کاربردهای کلان داده‌ها

چکیده

در طول چند دهه‌ی گذشته، سازمان‌های بزرگ تجاری در زمینه‌های مختلف اقدام به جمع‌آوری داده‌ها از بخش‌های متفاوت و در قالب‌های متعدد کرده‌اند و تلاش نموده‌اند تا مجموعه‌دادگان را به هم ارتباط داده و بر اساس آنها تصمیمات با ارزش تجاری اتخاذ نمایند. مانع کلیدی در اجرای این امر، ناتوانی سیستم‌های موجود برای پردازش داده‌های بزرگی است که بخشی از این داده‌ها دارای ساختار و بخشی دیگر بدون ساختار هستند. همانطور که در فصل‌های قبلی مشاهده شد، گام‌های بلند فناوری در طول چند سال گذشته توانسته است تا ناتوانی پردازش مجموعه‌دادگان بزرگ را رفع کرده و توانایی کاوش و تحلیل داده‌های بزرگ را فراهم نماید. شرکت‌هایی که در حوزه‌ی انبار داده‌ها هستند، این روند را به عنوان فرصت بزرگی در جهت کمک به کاربران خود دیده‌اند تا کاربران بتوانند به کاوش پیشینه‌ی داده‌های خود بپردازند و بر اساس دیدی که از کاوش داده‌های جمع شده‌ی خود در طول دهه‌ها به دست می‌آورند، به کسب و کار خود ارزش‌های تاکتیکی و استراتژیک بیفزایند. در این فصل، ما نمونه‌های کلی را خواهیم دید که چگونه کسب و کارهای مختلف به تحلیل داده‌های خود می‌پردازند و با استفاده از آنها اهداف تجاری خود را ارتقا می‌دهند. ما چندین مثال در زمینه‌های خدمات مالی، خرده‌فروشی، ساخت و تولید، ارتباطات، رسانه‌های اجتماعی، و مراقبت از سلامتی ارائه خواهیم داد.

کلمات کلیدی: تحلیل سبد؛ تشخیص کلاهبرداری؛ ریزش مشتری؛ تحلیل مسیر؛ تحلیل پیش‌بینی؛ تحلیل احساسات؛ تحلیل شبکه‌های اجتماعی؛ ایجاد نشست؛ تحلیل نموداری؛ تجسم و بصری‌سازی داده‌ها؛ خوشه‌بندی K-means

۱. مقدمه

تمام شرکت‌های بزرگ با محیط‌های به شدت رقابتی و با فشار ثابت روبرو هستند تا سودآوری را با استفاده از شناسایی راهکارهای عملیاتی افزایش دهند و در عین حال خطر کسب و کار را نیز به حداقل برسانند. تمام کسب و کارهای بزرگ به اهمیت تحلیل پیشینه و سوابق داده‌ها پی برده‌اند، پیشینه‌ی داده‌ها در واقع داده‌هایی

هستند که این شرکت‌ها در طی سالیان دراز آنها را جمع‌آوری کرده‌اند، و تحلیل این داده‌ها به بخشی جدایی‌ناپذیر برای گرفتن تصمیمات استراتژیک در این شرکت‌ها تبدیل شده است. بنابراین انگیزه‌ی زیادی وجود دارد که سیستم‌های یکپارچه‌ای برای مدیریت داده‌ها راه‌اندازی شده و از هوش تجاری و روش‌های تحلیل برای بهبود کسب و کار آنها استفاده شود.

در سال‌های گذشته، تحلیل‌های کلان داده‌ها توجه گسترده‌ای را در کاربردهای متعدد و در حوزه‌های مختلف، هم در صنعت و هم در دانشگاه به خود جلب نموده است. اگر چه این حوزه در دهه‌ی گذشته پیشرفت چشمگیری داشته است، با این حال همچنان مشکلات چالش‌برانگیزی وجود دارد و باید برای مسائل جدید و پیچیده در بازار رو به رشد این حوزه راه‌حل‌هایی یافته شوند. روش‌های مختلفی در مدلسازی، تحلیل آماری، داده‌کاوی، و یادگیری ماشین برای پیش‌بینی رویدادهای بعدی در آینده و پیش‌بینی رفتارهای مشتری مورد استفاده قرار می‌گیرند تا پس از آن بر اساس این موارد، اقدامات فعالانه‌ای برای حفاظت و ارتقای اهداف کسب و کار انجام شوند.

در بخش‌های زیر، ما بررسی سطح بالایی بر چالش‌های مطرح شده توسط صنایع مختلف ارائه خواهیم داد و همچنین در این مورد بحث می‌کنیم که کلان داده‌ها چگونه برای حل این چالش‌ها در بخش‌های مربوط به کسب و کار آنها مورد استفاده قرار می‌گیرد. اگر چه تحلیل‌های کلان داده‌ها پتانسیل استفاده شدن در صنایع حوزه‌های مختلفی را دارند، ولی ما بررسی خود را به تعداد کمی از این حوزه‌ها، یعنی حوزه‌ی بانکی و مالی (بخش ۲)، خرده‌فروشی (بخش ۳)، ساخت و تولید (بخش ۴)، مخابرات (بخش ۵)، رسانه‌های اجتماعی (بخش ۶)، و مراقبت از سلامتی (سلامتی ۷) محدود می‌کنیم.

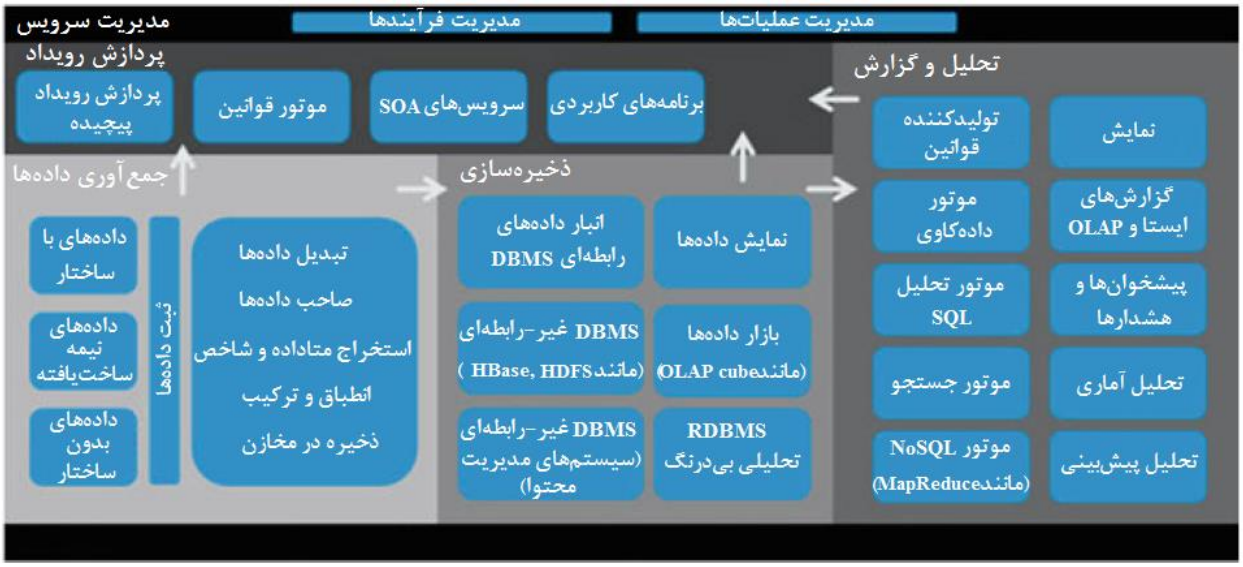
۲. معماری مرجع کلان داده‌ها

شکل ۱ چارچوب معماری سطح بالایی [۱] از یک سیستم معمولی کلان داده‌ها را نشان می‌دهد که شامل اجزای زیر است:

۱. جمع‌آوری داده‌ها از منابع مختلف،
۲. زیرساختی برای انجام تبدیل‌های گوناگون بر روی داده‌ها،
۳. ذخیره‌سازی داده‌ها در مخازن مختلف،

۴. اجرای موتورهای تحلیلی با عملکرد بالا،

۵. مجموعه ابزار گزارش‌دهی و نمایش نتایج و فرآیندها.



شکل ۱. معماری زیرساخت کلان داده‌ها

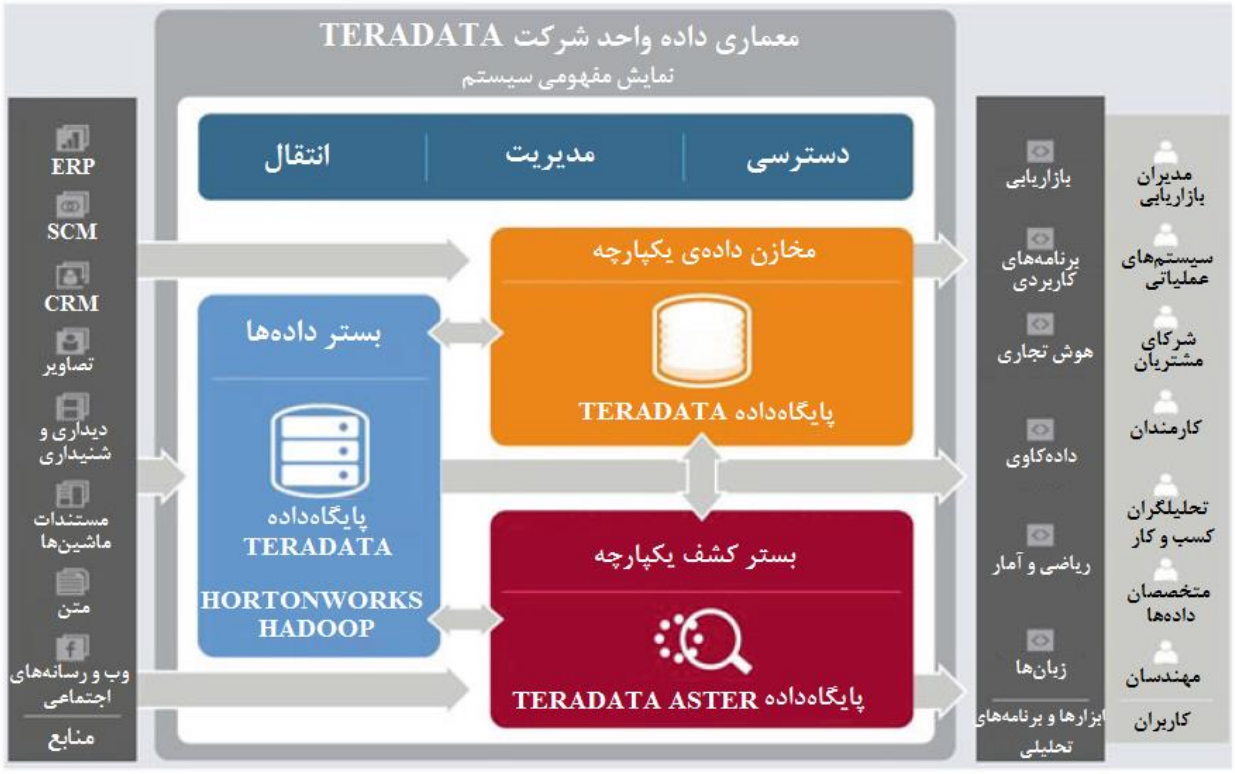
منابع داده‌ها می‌توانند برگرفته از سیستم‌های عملیاتی باشند که ساختار خوبی دارند (مانند طرح‌ها/ جداول/ ستون‌ها/ غیره) یا می‌توانند بدون ساختار باشند مانند داده‌های رسانه‌های اجتماعی، داده‌های جریان کلیک، رویدادهای ثبت شده، و داده‌های چندرسانه‌ای. اکثر داده‌های با ساختار[1] (ساخت‌یافته) در محیط‌های معمولی برای ذخیره‌سازی داده‌ها ذخیره می‌شوند و داده‌های نیمه‌ساخت‌یافته[2] و بدون ساختار[3] (غیر ساخت‌یافته) نیز در خوشه‌های Hadoop ذخیره می‌شوند. داده‌ها در سیستم‌های جمع‌آوری‌کننده‌ی داده‌ها از قبیل بازار داده‌ها و انواع مختلف موتورهای تحلیلی توزیع می‌شوند، کاربران در این اماکن می‌توانند با استفاده از ابزارهای تحلیلی و گزارش‌دهی بر اساس SQL به پرس و جو (کوئری) بر روی این داده‌ها بپردازند و اطلاعات موردنیاز خود را بیابند. بسته به کاربرد مورد نظر، روش‌های تحلیلی مختلفی از قبیل تحلیل همبستگی، تحلیل روند و الگو، فیلترسازی مشارکتی، تحلیل سری‌های زمانی، تحلیل گراف، تحلیل مسیر، و تحلیل متن بر روی داده‌ها اجرا می‌شوند و این روش‌های تحلیلی پیش از نمایش داده‌ها انجام می‌گیرند، نمایش داده‌ها بر روی

[1] structured data
[2] semi-structured data
[3] non-structured data

پیشخوان با استفاده از روش‌های متعدد بصری‌سازی و نمایش صورت می‌گیرد. بحث دقیقی بر روی این اجزاء در فصول قبلی ارائه شده است.

شرکت‌های فروشنده‌ی زیادی هستند که راه‌حل‌هایی را بر اساس معماری مرجع فوق ارائه می‌دهند، که Teradata و IBM دو نمونه از این شرکت‌ها می‌باشند. شکل ۲ بستر تحلیل کلان‌داده را نشان می‌دهد که متعلق به شرکت Teradata است و بستر معماری داده واحد نامیده می‌شود [۲] و قابلیت‌های زیر را دارد:

۱. ضبط و تبدیل داده‌ها از انواع منابع مختلفی که به صورت با ساختار، نیمه‌ساخت‌یافته، یا بدون ساختار هستند.
۲. توانایی پردازش حجم عظیمی از داده‌ها با استفاده از Hadoop به همراه کشف داده‌ها و ادغام ذخایر داده‌ها.
۳. پشتیبانی از توابع تحلیلی از پیش آماده در دسته‌هایی به صورت تحلیل مسیر، تحلیل خوشه، تحلیل آماری، تحلیل پیش‌بینی، تحلیل متن، تحلیل رابطه‌ای، و تحلیل گراف.
۴. مقیاس‌پذیری و عملکرد بالا.



شکل ۲. معماری داده‌ی واحد شرکت Teradata

جزئیات بیشتر بر روی راه‌حل‌های کلان‌داده‌ها در مرجع [۲] ارائه شده است.

اگر چه معماری مرجع نشان داده شده در شکل ۱ به ارائه‌ی مجموعه‌ی کاملی از قابلیت‌هایی می‌پردازد که در هر برنامه‌ی کاربردی کلان‌داده‌ها موردنیاز هستند، با این حال لازم به ذکر است که تمام زیرسیستم‌های نشان داده شده در این مرجع لازم نیست که در هر برنامه‌ی کاربردی حضور داشته باشند و داشتن تمام این اجزاء برای تمام کاربردها الزامی نیست. در ادامه در بخش‌های زیر به ارائه‌ی چارچوب‌ها و اجزای آن برای کاربردهای خاص صنعت می‌پردازیم.

۳. کاربردهای کلان‌داده‌ها در بانکداری و صنایع مالی

مقادیر عظیمی از داده‌ها توسط صنایع مالی و بانکداری در حال تولید هستند، این داده‌ها از طریق سرویس‌های مختلفی تولید می‌شوند، از قبیل حساب‌های پس‌انداز/حسابرسی، بانکداری همراه، کارت‌های اعتباری و بدهی، وام‌ها، بیمه، و سرویس‌های سرمایه‌گذاری که همگی این سرویس‌ها توسط این صنایع ارائه می‌شوند. اکثر این داده‌های تولید شده به صورت داده‌های با ساختار (ساخت‌یافته) هستند. همچنین، اکثر این سازمان‌ها دارای شعب آنلاین نیز می‌باشند تا سرویس‌دهی و بازاریابی بهتری را ارائه دهند که مقادیر زیادی از داده‌ها نیز از این طریق جمع‌آوری می‌شوند. همانطور که در شکل ۳ نشان داده شده است، برخی از کانال‌های جمع‌آوری داده‌ها عبارتند از:

- تعاملات مشتری از طریق پست‌های الکترونیکی و چت‌های ثبت شده؛
- شبکه‌های اجتماعی از طریق توییت‌ها و پُست‌ها/فیدهای Facebook؛ و
- داده‌های نیمه-ساخت‌یافته از طریق logهای ثبت شده از وب و عقاید مشتری‌ها.

شکل ۳. تحلیل‌های کلان‌داده‌ها در صنعت بانکداری [۳].

اکثر داده‌های جمع‌آوری شده مورد استفاده قرار نمی‌گیرند، و صنعت بـه دنبـال فنـاوری جدیـدی در زمینه‌ی داده‌کاوی و تحلیل تجاری است تا به درک و شناسایی نیازهای مشتری و پیشنهاد سرویس‌هـای جدیـد کمک کند، که این امر فرصت‌های کسب و کار آنها را ارتقا داده و سود خـالص و میـزان سودبخشـی را افـزایش خواهند داد. همچنین صنعت مالی به دنبال راه‌حل‌هایی در زمینه‌ی مدیریت خطر و تشخیص کلاهبـرداری نیـز می‌باشد تا افشای اطلاعات محرمانه‌ی تجاری را به حداقل برساند. یکی دیگـر از زمینـه‌هـای مـورد علاقـه بـرای صنعت در استفاده از تحلیل‌های کلان‌داده‌ها، یافتن استراتژی‌هایی برای حفظ مشتری‌ها می‌باشد.

در بخش‌های زیر، ما در مورد نحوه‌ی استفاده از تحلیل‌های کلان‌داده‌ها در برخی از این حوزه‌های مهم با جزئیات بیشتری بحث می‌کنیم.

۳-۱. تشخیص کلاهبرداری

بررسی‌ها و مطالعات متعدد [۴] نشان می‌دهند که صنعت سرویس‌های مالی و بانکداری در میان صنایع مختلف، قربانی بسیاری از موارد کلاهبرداری می‌باشد. برخی از کلاهبرداری‌هایی که به طور گسترده در صنعت بانکداری شناخته می‌شوند، عبارتند از:

۱. کلاهبرداری آنلاین بانکی: این نوع کلاهبرداری شامل کلاهبردارانی است که دسترسی به حساب قربانیان را در دست گرفته و تراکنش‌هایی را انجام می‌دهند تا وجوه بانکی را از حساب‌های آنها خارج نمایند.
۲. کلاهبرداری در کارت: این نوع کلاهبرداری شامل کلاهبردارانی است که اطلاعات کارت بانکی را ربوده و تراکتش‌های تقلبی را انجام می‌دهند.
۳. کلاهبرداری در داخل سیستم بانکی: این نوع شامل کلاهبرداری‌هایی است که توسط کارکنان بانک انجام می‌شوند.
۴. پولشویی: جُرمی است که شامل تراکنش‌هایی عمدتا با بانک‌های خارجی می‌باشد تا از این طریق ریشه‌های ثروت‌های غیرقانونی را پنهان کنند.

رویکرد رایج و معمولی برای غربال این موارد، گزارش به صورت دستی و استفاده از قوانین مختلف است که این رویکرد تنها برای روند پذیرش عملیات بانکی مفید بوده و برای تشخیص کلاهبرداری و متوقف کردن خسارت مفید نمی‌باشد. صنعت مالی نیاز دارد که تشخیص کلاهبرداری به صورت بی‌درنگ[1] انجام شود تا تراکنش‌های مربوط به کلاهبرداران به صورت بی‌درنگ شناسایی شده و اجرای آنها متوقف شود [۵].

استفاده از تحلیل‌ها برای تشخیص الگوهای رفتار کلاهبرداری، عنصر کلیدی در تشخیص کلاهبرداری است. این امر نیاز به درک روشنی از رفتار گذشته‌ی مشتری از لحاظ ماهیت تراکنش‌ها دارد تا بتوان تراکنش‌های سالم و تراکنش‌های مربوط به کلاهبرداری را به طور موثری تفکیک نمود، این کار با تحلیل تراکنش‌ها بر اساس پروفایل مشتری صورت می‌گیرد، تراکنش‌هایی که ممکن است شامل یک نمره منفی (امتیاز خطر) باشد. این فرآیند امتیاز دادن به تراکنش‌ها نیاز دارد که ماهیت غیرقابل پیش‌بینی بودن تراکنش‌ها بر

[1] real-time

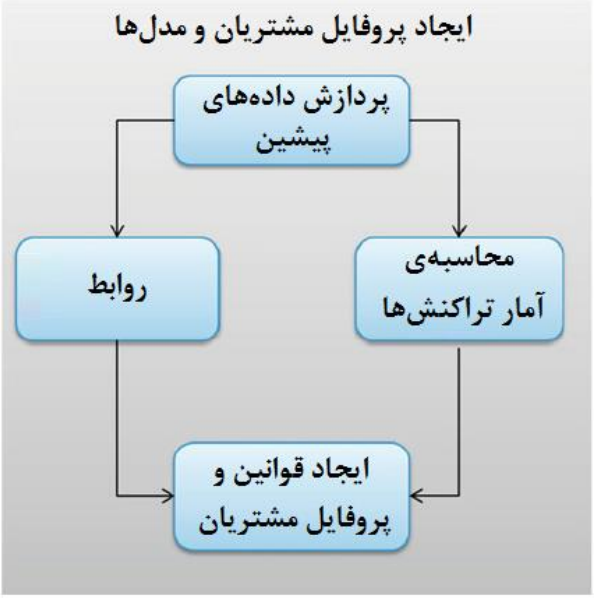
اساس رفتار مشتری‌های مختلف در نظر گرفته شود، چرا که مجموعه‌ی مشتری‌ها شامل مشتریان عادی و مُجرمان است.

از این رو تشخیص کلاهبرداری شامل یک فرآیند دو مرحله‌ای است که عبارتند از:

ایجاد پروفایل مشتریان بر اساس سوابق تراکنش‌ها و شناسایی الگوی تراکنش‌هایی که منجر به کلاهبرداری می‌شوند.

پروفایل مشتریان برای شناسایی هر گونه نمونه‌ی پَرت[1] یا تطابقی از توالی‌ها / رویدادها با الگوهای از پیش‌تعریف شده‌ی کلاهبرداری‌ها مورد استفاده قرار گرفته و از انجام تراکنش‌های احتمالا مرتبط با کلاهبرداری پیشگیری شود.

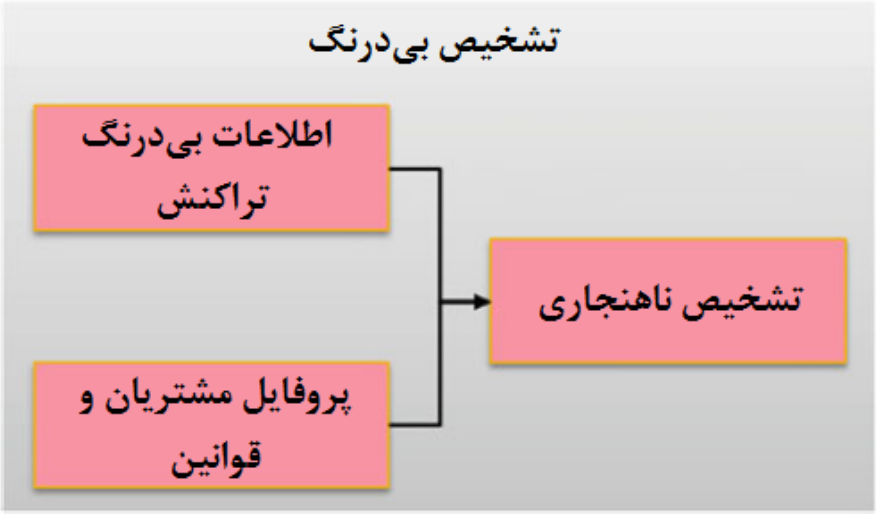
ایجاد پروفایل مشتریان به صورت استفاده از روش‌های آماری با استفاده از محاسبه‌ی میانگین آماری، مقادیر بیشینه و کمینه، انحراف معیار و غیره بر روی سوابق تراکنش‌ها است تا ترکیبی از تراکنش‌های معمولی به دست آید. شکل دیگری از ایجاد پروفایل مشتریان به صورت به دست آوردن روابطی است که تراکنش‌ها میان چه کسانی انجام شده‌اند. روش‌های گرافیکی [۶] برای ثبت روابط موجود در شبکه مورد استفاده قرار می‌گیرند و این کار را با استفاده از انطباق تراکنش‌ها بین مشتریانی انجام می‌دهند که از روش‌های پرداخت استفاده کرده‌اند. شکل ۴ جریان ایجاد این الگوها و پروفایل مشتریان را نشان می‌دهد.



شکل ۴. ایجاد پروفایل مشتری برای تشخیص کلاهبرداری

[1] outlier

شکل ۵ جریان تشخیص بی‌درنگ کلاهبرداری و جداسازی آن در حین اجرای یک تراکنش را نشان می‌دهد. اگر تراکنش انجام شده از نظر مقدار تراکنش، اتصال تراکنش، و غیره با پروفایل مشتری مطابقت ندارد، آنگاه برای بررسی‌هایی در سطوح بالاتر تشخیص داده می‌شود. تشخیص آماری نمونه‌ی پرت بر اساس سوابق آماری در پروفایل مشتری یکی از روش‌های تشخیص تراکنش مشکوک است.



شکل ۵. تشخیص بی‌درنگ کلاهبرداری با استفاده از پروفایل مشتری

تحلیل الگو [۷] در رویداد تراکنش‌ها و مقایسه‌ی آن با الگوهای از پیش‌تعریف شده‌ی فعالیت‌های کلاهبرداری، روش محبوبی است که در شناسایی بی‌درنگ هر گونه مشتری کلاهبردار مورد استفاده قرار می‌گیرد. روش‌های تحلیل سری‌های زمانی [۸] نیز برای شناسایی این مورد به کار گرفته می‌شود که آیا فعالیت مشتری با قوانین تجاری که عنوان کلاهبرداری تعریف شده‌اند، مطابقت دارد یا خیر.

۳-۲. *پولشویی*

پولشویی یک نوع کلاهبرداری پیچیده‌تری است، و تشخیص آن نیاز به راه‌اندازی مراحل پیچیده‌تر و ادغام سیستم‌های چند-بُعدی پایگاه‌داده‌ها را دارد که داده‌های هر یک از این پایگاه‌داده‌ها نیز از منابع مختلف مانند پایگاه‌داده‌های تراکنش‌های بانکی و اجرای قانون و غیره جمع‌آوری شده‌اند.

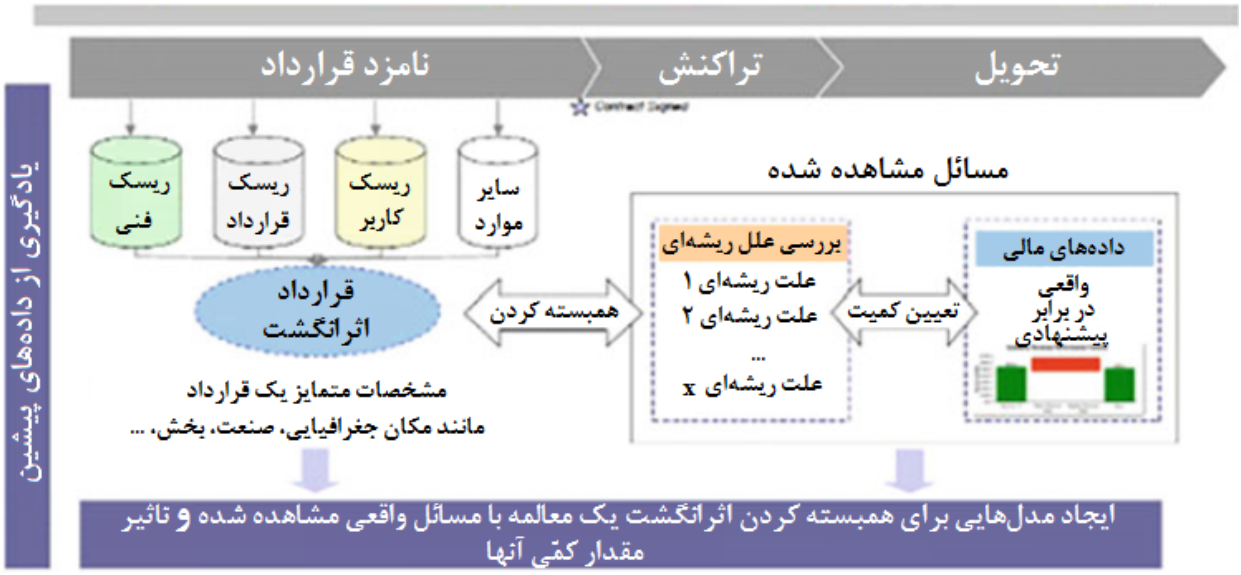
شبکه‌های پیچیده‌ای از روابط میان اجزای مختلف با استفاده از پیوند دادن داده‌های تولید شده بر روی منابع مختلف از قبیل تلفن، ایمیل، مرورگر وب، سوابق سفر و غیره شناسایی می‌شوند و بدین ترتیب پیوندهای موجود بین بازیگران آشنا و ناشناس تشخیص داده می‌شوند. گراف‌هایی از موسسات متصل بانکی،

حساب‌های بانکی مشتری‌ها، و تراکنش‌های انجام شده‌ی بانکی در زمان‌های خاص با استفاده از وسایل خاص مورد استفاده قرار می‌گیرند تا به روش‌های شناسایی بالقوه پولشویی کمک نمایند. روش‌های تحلیل داده‌ها از قبیل خوشه‌بندی، دسته‌بندی، شناسایی داده‌های پَرت، و ابزارهای بصری‌سازی داده‌ها [۹] نیز می‌توانند برای تشخیص الگوها در تراکنش‌ها مورد استفاده قرار گیرند، تراکنش‌هایی که شامل حجم عظیمی از جابه‌جایی پول بین مجموعه‌ی خاصی از حساب‌ها هستند. این روش‌ها پتانسیل شناسایی الگوها و روابط بین فعالیت‌های کلیدی را دارند که این امر می‌تواند به شناسایی موارد مشکوک جهت تحقیق و رسیدگی بیشتر کمک نماید.

۳-۳. *تحلیل خطر*

به طور کلی، بانک‌ها و موسسات مالی روش‌هایی برای اندازه‌گیری میزان خطر و کاهش آن دارند. نیروهای مختلف بازاری با انواع مختلف خطرها روبرو هستند، و درک درستی از ضررهای احتمالی برای تمام شرایط ممکن مورد نیاز است.

علاوه بر انواع مختلف خطرها در صنایع مالی [۱۰]، پیش‌بینی اعتبار دادن وام و حساب‌های کارت اعتباری نیز یکی از حوزه‌های مهمی است که به علت گستردگی این گونه حساب‌ها و کاهش ضررهای ناشی از آنها، به یکی از مسائل اساسی در کسب و کارها تبدیل شده است. پیش‌بینی عوامل مختلفی که در دادن اعتبار نقش اساسی دارند با استفاده از روش‌های داده‌کاوی انجام می‌شوند، روش‌هایی که مربوط به انتخاب ویژگی و ارتباط ویژگی هستند (شکل ۶). بر اساس نتایج به دست آمده از تحلیل‌ها، بانک‌ها می‌توانند مشتریانی را شناسایی کنند که متعلق به دسته‌ی کم-خطر هستند یا مبالغ پرداخت مناسب و قابل قبولی را به مشتریان پیشنهاد دهند.



شکل ۶. چارچوب تحلیل خطر مالی

۴. کاربردها در صنعت خرده‌فروشی

اکثر خرده‌فروشان بزرگ مانند Costco، Wal-Mart، Target، و Amazon از کلان داده‌ها برای عملیات مختلف خود از جمله مدیریت اموال، توصیه‌ی محصولات، ردیابی جمعیت‌شناسی مشتریان، و همچنین ردیابی و مدیریت تاثیرات منفی یادآوری محصولات استفاده می‌کنند. برخی از خرده‌فروشان نیز از داده‌های مرتبط با مشتری‌ها برای بهبود کیفیت سرویس خود و ارتقای وفاداری مشتریان استفاده نموده‌اند.

۴-۱. پیشنهاد محصولات

یکی از استراتژی‌های شناخته شده‌ای که شرکت‌های خرده‌فروشی برای افزایش درآمد خود از آن استفاده می‌کنند، پیشنهاد محصولات به مشتریان است، این محصولات پیشنهاد شده بر اساس محصولاتی توصیه می‌شوند که مشتری در حال حاضر مشغول خرید آنها می‌باشد و از این رو ممکن است علاقمند به خرید محصولات پیشنهادی نیز باشد. این مورد نوعی از یک خرده‌فروشی الکترونیکی است که در آن، سیستم‌های پایانی به اجرای موتورهای توصیه‌کننده‌ی محصولات می‌پردازند، این موتورها با استفاده از ارجاعات متقابل میان کالاهای فروخته شده به مشتریان مختلف، پیشنهاد خرید همان کالاها را به مشتریان دیگری می‌دهند که قصد خرید کالاهای مشابهی را دارند.

خرده‌فروشانی که حضور آنلاین و آفلاین (تجارت خشت و ملات[1]) دارند، می‌توانند از داده‌های جمع‌آوری شده از کانال‌های مختلف استفاده کنند و الگوهای خرید را یافته و محصولات را به صورت آنلاین پیشنهاد دهند. تحلیل مسیر و الگو برای تحلیل پیشینه‌ی رفتار خرید مشتری در کانال‌های مختلف مورد استفاده قرار می‌گیرد تا توصیه‌هایی با کیفیت بالا تولید شوند. روش‌های فیلتر کردن مشارکتی بر روی پیشینه‌ی خریدهای یک مشتری و الگوهای جستجو به کار گرفته می‌شوند، تا با مقایسه با رفتار دیگر مشتریان، توصیه‌های بعدی پیش‌بینی شود.

روش‌های فیلتر کردن مشارکتی در سیستم‌های توصیه‌گر توسط خرده‌فروشان الکترونیکی [۱۱] از قبیل Amazon برای پیشنهاد محصولات مورد استفاده قرار می‌گیرند، همچنین روش‌های فیلتر کردن مشارکتی توسط موتور توصیه‌گر فیلم نیز به کار گرفته می‌شود، همانطور که Netflix از آن استفاده می‌کند. از این روش‌ها به صورت آفلاین نیز به کار گرفته می‌شوند تا تراکت‌های آخر هفته، تبلیغات بر روی رسید فروش‌ها، یا تبلیغ بر روی بسته‌ی محصولات به منظور ارتقاء فروش به درستی ایجاد شوند.

۴-۲. *پیش‌بینی گرایشات*

خرده‌فروشان اطلاعات بسیار زیادی در مورد مشتری‌های خود از جمله مکان، جنسیت، و سن آنها را در حین تراکنش‌های مختلف معاملات خود جمع‌آوری می‌کنند. کاوش داده‌های خرده‌فروشی می‌تواند به شناسایی الگوهای خرید مشتریان و گرایشات آنها کمک کند که این امر به نوبه‌ی خود به شناسایی نیازهای مشتری برای برنامه‌ریزی موثر در جهت تبلیغ محصولات و جذب مشتریان بیشتر و افزایش درآمد/سود کمک می‌نماید [۱۲]. تحلیل چند-بُعدی و ابزارهای بصری‌سازی مجموعه‌دادگان می‌تواند برای پیش‌بینی مورد استفاده قرار گیرند تا بتوانند به برنامه‌ریزی تدارکات/حمل و نقل کالاهای مورد نیاز شرکت کمک کنند.

[1] brick and mortar

۵. کاربردهای کلان داده‌ها در ساخت و تولید

شرکت‌های تولیدی در سراسر جهان به شدت رقابتی شده‌اند و سود خالص نهایی[1] در کسب و کار تجاری روز به روز کاهش می‌یابد. تولیدکنندگان همیشه به دنبال بهینه‌سازی هزینه‌های در حال اجرای کارخانه‌ها هستند و در نتیجه هزینه‌های نهایی افزایش می‌یابند. همانطور که در ادامه شرح داده می‌شود، تحلیل‌های کلان داده‌ها می‌توانند در چندین حوزه مورد استفاده قرار گیرند [۱۳].

۵-۱. تعمیر و نگهداری پیشگیرانه

در دنیای خودکار ساخت و تولید، حسگرها در هر جای ممکن برای نظارت بر خط مونتاژ مورد استفاده قرار می‌گیرند و بدین ترتیب خرابی سیستم‌ها می‌توانند به سرعت شناسایی و تعمیر شوند تا مدت زمانی که خط تولید کار نمی‌کند، به حداقل کاهش یابد. علت اصلی خرابی سیستم‌ها می‌تواند به یک یا چند پارامتر متعدد بستگی داشته باشد که در زیرسیستم‌های کوچکتری پخش شده‌اند که این زیرسیستم‌ها به خط مونتاژ متصل هستند. مقدار زیادی از داده‌های حسگر همگی داده‌های بدون ساختاری هستند که از سطح کارخانه‌ی در حال اجرا و تولید جمع‌آوری شده‌اند. سوابق تعمیر و نگهداری از زیرسیستم‌های مختلف نیز به صورت داده‌های نیمه‌ساخت‌یافته جمع‌آوری می‌شوند. مستنداتی مربوط به بهره‌وری نسبت به حداکثر ظرفیت نیز به همراه سوابق تعمیر و نگهداری و داده‌های حسگرها جمع‌آوری می‌شوند.

تحلیل سری‌های زمانی بر روی زیرسیستم‌های مختلف بر اساس داده‌های حسگرهای مربوط به آن زیرسیستم‌ها انجام می‌گیرد و عمل تطابق الگو برای یافتن خرابی‌های احتمالی بر روی این داده‌ها اجرا می‌شود. همچنین، تحلیل مسیر و روش‌های ایجاد نشست نیز برای ضبط رویدادهای بحرانی مورد استفاده قرار می‌گیرند تا بر اساس همبستگی‌های موجود بین داده‌های خوانده شده توسط حسگرها، سوابق تعمیر و نگهداری، و مستندات جمع‌آوری شده به پیش‌بینی خرابی‌های احتمالی بپردازند. این امر کمک می‌کند تا اقدامات پیشگیرانه‌ای انجام شود تا خط مونتاژ برای مدت زمان طولانی و بدون وقفه در حال اجرا باشد و همچنین به بهبود ایمنی عملیات‌های در حال اجرا نیز کمک می‌کند.

[1] margins

۵-۲. *پیش‌بینی تقاضا*

با توجه به این که سفارشات روز-به-روز به صورت پویا در حال تغییر هستند، مهمترین عامل در کسب و کارهایی که با صنعت تولید در ارتباط هستند، این است که از منابع به صورت بهینه استفاده شود. وقتی که پیش‌بینی فروش و زمانبندی به درستی انجام گیرد، آنگاه به برنامه‌ریزی برای مواردی از قبیل به دست آوردن به موقع مواد اولیه، افزایش یا کاهش تولید، مدیریت انبار، و تدارکات حمل و نقل کمک خواهد کرد. در کوتاه مدت، اگر تخمین تقاضا بیش از حد زیاد در نظر گرفته شود، آنگاه سازنده را با محصولات به فروش نرفته‌ای مواجه می‌کند که می‌تواند تخلیه و خسارت مالی شدیدی به وی وارد نماید، همچنین اگر تخمین تقاضا بیش از حد کم در نظر گرفته شود، آنگاه منجر به از دست رفتن فرصت‌های فروش زیادی خواهد شد. در بلند مدت، پیش‌بینی تقاضا نیازمند برنامه‌ریزی بر روی سرمایه‌گذاری‌های استراتژیک و رشد کسب و کار است. از این رو، اجرای موثر یک کسب و کار با حداکثر سودآوری به یک سیستم جامع پیش‌بینی نیاز دارد.

سری‌های زمانی یک روش پیش‌بینی مشهور است که برای پیش‌بینی تقاضاها در آینده مورد استفاده قرار می‌گیرد و بر اساس داده‌های سوابق فروش می‌باشد. وقتی که محیط با عواملی مانند نیازهای در حال تغییر مشتری و تاثیر رقابت به صورت پویا است، روش ساده‌ی سری‌های زمانی نمی‌تواند پیش‌بینی درستی از آینده داشته باشد.

مدلسازی پیش‌بینی‌کننده[1] [۱۴] یک روش پیشرفته‌تر و دقیق‌تر است که توانایی در نظر گرفتن تمام متغیرهایی را دارد که بر روی تقاضاهای آینده تاثیر می‌گذارند. این روش همچنین آزمایش سناریوهای متنوع را نیز ممکن ساخته و به درک روابط بین عوامل تاثیرگذار و نحوه تاثیر آنها بر روی تقاضای پایانی نیز کمک می‌کند.

۶. کاربردهای کلان‌داده‌ها در مخابرات

با گسترش سرویس‌های مخابراتی در سراسر جهان، صنعت مخابرات در تلاش است با ارائه‌ی سرویس‌های گوناگون در زمینه‌ی صدا، ویدئو، و داده‌ها به بازارهای مختلف وارد شود. با توسعه‌ی فناوری‌ها و سرویس‌های

[1] Predictive modeling

جدید در میان کشورهای مختلف، بازار این صنعت نیز به سرعت در حال رشد است و بین فراهم‌کنندگان مختلف سرویس به شدت رقابت ایجاد شده است.

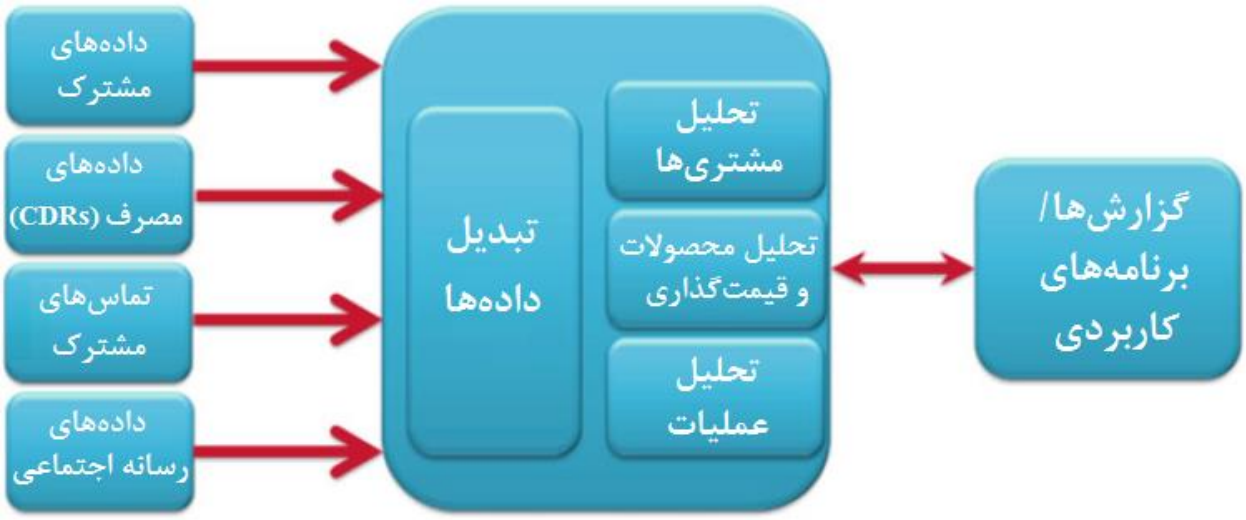
شکل ۷ چارچوبی از تحلیل کلان‌داده‌ها را برای حوزه‌ی مخابراتی نشان می‌دهد، که به عنوان پایه‌ای برای فرموله کردن استراتژی‌ها برای کسب و کار بهتر مورد استفاده قرار می‌گیرد. دیدگاه‌های تجاری برای بخش‌های مختلف کسب و کار بر اساس داده‌هایی استخراج می‌شود که از بسترهای متنوع جمع‌آوری شده‌اند. برخی از این موارد عبارتند از:

داده‌های مشتری / مشترک: اطلاعات و پیشینه‌ی رابطه با فراهم‌کننده.

الگوهای مصرف.

سوابق سرویس مشتری: شکایات مربوط به سرویس یا درخواست برای سرویس‌های اضافی و بازخورد[1].

اظهارنظرهای نوشته شده در رسانه‌های اجتماعی.



شکل ۷. چارچوب کلان‌داده‌ها در حوزه‌ی مخابرات

در بخش‌های زیر، ما به بررسی حوزه‌هایی خواهیم پرداخت که در آنها، صنعت با استفاده از کلان‌داده‌ها در تلاش است تا راه‌هایی را برای حفظ و تولید درآمد شناسایی کند.

[1] feedback

۶-۱. ریزش مشتری

به خوبی مشخص شده است که ریزش مشتری یک مشکل بزرگ برای تمام فراهم‌کنندگان سرویس مخابراتی می‌باشد. مشتریان فراهم‌کننده‌ی سرویس موجود را ترک نموده و در شرکت رقیب ثبت‌نام می‌کنند که باعث خسارت مالی و سوددهی می‌شود. به دست آوردن مشتریان جدید با استفاده از تبلیغات جدید، یک کار پُر هزینه است و بر روی افزایش هزینه‌های بازاریابی تاثیر دارد که به نوبه‌ی خود بر روی سودآوری شرکت تاثیر خواهد داشت.

مطالعات نشان داده است که شناسایی عوامل کلیدی ریزش مشتری به صورت فعال و توسعه‌ی استراتژی‌های حفظ مشتری، برای به حداقل رساندن کاهش درآمد و سوددهی کمک می‌کند. پس از آن، فراهم‌کننده‌ی سرویس می‌تواند بر روی ارتقای زیرساخت شبکه در جهت کیفیت بهتر سرویس و پشتیبانی بهتر از سرویس‌ها برای حفظ و رشد پایگاه مشتریان تمرکز کند.

به طور معمول برای شناسایی عوامل تحریک‌آمیز در ریزش مشتری‌ها و اعمال این عوامل بر روی مشترکان موجود و ارزیابی فرصت‌های لغو و انتقال سرویس آنها به فراهم‌کننده‌ی دیگر، روش‌های متعدد تحلیل داده‌های آماری [۱۵] مورد استفاده قرار می‌گیرند. با استفاده از داده‌های رفتار مشتری که از کانال‌های مختلف مانند پروفایل‌های تماس، تماس‌های شکایات مشتریان با مراکز تماس، اظهارنظرها از طریق پست‌الکترونیک، و بررسی بازخوردها جمع‌آوری شده‌اند، می‌توان پیش‌بینی ریزش بهتری برای شناسایی مشتریان با ریسک بالا انجام داد. به منظور تشخیص الگوهای رویدادهایی که منجر به ریزش مشتری می‌شوند، روش‌های تحلیل مسیر مورد استفاده قرار می‌گیرند. با استفاده از دسته‌بند نایو بیزین برای تحلیل متن، یک مدل ایجاد می‌شود تا مشتریان با ریسک بالا شناسایی شود.

یک روش مشهور دیگر از موتورهای گراف [۱۶] استفاده می‌کند تا اتصالات بین کاربران را بر اساس سوابق جزئیات تماس نمایش دهد و سپس جوامع و افراد تاثیرگذار را در میان جوامع کاربران شناسایی کند. یکی از اقدامات اصلاحی برای مقابله با این مشکل، متعهد کردن مشتریانی است که احتمال ریزش آنها بالا می‌باشد و همچنین ایجاد انگیزه و تمدید قرارداد آنها برای مدت زمان بیشتر است.

۶-۲. *ارتقاء سرویس‌ها*

فراهم‌کنندگان سرویس مخابراتی به طور مداوم به دنبال افزایش درآمد خود هستند و این کار را با پیشنهاد سرویس‌های کمکی و بیشتر به مشتریان انجام می‌دهند، و این پیشنهاد بر این اساس صورت می‌گیرد که مشتریان بر اساس طرح اشتراک فعلی آنها ممکن است به سرویس‌های پیشنهاد شده نیز علاقمند باشند. این پیشنهاد همچنین بر اساس تحلیل ارجاع متقابل میان مشتریانی صورت می‌گیرد که پرفایل‌های مشابهی دارند. استراتژی دیگر، ارتقاء سرویس به بهترین طرح ممکن با افزایش اندک قیمت است. روش‌های تحلیل داده‌هایی که برای این موتورهای توصیه‌گر مورد استفاده قرار می‌گیرند [۱۷]، به طور اساسی مشابه روش‌هایی است که برای کسب و کار خرده‌فروشی الکترونیکی (e-tailing) به کار گرفته می‌شوند.

۷. کاربردهای کلان‌داده‌ها در رسانه‌های اجتماعی

رسانه‌های اجتماعی آنلاین در حال رشد و توسعه هستند، این امر را می‌توان از رشد پایگاه کاربران فعال و مقدار داده‌هایی مشاهده نمود که توسط این رسانه‌ها تولید می‌شوند. در این روزها سایت‌هایی از قبیل Facebook، Twitter، +Google، LinkedIn، Reddit و Pinterest برخی از امکان اجتماع کاربران آنلاین هستند. حتی شرکت‌های بزرگ نیز استفاده از رسانه‌های اجتماعی را به عنوان یک کانال کسب و کار آغاز نموده‌اند و ایجاد حساب‌های Facebook، حساب‌های Twitter، کانال‌های TouTube، و وبلاگ شرکت‌ها برخی از این موارد است. باز بودن ذاتی رسانه‌های اجتماعی برای همه جهت شنیدن و گوش دادن به نظرات آنها و ایجاد روابط جدید، مسیر همواری را برای ایجاد داده‌های باارزش به وجود آورده است. این امر توجه متخصصان داده‌ها را در بررسی استفاده از رسانه‌های اجتماعی در حوزه‌های مختلف به خود جلب نموده است.

شکل ۸ یک چارچوب معمولی را برای کاربردهای شامل تحلیل رسانه‌های اجتماعی [۱۸] با اجزای اصلی مختلف نشان می‌دهد. تحلیل رسانه‌های اجتماعی شامل جمع‌آوری و تحلیل داده‌های عظیمی است که از طریق رسانه‌های اجتماعی تولید شده‌اند تا تصمیمات تجاری مناسبی بر اساس این تحلیل گرفته شود. اهداف این تحلیل شامل استراتژی‌های بازاریابی محصول، ارتقاء برند، شناسایی مسیرهای جدید فروش، توجه به مشتری، پیش‌بینی رویدادهای آینده، توسعه‌ی کسب و کار جدید، و غیره است.

شکل ۸. یک چارچوب معمول برای تحلیل رسانه‌های اجتماعی

جریان کاری شامل مرحله‌ی جمع‌آوری داده‌ها، مرحله‌ی تبدیل داده‌ها، مرحله‌ی تحلیل، و داشبورد نمایشی است. داده‌های رسانه‌های اجتماعی عمدتا شامل داده‌های بدون ساختاری هستند که از پُست‌های نوشته شده در وبلاگ‌ها و نظرات نوشته شده برای آنها، پیوند دوستان در Facebook، توییت‌ها و توییت‌های مجدد[1] در Twitter، و غیره به دست آمده‌اند. با توجه به هدف خاصی که قرار است تحلیل بر اساس آن صورت گیرد، داده‌های خام فیلتر می‌شوند و سپس برای درک و پیش‌بینی ساختار و پویایی‌های تعاملات جامعه مورد تحلیل قرار می‌گیرند. تحلیل احساسات [۱۹] و مدیریت شهرت، برخی از این کاربردها هستند که برای انجام آنها، از پردازش زبان طبیعی برای کاوش وبلاگ‌ها و نظرات استفاده می‌شود. روش‌های تحلیل گراف نیز برای شناسایی جوامع و افراد با نفوذ در داخل این جوامع مورد استفاده قرار می‌گیرند.

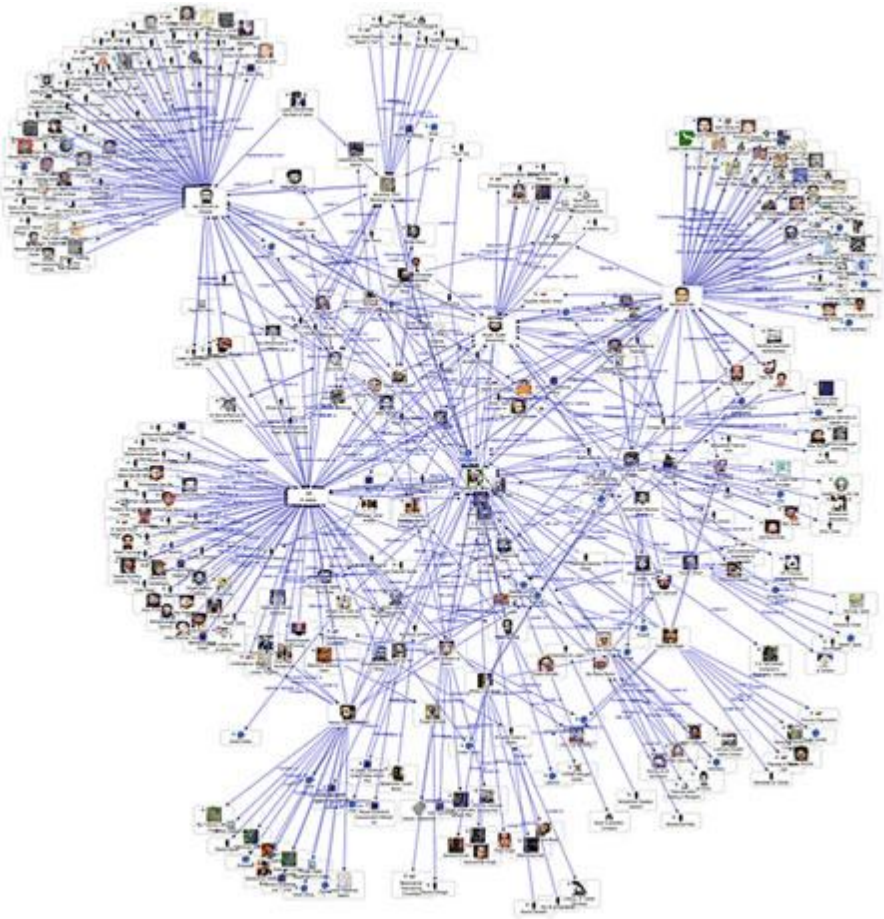
۷-۱. بازاریابی از طریق رسانه‌های اجتماعی

در رسانه‌های اجتماعی، به خوبی مشخص است که افراد مختلف به دلیل عوامل متعددی در سطوح مختلفی از تاثیرگذاری بر دیگران قرار دارند (یعنی صفحه‌ی افراد مختلفی به دلایل متعددی در رسانه‌های اجتماعی مورد تعقیب دیگران قرار می‌گیرد) و این مورد با استفاده از تعداد اتصالاتی مشخص می‌شود که فرد در رسانه‌های

[1] tweets/ retweets

اجتماعی با دیگران دارد (یعنی با تعداد افرادی که صفحه‌ی وی را دنبال می‌کنند). اتصالات کاربر-به-کاربر در یک گراف در شکل ۹ نشان داده شده است که به شناسایی افراد تاثیرگذار کمک می‌کند [۲۰]، سپس از این افراد تاثیرگذار می‌توان برای تبلیغ محصولات استفاده نمود. از این مورد می‌توان برای اطلاع‌رسانی در مورد نام تجاری (برند[1]) و کمک به بازاریابی ویروسی محصولات جدید استفاده نمود.

همچنین، با ایجاد انگیزه در میان مشتری‌هایی که تاثیر و نفوذ زیادی در جامعه دارند، و با استفاده از تاثیر و نفوذ آنها می‌توان ریزش مشتری را محدود یا از آن پیشگیری نمود.



شکل ۹. شناسایی افراد تاثیرگذار با استفاده از روش‌های خوشه‌بندی K-means

[1] brand

۷-۲. توصیه‌های اجتماعی

تحلیل گراف و پیوند به طور گسترده‌ای در سایت‌های حرفه‌ای شبکه‌های اجتماعی از قبیل LinkedIn مورد استفاده قرار می‌گیرند تا افراد حرفه‌ای دیگری را که کاربر ممکن است علاقمند به برقراری ارتباط با آنها باشد، را از طریق ترکیب اتصالات موجود شناسایی نموده و به کاربر پیشنهاد دهند.

سایت Reddit به منظور پیشنهاد مقالات و نوشته‌های جدیدی که شاید مورد علاقه‌ی کاربر باشند، به بررسی و تحلیل شباهت موجود میان گراف‌های ایجاد شده با استفاده از مقالات/ نوشته‌ها و مواردی می‌پردازد که کاربر علاقه‌ی خود را به مطالعه‌ی مقالات و نوشته‌های آن حوزه اعلام نموده است. لیست مقالات موجود در پایگاه‌داده، پروفایل کاربر، و پروفایل علایق کاربر مورد بررسی و تحلیل قرار می‌گیرند تا پیشنهادات و توصیه‌هایی به کاربران مختلف ارائه شود. الگوریتم خوشه‌بندی K-means در واقع روش تحلیلی استفاده شده برای سازماندهی داده‌ها به صورت گروه‌ها یا خوشه‌هایی هستند که بر اساس ویژگی‌های به اشتراک گذاشته شده توسط کاربران می‌باشند.

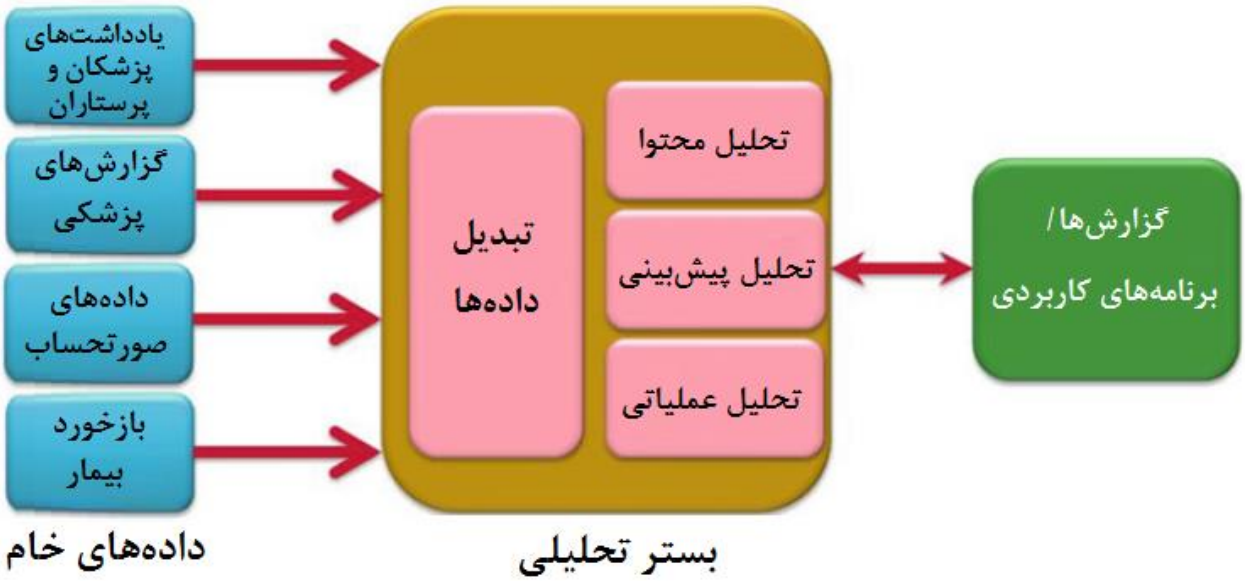
۸. کاربردهای کلان‌داده‌ها در مراقبت از سلامتی

استفاده از تحلیل کلان‌داده‌ها اهمیت زیادی در صنعت مراقبت از سلامتی به دست آورده است، علت این اهمیت نیز به مشخصات این حوزه از قبیل مجموعه‌داده‌ی عظیمی از پرونده‌های الکترونیکی افراد، ارائه‌ی سرویس با کمترین هزینه، نیاز به پشتیبانی از اخذ تصمیمات مهم و بحرانی، و غیره مربوط می‌شود.

شکل ۱۰ یک چارچوب معمول را برای کاربردهای کلان‌داده‌ها در صنعت مراقبت از سلامتی نشان می‌دهد که اجزای مختلف یک بستر معمولی در آن نمایش داده شده است. مقدار بسیار زیادی از داده‌های جمع‌آوری شده از حوزه‌ی مراقبت از سلامتی شامل داده‌های بالینی از قبیل پرونده‌های آزمایشگاهی، نسخه‌ی دکترها، مکاتبات پزشکی، پرونده‌های الکترونیکی پزشکی[1] (EMRها)، درخواست‌ها، و هزینه‌ها می‌باشند. تحلیل‌های پیشرفته بر رو این داده‌ها برای بهبود حفظ مشتری و نتایج، افزایش کارایی، و حفظ هزینه‌ها در کمترین سطح ممکن مورد استفاده قرار می‌گیرند. این تحلیل‌ها همچنین برای انجام تحقیقات کامل و تشخیص

[1] Electronic Medical Records (EMRs)

عوارض جانبی داروها مورد استفاده قرار می‌گیرند که این امر به رد صلاحیت و جمع‌آوری داروهای مضر سرعت می‌بخشد.



شکل ۱۰. چارچوب تحلیلی برای مراقبت از سلامتی

در ادامه چند مثال از تحلیل‌های کلان‌داده‌ها در صنعت مراقبت از سلامتی بیان می‌شود:

یافتن روش‌های جدید درمان

موسسات ملی بهداشت [۲۱] در ایالات متحده پایگاه‌داده‌هایی را نگه‌داری می‌کنند که شامل تمام مقالات پزشکی منتشر شده در حوزه‌های مختلف مراقبت از سلامتی می‌باشند و این مقالات را در دسترس تمام محققان علاقمند قرار می‌دهند. حجم این مجموعه‌داده‌ی متشکل از اسناد، بسیار بزرگ است و کاوش اطلاعات معنی‌دار و مفید در آن به یک چالش تبدیل شده است.

محققان از جستجوهای معنایی بر روی این پایگاه‌داده استفاده کرده‌اند تا روابط جدیدی را بین درمان‌ها و نتایج آنها به دست آورند. تحلیل گراف [۶] توسط محققانی مورد استفاده قرار می‌گیرد که بر روی سرطان تمرکز دارند و به این حقیقت پی برده‌اند که ایمونوتراپی بهتر از شیمی‌درمانی در برخی از موارد خاص سرطان عمل می‌کند. روش‌های نمایش بصری[1] [۲] نیز برای یافتن سریع همبستگی‌ها مورد استفاده قرار می‌گیرد.

[1] Visualization

مسیر چند-رویدادی برای جراحی

استفاده از روش‌های تحلیل مسیر و الگو بر روی داده‌های به دست آمده از پرونده‌ی بیمارانی با رویه‌های درمانی مختلف، این امکان را فراهم می‌آورد تا توالی رویدادها برای جراحی‌های گران قیمت شناسایی شود (یعنی اینکه مشخص شود کدام عمل در ابتدا و کدامیک باید در ادامه انجام شود، و بدین صورت ترتیب رویدادهایی که باید انجام شوند، مشخص گردد). با استفاده از این اطلاعات، مراقبت‌های پیشگیرانه‌ی بهتری برای جلوگیری از جراحی‌های پر هزینه و کمک به کاهش هزینه‌های پزشکی می‌توانند ارائه شوند.

کاهش بازبینی ادعا

ارزیابی ادعاهای پزشکی شامل بررسی نسخه‌ی دکترها، مدارک پزشکی، و مستندات رویه‌ای صورتحساب است، به خصوص در مواردی که رویه‌ی درمان پیچیده و شامل پروسه‌های متعدد است، انجام این کار بسیار وقت‌گیر و فرآیند دشواری می‌باشد. به منظور کاهش چنین تلاشی که باید به صورت دستی انجام شود، روش‌های تحلیل متن، یعنی نگاشت فازی (FuzzyMatch) برای تعیین شیوه‌های پرداخت نادرست و همچنین سوءاستفاده‌های احتمالی، کلاهبرداری، یا فعالیت‌های ناخواسته به کار گرفته می‌شوند.

۹. توسعه‌ی برنامه‌های کاربردی تحلیل کلان داده‌ها

چارچوب موردنظر برای برنامه‌های کاربردی تحلیل کلان‌داده‌ها از نظر مفهومی مشابه برنامه‌های کاربردی هوشمند تجاری است ولی تفاوت‌های زیر را دارند:

- تفاوت اصلی بر روی این مورد است که داده‌های با ساختار و بدون ساختار چگونه به صورتی که در فصل چارچوب کلان‌داده‌ها (فصل ۲) نشان داده شده است، ذخیره و پردازش شوند. برخلاف مدل‌های رایج و معمول که در آنها ابزار BI[1] (ابزار هوش تجاری) بر روی داده‌های با ساختار و عمدتا بر روی یک گره‌ی مستقل اجرا می‌شوند، برنامه‌های کاربردی کلان‌داده‌ها به منظور پردازش مقیاس بزرگی از داده‌ها، پردازش و اجرای تحلیل‌ها را بر روی گره‌های متعدد تقسیم می‌کنند، گره‌هایی که به طور محلی به داده‌های موجود دسترسی دارند.

[1] Business Intelligent (BI)

- برخلاف ابزار کلاسیک BI (هوش تجاری)، ابزارهای تحلیل کلان‌داده‌ها پیچیده و نیازمند برنامه‌نویسی زیاد هستند و همچنین باید توانایی تحلیل فرمت‌های مختلف داده‌ها را داشته باشند.
- یک چارچوب متفاوت برای توسعه‌ی برنامه کاربردی است که از مزیت اجرای تعداد زیادی از وظایف به صورت موازی بر روی گره‌های مختلف بهره می‌برد.

توسعه‌ی برنامه‌های کاربردی کلان‌داده‌ها نیازمند آگاهی از مشخصات خاص بسترهای مختلف است از قبیل:

- بستر محاسباتی – یک بستر با عملکرد بسیار بالا است که شامل چندین گره‌ی پردازشی می‌باشد که از طریق یک شبکه‌ی با سرعت بالا به هم متصل شده‌اند.
- سیستم ذخیره‌سازی – یک سیستم ذخیره‌سازی مقیاس‌پذیر برای رسیدگی به مجموعه‌دادگان بسیار بزرگ در انجام اموری مانند ثبت، تبدیل، و تحلیل داده‌ها است؛
- سیستم مدیریت پایگاه‌داده؛
- الگوریتم‌های تحلیلی – توسعه و ایجاد این الگوریتم‌ها از ابتدا یا استفاده از الگوریتم متن‌باز موجود یا مجموعه نرم‌افزارهای تجاری؛
- نیازهای عملکردی و مقیاس‌پذیری

توسعه‌دهندگان برنامه‌های کاربردی کلان‌داده‌ها به غیر از شناختی که باید در مورد معماری کلی بستری داشته باشند که کاربردهای موردنظر کلان‌داده‌ها قرار است در آن بستر پیاده‌سازی شوند، این توسعه‌دهندگان باید با چارچوب‌های مشهور برنامه‌های کاربردی کلان‌داده‌هایی که توسط بستر پشتیبانی می‌شوند، نیز آشنایی داشته باشند. مشهورترین چارچوب/ مجموعه‌ی نرم‌افزاری که توسعه‌ی کلان‌داده‌ها را ممکن می‌سازد، Apache Hadoop است که مجموعه‌ای از چندین پروژه‌ی متن‌باز می‌باشد. چارچوب Hadoop شامل قابلیت‌های مختلفی است که بر پایه‌ی سیستم‌های فایل توزیع شده‌ی Hadoop[1] (HDFS) و یک مدل برنامه‌نویسی به نام MapReduce و دیگر اجزای مختلف زیرساخت می‌باشد که چارچوب را پشتیبانی می‌کنند. این موارد شامل PIG، HIVE، JAQL، و HBase هستند.

[1] Hadoop Distributed File Systems (HDFS)

ایجاد برنامه‌های کاربردی تحلیلی پیچیده نیاز به تخصص بالا در به کارگیری روش‌ها و الگوریتم‌های داده‌کاوی بر روی چارچوب و معماری بستری است که این برنامه‌های کاربردی برای آن چارچوب و معماری در نظر گرفته شده‌اند. پیاده‌سازی‌های الگوریتم‌های مشهور به صورت متن‌باز[1] و برخی نیز به صورت پیاده‌سازی اختصاصی (با مالکیت معنوی) در دسترس هستند. نمونه‌هایی از پیاده‌سازی‌های متن‌باز عبارتند از:

- R برای تحلیل آماری،
- Lucene برای جستجو و تحلیل متن،
- کتابخانه‌ی Mahout – مجموعه‌ای از الگوریتم‌های تحلیلی است که به طور گسترده مورد استفاده قرار می‌گیرند و با استفاده از الگوی کاهش/ نگاشت[2] بر روی بسترهای Hadoop پیاده‌سازی شده‌اند تا برای ایجاد برنامه‌های کاربردی به کار گرفته شوند. این موارد شامل روش‌های فیلتر کردن مشارکتی، الگوریتم‌های خوشه‌بندی، دسته‌بندی، متن‌کاوی، و تحلیل سبد خرید می‌باشند.

پیاده‌سازی توابع تحلیل توسط فروشندگان شخص-ثالث یا متن‌باز ارائه می‌شود و واسطی مخصوص برنامه‌نویسان را دارند. یکی از چالش‌های اصلی توسعه‌دهندگان برنامه‌ی کاربردی، پیچیدگی موجود در برخی از عناصر کلیدی در استفاده از APIها در برنامه‌ی کاربردی است. این چالش‌ها عبارتند از:

- ادغام بسته‌های نرم‌افزاری متن‌باز با سیستم کلی و این که چگونه کتابخانه‌ها در معرض توسعه‌دهندگان قرار گیرد.
- پشتیبانی از جمع‌آوری داده‌های ورودی موردنیاز برای توابع از جداول پایگاه‌های داده، فایل‌های خام، و غیره.
- پشتیبانی از ذخیره‌ی نتایج توابع در جداول، بافرهای موقتی، فایل‌ها، و غیره،
- توانایی پیوند دادن روش‌های تحلیل متعدد به صورت یک توالی زنجیری (یعنی توانایی اجرای روش‌های تحلیل به صورت پشت سر هم)، به طوری که خروجی یک تابع به عنوان ورودی تابع بعدی در نظر گرفته شود و بدین ترتیب پیاده‌سازی برنامه‌ی کلی و نهایی ساده‌تر شود.

راه‌حل‌های بستر تجاری کلان‌داده‌ها توسط شرکت‌هایی مانند IBM [۲۳] و Teradata [۲] ارائه شده‌اند و شامل چارچوب‌های اختصاصی متعلق به این شرکت‌ها هستند. ادغام بسته‌های نرم‌افزاری متن‌باز مختلف و

[1] open source
[2] map/reduce

همچنین پیاده‌سازی/ پشتیبانی از بسته‌های اختصاصی در جایی که کتابخانه‌ی متن‌باز فاقد قابلیت موردنظر است، از عوامل کلیدی برای توانایی فروش بستر هستند. این راه‌حل‌های ادغام شده‌ی تجاری با توجه به چالش‌های ذکر شده، باعث سادگی در استفاده از بستر می‌شوند و همچنین در هنگام بازاریابی بسترها نیز استفاده از راه‌حل‌های متن‌باز به عنوان یکی از نقاط قوت مطرح می‌شود.

۱۰. نتیجه‌گیری

اکثر شرکت‌های بزرگ با مشکل پیدا کردن مقادیر در میان حجم زیادی از داده‌ها روبرو هستند، حجم عظیمی از داده‌ها که در طول سال‌ها جمع‌آوری کرده‌اند. بسته به بخش کسب و کاری که تجارت مورد نظر با آن سروکار دارد، روش‌های مختلفی برای تحلیل داده‌ها جهت شناسایی کسب و کارهای جدید، بهینه‌سازی بهره‌وری عملیاتی و غیره مورد استفاده قرار گرفته‌اند تا فروش کل را افزایش دهند.

ما در این مقاله سعی کردیم تا چندین حوزه از صنایع مختلفی را ارائه دهیم که کاربردهای کلان داده‌ها و تحلیل‌های آن به طور موثری در این حوزه‌ها مورد استفاده قرار گرفته‌اند. شناسایی حوزه‌های جدید و بررسی راه‌حل‌های جدید می‌تواند مسیر تمرکز مطالعاتی برای آینده باشد. شرکت‌ها ارزش را در سرمایه‌گذاری بر روی استراتژی‌های داده-محور دانسته و پی برده‌اند که برای پیشرفت در میدان رقابت، استراتژی کلان داده‌ها یک جزء کلیدی در کسب و کار است.