر می رسد تا نسبت داده ساخت یافته از پتانسیل تجارت محور بیشتر برخوردار باشد. معمولا، تحلیل متن یک فرایند برای استخراج اطلاعات سودمند و دانش از متن ساخت نیافته می باشد. متن کاوی به صورت بین رشته ای می باشد و شامل بازیابی اطلاعات، فراگیری ماشین امار ها، زبانشناسی رایانش و داده کاوی به طور ویژه می باشد. بخش عمده سیستم های متن کاوی بر اساس اظهارات متن و پردازش زبان طبیعی (NLP) می باشند که بر NLP تاکید بیشتری شده است. NLP اجازه تحزیه و تحلیل، تفسیر و حتی تولید متن را به کامپیوتر ها می دهد. تعدادی از روش های متداول NLP عبارتند از اکتساب لغوی، شفافیت مفهوم کلمه، برچسب زنی بخشی از گفتار و گرامر رایگان محتوی احتمالی. تعدادی از تکنیک های مبتنی بر NLP برای متن کاوی نظیر استخراج اطلاعات، مدل های موضوع، خلاصه سازی متن، دسته بندی، خوشه بندی، پاسخ به پرسش و ایده کاوی استفاده شده اند.

۳-۸-٤ تجزیه و تحلیل داده وب

تحزیه و تحلیل داده وب همانند یک رشته تحقیق فعال ظهور کرده است. بازیابی خودکار، استخراج و ارزیابی اطلاعات از اسناد وب و سرویس ها برای کشف دانش سودمند از جمله اهداف تحلیل داده وب محسوب می شوند. تحلیل وب به چندین فیلد تحقیاتی نظیر پایگاه داده، بازیابی اطلاعات، NLP و متن کاوی ربط دارد. ما مطابق با بخش های مختلفی که قرار است کاوش گردند. تحلیل داده وب را به سه فیلد مرتبط دسته بندی می نمایند: کاوش محتوی وب، کاوش ساختار وب و کاوش کاربرد وب.

کاوش ساختار وب شامل مدل هایی برای کشف ساختار های لینک وب می باشد. در اینجا، ساختار به نمودار های شماتیک لینک شده به وب سایت یا میان وب سایت های چند تایی گفته می شود. مدل ها بر اساس ساختار های مکان نگر ساخته می شوند که با هایپر لینک های دارای توصیف لینک یا بدون توصیف لینک شده اند. شباهت ها و همبستگی ها در میان وب سایت های متفاوت در این قبیل مدل ها آشکار می شوند و برای دسته بندی صفحات وب استفاده می شوند. استفاده کامل از مدل ها برای جستجو صفحات وب مربوطه توسط Page Rankو CLEVER صورت گرفته است. خزنده های موضوع محور یک مورد موفق دیگر با استفاده از مدل ها می باشد.

کاوش استفاده وب در نظر دارد تا داده های کمکی تولید شده از طریق گفتمان ها یا فعالیت ها را کاوش نماید. کاوش محتوی وب و کاوش ساختار وب از داده وب جامع استفاده می کنند. داده استفاده از وب شامل ثبت های دسترسی در سرورهای وب و سرور های پروکسی ، ثبت های تاریخی مرورگر ها، پروفایل های کاربر، داده ثبت ، جلسات یا مبادلات کاربر، ذخیره گاه، پرس و جو های کاربر، داده بوک مارک، کلیک های موس و مرور ها و هر نوع داده تولید شده از طریق فعل و انفعال با وب می باشد . وقتی سرویس های وب و وب ۲,۰ بالغ و رایج می شوند، داده های استفاده وب به طور فزاینده ای از تنوع بالا برخوردار خواهند شد. کاوش استفاده وب نقش های کلیدی در فضای شخصی شده، تجارت الکترونیکی، امنیت/محرمانگی شبکه و دیگر حوزه های در حال ظهور ایفاء می کند. برای مثال، سیستم های پیشنهاد گرد همکارانه می توانند تجارت الکترونیک را را طریق بکار گیری ترجیحات مختلف کاربران شخصی سازند.

٤-۸-٤ تجزیه و تحلیل داده چند رسانه ای

داده های چند رسانه ای (عمدتا شامل عکس، صوت و ویدیو) با سرعت خیره کننده ای رشد کرده اند که دانش سودمند را استخراج می کنند و معانی را از طریق تحلیل درک می نمایند. نظر به این که داده چند رسانه ای ناهمگن می باشد و بخش عمده چنین داده هایی حاوی اطلاعات غنی تر نسبت به داده ساخت یافته یا داده متن می باشد، استخراج اطلاعات با چالش بزرگ تفاوت های معنایی مواجه شده است. تحقیق در مورد تجزیه و تحلیل چند رسانه ای چندین قاعده را پوشش می دهد. تعدادی از اولویت های تحقیقاتی عبارتند از خلاصه

سازی چند رسانه ای، نشانه گذاری چند رسانه ای، فهرست سازی و بازیابی چند رسانه ای، پیشنهاد چند رسانه ای و کشف رویداد چند رسانه ای و غیره.

خلاصه سازی صوتی را می توان از طریق استخراج کلمات برجسته یا عباراتی از ابر داده ها یا سنتز کردن نمایش جدید تکمیل نمود. خلاصه سازی ویدیویی برای تفسیر توالی محتوی ویدیویی معرف یا بسیار مهم می باشد و می تواند دینامیک یا استاتیک باشد. روش های خلاصه سازی ویدیویی استاتیک از توالی فریم کلیدی یا فریم های کلید حساس– محتوی استفاده می کند تا ویدیو را نمایش دهد. چنین روش هایی ساده هستند و برای چندین کاربرد کسب و کار استفاده شده اند (برای مثال یاهو، آلتا ویستا و گوگل) اما عمکرد اشان ضعیف است. روش های خلاصه سازی دینامیک از سری های فریم ویدیویی برای نمایش فیلم استفاده می کنند و سنجش های هموار را برای طبیعی تر کردن منظر خلاصه سازی نهایی بر می گزینند. یک سیستم خلاصه سازی چند رسانه ای عنوان محور (TOMS) را پیشنهاد می دهند که می توانند اطلاعات مهم را به طور خودکار در یک ویدیو متعلق به حوزه موضوع معین و برا اساس مجموعه معین خصیصه های استخراج شده از ویدیو خلاصه می کنند.

نشانه گذاری چند رسانه ای یک سری برچسب ها را وارد می کند تا محتویات تصاویر و ویدیو ها را در سطوح نحوی و معنایی توصیف نماید. مدیریت، خلاصه سازی و بازیابی داده های چند رسانه ای را با این قبیل برچسب ها می توان به راحتی اجراء نمود. چون نشانه گذاری دستی متمرکز بر زمان و کار می باشد، نشانه گذاری خودکار بدوت مداخلات انسانی بی نهایت جذاب می شود. چالش اصلی نشانه گذاری چند رسانه ای خودکار یک تفاوت معنایی می باشد. هر چند پیشرفت بسیار زیاد صورت گرفته است، عملکرد روش های نشانه گذاری خودکار موجود هنوز به بهبود نیاز دارد. بعضی تلاش ها در زمان های اخیر صورت گرفته اند تا نشانه گذاری چند رسانه های دستی و خودکار را به طور همزمان کشف نمایند.

فهرست بندی چند رسانه ای و بازیابی شامل توصیف، ذخیره سازی و سازماندهی اطلاعات چند رسانه ای و کمک به کاربران برای مراجعه راحت و سریع منابع چند رسانه ای می باشند. معمولا فهرست بندی چند رسانه ای و بازیابی از پنج تکنیک تشکیل می شوند که عبارتند از: تحلیل ساختاری، استخراج خصیصه، داده کاوی، دسته بندی و نشانه گذاری، پرس و جو و بازیابی تحلیل ساختاری با هدف قطعه بندی ویدیو به اجزای ساختاری معنایی از جمله کشف مرز لنز ها، استخراج فریم کلید و قطعه بندی صحنه صورت گیرد. دومین تکنیک مطابق با نتیجه تحلیل ساختاری عبارتست از استخراج خصیصه که عمدتا شامل کاوش بیشتر خصیصه های فریم های کلید، اشیاء، متون و حرکت ها که بنیاد شاخص سازی ویدیویی و بازیابی هستند. داده کاوی، دسته بندی و نشانه گذاری قرار است برای خصیصه های استخراج شده برای یافتن شیوه های محتویات ویدیویی استفاده شوند و ویدیو ها را بر طبق مقوله های زمان بندی شده مطرح می کنند تا شاخص های

ویدیویی را تولید نمایند. سیستم به محض دریافت پرس و جو از روش سنجش شباهت استفاده می کند تا به ویدیو کاندید رجوع کند. نتیجه بازیابی باعث بهبود بازخورد مربوطه می شود.

پیشنهاد چند رسانه ای قرار است محتویات چند رسانه ای را مطابق با ترجیحات کاربران پیشنهاد دهد. پیشنهاد چند رسانه ای ثابت کرده است که یک رویکرد موثر برای ارایه خدمات شخصی شده می باشد. بخش عمده سیستم های پیشنهاد موجود را می توان به سیستم های محتوی محور و سیستم های مبتنی بر فیلترینگ– همکارانه دسته بندی نمود. خصیصه های عمومی کاربران یا علایق اشان و کاربران پیشنهادشده برای دیگر محتویات با خصیصه های مشابه در روش های مبتنی بر محتوی شناسایی می شوند. این روش ها تا حد زیادی به سنجش شباهت محتوی تکیه می کنند اما اکثر انها بواسطه محدودیت تحلیل و مشخصه های مازاد دچار مشکل می شوند. روش های مبتنی بر فیلترینگ همکارانه یک سری گروه ها را با علایق مشابه شناسایی می کنند و محتویان را مطابق با رفتار اشان برای اعضای گروه پیشنهاد می دهند. اخیرا، روش ترکیبی معرفی می گردد که مزیت های دو نوع روش فوق الذکر را یکچارچه می سازد تا کیفیت پیشنهاد را بهبود بخشد.

ارزیابی بازیابی ویدیو TREC برای کشف رخداد حادثه در کلیپ های ویدیویی بر اساس Event Kit توسط موسسه ملی استاندارد ها و فناوری (NIST) ایالات متحده شروع شده بود که حاوی بعضی توصیفات متنی مرتبط با مفاهیم و مثال های ویدیویی می باشد. محققان در یک الگوریتم جدید را در مورد کشف رویداد چند رسانه ای ویژه با استفاده از مثال های اموزش مثبت پیشنهاد داده است . تحقیق در مورد کشف رویداد ویدیویی هنوز در مرحله ابتدایی اش می باشد و عمدتا بر ورزش ها یا حوادث خبری، راه اندازی یا رویداد های غیر عادی در پایش ویدیو ها و دیگر حوادث مشابه با الگو های تکراری تمرکز می کند.

٥–٨–٤ تجزیه و تحلیل داده شبکه

تجزیه و تحلیل داده شبکه از تحلیل کمیتی اولیه و تجزیه و تحلیل شبکه جامعه شناختی بر طبق تحلیل شبکه اجتماعی آنلاین در شروع قرن بیست و یکم ظاهر شده اند. چندین سرویس شبکه بندی اجتماعی که شامل تویتر، فیسبوک و لینکدین و موارد دیگر هستند به طور فزاینده دای در عرض چند سال محبوب شده اند. این قبیل سرویس های شبکه اجتماعی به طور معمول شامل داده های لینک شده انبوه و داده محتوی هستند . داده لینک شده عمدتا در شکل ساختار های گرافیکی می باشد و ارتباطات بین دو هویت توصیف می شوند. داده محتوی حاوی متن، عکس و دیگر داده های چندد رسانه ای شبکه می باشد. محتوی غنی در چنین شبکه هایی سبب چالش های غیر منتظره و فرصت هایی برای تحلیل داده می شوند. تحقیق موجود در بافت های سرویس شبکه بندی اجتماعی مطابق با چشم انداز داده محور را می توان به دو مقوله دسته بندی نمود که عبارتند از : تحلیل ساختاری مبتنی بر لینک و تحلیل مبتنی بر محتوی.

تحقیق در زمینه تجزیه و تحلیل ساختاری مبتنی بر لینک همیشه به پیش بینی لینک، کشف اجتماع، تکامل شبکه اجتماعی و تحلیل تاثیر اجتماعی و موارد دیگر تعهد داشته است. SNS ممکن است همانند گراف ها بصری سازی شود که هر راس در آن گراف ها با کاربر شباهت دارد و لبه ها با همبستگی ها در میان کاربران شباهت دارند. چون SNS ها یک سری شبکه های دینامیک هستند، راس ها و لبه های جدید به طور دایمی به گراف ها افزوده می شوند. پیش بینی لینک قرار است احتمال اتصال آینده بین دو راس را پیش بینی نماید . تعدادی از تکنیک ها را می توان برای پیش بینی لینک یعنی دسته بندی مبتنی بر خصیصه، روش های احتمالی و جبر خطی استفاده نمود. دسته بندی مبتنی بر خصیصه قرار است تا یک گروه از خصیصه ها را برای راس انتخاب نماید و از اطلاعات لینک موجود برای تولید دسته کننده های دوتایی برای پیش بینی لینک آینده استفاده می کند. روش های احتمالی در نظر دارد تا مدل هایی رابرای احتمالات اتصال در میان راس ها در SNS می سازند. شباهت بین دو راس مطابق با ماتریس مشابه تکی در جبر خطی محاسبه می گردد. جامعه از طریق ماتریس گرافیک فرعی نمایش داده می شود که لبه ها در این ماتریس یک سری راس ها را در دانسیته بسیار کمتر خصیصه گراف های فرعی وصل می کنند در حالی که لبه های بین دو گراف فرعی یک دانسیته بسیار کمتر را نشان می دهند.

تعدادی از روش ها برای کشف اجتماع پیشنهاد شده و بررسی شده اند که اکثر انها یک سری کار های هدف مبتنی بر جانمایی هستند و بر مفهوم درک ساختار اجتماعه تکیه می کنند. دو و همکارانش از ویژگی هم پوشانی جوامع در زندگی واقعی استفاده کرده اند تا روش کشف جامعه SNS مقیاس بزرگ موثر را پیشنهاد دهند. تحقیق در مورد SNS در نظر دارد تا بدنبال قانون و مدل استنتاح قیاسی باشد تا تکامل شبکه تفسیر شود. بعضی بررسی های تجربی مشخص کرده اند که بایاس نزدیکی، محدودیت های جغرافیایی و دیگر فاکتور ها نقش های مهمی در تکامل SNS ایفاء می کنند و بعضی روش های تولید برای کمک به شبکه و طراحی سیستم پیشنهاد می شوند.

تاثیر اجتماعی به موردی گفته می شود که در ان زمان افراد رفتار اشان را تحت تاثیر دیگران تغییر می دهند. استحکام تاثیر اجتماعی به رابطه میان افراد، فواصل شبکه، اثر زمانی و مشخصه های شبکه ها و افراد بستگی دارد. بازاریابی، تبلیغ، پیشنهاد و دیگر کاربرد ها می توانند از طریق سنجش کمیتی و کیفی تاثیر افراد بر دیگران از تاثیر اجتماعی سود ببرند. معمولا اگر تکثیر محتویات در SNS موردتوجه قرار می گیرد ، عملکرد تحلیل ساختاری مبتنی بر لینک ممکن است بیشتر بهبود یابد.

تحلیل مبتنی بر محتوی در SNS همانند تحلیل رسانه اجتماعی شناخته می شود. رسانه اجتماعی شامل متن، چند رسانه ای، موقعیت یابی و نظرات می باشد. از اینرو، تحلیل رسانه اجتماعی با چالش های بی سابقه مواجه می گردد. اولا، داده های رسانه اجتماعی در حال رشد دایمی و انبوه بایستی بر طبق پنجره زمانی منطقی به طور خودکار تجزیه و تحلیل شوند. ثانیا، داده رسانه اجتماعی حاوی نویز بسیار زیاد می باشد . برای مثال، فضای بلاگ حاوی تعداد زیاد بلاگ های اسپم (spam) می باشد و از اینرو توییت های ناچیز در توییتر انجام می دهد.

ثالثا، SNS ها از نوع شبکه های دینامیک هستند که مکررا و دایما فرق دارند و به روز می شوند. تحقیق موجود در زمینه تحلیل رسانه اجتماعی هنوز در مرحله اولیه اش می باشد. با در نظر گرفتن این موضوع که SNS حاوی اطلاعات انبوه می باشد، انتقال یادگیری در شبکه های ناهمگن با هدف جابجایی اطلاعات دانشی در میان رسانه های مختلف صورت می گیرد.

٦-٨- ٤ تجزیه و تحلیل داده موبایل

Android Apps تا اوریل ۲۰۱۳ بیش از ٦٥٠٠٠٠ اپلیکیشن را مهیاء کرده است و تقریبا کل دسته ها را پوشش می دهد. جریان داده موبایل ماهانه تا انتهای سال ۲۰۱۲ به ۸۸٥ پتابایت رسیده است. داده انبوه و کاربرد های فراوان در جستجو تحلیل موبایل هستند اما همچنین سبب چالش های کمی می شوند. در کل، داده موبایل دارای مشخصه های منحصر به فرد یعنی حسگری موبایل، انعطاف پذیری حرکت، نویز و مقدار زیاد افزونگی می باشد. اخیرا، تحقیق جدید در مورد تحلیل موبایل در حوزه های مختلف شروع شده است. چون تحقیق در مورد تحلیل موبایل دقیقا شروع شده است، ما تنها کاربرد های تحلیل نمونه و اخیر را در این بخش معرفی خواهیم کرد.

تلفن های موبایل با رشد تعداد کاربران موبایل و بهبود عملکرد در زمان حال برای ساخت و حفظ جوامعی نظیر جوامع با مکان های جغرافیایی و جوامع مبتنی بر پس زمینه های فرهنگی و علایق متفاوت سودمند هستند (برای مثال جدید ترین وب چت). جوامع شبکه سنتی یا جوامع SNS بطور مختصور یک تعامل انلاین در میان اعضاء هستند و جوامع تنها در زمانی فعال هستند که اعضاء قبل از کامپیوتر ها می نشینند .بطور عکس، تلفن های موبایل می توانند از فعل و انفعال غنی در هر زمان و هر مکانی پشتیبانی نمایند. جوامع موبایل همانند یک گروه از افراد با سرگرمی های یکسان (یعنی بهداشت، ایمنی و سرگرمی) تعریف می شوند که به همراه هم در شبکه ها جمع می شوند، برای مطرح کردن هدف مشترک ملاقات می کنند و برای اجرای طرح اشان دست به کار می شوند. محققان یک مدل کیفی از جامعه موبایلی را پیشنهاد داده اند. اکنون این باور تا حد زیادی وجود دارد که اپلیکیشن های جامعه موبایل تا حد زیادی توسعه صنعت موبایل را بهبود خواهد بخشید.

اخیرا پیشرفت در زمینه حسگر های بی سیم، فناوری ارتباط موبایلی و پردازش جریان به مردم این قدرت را می دهد تا شبکه منطقه بدن را بسازند تا سلامت اشان را به طور انی پایش نمایند. معمولا، داده پزشکی از انواع حسگر ها دارای مشخصه های متفاوت بر حسب ویژگی ها، روابط زمان و فضا و همچنین روابط روانشناختی و موارد دیگر می باشند.

بعلاوه، چنین محموعه های داده هایی شامل محافظت محرمانگی و ایمنی هستند. گراگ و همکارانش یک مکانیزم تجزیه و تحلیل حمل و نقل چند نمایی داده خام را برای پایش آنی بهداشت معرفی می کنند. تحت

شرایطی که تنها مشخصه های بی نهایت جامع مرتبط با بهداشت در دسترس هستند، پارکو همکارانش به بررسی رویکرد ها برای استفاده بهتر پرداخته اند.

محققان کالج دانشگاه گجویچ نروژ و Derawi Biometrics برای توسعه کاربرد تلفن های هوشمند همکاری کرده اند که گام ها را در زمانی که مردم پیاده روی می کنند و از اطلاعات گام برای باز کردن قفل سیستم ایمنی استفاده می کند، تجزیه و تحلیل می نمایند. در ضمن ، روبرت دلانو و بریان پاریسه از موسسه فناوری گرجستان یک اپلیکیشن را توسعه داده اند که iTrem نامیده شد که لرزش بدن انسان را با لرزه نگار در داخل تلفن موبایل پایش می کند تا بر بیماری های سیستم عصبی و پارکینسون غلبه نماید.

٤-٩ کاربرد های کلیدی کلان داده

٤-٩-١ کاربرد کلان داده در شرکت ها

اکنون کلان داده عمدتا از شرکت ها ظاهر می شود و عمدتا در انها استفاده می شود در حالی هوش تجاری و OLAP را می توان نسل های قبلی اپلیکیشن کلان داده به حساب اورد. اپلیکیشن کلان داده در شرکت ها می تواند راندمان تولید و رقابت پذیری اشان را در بعضی جنبه ها ارتقاء بخشد. بویژه، شرکت ها در بازاریابی با تحلیل همبستگی کلان داده می توانند رفتار مشتری را با دقت پیش بینی نمایند و شیوه های کسب و کار جدید را بیابند. شرکت ها در برنامه ریزی فروش ها بعد از مقایسه داده انبوه می توانند قیمت های کالا اشان را بهبود بخشند. شرکت ها در عملیات می توانند راندمان و رضایتبخشی فعالیت اشان را بهبود بخشند، نیروی کار را بهینه سازند، با دقت الزامات تخصیص پرسنل را پیش بینی نمایند، از ظرفیت تولید مازاد جلوگیری می کنند و هزینه کارگر را کاهش می دهند . شرکت ها در زنجیره تامین با استفاده از کلان داده می توانند بهینه سازی فهرست موجودی، بهینه سازی لجستیکی و هماهنگی تامین کننده را انجام می دهند تا شکاف بین تامین و تقاضا را کاهش دهند، بودجه ها را کنترل نمایند و خدمات را بهبود بخشند.

کاربرد کلان داده در شرکت های امور مالی به سرعت توسعه یافته است. برای مثال، بانک مرچانتس چین (CMB) از تحلیل داده استفاده می کند تا تشخیص دهد که چنین فعالیت هایی همانند تجمیع امتیاز چند مرتبه ای و مبادله امتیاز در مغازه ها برای جذب مشتریان کیفی موثر هستند. بانک از طریق ایجاد مدل هشدار حذف تصادفی مشتری می تواند محصولات مالی پربازده را به ۲۰ درصد مشتریان برتر می فروشند که به احتمال زیاد برای حفظ انها به طور تصادفی حذف می شوند. در نتیجه، نسبت حذف تصادفی مشتریان با کارت های طلایی و کارت های گل آفتابگردان به ترتیب تا ۱۵ و ۷ درصد کاهش یافته اند. مشتریان کسب و کار کوچک بالقوه از طریق تحلیل ثبت های تراکنس مشتریان می توانند به طور کارآمد شناسایی شوند. بهره های قابل توجه عملکرد از طریق استفاده از بانکداری راه دور و پلت فرم ارجاعی ابری برای اجرای فروش جانبی کسب گردیدند.

بخش عمده کاربرد کلاسیک به طور واضح در تجارت الکترونیک می باشد. ده ها هزار تراکنش در سایت Taobao انجام می گیرند و زمان تراکنش مشابه، قیمت های کالا و کمیت های خرید هر روز ثبت می شوند . مکعب داده Taobao یک اپلیکیشن کلان داده در پلت فرم Taobao می باشد که بازرگان ها از طریق ان می توانند از وضعیت قابل رویت صنعتی پلت فرم Taobao، شرایط بازار برند هایشان و رفتار های مصرف کنندگان مطلع شوند. در واقع، مشتریان بیشتری می توانند کالا های دلخواه اشان را با قیمت های قابل ترجیح بیشتر بخرند. وام اعتباری علی بابا به طور خودکرا تجزیه و تحلیل می گردد و قضاوت می کند که آیا دادن وام به شرکت ها از طریق داده تراکنش شرکت در ظاهر فناوری کلان داده کسب شده است در حالی که مداخله دستی در کل فرایند رخ نمی دهد. این موضوع افشاء می گردد که علی بابا تا اینجا بیشتر از ۳۰ میلیارد بوان RMB با تنها حدود ۰,۳ درصد وام های بد قرض داده است که تا حد زیادی کمتر از وام های دیگر بانک های تجاری می باشد.

۲-۹-٤کاربرد کلان داده مبتنی بر IoT

IoT نه تنها منبع مهم کلان داده می باشد بلکه همچنین یکی از بازار های مهم اپلیکیشن های کلان داده می باشد. اپلیکیشن های IoT به دلیل تنوع بالااشیاء به طور دایمی ظاهر می شوند .

شرکت های لجستیک ممکن است کاربرد کلان داده IoTرا زیاد تجربه کرده اند. برای مثال، کامیون های UPS به حسگر ها، آداپتورهای بی سیسم و جی پی اس مجهز می شوند از اینرو ستاد مرکزی می تواند موقعیت های کامیون را ردیابی نماید و از شکست های موتور جلوگیری می کند. در واقع، همچنین این سیستم به UPS کمک می کند تا برکارکنان اش نظارت کرده و مدیریت نماید و مسیر های تحویل را بهینه سازد. مسیرهای تحویل بهینه که برای کامیون های UPS مشخص شده اند از تجربه رانندگی قبلی اشان استخراج می شوند. راننده های UPS در سال ۲۰۱۱ به طور تقریبی ٤٨,۲٨ میلیون کیلومتر کمتر رانندگی کرده اند.

شهر هوشمند یک منطقه تحقیق داغ مبتنی بر کاربرد داده IoT می باشد. برای مثال، همکاری پروژه شهر هوشمند بین بخش Miami-Dade در فلوریدا و IBM به طور نزدیکی ۳۵ نوع دپارتمان دولتی محله کلیدی و شهر میامی را وصل می کند و به رهبران دولت کمک می کند تا به پشتیبانی اطلاعات بهتر در تصمیم گیری برای مدیریت بر منابع آب، کاهش پارازیت ترافیک و بهبود ایمنی عمومی دست یابند. کاربرد شهر هوشمند سبب سود هایی در جنبه های مختلف برای منطقه Dade می گردد. برای مثال، دپارتمان مدیریت پارک منطقه Dade به دلیل شناسایی بموقع و تعمیر لوله های آب که در این سال نشتی داشته اند یک میلیون دلار در صورتحساب های آب صرفه جویی می کند.

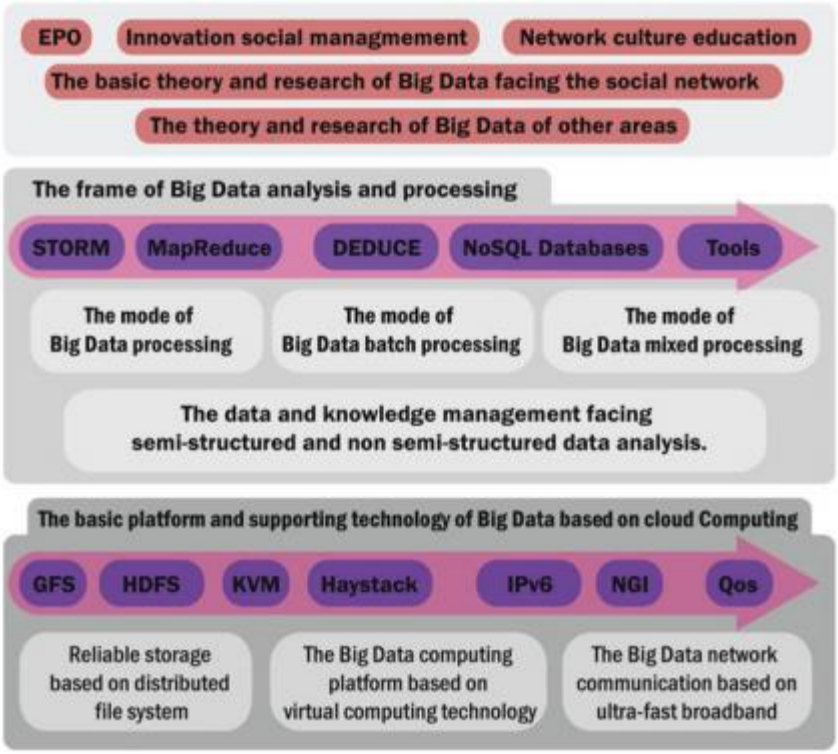
۳-۹-٤ کاربرد کلان داده شبکه محور اجتماعی آنلاین

SNS انلاین یک ساختار اجتماعی می باشد و از افراد اجتماعی و ارتباطات بین افراد بر اساس شبکه اطلاعات تشکیل می گردد. کلان داده SNS آنلاین عمدتا از پیام های آنی، اجتماعی آنلاین، میکرو بلاگ و فضای مشترک ناشی می گردد که انواع فعالیت های کاربر را معرفی می کند. تحزیه و تحلیل کلان داده از SNS انلاین از روش های محاسباتی تحلیلی استفاده می کند که ظاهرا از طریق فرضیه ها و روش هایی برای درک روابط در جامعه انسانی فراهم شده است و شامل ریاضیات، اطلاع رسانی، جامعه شناحتی و علم مدیریت و موارد دیگر از سه بعد می باشد و ساختار شبکه، فعل و انفعال گروه و گسترش اطلاعات را منظور می کند. اپلیکیشن شامل تجزیه و تحلیل ایده عمومی شبکه، جمع اوری و تحلیل هوش شبکه و تحصیل انلاین می باشد. چارچوب فنی اپلیکیشن کلان داده SNS انلاین در تصویر٤ توضیح داده می شود. اپلیکیشن های کلاسیک کلان داده از SNS انلاین درذیل معرفی می شوند که عمدتا اطلاعات محتوی و اطلاعات ساختاری برای کسب مقادیر را کاوش کرده و تحلیل می نماید.

اپلیکیشن های مبتنی بر محتوی : زبان و متن دو شکل بسیار مهم ارایه در SNS هستند. ترجیح، احساس، علاقه و تقاضا کاربر از طریق تحلیل زبان و متن ممکن است آشکار گردند.

اپلیکیشن های مبتنی بر ساختار : کاربران در SNS همانند گره نشان داده می شوند در حالی که رابطه اجتماعی ، علاقه و سرگرمی ها و مواردیگر یک سری روابط را درمیان کاربران درون ساختار خوشه ای تجمیع می شوند. چنین ساختاری با روابط نزدیک میان افراد داخلی اما روابط بیرونی بی قاعده را جامعه نامیده اند. تحلیل مبتنی بر جامعه دارای اهمیت حیاتی برای بهبود پخش اطلاعات و برای تحلیل رابطه بین فردی دارد .

آزمایش دپارتمان پلیس سانتا کروز ایالات متحده از طریق اعمال داده برای تحلیل پیشگویانه تجربه شده است . دپارتمان پلیس از طریق تحلیل SNS می تواند روندها و شیوه های جرم را کشف نماید و حتی نرخ های جرم را در مناطق مهم کشف نماید..



تصویر٤- فناوری های فعال برای داده های بزرگ آنلاین شبکه اجتماعی

شرکت موتور جستجو و رایانش ولفرام آلفا دراوریل ۲۰۱۳ در مورد قانون رفتار اجتماعی از طریق تجزیه و تحلیل داده اجتماعی بیش از یک میلیون کاربرامریکایی فیسبوک بررسی انجام داده است. شرکت مطابق با تحلیل متوجه شده بود که بخش عمده کاربران فیسبوک در ابتدای ۲۰ سالگی در دام عشق گرفتار می شوند و وقتی انها حدود ۲۷ سال سن دارند نامزد می کنند سپس وقتی حدود سی سال سن دارند ازدواج می کنند. در نهایت ، روابط ازدواج اشان یک سری تغییرات کند را بین ۳۰ تا ٦۰ سالگی نشان می دهد. این قبیل نتایج تحقیق با داده سرشماری جمعیت شناختی ایالات متحده هماهنگ هستند. بعلاوه، شرکت گلوبال پالس یک تحقیقی را انجام داده بود که بعضی قوانین را در فعالیت های اجتماعی و اقتصادی با استفاده از داده SNS آشکار کرده بود. این پروژه از پیام های تویتری به زبان های انگلیسی، ژاپنی و اندونزیایی از جولای ۲۰۱۰ تا اکتبر ۲۰۱۱ که در دسترس عموم می باشد، استفاده کرده است تا موضوعات مرتبط با غذا ، سوخت ، مسکن و وام را تجزیه و تحلیل نماید . درک بهتر رفتار عمومی و نگرانی ها می تواند هدف این تحلیل باشد. این پروژه به تحلیل کلان داده SNS از جنبه های مختلف پرداخته است:

(۱) پیش بینی رخداد حوادث غیر عادی از طریق کشف رشد تند یا افت مقدار عناوین،

(۲) مشاهده روند های گفتمان هفتگی و ماهانه در تویتر ؛ پدید اوری مدل ها برای تغییر پذیری در سطح توجه در موضوعات خاص با گذشت زمان؛

(۳) درک روند های تبدیل موضوعات فرعی جانبی متفاوت و

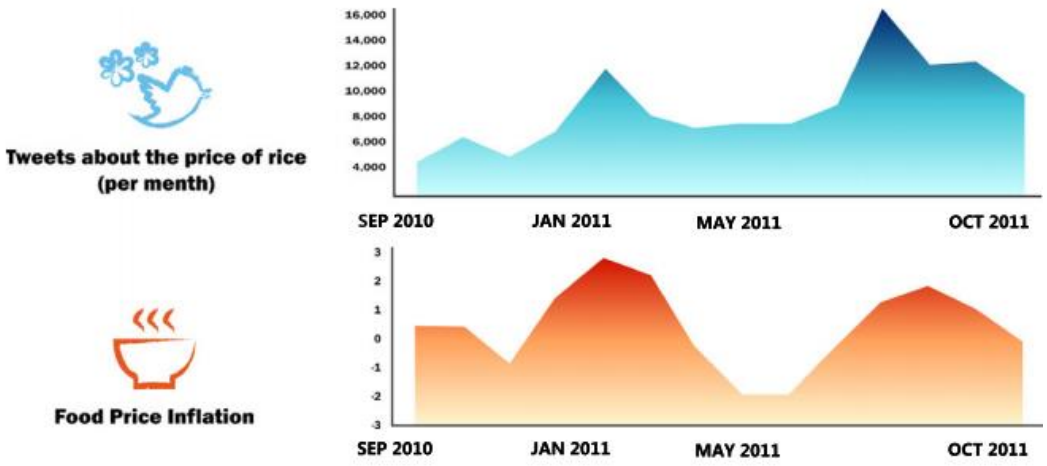
(٤) پیش بینی روند ها با شاخص های بیرونی که در گفتمان های تویتر دخیل بوده اند. پروژه به عنوان مثال کلاسیک کشف کرده است که تغییر تورم قیمت غذا از امار های آفلاین اندونزی با تعداد توییت ها با قیمت برنج در تویتر مطابقت دارد که درتصویر۵ نشان داده می شود.

به طور کلی، کاربرد کلان داده از SNS انلاین ممکن است به درک بهتر رفتار کاربر و تسلط قوانین فعالیت های اقتصادی و اجتماعی از سه جنبه زیر کمک نماید:

هشدار اولیه: برای غلبه سریع بر بحران اگر از طریق کشف موارد غیر عادی در استفاده از دستگاه ها و خدمات الکترونیکی مشخص گردد.

پایش انی: ارایه اطلاعات دقیق برای تدوین سیاست ها و طرح از طریق پایش رفتار جاری، احساس و ترجیح کاربران.

بازخورد آنی: کسب باخورد های گروها در برابر بعضی فعالیت های اجتماعی بر اساس پایش انی.



تصویر ۵- ارتباط بین تویت های مربوط به قیمت برنج و تورم قیمت مواد غذایی

٤-٩-٤ کاربرد های کلان داده مراقبت بهداشتی و پزشکی

داده های مراقبت بهداشتی و پزشکی به طور دایمی و خیلی سریع در حال رشد داده پیچیده هستند و حاوی مقادیر اطلاعات متنوع و فراوان می باشند. کلان داده دارای پتانسیل نامحدود ذخیره سازی موثر، پردازش، پرس و جو و تحلیل داده پزشکی بوده است. کاربرد کلان داده پزشکی تا حد زیادی بر کسب و کار مراقبت بهداشتی تاثیر داشته است.

برای مثال، شرکت بیمه اتنا لایف تعداد ۱۰۲ بیمار را از مخزنی با هزار بیمار انتخاب کرده بود تا آزمایش را با هدف کمک به پیش بینی ریکاوری بیماران سندروم متابولیک را تکمیل نماید. شرکت در آزمایش مستقل تعداد ٦۰۰۰۰۰ نتیجه تست آزمایشگاهی و ۱۸۰۰۰۰ ادعا را از طریق سری های نتایج تست کشف سندروم متابولیک

بیماران در سه سال متوالی اسکن کرده بود. بعلاوه، شرکت نتیجه نهایی را به طرح درمان شخصی شده بی نهایت خلاصه کرده است تا فاکتور های خطرناک و طرح های درمان مهم بیماران را براورد نماید. سپس، پزشکان ممکن است با تجویز استاتین ها و کمک به بیماران برای کاهش وزن تا پنج چوند یا پیشنهاد به بیماران برای کاهش تری گلسیرید کلی در بدن هایشان ناخوشی را تا ۵ درصد در ده سال اینده کاهش می دهند اگر محتوی قند در بدن هایشان بیش از ۲۰ می باشد.

مرکز پزشکی مونت سیانی در ایالات متحده از فناوری های شرکت کلان داده Ayasdi استفاده می کند تا کل توارث ژنتیکی Escherichia Coli از جمله بیش از یک میلیون گونه DNA را تجزیه تحلیل نماید تا بررسی نماید چرا رشته های باکتریایی در انتی بیوتیک ها مقاومت می کنند. شرکت Ayasdi از تحلیل داده جانمایی یا روش تحقیق ریاضی برند جدید استفاده می کند تا مشخصه های داده را درک نماید.

برنامه Health Vault شرکت مایکروسافت که در سال ۲۰۰۷ راه اندازی شده است یک اپلیکیشن عالی کلان داده پزشکی می باشد که در سال ۲۰۰۷ راه اندازی شده است. هدف اش این است تا اطلاعات بهداشتی انفرادی را در دستگاه های پزشکی انفرادی و خانواده مدیریت نماید. اخیرا، اطلاعات بهداشتی را می توان وارد کرد و با دستگاه های موبایل هوشمند آپلود نمود و از ثبت های پزشکی انفرادی از طریق آژانس طرف ثالث وارد شده اند. بعلاوه، ان را می توان با اپلیکیشن طرف سوم با کیت توسعه نرم افزاری و واسط باز یکپارچه نمود.

۵-۹-٤ هوش جمعی

تلفن های موبایل و تبلت ها با توجه به توسعه سریع ارتباط بی سیم و فناوری های حسگر به طور فزاینده ای دارای ظرفیت های حسگری و رایانش قوی تر شده اند. در نتیجه، حسگری جمعیت به موضوع کلیدی در رایانش موبایل تبدیل می شود. تعداد زیادی از کاربران در حسگری جمعیت از دستگاه های موبایل به عنوان واحد های حسگری اصلی برای انجام هماهنگی با شبکه های موبایل برای توزیع وظایف حس شده و جمع اوری و استفاده از داده های حس شده استفاده می کنند. این حسگری به ما کمک می کند تا وظایف حسگری اجتماعی پیچیده و مقیاس بزرگ را تکمیل نماییم. شرکت کننده ها یی که در حسگری جمعیت یک سری وظایف حسگری پیچیده را تکمیل می کنند به مهارت های حرفه ای نیاز ندارند. حسگری جمعیت در شکل منبع یابی جمعیت به طور موفقیت امیزی برای عکس دارای برچسب جغرافیایی، موقعیت یابی و ناوبری، حسگری ترافیک شهری، پیش بینی بازار، ایده کاوی و دیگر کاربرد های متمرکز بر کار استفاده شده است.

منبع یابی جمعیت که یک رویکرد جدید برای حل مسئله می باشد یک تعداد کاربرات عمومی را همانند زیرساخت به حساب می اورد و وظایف را در وضعیت آزاد و داوطلبانه توزیع می کند. در واقع، منبع یابی جمعینی از طریق چندین شرکت قبل از ظهور کلان داده استفاده شده است. برای مثال، P&G، BMW و Audio کار تحقیق و توسعه و ظرفیت های طراحی اشان را در ظاهر از طریق منبع یابی جمعیت بهبود بخشیده اند. ایده

اصلی منبع یابی جمعیت این است تا وظایف را برای کاربران عمومی توزیع نماید و وظایفی را تکمیل نماید که کاربارن انفرادی نمی توانستند یا نمی خواهند تکمیل نمایند. منبع یابی جمعیت بدون نیاز به بکار گیری عمدی ماژول های حسگری و بکار گیری حرفه ای می تواند گستره سیستم حسگری را گسترش بخشد تا به مقیاس شهر و حتی مقیاس های بزرگ تر برسد.

منبع یابی جمعیت فضایی در عصر کلان داده به یک موضوع داغ تبدیل می شود. چارجوب کارکردی منبع یابی جمعیت فضایی در زیر نشان داده می شود. کاربر ممکن است درخواست سرویس و منابع مرتبط با مکان مشخص شده را نماید. سپس کاربران موبایل که قصد دارند تا در وظیفه شرکت نمایند، به مکان مشخص شده برای کسب داده مرتبط (نظیر ودییو ، صوت یا تصاویر) دست می یابند. در نهایت، داده کسب شده برای درخواست کننده سرویس فرستاده خواهد شد . با توجه به رشد سریع دستگاه های موبایل و عملکرد های قدرتمند ارایه شده از طریق این دستگاه های موبایل می توان پیش بینی نمود که منبع یابی جمعیت فضایی نسبت به منبع یابی جمعیت سنتی نظیر آمازون Turk و Crowdflower شایع تر خواهد شد .

٦-٩-٤ شبکه هوشمند

شبکه هوشمند در واقع شبکه قدرت نسل بعدی می باشد که از شبکه های انرژی سنتی یکی شده با رایانش، ارتباطات و کنترل برای تولید بهینه، تامین و مصرف انرژی الکتریکی تشکیل می گردد. شبکه هوشمند مرتبط با کلان داده از انواع منابع نظیر موارد زیر تولید می شوند که عبارتند از:

(١) عادات استفاده از توان توسط کاربران،

(٢) داده سنجش فازور که از طریق واحد سنجش فازور بکار رفته در گستره ملی اندازه گیری می شوند،

(٣) داده مصرف انرژی که از طریق متر های هوشمند در زیرساخت اندازه گیری پیشرفته (AMI) اندازه گیری شده اند ،

(٤) قیمت گذاری بازار انرژی و داده مزایده ،

(٥) داده مدیریت، کنترل و نگهداری دستگاه ها و تجهیزات در تولید، انتقال و شبکه های توزیع قدرت .

شبکه هوشمند سبب چالش های زیر در بهره برداری زا کلان داده می شود:

برنامه ریزی شبکه: مناطقی را از طریق تجزیه و تحلیل داده در شبکه هوشمند می توانند شناسایی نمود که دارای باز الکتریکی بالا اضافی یا فرکانس های قطع برق قدرت بالا هستند. حتی خطوط انتقال با احتمال شکست بالا را می توان شناسایی نمود. چنین نتایجی تحلیلی ممکن است در به رزو رسانی شبکه، انتقال و تعمیر و نگهداری و موارد دیگر مشارکت داشته باشند. برای مثال، محققان دانشگاه کالیفرنیا واقع در لس آنجلس یک نقشه الکتریکی را مطابق با فرضیه کلان داده طراحی کرده اند و نقشه کالیفرنیا را از طریق یکی کردن اطلاعات سرشماری و اطلاعات استفاده آنی از برق که از طریق شرکت های برق الکتریکی تهیه شده اند، ترسیم کرده بودند. نقشه هر بلوک را همانند واحدی برای نمایش مصرف برق هر بلوک در همان لحظه به حساب می اورد. نقشه حتی می تواند مصرف برق بلوک را با میانگین سرانه درامد و نوع ساختمان مقایسه نماید تا عادات مصرف دقیق تر برق انواع گروه ها در جامعه را آشکار سازد. این نقشه یک

پیش بار بصری و موثر را برای برنامه ریزی شبکه برق در شهر فراهم می کند. تغییر ترجیحی در تاسیسات شبکه برق در بلوک های با فرکانس های قطعی قدرت بالا و بار های اضافی جدی ممکن است انجام گیرد که همینطور در نقشه نمایش داده شده است.

فعل و انفعال بین تولید قدرت و مصرف قدرت: شبکه قدرت ایده آل بایستی تولید و مصرف قدرت را متوازن سازد. از اینرو، شبکه قدرت سنتی بر اساس رویکرد تک مسیری انتقال – تبدیل – توزیع – مصرف ساخته می شود که تطبیق ظرفیت تولید را مطابق با تقاضا مصرف قدرت مجاز نمی سازد از اینرو به افزونگی و اتلاف انرژی الکتریکی منجر می شود. بنابراین، متر های الکتریکی هوشمند برای بهبود راندمان تامین برق توسعه یافته اند. TXU Energy دارای چندین کاربرد موفق مترهای الکتریکی هوشمند بوده است که می توانند به داده بهره برداری قدرت خواندن تامین کننده در هر ۱۵ دقیقه بجای هر ماه در گذشته کمک نمایند. هزینه کارگر برای خواندن متر تا حد زیادی کاهش می یابد چون داده های استفاده برق به کرات و به سرعت تهیه و تحلیل می شوند، شرکت های تامین برق می توانند قیمت الکتریسیته را مطابق با راس و دوره های کم مصرف برق تنظیم نمایند. TXU Energy از این سطح قیمت برای پایداری نوسانات راس و کم مصرف های برق استفاده کرده است. در واقع کاربرد کلان داده در شبکه هوشمند می تواند به تحقق قیمت گذاری دینامیک تقسیم زمانی کمک نماید که یک موقعیت برد– برد برای تامین کننده های انرژی و کاربران می باشد.

دسترسی انرژی تجدید پذیر متناوب: اکنون چندین منبع انرژی نظیر باد و خورشید را می توان به شبکه های قدرت وصل نمود. از اینرو، چون ظرفیت های تولید قدرت منابع انرژی جدید به طور نزدیک با شرایط جوی ربط دارند که تصادفی بودن و تناوبی بودن را نشان می دهند، اتصال انها به شبکه های برق چالش برانگیز می باشد. اگر کلان داده شبکه های برق به طور موثر تجزیه و تحلیل گردد، این قبیل منابع انرژی جدید تجدید پذیر متناوب را می توان به طور موثری مدیریت نمود: الکتریسیته تولیدی منابع انرژی جدید رامی توان به مناطق با کمبود الکتریسیته تخصیص داد . چنین منابع انرژی می توانند مکمل قوه محرکه مولد برق و تولید های برق حرارتی باشند .

۱۰-٤ جمع بندی

ما در مقاله حاضر به بازنگری پس زمینه و جدید ترین فناوری کلان داده می پردازیم. اولا، ما دورنمای کلی کلان داده را معرفی می کنیم و فناوری های مرتبط نظیر رایانش ابری، اینترنت اشیاء، مراکز داده و هادوپ را بازنگری می نماییم. سپس ما بر چهار فاز زنجیره ارزش کلان داده یعنی تولید داده، اکتساب داده، ذخیره سازی داده و تحلیل داده تمرکز می کنیم. ما برای هر فاز یک دورنمای کلی را معرفی می کنیم، در مورد چالش های فنی بحث می نماییم و جدید ترین پیشرفت ها را بازنگری می کنیم. در نهایت چندین کاربرد نمونه کلان داده از جمله مدیریت شرکت، اینترنت اشیاء، شبکه های اجتماعی، کاربرد های پزشکی، هوش جمعی و شبکه هوشمند

را بازبینی می نماییم. این بحث ها در نظر دارند تا بازنگری جامع و تصویر بزرگ برای خواننده های این حوزه مهیج فراهم گردد.

ما درمابقی این بخش یک سری نقاط تحقیق داغ را خلاصه می سازیم و مسیر های تحقیق احتمالی کلان داده را پیشنهاد می دهیم. همچنین در مورد روند های توسعه بالقوه در این حوزه کاربرد و تحقیق گسترده بحث می نماییم .

فصـــل پنجم

جمع‌بندی و پیشنهادها

مقدمه

پژوهش حاضر با هدف **تجزیه و تحلیل کاربرد Big Data** انجام شد. گزارش حاضردر پنج فصل تدوین گشت.
فصل اول به بیان کلیاتی در خصوص ضرورت انجام تحقیق، اهداف و سوالات تحقیق و توصیف اجمالی واژگان تخصصی اختصاص یافت. در فصل دو مبانی نظری و پیشینه تحقیق و مطالعات مرتبط شرح داده شد.روش اجرای تحقیق و توضیحاتی در باب روش و ابزار گردآوری اطلاعات در فصل سوم ارائه شد .نتایج حاصل از تجزیه وتحلیل اطلاعات نیز در قالب جداول و نمودارهایی در فصل چهار ارائه شد .فصل حاضر با عنوان فصل پنجم، به بیان خلاصه ای از نتایج تحقیق و نیز بحث در خصوص نتایج حاصله پرداخته و پس از یک نتیجه گیری کلی به بیان محدودیت هایی می پردازد که محقق طی اجرای این پژوهش با آنها مواجه بوده است .
نهایتاً پیشنهاداتی در قالب پیشنهادات علمی وکاربردی ارائه میگردد.

۵-۱ نتایج حاصل از تحقیق

در مقاله حاضر به بازنگری پس زمینه و جدید ترین فناوری کلان داده می پردازیم. اولا، ما دورنمای کلی کلان داده را معرفی کردیم و فناوری های مرتبط نظیر رایانش ابری، اینترنت اشیاء، مراکز داده و هادوپ را بازنگری می نماییم. سپس ما بر چهار فاز زنجیره ارزش کلان داده یعنی تولید داده، اکتساب داده، ذخیره سازی داده و تحلیل داده تمرکز می کنیم . ما برای هر فاز یک دورنمای کلی را معرفی می کنیم، در مورد چالش های فنی بحث می نماییم و جدید ترین پیشرفت ها را بازنگری می کنیم. در نهایت چندین کاربرد نمونه کلان داده از جمله مدیریت شرکت، اینترنت اشیاء، شبکه های اجتماعی، کاربرد های پزشکی، هوش جمعی و شبکه هوشمند را بازبینی می نماییم. این بحث ها در نظر دارند تا بازنگری جامع و تصویر بزرگ برای خواننده های این حوزه مهیج فراهم گردد.
ما یک سری نقاط تحقیق داغ را خلاصه می سازیم و مسیر های تحقیق احتمالی کلان داده را پیشنهاد می دهیم.
همچنین در مورد روند های توسعه بالقوه در این حوزه کاربرد و تحقیق گسترده بحث می نماییم.

۵-۲ پیشنهادها

نتایج حاصل از هر پژوهشی به امید ادامه یافتن راه تحقیق و پژوهش در خصوص آن موضوع و بهره برداری از نتایج آن به جامعه پژوهشگران و مسئولین ذی صلاح آن موضوع ارائه میگردد. از اینرو ارائه هر نوع پیشنهادی در

این گزارشات می تواند راه را برای مطالعات بعدی و نیز تصمیمگیریهای اجرایی در آن خصوص هموار سازد. در این پژوهش نیز پیشنهاداتی در قالب پیشنهادات پژوهشی جهت کارهای مطالعاتی بعدی و نیز پیشنهادات کاربردی برای تصمیمگیریهای اجرایی و توجه مسئولین ذیربط به آن حوزه ارائه شده است.

تجزیه و تحلیل کلان داده با چندین چالش مواجه است اما تحقیق اخیر هنوز در مرحله اولیه می باشد . تلاش های تحقیق قابل توجهی برای بهبود کارامدی نمایش ، ذخیره سازی و تحلیل کلان داده مورد نیاز هستند .

۵-۲-۱ پیشنهادات کاربردی

بر اساس یافته های پژوهش حاضر، در این قسمت سعی شده است تا به بیان پیشنهاداتی کاربردی پرداخته شود. امید است دستگاه های مسئول و سازمانهای ذیربط به این پیشنهادات به دیده دقت بنگرند.

- ✓استفاده از مدل های مناسب مدیریت پایگاه داده در فرآیندهای بازاریابی به منظور بهبود فرآیند کسب و کار
- ✓آشنا نمودن مدیران و کارشناسان نسبت به مزایای ناشی از کاربرد BigData در فرآیندهای بازاریابی از طریق برگزاری سمینار و همایش ها
- ✓بالا بردن توانایی مدیران در بکارگیری سیستم BigData به منظور بهبود فرآیندهای بازاریابی
- ✓استفاده از تکنیک های پیشرفته ارزیابی و تحلیل داده ها جهت بهبود فرآیندهای بازاریابی
- ✓تخصیص بودجه از سوی دولت و ارگان های مربوطه به منظور استفاده BigData در شرکت های کسب و کار به منظور بهبود فرآیند بازاریابی

۵-۲-۲ پیشنهادات آتی

- ✓بررسی مزایا و معایب ناشی از بکارگیری Big Data در طی پژوهش های آتی
- ✓بررسی موانع ناشی از بکارگیری Big Data در طی پژوهش های آتی

۵-۳ موضوعات باز

تجزیه و تحلیل کلان داده با چندین چالش مواجه است اما تحقیق اخیر هنوز در مرحله اولیه می باشد. تلاش های تحقیق قابل توجهی برای بهبود کارامدی نمایش، ذخیره سازی و تحلیل کلان داده مورد نیاز هستند .

۵-۳ تحقیق تئوریک

هر چند کلان داده یک حوزه تحقیق داغ با پتانسیل زیاد در هر دو بخش دانشگاهی و صنعتی نیست ، یک سری مشکلات مهم وجود دارند که بایستی حل شوند و این مشکلات در ادامه مورد بحث قرار می گیرند.

۵-۳-۱ مشکلات اساسی کلان داده: تعریف کلی و دقیق برای کلان داده، مدل ساختاری کلان داده، توصیف رسمی کلان داده و سیستم تئوریک علم داده الزامی و اجباری می باشد. اکنون چندین بحث در مورد کلان داده بیشتر همانند تصریح تجاری نسبت به تحقیق علمی می باشد. این موضوع بدان دلیل است که کلان داده به طور رسمی و ساختاری تعریف نمی شود و مدل های موجود به طور دقیق تایید نمی شوند.

۵-۳-۲ استاندارد سازی کلان داده: سیستم ارزیابی کیفیت داده و استاندارد ارزیابی کارامدی رایانش ابری بایستی توسعه یابد. تعدادی از راه حل های کاربرد های کلان داده مدعی هستند می توانند پردازش داده و تحلیل ظرفیت ها را در همه جنبه ها بهبود بخشند اما هنوز در این حوزه استاندارد و معیار ارزیابی یک شکل برای متوازن سازی راندمات رایانش کلان داده با روش های ریاضی سخت وجود ندارد. عملکرد را تنها می توان با رد زمانی ارزیابی نمود که سیستم اجراء گردیده و مستقر می شود که نمی توان به طور افقی مزیت ها و معایب انواع راه حل های جایگزین را حتی قبل و بعد از اجرای کلان داده مقایسه نمود. بعلاوه، چون کیفیت داده یک مبنای مهم برای پیش پردازش، ساده سازی و غربالگری داده است، همچنین یک مشکل فوری برای ارزیابی موثر و دقیق کیفیت داده نیز می باشد. **تکامل شیوه های رایانش کلان داده:** این کار شامل شیوه حافظه ، شیوه جریان داده، شیوه PRAM و شیوه MR و غیره می باشد. ظهور کلان داده باعث جرقه ای در پیشرفت های طراحی الگوریتم می گردد که از رویکرد متمرکز بر رایانش به رویکرد متمزکر بر داده تبدیل شده است. انتقال داده یک گلوگاه مهم رایانش کلان داده بوده است. از اینرو، چندین مدل رایانش جدید که برای کلان داده متناسب شده اند، پدیدار شده اند و مدل های بیشتری در افق قرار دارند.

۵-٤ پیچیدگی های عملی

۵-٤-۱ مدیریت کلان داده

ظهور کلان داده سبب چالش هایی جدید برای مدیریت داده سنتی می شود. اکنون، بعضی تلاش های تحقیقاتی در زمینه پایگاه داده مبتنی بر کلان داده و فناوری های اینترنت، مدل های ذخیره سازی و پایگاه های داده مناسب برای سخت افزار جدید، یکپارچه سازی داده ناهمگن و چند ساخت یافته، مدیریت داده رایانش پیش گستر و موبایل، مدیریت داده SNS و مدیریت داده پراکنده صورت گرفته اند.

۵-٤-۲ جستجو ، کاوش و تحلیل کلان داده

پردازش داده همیشه یک نقطه داغ تحقیقانی در کلان داده می باشد. مشکلات مرتبط عبارتند از جستجو و کاوش مدل های SNS، الگوریتم های جستجو کلان داده، جستحو پراکنده، جستجو P۲P، تجزیه و تحلیل بصری سازی کلان داده، سیستم عای پیشنهاد انبوه، سیستم های رسانه اجتماعی، کاوش آنی کلان داده، کاوش تصویر، کاوش معنایی، کاوش متن، کاوش داده چند ساخت یافته و فراگیری ماشین و مواردیگر.

۳-٤-۵ یکپارچه سازی و منشاء کلان داده

همانطورکه بحث شده است، مقدار کسب شده از بصری سازی جامع مجموعه های داده چند تایی بسیار بالاتر از مقدار کل مجموعه داده انفرادی می باشد. از اینرو، یکپارچه سازی منابع داده مختلف یک مشکل بموقع می باشد. یکپارچه ساز یداده با چالش های متعدد نظیر الگو های داده متفاوت و مقدار زیاد داده افزونه مواجه می باشد. منشاء داده عبارتست از فرایند تولید و تکامل داده با گذشت زمان و عمدتا برای بررسی مجموعه های داده چند تایی بجای مجموعه داده تکی استفاده شده است. بنابراین، برررسی این موضوع ارزشمند است که چگونه اطلاعات منشاء داده را که استاندارد های مختلف و از مجموعه های داده متفاوت را به طور برجسته نشان می دهند، یکپارچه سازیم.

٤-٤-۵ کاربرد کلان داده

اکنون کاربرد کلان داده دقیقا شروع می شود و ما بایستی روش های کارامد تر را برای استفاده کامل از کلان داده کشف نماییم. از اینرو ، کاربرد های کلان داده در علم، مهندسی، پزشکی، مراقبت بهداشتی، فایننس، کسب و کار، اجرای قانون، تحصیل، حمل و نقل، خرده فروشی و ارتباط راه دور، کاربرد های کلان داده در تجارت های اندازه کوچک و متوسط، کاربرد های کلان داده در دپارتمان های دولت، خدمات کلان داده و فعل وا نفعال انسان – کامپیوتر کلان داده و موارد دیگر همگی از مشکلات مهم تحقیقاتی محسوب می شوند.

۵-٤-۵ امنیت داده

همیشه امنیت و محرمانگی در فناوری اطلاعات دو نگرانی کلیدی هستند . وقتی حجم داده در عصر کلان داده با سرعت زیاد رشد می کند ، یک سری ریسک های ایمنی شدید تر وجود دارند در حالی که روش های محافظت داده سنتی از قبل برای کلان داده کاربرد پذیر نشان داده نشده اند. بویژه، ایمنی کلان داده با چالش های مرتبط با امنیت زیر مواجه است:

٦-٤-۵ محرمانگی کلان داده

محرمانگی کلان داده شامل دو جنبه می باشد:

(۱) محافظت از اطلاعات محرمانه شخصی در طول اکتساب دادن: علایق شخصی، سرگرمی ها و ویژگی های بدنی و موارد دیگر کاربران ممکن به راحتی در دسترس قرار گیرد و کاربران ممکن است اگاه شوند.

(۲) همچنین داده های محرمانه شخصی ممکن است در طول ذخیره سازی، انتقال و استفاده فاش گردند حتی اگر با مجوز کاربران برداشته شوند. برای مثال، فسیبوک همانند یک شرکت کلان داده با داده SNS بسیار دقیق فرض می گردد. رون بوئیس محقق Skull Security بر اساس تحقیق به داده هایی در صفحات عمومی کاربران

فیسبوک دست یافته بود که در اصلاح تنظیم اطلاعات شخصی اشان از طریق ابزار کسب اطلاعات موفق نبودند. رون بوئیس این قبیل داده ها را دروت بسته ۲٫۸ گیگابایتی پک کرده بود و بذر BitTorrent را برای دیگران ایجاد کرده بود تا دانلود نمایند. ظرفیت تحلیل کلان داده ممکن است به کاوش اسرار شخصی از اطلاعات ظاهرا ساده منجر گردد. از اینرو، محافظت از اطلاعات محرمانه شخصی به یک مشکل چالش برانگیز و جدید تبدیل خواهد شد.

۵-٤-۷ کیفیت داده

کیفیت داده بر استفاده از کلان داده تاثیر می گذارد. داده کیفیت پایین باعث اتلاف در انتقال و منابع ذخیره سازی با کاربرد پذیری ضعیف می گردد. از اینرو، مقدار زیادی فاکتور ها وجود دارند که ممکن است کیفیت داده را محدود سازند و برای مثال تولید، اکتساب و انتقال ممکن است همگی بر کیفیت داده تاثیر بگذارند. کیفیت داده عمدتا در دقت، تمامیت، افزونگی و پایداری اش آشکار می گردد. ولو این که مقدار زیادی از سنجش ها صورت گرفته اند تا کیفیت داده را بهبود بخشند ولی مشکلات مربوطه به خوبی مورد رسیدگی قرار نگرفته اند. از اینرو، روش های موثر برای کشف خودکار کیفیت داده و تعمیر بعضی داده های آسیب دیده بایستی مورد بررسی قرار گیرند.

۵-٤-۸ مکانیزم ایمنی کلان داده

کلان داده به دلیل مقیاس بزرگ و تنوع بالایش سبب چالش هایی برای رمز گذاری داده می شود. عملکرد روش های رمزگذاری قبلی بر داده های مقیاس کوچک و متوسط نمی توانست تقاضاهای کلان داده را براورده سازد از اینرو رویکرد های رمز نویسی کارامد کلان داده بایستی توسعه یابند. الگو های موثر مدیریت ایمنی، کنترل دسترسی و ارتباطات ایمنی بایستی برای داده های ساخت یافته، نیمه ساخت یافته و ساخت نیافته مورد بررسی قرار گیرند. بعلاوه، تفکیک، محرمانگی، تمامیت، دسترس پذیری، کنترل پذیری و قابل ردیابی بودن داده های مستاجر تحت شیوه چند مستاجری بایستی با تضمین کارامدی فرضی قبلی میسر گردند.

۵-٤-۹ کاربرد کلان داده در امنیت اطلاعات

کلان داده نه تنها سبب چالش هایی برای امنیت اطلاعات می گردد بلکه همچنین فرصت های جدید را برای توسعه مکانیزم های امنیت اطلاعات ارایه می نماید. برای مثال، ما می توانیم روزنه های ایمنی بالقوه و تهدید پایدار پیشرفته (APT) را بعد از تحلیل کلان داده در شکل پرونده های ثبت سیستم کشف نفوذ کشف نماییم. بعلاوه، مشخصه های ویروس، مشخصه های راه نفوذ و مشخصه های حمبه ممکن است به راحتی از طریق تحلیل کلان داده شناسایی شوند.

ایمنی کلان داده باعث جلب توجه زیاد محققان شده است. از اینرو، تنها تحقیق محدود در مورد نمایش کلان داده ناهمگن چند منبعی، سنجش و روش های جامع معنایی، تئوری های مدل سازی و مدل های رایانش، ذخیره سازی توزیعی بهینه سازی بهره وری انرژی و معماری های سیستم نرم افزاری و سخت افزاری پردازش شده وجود دارد. امنیت کلان داده از جمله اعتبار، بک اپ و ریکاوری، تعمیر و نگهداری کامل و امنیت به طور ویژه بایستی بیشتر بررسی گردند.

۵-۵ چشم انداز

ظهور کلان داده یک سری فرصت های بزرگ را باز می کند. فناوری در عصر فناوری اطلاعات یک نگرانی مهم بوده است در حالی که فناوری می تواند محرک توسعه داده باشد. داده در عصر کلان داده با برتری ارزش داده و پیشرفت ها در زمینه اطلاعات باعث هدایت پیشرفت فناوری ها در آینده نزدیک می شود. کلان داده نه تنها دارای تاثیر اجتماعی و اقتصادی خواهد بود بلکه همچنین بر روش های زندگی و تفکر هر فرد تاثیر می گذارد که دقیقا اتفاق نمی افتد. ما نمی توانستیم آینده را پیش بینی نماییین اما می توانیم اقدامات احتیاطی را برای رویداد های احتمالی اتخاذ نماییم که در آینده رخ می دهند.

داده با مقیاس بزرگ تر، تنوع بیشتر و ساختار های پیچیده تر: هر چند فناوری های معرفی شده از طریق هادوپ به موفقیت زیادی دست یافته اند، چنین فناوری هایی بر حسب انتظار عقب می افتند و با در نظر گرفتن توسعه سریع کلان داده تعویض خواهند شد. مبنای تئوریک هادوپ در ابتدای سال ۲۰۰۶ ظاهر شده است. بعضی محققان در مورد روش های بهتر غلبه بر داده های ساخت یافته پیچیده تر، مقیاس بزرگ تر و متنوع تر نگران بوده اند. این تلاش ها توسط پایگاه داده توزیعی – جهانی گوگل و پایگاه داده سنتی توزیعی، توسعه پذیر و تحمل کننده عیب F۱ نشان داده می شوند . فناوری ذخیره سازی کلان داده در آینده از پایگاه های داده توزیعی استفاده خواهد کرد ، از مکانیزم های تراکنش مشابه با پایگاه های داده رابطه ای پشتیبانی می کند و به طور موثر به داده از طریق گرامر های مشابه با SQL رسیدگی می کند.

۵-۵-۱ عملکرد منبع داده

نظر به این که کلان داده حاوی مقادیر حجیم می باشد، کلان داده برتر به معنی منابع برتر می باشد. از طریق تحلیل زنجیره ارزش کلان داده می توان مشاهده نمود که ارزش ان از خود داده، فناوری ها و ایده ها ناشی می گردد و منابع داده را می توان هسته ان محسوب نمود. سازماندهی مجدد و یکپارچه سازی مجموعه های داده متفاوت می توانند ارزش های بیشتر را خلق نمایند. شرکت هایی که بر منابع کلان داده تسلط دارند ممکن است به سود های عظیم از طریق اجاره و تخصیص حقوق استفاده از داده هایشان دست یابند.

کلان داده باعث بهبود هم جوشی متقابل علم می گردد: کلان داده نه تنها هم جوشی جامع رایانش ابر، اینترنت اشیاء، مرکز داده و شبکه های موبایل را بهبود می بخشد بلکه همچنین هم جوشی متقابل قواعد مختلف را اجباری می سازد. توسعه کلان داده بایستی فناوری های نواورانه و روش ها را بر حسب اکتساب داده، ذخیره سازی، پردازش، تحلیل و امنیت اطلاعات کشف نماید. سپس تاثیرات کلان داده بر مدیریت تولید، عملیات کسب و کار و تصمیم گیری بایستی برای شرکت های مدرن از چشم انداز مدیریت بررسی گردند. علاوه بر این ، کاربرد کلان داده با فیلد های خاص به مشارکت استعداد های بین رشته ای نیاز دارد.

بصری سازی: اصل انچه شما می بینید در واقع ان چیزی است که کسب می کنید در بعضی سناریو های فعل وا نفعال انسان – کامپیوتر مطرح می شود. داده ترکیبی در کاربرد های کلان داده برای تصمیم گیری بسیار سودمند می باشد. تنها زمانی که نتایج تحلیلی با مساعدت نمایش داده شداند، این نتایج به طور موثر توسط کاربران استفاده شده اند. گزارش ها، هیستوگرام ها، نمودار های کلوچه ای و منحنی های رگرسیون و موارد دیگر به کرات استفاده می شوند تا نتایج تحلیل داده را بصری سازند. شکل های نمایش جدید نظیر Microsoft renlifang، موتور جستحو اجتماعی، استفاده از نمودار های رابطه ای برای بیان رابطه بین فردی در آینده رخ خواهند داد.

۲-۵-۵ داده محور

معروف است کمه برنامه ها از ساختار های داده و الگوریتم ها تشکیل می شوند و ساختار های داده برای ذخیره سازی داده استفاده می شوند. در تاریخ طراحی برنامه مشاهده شده است که نقش داده به طور فزاینده ای مهم تر می شود. در عصر داده های مقیاس کوک که منطق در این عرصه پیچیده تر از داده می باشد، طراحی برنامه عمدتا فرایند محور می باشد. چون دادع کسب و کار پیچیده تر می شود ناچارا روش های طراحی هدف (شی) محور توسعه می یابند. این روز ها، پیچیدگی داده کسب و کار بسیار برتر از منطق کسب و کار بوده است. متعاقبا برنامه ها به تدریج از حالت متمرکز بر الگوریتم به حالت متمرکز بر داده تغییر شکل می دهند. پیش بینی می گردد که روش های طراحی برنامه داده محور با طور معین ظاهر می شوند و این که تاثیر گسترده بر توسعه فناوری اطلاعات در مهندسی نرم افزار، معماری و طراحی مدل در میان موارد دیگر دارند.

کلان داده باعث جرقه ای در تحول فکری می شود: کلان داده و تحلیل ان به تدریج بر روش های تفکر امان تاثیر برجسته ای دارند. تحول تفکر که از طریق کلان داده جرقه زده شده است خلاصه می گردد و از قرار زیر می باشد:

- ما در طول تحلیل داده تلاش خواهیم کرد تا از کل داده ها بجای تحلیل مجموعه کوچکی از داده نمونه استفاده کنیم .
- ما در مقایسه با داده دقیق مایل خواهیم بود تا داده های پیچیده و بیشمار را بپذیریم.

- ما بایستی توجه بیشتری را بر همبستگی های بین چیز ها بجای کشف رابطه سببی معطوف نماییم.
- الگوریتم های ساده کلان داده موثر تر از الگوریتم های پیچیده تر داده کم هستند.
- کاهش فاکتور های ذهنی و عجولانه در طول تصمیم گیری از نتایج تحلیلیل کلان داده محسوب می شوند و دانشمندان داده با کارشناسان عوض خواهند شد.

تقاضاها و اراده موجودات بشر در کل تاریخ جامعه انسانی همیشه قدرت های منبع برای بهبود پیشرفت فناورانه و علمی هستند. کلان داده ممکن است پاسخ های رفرنس را برای موجودات بشر فراهم نماید تا از طریق کاوش و پردازش تحلیلی تصمیم گیری نماید اما نمی توانست جایگزین تفکر انسانی گردد. در واقع این تفکر انسان است که استفاده های گسترده از کلان داده را بهبود می بخشد. کلان داده بیشتر شبیه مغز توسعه پذیر و قابل پیشرفت انسان و نه جایگزین مغز انسانی می باشد. مردم با ظهور ایترنت اشیاء، توسعه فناوری حسگری موبایل و پیشرفت فناوری اکتساب داده نه تنها کاربران و مصرف کننده های کلان داده بلکه همچنین تولید کننده ها و شرکت کننده های در ان می باشند. حسگری رابطه اجتماعی، منبع یابی جمعیت، تحلیل کلان داده در SNS و دیگر کاربرد ها که با فعالیت های انسانی مبتنی بر کلان داده رابطه بسیار نزدیکی داسته اند، به طور فزاینده ای مورد نگرانی خواهند بود و قطعا سبب تبدیل های بیشمار فعالیت های اجتماعی در جامعه آینده خواهند شد.

مراجع

پایان نامه فارسی:

جلالی، شبنم، « تجزیه و تحلیل کاربرد Big Data در فرآیند های بازاریابی»، استاد راهنما :جناب آقای دکتراصلانی، ۱۳۸٦،دانشگاه تهران (پردیس البرز)، ویرایش دی ماه سال

مقاله اینترنتی / آنلاین: **Research paper**

-Big Data: A Survey,
Min Chen · Shiwen Mao · Yunhao Liu
Mobile Netw Appl (۲۰۱٤) ۱۹:۱۷۱–۲۰۹
DOI ۱۰,۱۰۰۷/s۱۱۰۳٦-۰۱۳-۰٤۸۹-۰

-E-commerce logistics distribution mode in big-data context: A case analysis of JD.COM,
Kangning Zhenga,b, Zuopeng Zhang (Justin)c,∗, Bin Songd,
Received ۱٦ July ۲۰۱۸; Received in revised form ۱ October ۲۰۱۹; Accepted ۱٦ October ۲۰۱۹
Elsevier

-Big data analytics as an operational excellence approach to enhance
sustainable supply chain performance
Surajit Baga, Lincoln C. Woodb,c,*, Lei Xud,e, Pavitra Dhamijaf, Yaşanur Kayikcig

Abstract

Abstract In this paper, we review the background andstate-of-the-art of big data. We first introduce the generalbackground of big data and review related technologies, such as could computing, Internet of Things, data centers,and Hadoop. We then focus on the four phases of the valuechain of big data, i.e., data generation, data acquisition, data storage, and data analysis. For each phase, we introduce thegeneral background ; discuss the technical challenges, andreview the latest advances. We finally examine the several representative applications of big data, including enterprisemanagement, Internet of Things, online social networks,medial applications, collective intelligence, and smart grid.
These discussions aim to provide a comprehensive overview and big-picture to readers of this exciting area. This survey is concluded with a discussion of open problems and future directions.

Keywords Big data · Cloud computing · Internet of things · Data center · Hadoop · Smart grid · Big data analysis

Payam Noor University
Faculty of Engineering

Seminar Report
Department of Computer Engineering and Information Technology

Big Data

Vida Sepasi

Supervisor:
Dr. Razavi

February,۲۰۲۲