Big Data Related Technologies, Challenges and Future Prospects, Min Chen, Shiwen Mao, Yin Zhang, Victor C.M. Leung, SPRINGER BRIEFS IN COMPUTER SCIENCE, 2014

کلان داده فن‌آوری‌های وابسته، چالش‌ها و چشم انداز آینده

ترجمه: رضا سعیدی نیا

تقدیم به روح پدرم حاجی نورالله سعیدی نیا

روحش شاد و یادش گرامی

شادی همه درگذشتگان صلوات

فهرست مطالب

فصل اول: مقدمه

چکیده: اصطلاح کلان داده[1] به خاطر انفجار افزایشی داده‌های عمومی ابداع شد و عمدتاً برای توصیف مجموعه داده‌های[2] زیاد استفاده می‌شود. در این فصل، تعاریف کلان داده را معرفی می‌کنیم و تکامل آن‌را در ۲۰ سال اخیر مرور می‌کنیم. به طور خاص، ویژگی‌های کلان داده ، همچنین ویژگیهای 4V آن شامل حجم[3]، تنوع، سرعت و ارزش را معرفی می‌کنیم. چالش‌های مربوط با کلان داده نیز در این فصل معرفی می شوند.

۱-۱- طلوع دوران کلان داده

در ۲۰ سال اخیر، داده‌ها در مقیاس بزرگی در رشته‌های مختلف افزایش یافته‌اند. براساس یک گزارش از شرکت داده‌های بین المللی (IDC[4]) در سال ۲۰۱۱ کل داده‌های تولید شده و کپی شده در دنیا 1.8ZB حدود 10^{21}B می‌باشد که در طول ۵ سال ۲ برابر شده است. این رقم در دو سال آینده دو برابر خواهد شد.

اصطلاح کلان داده به خاطر انفجار افزایشی داده‌های عمومی ابداع شد و عمدتاً برای توصیف مجموعه داده‌های زیاد استفاده می‌شود. در مقایسه با مجموعه داده‌های سنتی، کلان داده عموماً نیازمند انبوه داده‌های ساختارنیافته می‌باشد که نیازمند تجزیه تحلیل زمان-قطعی[5] بیشتری می‌باشد. مجلات زیادی مثل Economist, New York Time و غیره راجع به کلان داده مطلب می‌نویسند و در ژورنال‌های علمی معتبر مطلب در مورد آن چاپ می‌شود. بسیاری از آژانس‌های دولتی نقشه‌هایی برای شتاب در تحقیقات و کاربردهای کلان داده دارند و صنایع نیز در مورد پتانسیل کلان داده علاقه مند هستند.

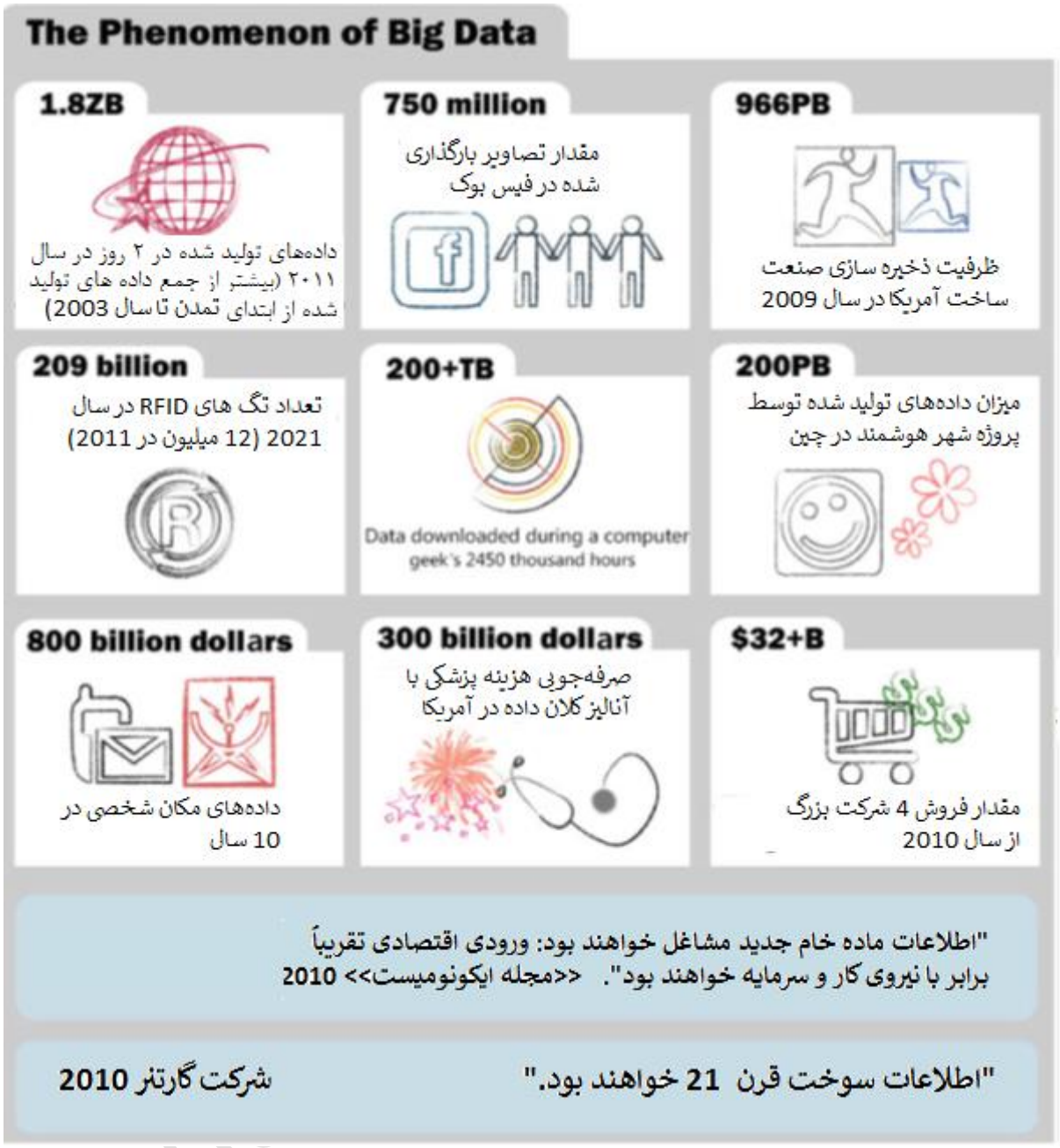
رشد حجم داده‌های تولید شده بسیار زیاد می‌باشد و این چالش‌هایی در پی دارد و متقاضی راه‌حل‌های سریع است:

۱- توسعه‌های اخیر در فنآوری IT به راحتی داده تولید می‌کنند، بطور مثال در هر دقیقه ۷۲ ساعت ویدیو در یوتیوب بارگذاری می‌شود که با چالش جمع‌آوری و تجمع حجم عظیمی از داده‌ها از منابع داده توزیع شده مواجه هستیم.
۲- داده‌های جمع‌آوری شده بطور افزایشی در حال زیاد شدن هستند که باعث مساله چگونگی ذخیره و مدیریت مجموعه داده نامتجانس حجیم هستیم که تیازمند نیازها و زیرساخت‌های سخت افزاری و نرم‌افزاری می‌باشد.
۳- با در نظر گرفتن عدم تجانس، مقیاس‌پذیری، زمان قطعی، پیچیدگی، و محرمانگی چنین داده‌های کلانی نیاز به داده‌کاوی موثر در سطوح مختلف تجزیه تحلیل، مدل‌سازی، پیش‌بینی، و روش‌های بهینه سازی دارند تا بتوانیم تصمیم گیری را بهبود دهیم.

رشد سریع رایانش ابری و اینترنت اشیاء، IOT[6] باعث تسریع رشد داده‌ها می‌شود. رایانش ابری، محافظت امن، سایت‌های دسترسی، و کانال‌ها را برای مجموعه داده‌ها فراهم می‌کند. در IOT سنسورها در همه جهان در حال جمع‌آوری و ارسال داده‌ها هستند که باید در ابر پردازش و ذخیره شوند، شکل ۱-۱ توسعه حجم داده‌های عمومی را نشان می‌دهد.

---

[1] Big Data
[2] Dataset
[3] Volume-Variety- Velocity, and Value= 4V
[4] International Data Corporation
[5] Real- Time
[6] Internet of things



شکل ۱-۱ نمایش افزایش مداوم داده‌ها

۱-۲ تعریف و ویژگی‌های کلان داده

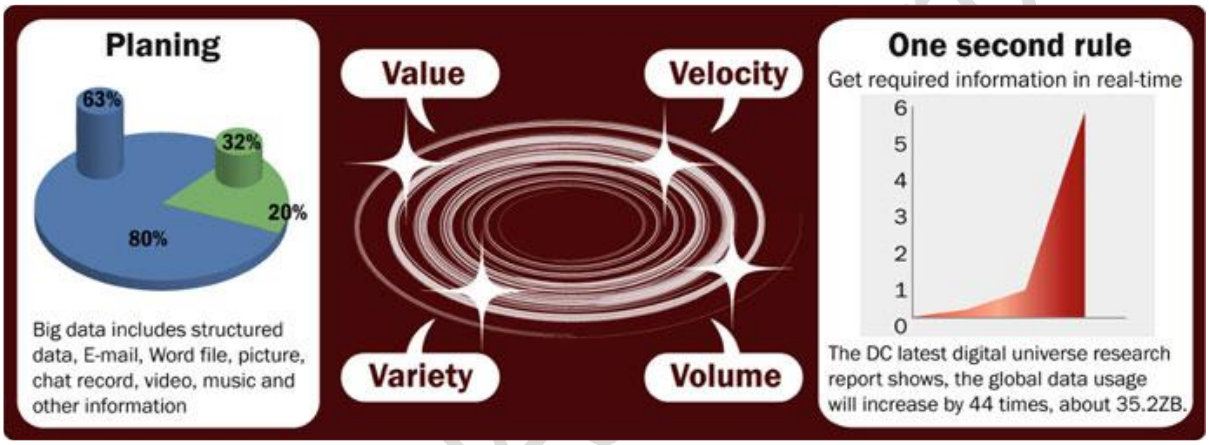
کلان داده یک مفهوم انتزاعی است و تعاریف مختلفی برای آن مطرح شده است مهمترین آن این است: ***داده‌هایی هستند که نمی‌توانند به روش‌های IT سنتی مشاهده، جمع‌آوری، مدیریت و پردازش شوند***. از دید حجم داریم:

۱- حجم‌های مجموعه داده که با استاندارد کلان داده تطابق دارد در حال تغییر هستند و ممکن است در زمان و با توسعه فن‌آوری تغییر کنند.

۲- حجم‌های مجموعه داده که با استاندارد کلان داده تطابق دارند ممکن است در کاربردهای مختلف تغییر کنند.

برای کلان داده مدل‌های مختلفی ارائه شده است مثل مدل 3V که شامل Volume: حجم رو به افزایش است، Velocity: سرعت پردازش باید سریع باشد، Variety: تنوع مشخص کننده انواع مختلف داده‌ها است مثل ساختار یافته، نیمه ساختار یافته و ساختارنیافته مثل صدا، تصویر، ویدیو، صفحه وب، متن و داده‌های سنتی ساختار یافته.

تعاریف مختلفی برای کلان داده مطرح شده است از جمله تعریف IDC: ***فن آوری کلان داده نسل جدیدی از فن‌آوری و معماری‌ها را معرفی می‌کند و طوری طراحی شده است تا بطور مقتصدانه ارزش را از حجم زیاد و متنوع داده استخراج کند با تشخیص با سرعت بالا و/یا تجزیه تحلیل***. در این تعریف کلان داده بصورت مدل 4V شامل Volume:حجم، Variety: تنوع، Velocity: سرعت تولید و Value: ارزش تعریف می‌شود. این تعریف بحرانی‌ترین مساله را در کلان داده مشخص می‌کند: یعنی بدست آوردن ارزش از مجموعه داده‌ها در مقیاس فوق العاده زیاد با انواع متفاوت و سرعت تولید بالا. شکل ۱-۲ این مدل را نشان می‌دهد.



شکل ۱-۲: مدل 4V از ویژگی‌های کلان داده.

۳-۱ ارزش کلان داده: تحقیقات روی ۵ مرکز کلیدی نشان می‌دهد:

- در حوزه سلامت با استفاده از کلان داده ۸٪ بهبود به دست آمد.
- در حوزه خرده فروشی‌ها ۶۰٪ بهبود به دست می‌آید.
- کلان داده تأثیر در بهبود عملکرد دولت دارد.
- در اروپا ۱۰۰ میلیارد یورو صرفه جویی در حوزه اقتصاد دارد.

۴-۱ چالش‌های کلان داده

چالش‌ها در گردآوری داده‌ها[1]، ذخیره‌سازی[2]، مدیریت[3] و تجزیه تحلیل[4] می‌باشد. چالش‌هایی کهحوزه کلان داده با آن مواجه می‌باشد عبارتند از:

[1] data acquisition
[2] Storage
[3] Management
[4] Analysis

- **نمایش داده**[1]: بسیاری از مجموعه داده‌ها شامل سطوح خاصی از عدم تجانس در نوع، ساختار، معنی، سازمان‌دهی، دانه‌بندی، و قابلیت دسترس‌پذیری دارند. هدف نمایش داده این است که داده‌ها برای تجزیه تحلیل کامپیوتری و تفسیر کاربر با معنی‌تر به نظر برسند. به علاوه، نمایش نامناسب داده‌ها، ارزش داده‌ها را کاهش می‌دهد و ممکن است مانع تجزیه تحلیل موثر داده‌ها شوند. نمایش موثر داده‌ها ساختار داده‌ها، کلاس، و نوع آنها و تجمع فن‌آوری‌ها را انعکاس خواهد داد.
- **کاهش افزونگی و فشرده‌سازی داده‌ها**[2]: معمولاً سطح قابل توجهی از افزونگی در مجموعه داده‌ها وجود دارند. کاهش افزونگی و فشرده‌سازی داده‌ها برای کاهش هزینه غیرمستقیم کل سیستم موثر است بالاخص اگر ارزش داده‌ها تحت تاثیر قرار نگیرد. به عنوان مثال اکثر داده‌های تولید شده توسط شبکه سنسورها خیلی افزونه هستند که ممکن است در یک سطحی فشرده سازی و فیلتر شوند.
- **مدیریت چرخه عمر داده‌ها**[3]: در مقایسه با پیشرفته‌های نسبتاً کُند در سیستم های ذخیره‌سازی، سنسورهای فراگیر شده و محاسبات، داده‌ها را در مقیاس و سرعت بالا تولید می‌کنند. ما با چالش‌های زیادی مواجه هستیم که یکی از آنها این است که سیستم ذخیره سازی موجود نمی‌تواند چنین حجمی از داده‌ها را پشتیبانی کند. بطور کلی ارزش مخفی در کلان داده وابسته به تازگی داده‌ها می‌باشد. بنابراین مفاهیم مهمی مرتبط با ارزش روش‌های تجزیه تحلیل[4] باید ابداع شود تا تصمیم بگیرد چه داده‌هایی باید ذخیره شوند و چه داده‌هایی باید حذف شوند.
- **مکانیزم‌های روش تجزیه تحلیل:** سیستم‌ها تحلیلی کلان داده باید انبوهی از داده‌های نامتجانس را در زمانی محدود پردازش کنند. به هر حال RDBMSهای سنتی با کمبود مقیاس‌پذیری و قابلیت توسعه طراحی شده‌اند که نمی‌توانند نیازهای کارایی را برآورده کنند. پایگاه داده‌های غیر رابطه‌ای در تجزیه تحلیل داده‌های غیرساختار یافته مزایایی دارند و در کلان داده‌ها استفاده می‌شوند. البته این پایگاه داده‌ها کارایی پایگاه داده‌های رابطه ای را ندارند و بعضی سرمایه گذاران از پایگاه داده‌های ترکیبی استفاده می‌کنند تا مزایای هر دو را استفاده کنند. (مثل فیس بوک و تابو). تحقیقات بیشتری برای پایگاه داده در-حافظه [5]و داده‌ها براساس تجزیه تحلیل تقریبی مورد نیاز است.
- **محرمانگی داده‌ها**[6]: بسیاری از تهیه کنندگان سرویس کلان داده‌ها در حال حاضر نمی‌توانند بطور موثری چنین حجم انبوهی از داده‌ها را نگهداری و تجزیه تحلیل کنند به خاطر ظرفیت محدود شده آنها. آنها باید به افراد خبره یا ابزارهایی تکیه کنند تا داده‌ها را تجزیه تحلیل کنند، که این ریسک‌های امنیتی را افزایش می‌دهد. بنابراین باید واحدهای خوبی برای تضمین امنیت داده‌ها ایجاد شوند.
- **مدیریت انرژی:** انرژی مصرفی سیستم‌های محاسباتی مین فریم‌ها هم از دید اقتصادی و هم محیطی توجه زیادی را به خود جلب کرده است. با افزایش حجم داده‌ها و نیازهای تجزیه تحلیلی، پردازش، ذخیره سازی، و انتقال کلان داده مصرف توان و مدیریت برق بیشتری نیاز دارند. بنابراین باید مکانیزم‌های کنترل مصرف توان سطح-سیستم و مدیریتی برای کلان داده‌ها برقرار شوند در حالیکه قابلیت توسعه و قابل دسترس بودن هر دو تضمین شود.

---

[1] Data Representation
[2] Redundancy Reduction and Data compression
[3] Data Life Cycle Management
[4] Analytic
[5] In- memory
[6] Data Confidentiality

- **قابلیت توسعه و مقیاس‌پذیری**[1]: سیستم‌های تجزیه تحلیلی کلان داده‌ها باید مجموعه داده‌های حال و آینده را پشتیبانی کنند. الگوریتم‌های روش تجزیه تحلیل باید قادر باشند مجموعه داده‌های پیچیده‌تر و رو به توسعه‌تر را پردازش کنند.
- **همکاری**[2]: تجزیه تحلیل کلان داده‌ها یک حوزه تحقیقاتی فراگیر است که نیازمند این است که خبرگان در رشته-های مختلف با هم کار کنند تا پتانسیل کلان داده را به خوبی استفاده کنند. یک معماری شبکه کلان داده جامع باید ایجاد شود تا به مهندسین و دانشمندان در رشته‌های مختلف کمک کند تا به انواع مختلف داده‌ها دسترسی داشته باشند و سرمایه‌شان را به خوبی بهره برداری کنند.

[1] Expendability and Scalability
[2] Cooperation

فصل دوم فن‌آوری‌های مرتبط با کلان داده‌ها

چکیده: برای فهم عمیق کلان داده‌ها، این فصل چندین فن‌آوری پایه که کاملاً به کلان داده مرتبط هستند را معرفی می‌کند، شامل رایانش ابری[1]، اینترنت اشیاء، دیتاسنتر، و هدوپ. برای هر فن‌آوری ابتدا ویژگی‌های کلیدی هر فن‌آوری معرفی می‌شود و سپس رابطه بین آن فن‌آوری و کلان داده را مطرح می‌کنیم.

۲-۱ رایانش ابری

۲-۱-۱ مقدمات رایانش ابری

در کلان داده، زیرساخت سخت افزاری قابل اطمینان برای تهیه حافظه ذخیره‌سازی قابل اطمینان بحرانی است. زیرساختار سخت افزاری شامل منابع فن‌آوری ارتباطی و ICT می‌باشد. در چند سال اخیر پیشرفت‌ها در رایانش ابری روشی که مردم منابع سخت افزاری و نرم‌افزاری را بدست می‌آورند را تغییر داده است.

رایانش ابری نتیجه رشد در حوزه‌های محاسبات توزیع شده، محاسبات موازی، و محاسبات گرید یا مفاهیم تجاری علم کامپیوتر می‌باشد. رایانش ابری به معنی استفاده از نوعی زیرساخت IT مثل بدست آوردن منابع لازم از طریق اینترنت برحسب نیاز یا به روشی قابل توسعه می‌باشد. این سرویس‌ها ممکن است به نرم‌افزار، اینترنت و غیره مرتبط باشند. بطور خلاصه به معنی حالتی است که کاربران به یک سرور از طریق شبکه در مکان دور دسترسی دارند و سرویس‌های تهیه شده توسط سرور را استفاده می‌کنند.

این مفهوم عمدتاً از مفاهیم درهم پیچیده مثل محاسبات و زیرساخت‌های عمومی مجازی تکامل یافته است. اجزاء اصلی رایانش ابری در شکل ۲-۱ نشان داده شده‌اند.

سرویس‌های فراهم شده از رایانش ابری می‌توانند با سه مدل سرویس و سه مدل توسعه توصیف شوند. چنین ترکیبی ویژگی‌های مهم زیادی دارد مثل، سلف-سرویس در صورت نیاز، دسترسی به شبکه گسترده، استخر منابع[2]، سرعت، کشسان بودن[3]، و مدیریت سرویس بنابراین نیازهای کاربردهای زیادی را برآورده می‌کنند. بنابراین رایانش ابری برای تجزیه تحلیل و کاربردهای کلان داده سودمند هستند.

۲-۱-۲ رابطه بین رایانش ابری و کلان داده‌ها

رایانش ابری با کلان داده ارتباط تنگاتنگی دارد. اجزاء اصلی رایانش ابری در شکل ۲-۱ نشان داده شده است. کلان داده مفعول عملیات محاسباتی است و بر ظرفیت ذخیره سازی و ظرفیت محاسباتی یک سرور ابری تاکید دارد. هدف اصلی رایانش ابری استفاده از منابع محاسباتی خیلی زیاد و ظرفیت‌های محاسباتی تحت مدیریت متمرکز است تا برای کاربردها، اشتراک منابع را بوجود آورد و برای کاربردهای کلان داده ظرفیت محاسباتی ایجاد کند. توسعه رایانش ابری راه‌حل‌هایی برای ذخیره سازی و پردازش کلان داده فراهم می‌کند. به عبارت دیگر، ظهور کلان داده نیز به توسعه رایانش

---

[1] Cloud computing
[2] Resource pool
[3] Elasticity

ابری شتاب می‌بخشد. فن‌آوری ذخیره سازی توزیع شده براساس رایانش ابری به مدیریت موثر کلان داده کمک می‌کند. ظرفیت محاسباتی موازی رایانش ابری می‌تواند بهره‌وری بدست آوردن و تجزیه تحلیل کلان داده را بهبود دهد.

شکل ۲-۱: اجزاء کلیدی رایانش ابری

هرچندکه مفاهیم و فن‌آوری‌های هم‌پوش زیادی در رایانش ابری و کلان داده وجود دارند اما در دو جنبه عمده با هم متفاوت هستند: اولاً مفاهیم متفاوت هستند. رایانش ابری معماری IT را دگرگون می‌کند، اما کلان داده بر تصمیم‌گیری شغلی تأثیر دارد در حالیکه کلان داده وابسته به رایانش ابری به عنوان زیرساخت پایه برای عملیات می‌باشد. ثانیاً کلان داده و رایانش ابری مشتری‌های نهایی مختلفی دارند. رایانش ابری یک فن‌آوری است و محصول هدف دفاتر اطلاعات عمده (CIO[1]) به عنوان راه‌حل IT پیشرفته است و کلان داده محصول هدف دفاتر اجرای عمده (CEO[2]) است که روی عملیات مشاغل تمرکز دارد چون تصمیم گیران ممکن است مستقیماً فشار رقابت بازار را حس کنند، آنها ممکن است به روش‌های مختلفی کا را انجام دهند. با پیشرفت‌های رایانش ابری و کلان داده، این دو فن‌آوری بطور عمده و افزایشی به هم مرتبط و درهم پیچ شده‌اند. رایانش ابری با عملکردها شبیه کامپیوترها و سیستم‌های عامل است که منابع سطح پایین را فراهم می‌کند، کلان داده در سطح بالاتر عمل می‌کند که توسط رایانش ابری پشتیبانی می‌شود و عملکردهایی مشابه پایگاه داده و ظرفیت پردازش داده را فراهم می‌کند. کیسینگر رئیس EMC می‌گوید کاربرد کلان داده باید براساس رایانش ابری باشد.

[1] Chief Information Officers
[2] Chief Executive Officers

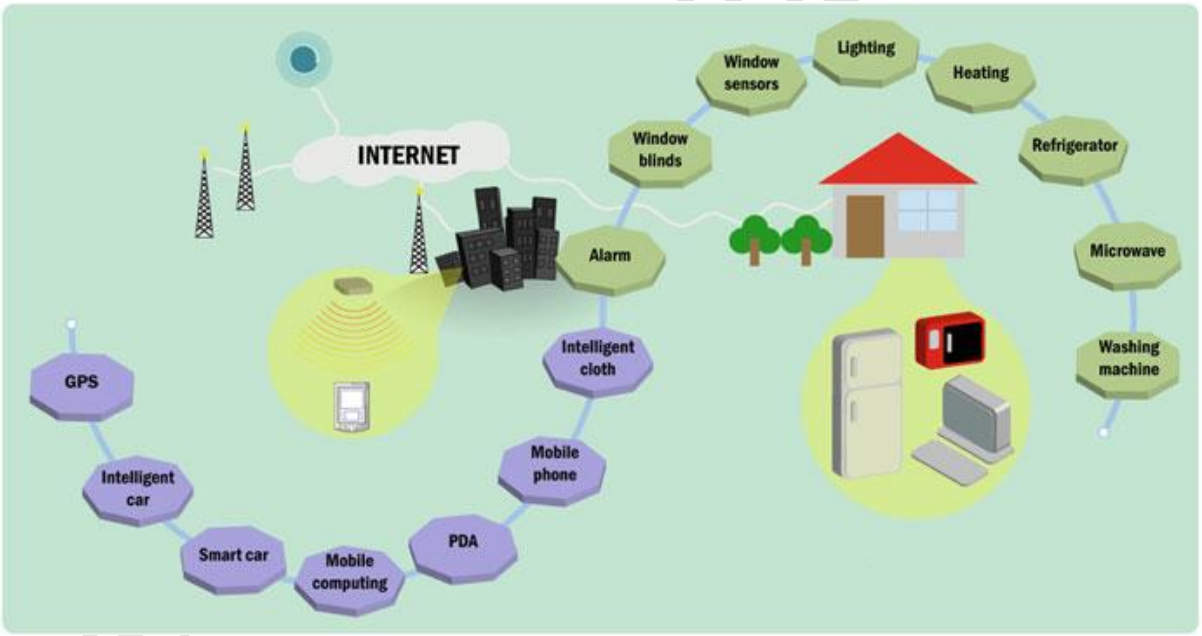
تکامل کلان داده به خاطر رشد سریع نیازهای کاربردی و رایانش ابری که از فن‌آوری‌های مجازی سازی توسعه یافته‌اند می‌باشد. بنابراین رایانش ابری هم محاسبات و پردازش را برای کلان داده فراهم می‌کند، هم خودش یک نوع سرویس است. به عبارت دیگر کلان داده و رایانش ابری لازم و ملزوم هم هستند و باعث ارتقاء یکدیگر می‌شوند.

۲-۲ اینترنت اشیاء

۲-۲-۱: مقدمات IOT: ایده اصلی IOT اتصال اشیاء مختلف در دنیای واقعی مثل RFID، bar code reader، سنسورها و گوشی‌های تلفن و غیره است. تا بتوانند در انجام یک کار مشترک با هم مشارکت کنند. معماری IOT در شکل ۲-۲ نشان داده شده است. به نظر می‌رسد که IOT توسعه یافته اینترنت است و یک بخش مهم آینده اینترنت خواهد بود. IOT عمدتاً برای کنترل، ارتباط و بررسی اشیاء دنیای واقعی استفاده می‌شود.

در مقایسه با اینترنت IOT ویژگی‌های کلیدی زیر را دارد:

- تجهیزات پایانه مختلف
- مالکیت خودکار داده‌ها
- ترمینال‌های هوشمند



شکل ۲-۲ نمایش معماری IOT

۲-۲-۲-: رابطه بین IOT و کلان داده

در IOT مقدار زیادی سنسورهای شبکه در دستگاههای واقعی جاسازی می‌شوند. چنین سنسورهایی در رشته‌های مختلف استفاده می‌شوند و می‌توانند انواع مختلفی داده مثل داده‌های محیطی، داده‌های جغرافیایی، داده‌های نجومی و داده‌های لوجیستیک را جمع‌آوری کنند. تجهیزات موبایل، امکانات انتقال، امکانات عمومی، موارد خانگی همه می‌توانند تجهیزات IOT باشند.

داده‌های انبوه تولید شده توسط IOT با کلان داده ویژگی‌های متفاوتی دارد زیرا از انواع متفاوتی داده جمع‌آوری شده‌اند و ویژگی‌های عدم تجانس، ویژگی غیرساخت یافتگی، نویز و رشد سریع دارند. هرچند که داده‌های IOT جاری عضو دائمی کلان داده نیستند، در سال ۲۰۳۰ تعداد سنسورها به بک تریلیون خواهد رسید و داده‌های IOT بر اساس پیش‌بینی HP مهمترین بخش کلان داده خواهند شد. اینتل اعلام کرد که کلان داده‌های IOT سه ویژگی دارد که با کلان داده تطابق دارد: الف) ترمینال‌های زیادی داده‌های انبوهی را تولید می‌کنند. (ب) داده‌های تولید شده با IOT غالباً نیمه ساختار یافته یا ساختار یافته هستند. (ج) داده‌های IOT زمانی مفید هستند که تجزیه تحلیل شوند.

موفقیت IOT بر اساس تجمع مفید کلان داده و رایانش ابری می‌باشد. توسعه گسترده IOT بسیاری از شهرها را به حوزه کلان داده وارد خواهد کرد. نیاز ضروری برای تطابق کلان داده برای کاربردهای IOT وجود دارد.

۳-۲ دیتاسنتر

در کلان داده، یک دیتاسنتر یک سازمان برای ذخیره متمرکز داده است و وظایفی مثل بدست آوردن داده‌ها، مدیریت داده‌ها، سازمان‌دهی داده‌ها و استفاده از مقادیر و توابع داده‌ها دارد. دیتاسنترها بیشتر با مفهوم داده سازگارند تا سنتر. یک دیتاسنتر حجم انبوهی از داده‌ها را بر اساس اهداف مرکزی آن مدیریت و سازمان‌دهی می‌کند. ظهور کلان داده، باعث توسعه زیاد و چالش‌های بزرگی در دیتاسنترها شده است:

- کلان داده به این نیاز دارد که دیتاسنتر از آن بطور خصوصی پشتیبانی کند. کلان داده نیازمندی‌های شدیدتری به ظرفیت ذخیره سازی و ظرفیت پردازش و ظرفیت انتقال شبکه دارد. شرکت‌ها باید دیتاسنترها را توسعه دهند تا ظرفیت سریع و موثر پردازش کلان داده را تحت نرخ محدود پول/کارایی بهبود دهند. دیتاسنتر باید زیرساخت را با تعداد زیادی گره فراهم کند، یک شبکه داخلی سرعت بالا بسازد، بطور موثری گرما را از بین ببرد، و بطور موثری داده‌ها را پشتیبان گیری کند. فقط وقتیکه دیتاسنترهای با مصرف انرژی مناسب، پایدار، امن، قابل توسعه و افزونه ساخته شوند عملیات نرمال کلان داده ممکن است برآورده شوند.
- رشد کاربردهای کلان داده، تکامل و نوآوری دیتاسنترها را شتاب می‌دهد. بسیاری از کاربردهای کلان داده معماری منحصر بفردشان را توسعه داده‌اند و مستقیماً واحدهای ذخیره، شبکه و فن‌آوری‌های محاسباتی مرتبط با دیتاسنتر را توسعه داده‌اند. با پیشرفت مداوم داده‌های ساختار یافته و ساختار نیافته و تنوع منابع تجزیه تحلیل داده‌ها، ظرفیت‌های پردازش و محاسبه دیتاسنترها خیلی توسعه خواهد یافت. در مجموع مقیاس دیتاسنتر بطور افزایشی در حال توسعه است، البته باید به عنوان یک مقیاس، قیمت عملیاتی توسعه دیتاسنترها کاهش یابد.
- کلان داده کارکردهای بیشتری را برای دیتاسنترها بوجود آورده است. در کلان داده یک مرکز داده باید ظرفیت‌های سخت افزاری و نرم‌افزاری مثل بدست آوردن، پردازش، سازمان‌دهی تجزیه تحلیل و کاربردهای کلان داده را افزایش دهد. دیتاسنتر به کارمندان کمک می‌کند داده‌های موجود را تجزیه تحلیل کنند، مسائل را کشف کنند و راه‌حل-هایی راجع به کلان داده را توسعه دهند.

کلان داده یک حوزه نوظهور است که رشد انفجاری زیرساخت‌ها و نرم افزارهای دیتاسنتر را توسعه می‌دهد. شبکه دیتاسنتر هسته پشتیبانی کلان داده است.

۴-۲ هدوپ[1]

۴-۲-۱ مقدمات هدوپ: هدوپ یک فن‌آوری مرتبط با کلان داده است که یک راه حل سیستماتیک کلان داده قدرتمند از طریق ذخیره سازی داده، پردازش داده، مدیریت سیستم و تجمع سایر ماژول‌ها شکل می‌دهد. چنین فن‌آوری لازم است تا بتواند بر چالش‌های کلان داده غلبه کند. هدوپ مجموعه‌ای از زیرساخت‌های نرم‌افزاری با مقیاس-بالا برای کاربردهای اینترنت مشابه با سیستم فایل گوگل[2] می‌باشد. در حال حاضر بزرگترین خوشه[3] هدوپ در یاهو کار می‌کند که شامل ۴۰۰۰ مجموعه گره[4] می‌باشد که برای پردازش و تجزیه تحلیل داده‌ها شامل آگهی های یاهو، داده‌های مالی، و لاگ‌های کاربر استفاده می‌شود.

هدوپ از دو بخش تسکیل شده‌است: HDFS, MR[5]. HDFS[6] یک منبع داده MR است که یک سیستم فایل توزیع شده است که روی سخت افزار تجاری اجرا می‌شود و براساس GFS گوگل طراحی شده است. HDFS پایه ذخیره‌سازی داده‌های کاربردهای هدوپ است، که فایل‌ها را در بلوک‌های داده 64MB توزیع می‌کند و چنین بلوک‌های داده را در گره‌های مختلف یک خوشه ذخیره می‌کند بطوریکه محاسبات موازی MR انجام‌پذیر باشد. یک خوشه HDFS شامل یک سیستم NAMENode است برای مدیریت متا دیتای سیستم فایل و DataNodes برای ذخیره داده‌های واقعی. یک فایل به یک یا چند بلوک تقسیم می‌شود و در DataNodes ذخیره می‌شود. کپی‌های بلوک‌ها در DataNodeهای مختلف ذخیره می-شوند تا از گم شدن داده جلوگیری شود. Apache HBase یک حافظه ستون‌گرا است که از Google BigTable تقلید می‌کند و بنابراین عملکردهای آپاچی HBase شبیه BigTable می‌باشد.

هدوپ مزایای زیادی دارد، اما موارد زیر به مدیریت و تجزیه تحلیل کلان داده مرتبط است:

- قابل توسعه بودن: هدوپ اجازه توسعه یا کوچک کردن زیرساخت سخت افزاری را بدون تغییر در فرمت داده‌ها می-دهد. سیستم بطور خودکار داده‌ها را باز توزیع می‌کند و کارهای محاسباتی با تغییرات سخت افزاری قابل تطابق است.
- بهره‌وری هزینه بالا: هدوپ محاسبات موازی مقیاس بالا را به سرورهای تجاری اعمال می‌کند، که هزینه به ازای هر TB مورد نیاز برای ذخیره سازی را خیلی کاهش می‌دهد. محاسبات مقیاس-بالا همچنین اجازه می‌دهد که با حجم داده رو به رشد متوالی تطابق پیدا کند.
- انعطاف پذیری قوی: هدوپ می‌تواند انواع زیادی از داده‌ها را از منابع مختلف مدیریت کند. به علاوه داده‌ها از منابع زیادی می‌توانند در هدوپ برای تجزیه تحلیل بیشتر سنتز شوند. بنابراین می‌تواند از عهده انواع چالش‌های کلان داده برآید.
- تحمل خطای بالا: گم شدن داده‌ها و محاسبات غلط ممکن است در طول تجزیه تحلیل کلان داده اتفاق بیافتد، اما هدوپ می‌تواند داده‌ها را بازیابی کند و خطاهای محاسباتی اتفاق افتاده با شکست‌های گره یا تصادم شبکه را تصحیح کند.

---

[1] Hadoop
[2] Google File System (GFS)
[3] Cluster
[4] Node
[5] Mapreduce
[6] Hadoop Distributed File System

۲-۴-۲ رابطه هدوپ و کلان داده

در حال حاضر هدوپ بطور گسترده در کاربردهای کلان داده و در صنعت استفاده می‌شود، مثل فیلتر کردن اسپم، جستجوی شبکه، تجزیه تحلیل Click stream، و توصیه نامه‌های اجتماعی. به علاوه تحقیقات دانشگاهی قابل توجهی هم‌اکنون براساس هدوپ می‌باشد. بعضی موارد قابل توجه در زیر ارائه می‌شود. همانطور که گفتیم در جون ۲۰۱۲، یاهو هدوپ را روی ۴۲۰۰۰ سرور در چهار دیتاسنتر اجرا کرد تا محصولات و سرویس‌هایش را پشتیبانی کند، مثل جستجو و فیلتر کردن اسپم. در حال حاضر بزرگترین خوشه هدوپ ۴۰۰۰ گره دارد اما تعداد گره‌ها با نسخه ۲,۰ می‌توانند به ۱۰۰۰۰ افزایش یابند. در همان ماه فیس بوک اعلام کرد که خوشه هدوپ آنها می‌تواند 100PB داده را پردازش کند، که می‌تواند در هر روز 0.5PB رشد کند. بعضی آژانس‌های مشهور هدوپ را استفاده می‌کنند. در مجموع شرکت های زیادی نسخه تجاری هدوپ را تهیه و استفاده کرده‌اند. مثل Cloudera, IBM, MapR, EMC, Oracle.

در کنار سیستم‌های تجاری مدرن، سنسورها بطور گسترده استفاده می‌شوند تا اطلاعات را برای مانیتور کردن محیط و پیش-بینی خطا و غیره توسعه دهند.

فصل سوم تولید و اکتساب کلان داده

چکیده: ما چند فن‌آوری کلیدی مرتبط با کلان داده ر ا معرفی کردیم. مثل رایانش ابری، IOT، دیتاسنتر، هدوپ. در ادامه روی زنجیره ارزش کلان داده تمرکز خواهیم کرد، که می‌تواند به چهار فاز عمده تقسیم شود: **تولید داده[1]، اکتساب داده[2]، ذخیره سازی داده[3]، و تجزیه تحلیل داده[4]**. اگر داده‌ها را مواد خام در نظر بگیریم، تولید و اکتساب پردازه اکتشاف است، ذخیره داده‌ها پردازه ذخیره‌سازی است، و تجزیه تحلیل داده یک پردازه تولید است که مواد خام را برای تولید ارزش جدید بهره‌برداری می‌کند.

۳-۱ تولید کلان داده

تولید داده اولین مرحله کلان داده است. بویژه دارای مقیاس بزرگ است و مجموعه داده‌های پیچیده از منابع داده توزیع شده و جغرافیایی گسترده تولید می‌شوند. منابع داده شامل سنسورها، ویدیوها، click streamها، و/یا همه انواع دیگر منابع داده قابل دسترس می‌باشد. در حال حاضر، منابع اصلی کلان داده، اطلاعات تجاری و عملیات در شرکت‌ها، اطلاعات حسگرها و لوجستیک در IOT، اطلاعات تبادل انسانی و اطلاعات مکانی در دنیای اینترنت و داده‌های تولید شده در تحقیقات علمی و ... می‌باشد. اطلاعات از ظرفیت‌های معماری‌های IOT و زیرساخت‌های شرکت‌های موجود پیش افتاده‌اند.

۳-۱-۱ داده‌های کسب و کار[5]

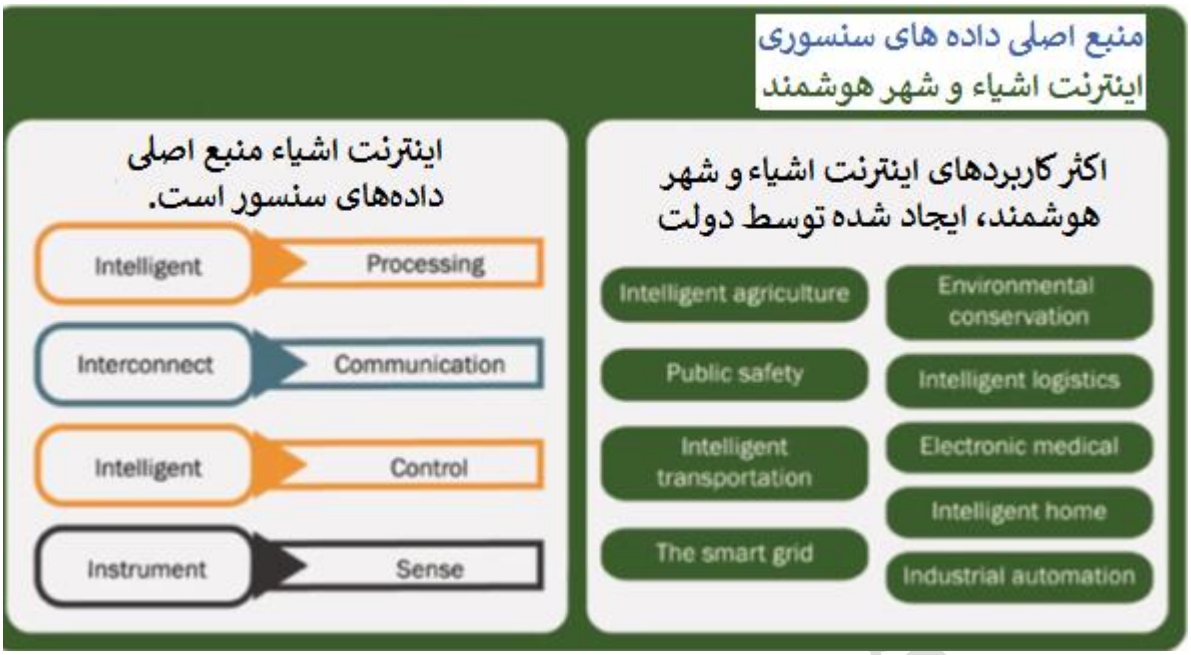
در سال ۲۰۱۳، IBM گزارشی منتشر کرد که نشان می‌دهد داده‌های داخلی شرکت‌های سرمایه گذار منابع اصلی کلان داده هستند. داده‌های داخلی شرکت‌های سرمایه گذاری عمدتاً شامل داده‌های تجاری برخط و تجزیه تحلیل برخط داده‌ها هستند، که اکثر آنها داده‌های ایستا هستند و توسط RDBMS به روش ساختار یافته مدیریت می‌شوند. به علاوه داده‌های محصولات، داده‌های فروش، داده‌های مالی و داده‌های تولیدی جدید، و ... در داده‌های داخلی شرکت‌ها مشارکت دارند.

در دهه‌های اخیر، داده‌های دیجیتال و IT پیشرفت زیادی داشته‌اند. تخمین زده می‌شود حجم داده‌های کسب و کار هر ۱,۲ سال دو برابر خواهد شد. افزایش متوالی حجم داده‌ها نیازمند تجزیه تحلیل آنی[6] است. به عنوان مثال آمازون میلیون‌ها عملیات ترمینال و بیشتر از ۵۰۰۰۰۰ پرس و جو از فروشندگان را در هر روز پردازش می‌کند. Walmart یک میلیون تجارت مشتری را در ساعت پردازش می‌کند و چنین داده‌های تجاری در یک پایگاه داده با ظرفیت بیش از2.5 PB وارد می‌شود. Akamai ۷۵ میلیون رخداد را برای آگهی‌های خود در روز تجزیه تحلیل می‌کند.

۳-۱-۲ داده‌های IOT: همانطور که گفتیم IOT یکی از منابع مهم کلان داده است. در شهرهای هوشمند که براساس IOT ساخته شده‌اند،کلان داده ممکن است از صنعت، کشاورزی، حمل و نقل و ترافیک، مراقبت‌های پزشکی، دپارتمان‌های عمومی، و محافظت از خانه‌ها بیایند. این داده‌ها در شکل ۳-۱ نشان داده شده‌اند.

---

[1] Data Generation
[2] Data Acquisition
[3] Data Storage
[4] Data Analysis
[5] Enterprise Data
[6] Real-Time



شکل ۳-۱ نمایش منبع اولیه داده‌های سنسوری

براساس پردازه‌های اکتساب داده‌ها و انتقال در IOT، معماری شبکه آن ممکن است به سه لایه تقسیم شود: لایه حسگر، لایه شبکه و لایه کاربرد. لایه حسگر مسئول اکتساب داده‌ها می‌شود و عمدتاً شامل شبکه‌های سنسور می‌باشد. لایه شبکه مسئول تبادل و پردازش اطلاعات است که انتقال و پردازش اطلاعات را برعهده دارد که انتقال ممکن است وابسته به شبکه‌های سنسور باشد و انتقال دور ممکن است وابسته به اینترنت باشد. در انتها لایه کاربرد، کاربردهای خاص IOT را پشتیبانی می-کند.

براساس ویژگی‌های IOT داده‌های تولید شده از IOT ویژگی‌های زیر را دارند:

- داده‌های در مقیاس بزرگ: در IOT تجهیزات گردآوری انبوه داده‌ها بطور توزیع شده استفاده می‌شوند که ممکن است داده‌های عددی ساده (مثل مکان) یا داده‌های چندرسانه‌ای پیچیده (مثل نظارت ویدیو) باشد. به منظور برآورده کردن نیازهای تجزیه تحلیل (پردازش) هم باید داده‌هایی که اخیراً بدست آمده‌اند و هم داده‌های تاریخچه در یک محدوده زمانی خاص، ذخیره شوند. بنابراین داده‌های تولید شده با IOT با مقیاس بزرگ توصیف می‌شوند.
- عدم تجانس: چون دستگاه‌های بدست آوردن داده‌ها متفاوت هستند داده‌ها نامتجانس خواهند بود.
- همبستگی فضا و زمان قوی[1]: در IOT، هر دستگاه اکتساب در مکان جغرافیایی خاصی قرار داده شده است و هر قطعه داده یک مهر زمانی[2] دارد. همبستگی زمان و فضا در ویژگی‌های داده‌های IOT مهم هستند. در طول تجزیه تحلیل داده‌ها و پردازش آن، فضا و زمان ابعاد مهمی برای تجزیه تحلیل آماری هستند.
- حساب‌های داده موثر برای بخش کوچکی از کلان داده[3]: در طول پردازه کسب و انتقال داده‌ها در IOT ممکن است میزان زیادی نویز بدست آید. در دیتاست‌های بدست آمده توسط دستگاه‌های اکتساب، فقط بخش کوچکی از داده-

[1] Strong Time and Space Correlation
[2] Time stamp
[3] Effective Data Accounts for Only a Small Portion of the Big Data

های ناهنجار با ارزش هستند. به عنوان مثال، در ویدئوهای ترافیکی بدست آمده فریم‌های ویدئوی کمی که از تنظیمات ترافیکی و تصادفات ترافیکی تخطی می‌کنند مهم‌تر از داده‌های بدست آمده از کل جریان ترافیکی عادی می‌باشند.

۳-۱-۳ داده‌های اینترنت

داده‌های اینترنت شامل ورودی‌های جستجو، لیست‌های گروه‌های اینترنتی، رکوردهای گفتگو، پیام‌های وبلاگ‌ها، که ویژگی مشابه دارند، ارزش بالا و چگالی کم دارند. چنین داده‌های اینترنتی، ممکن است بطور انفرادی کم ارزش باشند، اما با استخراج داده‌های کلان جمع‌آوری شده، اطلاعات مفید مثل عادات و سرگرمی‌های کاربران می‌توانند قابل شناسایی باشند و حتی امکان‌پذیر است تا رفتار و احساسات کاربران را پیش‌بینی کرد.

۳-۱-۴ داده‌های حیاتی

چون یک سری فن‌آوری‌های اندازه‌گیری حیاتی با کارایی بالا بطور خلاقانه‌ای در ابتدای قرن بیست و یک توسعه داده شده‌اند، مرز تحقیقاتی در رشته پزشکی-حیاتی به حوزه کلان داده وارد می‌شود. با مدل‌های تجزیه تحلیلی دقیق و مفید و سیستم‌های تئوری برای کاربردهای پزشکی، مکانیزم پشت پرده بیولوژیکی پیچیده ممکن است آشکار شود. در سال ۲۰۱۵ میانگین حجم داده‌های هر بیمارستان به 665TB رسید.

۳-۲ اکتساب و بدست آوردن کلان داده

فاز دوم سیستم کلان داده، اکتساب کلان داده شامل جمع‌آوری داده[1]، انتقال داده و پیش‌پردازش داده می‌باشد. در طول اکتساب داده وقتی‌که داده‌های خام جمع‌آوری شدند، یک مکانیزم انتقال موثر باید برای ارسال آن به سیستم مدیریت ذخیره مناسب استفاده شود تا کاربردهای تجزیه تحلیلی مناسب پشتیبانی شود. مجموعه داده‌های جمع‌آوری شده ممکن است شامل داده‌های غیر مفید یا افزونه باشند، که بطور غیرضروری فضای ذخیره‌سازی را افزایش می‌دهند و بر تجزیه تحلیل داده نهایی تأثیر دارند. به عنوان مثال، در مجموعه داده‌های جمع‌آوری شده توسط سنسورها افزونگی زیاد خیلی رایج می‌باشد. روش‌های فشرده‌سازی داده می‌تواند برای کاهش افزونگی اعمال شود. بنابراین عملیات پیش‌پردازش داده برای اطمینان از ذخیره داده موثر و اکتشاف آن ضروری هستند.

۳-۲-۱ جمع‌آوری داده

جمع‌آوری داده برای بهره‌وری از روش‌های جمع‌آوری داده‌ها می‌باشد تا داده‌های خام را از محیط‌های تولید داده ویژه بدست آورد. روش‌های جمع‌آوری داده برای داده‌های رایج در زیر نشان داده شده است:

- Log Files: این فایل‌ها شامل رکوردهایی هستند که بطور خودکار توسط سیستم منبع داده تولید می‌شوند تا فعالیت‌های مدنظر در فرمت‌های فایل را برای تجزیه تحلیل بعدی ضبط کنند. به عنوان مثال سرورهای وب در فایل لاگ تعداد کلیک‌ها، سرعت کلیک‌ها، ملاقات‌ها و سایر رکوردهای مربوط به کاربران وب را ذخیره می‌کنند.

[1] Data collection, Data transmission, Data preprocessing

- سنسورها: داده‌های سنسوری ممکن است به عنوان موج صدا، صدا، اتومبیل، شیمی، جریان برق، هوا، فشار، حرارت و غیره کلاس‌بندی شوند. اطلاعات حس شده از طریق سیم یا شبکه بی‌سیم منتقل می‌شوند. برای کاربردهایی که ممکن است به سادگی مدیریت و استفاده شوند، مثل سیستم نظارت ویدیو، شبکه سنسور سیمی یک راه حل رایج برای بدست‌آوردن اطلاعات مربوطه می‌باشد. گاهی مکان دقیق یک پدیده خاص ناشناخته است و گاهی محیط مانیتور شده انرژی یا زیرساخت ارتباطی را ندارد. در این حالت ممکن است ارتباط بی‌سیم برای انتقال داده از طریق گره‌های سنسور تحت انرژی محدود و ظرفیت ارتباطی کم، استفاده شود.
- روش‌های بدست آوردن داده‌های شبکه: در حال حاضر، بدست آوردن داده‌های شبکه از طریق خیزش در وب، سیستم تقسیم کلمه، سیستم کار و سیستم ایندکس و غیره انجام می‌شود.

۳-۲-۲ انتقال داده

بعد از تکمیل جمع‌آوری داده‌های خام، داده‌ها به یک زیرساخت ذخیره‌سازی داده برای پردازش و تجزیه تحلیل منتقل خواهند شد. همانطور که در بخش ۲-۳ بحث شد، داده‌های کلان در یک دیتاسنتر ذخیره می‌شوند، داده‌ها باید طوری لایه بندی شوند که بهره‌وری محاسباتی بهبود یابد یا نگهداری سخت افزار آسان باشد. به عبارت دیگر، انتقال داده داخلی ممکن است در دیتاسنترها اتفاق بیافتد. بنابراین، انتقال داده شامل دو فاز است: انتقال DCN بینابینی[1] و انتقال DCN داخلی[2].

انتقال داده DCN بینابینی از منبع داده به دیتاسنتر است که معمولاً با زیرساخت شبکه فیزیکی موجود بدست می‌آید. به خاطر رشد سریع تقاضاهای ترافیکی، زیرساخت شبکه فیزیکی در اکثر نواحی در دنیا با سیستم‌های انتقال فیبر نوری با هزینه خوب، سرعت و حجم بالا می‌باشد. در ۲۰ سال اخیر، تجهیزات مدیریت پیشرفته و فن‌آوری‌ها توسعه یافته‌اند، مثل معماری شبکه مالتی‌پلکس تقسیم طول موج براساس IP(WDM)، تا با کنترل و مدیریت شبکه‌های فیبر نوری تطابق داشته باشد.

انتقال داده داخلی، جریان‌های ارتباطی داده داخل دیتاسنترها می‌باشد. انتقال DCN داخلی وابسته به مکانیزم ارتباط داخل دیتاسنتر می‌باشد. یک دیتاسنتر شامل چندین رک سرور مجتمع شده و به هم متصل با شبکه اتصالی خودش می‌باشد.

۳-۲-۳ پیش‌پردازش داده

به خاطر تنوع زیاد منابع داده، مجموعه داده‌های جمع‌آوری شده نسبت به نویز، افزونگی، سازگاری و ... متغیر هستند و بدون شک ذخیره داده‌های بدون معنی اسراف می‌باشد. به علاوه بعضی روش‌های تجزیه تحلیل نیاز به کیفیت داده و دقت دارند. بنابراین داده‌ها باید تحت شرایط زیادی پیش‌پردازش شوند تا داده‌ها از منابع مختلف جمع‌آوری شوند تا تجزیه تحلیل موثر واقع شود. پیش‌پردازش داده‌ها نه تنها قیمت ذخیره‌سازی را کاهش می‌دهد بلکه دقت تجزیه تحلیل را بهبود می‌دهد. بعضی روش‌های پیش‌پردازش داده‌های رابطه‌ای در زیر بحث می‌شوند.

۱-۳-۲-۳: تجمع[3]

[1] Inter-DCN transmission
[2] Intra-DCN transmission
[3] Integration

تجمع داده ستون فقرات انفورماتیک‌های تجاری مدرن است که شامل ترکیب داده‌ها از منابع مختلف می‌باشد و برای کاربران نمای یکنواختی از داده‌ها ایجاد می‌کند. این یک رشته تحقیقاتی بالغ برای داده‌های سنتی می‌باشد. بطور تاریخی دو روش عمده تعیین می‌شود: ورهاوس داده و اتحادیه[1]. ورهاوس داده شامل یک پردازه به نام ETL[2] (استخراج، تغییر شکل دادن، و بارگذاری) می‌شود. استخراج شامل اتصال منابع سیستم، انتخاب، انباشت، تجزیه تحلیل و پردازش داده‌های لازم می‌شود. تغییرشکل دادن اجرای یک سری از قوانین می‌باشد تا داده‌های استخراج شده را به فرمت‌های استاندارد تبدیل کند. بارگذاری به معنی وارد کردن داده‌های استخراج شده و تغییر یافته به زیرساخت‌های هدف می‌باشد. بارگذاری پیچیده‌ترین روال از سه روال می‌باشد که شامل عملیاتی مثل تبادل، کپی، پاک‌سازی، استاندارد سازی، نمایش و سازمان‌دهی داده‌ها می‌باشد. یک پایگاه داده مجازی می‌تواند ساخته شود تا داده‌ها را از منابع مختلف جمع‌آوری و پرس و جو کند، اما چنین پایگاه داده‌ای نمی‌تواند شامل داده‌ها شود و شامل اطلاعات یا استانداردهای مرتبط با داده‌های واقعی است که در مکان خودشان قرار دارند. چنین روش‌های خواندن-ذخیره نمی‌تواند نیازهای کارایی بالای جریان داده یا برنامه‌های جستجو و کاربردها را برآورده کند. در مقایسه با پرس و جوها، داده در چنین روش‌هایی پویاتر است و باید در حال انتقال داده پردازش شود. به طور کلی روش-های تجمع داده، با موتورهای پردازش جریان و موتورهای جستجو همراه می‌شوند.

۲-۳-۲-۳ پاک سازی[3]

پاک سازی داده پردازه تشخیص داده‌های نادرست، غیرکامل، غیر مستدل و سپس اصلاح یا حذف چنین داده‌هایی به منظور کیفیت داده‌ها می‌باشد، در مجموع پاک‌سازی داده‌ها شامل ۵ مرحله مکمل هم است: ۱- تعریف و مشخص کردن انواع خطاها ۲- جستجو و شناسایی خطاها ۳-تصحیح خطاها ۴-مستندسازی مثال‌های خطا و انواع خطا ۵-اصلاح روال‌های ورودی داده تا خطاهای آینده کم شوند. در طول پاک سازی داده‌ها، فرمت‌های داده، مکمل سازی، رابطه سازی، و محدودیت‌ها باید برآورده شوند. پاک سازی داده‌ها اهمیت حیاتی دارد تا داده‌ها را سازگار نگه داریم، که در بسیاری از رشته‌ها مثل بانک داری، بیمه، صنعت خرده فروشی، تله ارتباطات، و کنترل ترافیک بطور گسترده، اعمال می‌شود.

در تجارت الکترونیک[4] اکثر داده‌ها بطور الکترونیکی جمع‌آوری می‌شوند که ممکن است مشکلات کیفیت داده جدی داشته باشند. مسائل کیفیت داده کلاسیک از عیب‌های نرم‌افزاری، خطاهای سفارشی شده، یا پیکربندی بد سیستم بدست می‌آیند.

۲-۴-۳ حذف افزونگی[5]

افزونگی داده اشاره به تکرار یا زیادتی داده‌ها دارد که معمولاً در بسیاری از مجموعه داده‌ها اتفاق می‌افتد. افزونگی می‌توانند قیمت تبادل داده‌های غیرضروری را افزایش دهد و باعث عیب‌هایی در سیستم‌های ذخیره سازی شود. مثلاً، اسراف ذخیره سازی، منجر به عدم سازگاری داده‌ها، کاهش قابلیت اطمینان داده‌ها و خرابی داده‌ها می‌شود. بنابراین روش‌های کاهش افزونگی مختلفی پیشنهاد شده‌اند، مثل تشخیص افزونگی، فیلتر کردن داده‌ها، و فشرده سازی داده‌ها. چنین روش‌هایی ممکن است به مجموعه داده‌های مختلف یا محیط‌های کاربردی اعمال شوند. به هر حال کاهش افزونگی ممکن است منجر به اثرات

---

[1] Federation
[2] Extract, Transform and Load
[3] Cleaning
[4] e-commerce
[5] Redundancy Elimination

منفی خاصی شود. به عنوان مثال، فشرده‌سازی و بازسازی منجر به بار محاسباتی اضافی می‌شود. بنابراین، مزایای کاهش افزونگی و هزینه باید با دقت موازنه شوند.

داده‌های جمع‌آوری شده از رشته‌های مختلف بطور افزونه‌ای در قالب‌های ویدیو یا تصویر ظاهر می‌شوند. مشهود است که تصاویر و ویدئوها شامل افزونگی‌های قابل توجهی هستند شامل افزونگی زمانی، افزونگی فضایی، افزونگی احتمالاتی، و افزونگی حس. فشرده سازی بطور عمده برای کاهش افزونگی در داده‌های ویدیو استفاده می‌شود که در بسیاری از استانداردهای کد ویدیو انجام می‌شود (MPEG-2, MPEG-4, H264/AVG, H.263). در یک مقاله مولفان مسائل فشرده سازی ویدئو را در سیستم‌های نظارت ویدئو با یک شبکه سنسور ویدیو مطرح کرده‌اند. مولفان یک روش براساس MPEG-4 جدید ارائه داده‌اند که مرتبط با افزونگی مرتبط با زمینه و پس زمینه یک صحنه است. پیچیدگی کد و نرخ فشرده سازی پایین روش پیشنهاد شده با نتایج ارزیابی نشان داده شده است.

فصل چهارم حافظه ذخیره سازی کلان داده

چکیده: در این فصل روی حافظه ذخیره‌سازی کلان داده تمرکز می‌کنیم. ما مطالب مهم شامل سیستم‌های ذخیره‌سازی[1] انبوه، سیستم‌های ذخیره‌سازی توزیع شده، و مکانیزم‌های ذخیره سازی کلان داده را مرور خواهیم کرد. به یک عبارت، زیرساخت ذخیره سازی مورد نیاز برای تهیه سرویس مخزن اطلاعات با فضای ذخیره سازی قابل اعتماد و به عبارت دیگر، باید دسترسی قدرتمند به رابط دسترسی برای پرس و جو و تجزیه تحلیل مقدار زیادی از داده‌ها تهیه کند. چنین زیرساخت ذخیره سازی معمولاً از زیرساخت سخت افزاری و مکانیزم‌های ذخیره سازی تشکیل شده است.

۴-۱ سیستم ذخیره‌سازی برای انبوه داده‌ها

ذخیره سازی داده‌ها اشاره به ذخیره‌سازی و مدیریت مجموعه داده‌های مقیاس بالا دارد در حالیکه دارای قابلیت اطمینان و دسترس‌پذیری است. زیرساخت سخت افزاری شامل تکنولوژی منابع ارتباط اطلاعات مشترک انبوه (ICT) است که برای بازخورد نیازهای فوری کارها بهره وری می‌شود، و چنین منابع ICT به روشی کشسان[2] سازماندهی می‌شوند. زیرساخت سخت افزاری باید کشسانی و قابل پیکربندی مجدد پویا را داشته باشد تا با محیط‌های کاربردی مختلف تطابق داشته باشد. روش‌های ذخیره‌سازی داده در بالای زیرساخت سخت افزاری استفاده می‌شوند تا مجموعه داده‌های در مقیاس بزرگ را نگه داری کنند. سیستم‌های ذخیره سازی باید با رابط‌های زیادی تجهیز شوند تا پرس‌و جوها سریع باشند.

کلان داده با رشد انفجاری داده‌ها مواجه است. داده‌ها با توجه به رشد پرشتاب نیازهای ضروری روی ذخیره سازی و مدیریت دارند. بطور رایج، تجهیزات ذخیره داده‌ها فقط تجهیزات جانبی سرورها هستند و اکثر آنها داده‌های ساختار یافته RDBMS را ذخیره، مدیریت، جستجو و تجزیه تحلیل می‌کنند. با توجه به رشد سریع کلان داده به GB, TB,PB تجهیزات ذخیره سازی سنتی برای مدیریت آنها کافی نیستند. اهمیت تجهیزات ذخیره سازی روز افزون است و هزینه ذخیره سازی مهمترین چالش بسیاری از شرکت‌های اینترنتی شده است. بنابراین تحقیق رو مخزن داده‌ها ضروری است.

تعداد زیادی سیستم ذخیره‌سازی برای برآورده کردن تقاضاهای کلان داده پدیدار شده‌اند. فن‌آوری‌های مخزن موجود می‌توانند به DAS[3] (مخزن الحاقی مستقیم) و مخزن شبکه کلاس‌بندی شوند و مخزن شبکه به NAS[4], SAN[5] تقسیم می‌شود.

در DAS، درایوهای دیسک مستقیماً به سرورها متصل می‌شوند. مخزن تجهیزات جانبی می‌باشد. DAS به تعداد کمی محیط سرور اعمال می‌شود، در حالیکه وقتی ظرفیت ذخیره‌سازی افزایش می‌یابد، بازدهی توان مخزن کاملاً کم خواهد بود و قابلیت بروزرسانی و توسعه بطور زیادی محدود می‌شود. وقتی سرور بد عمل می‌کند، داده‌ها قابل اکتساب نیستند و منابع ذخیره سازی داده‌ها نمی‌توانند به اشتراک گذاشته شوند. DAS عموماً در کامپیوترهای شخصی استفاده می‌شود و سرورهای با اندازه کم، که چنین کاربردهایی را پشتیبانی می‌کند و نیازمند ظرفیت ذخیره‌سازی کم است و مستقیماً مخزن مشترک چندکامپیوتر را پشتیبانی نمی‌کند. درایوهای TAP و RAID تجهیزات کلاسیک DAS هستند.

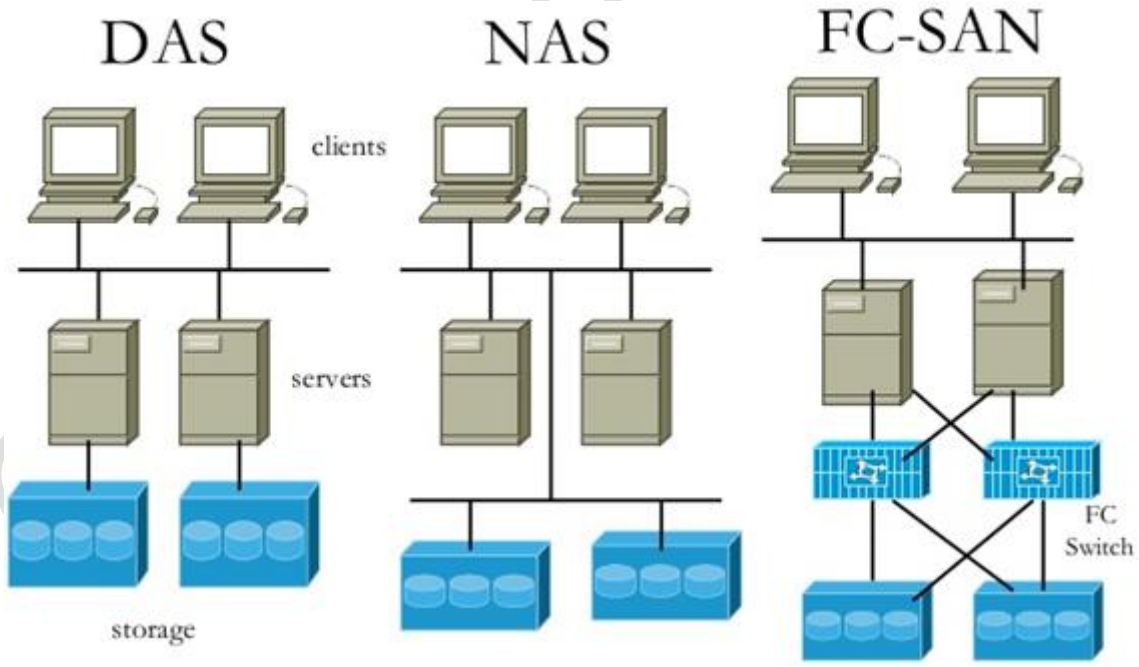
---

[1] Storage
[2] Elasticity
[3] Direct Attached Storage
[4] Network Attached Storage
[5] Storage Area Network

مخزن شبکه برای بهره‌وری از شبکه اصلی می‌باشند یا برای اینکه کاربران بتوانند به منابع مشترک سیستم اطلاعاتی بطور مشترک و دسترسی یکسان وصل شود طراحی شده است. تجهیزات مخزن شبکه شامل تجهیزات تبادل داده خاص، آرایه دیسک، کتابخانه tap، و سایر رسانه‌های ذخیره سازی و نرم افزار ذخیره‌سازی می‌باشند. مخزن شبکه با مخزن داده انبوه، اشتراک داده محدود شده، بهره‌وری کامل از داده‌کاوی و اطلاعات، قابلیت اعتماد داده‌ها، پشتیبانی داده‌ها و امنیت، و مدیریت داده یکنواخت توصیف می‌شود.

NAS تجهیزات ذخیره جانبی یک شبکه می‌باشد. NAS مستقیماً به شبکه از طریق هاب یا سوئیچ وصل می‌شود، و داده‌ها را به شکل فایل‌ها می‌فرستد. NAS دو ویژگی کلیدی دارد: اولاً در اتصال فیزیکی، مستقیما به تجهیزات ذخیره سازی شبکه وصل می‌شود. ثانیاً NAS اتصال مستقیم دیسک‌ها را کم می‌کند و تاخیر R/W را کاهش می‌دهد. به هر حال ساختار NAS جزو تجهیزات ذخیره‌سازی سرور قدیمی می‌باشد.

SAN روی ذخیره‌سازی داده، یک توپولوژی شبکه منعطف و اتصالات فیبر نوری سرعت-بالا تمرکز می‌کند، آن سوئیچینگ داده چند مسیره از طریق هر گره داخلی را مجاز می‌کند. مدیریت ذخیره داده در یک شبکه محلی ذخیره نسبتاً مستقل قرار دارد تا به حداکثر درجه اشتراک داده و مدیریت داده برسد. از سازمان یک سیستم ذخیره داده، DAS، SAN، NAS می-توانند به سه بخش تقسیم شوند: (الف) آرایه دیسک: آن پایه یک سیستم ذخیره سازی است و ضمانت برای ذخیره داده دارد. (ب) اتصال و شبکه زیرسیستم‌ها، که اتصال بین یک یا چند آرایه دیسک و سرور را فراهم می‌کند. (ج) نرم افزار مدیریت دیسک که اشتراک داده، بازایابی خرابی، و سایر کارهای مدیریت حافظه چند سرور را مدیریت می‌کند.



تفاوت بین سیستم‌های ذخیره سازی. SAN سریع‌تر و گرانتر است.

۴-۲ سیستم ذخیره سازی توزیع شده

اولین چالش راجع کلان داده این است که چگونه یک سیستم ذخیره توزیع شده با مقیاس بالا برای نگهداری استراتژیک داده-ها و پردازش و تجزیه تحلیل موثر داده‌ها توسعه دهیم. برای استفاده از یک سیستم توزیع شده برای ذخیره انبوه داده‌ها، فاکتورهای زیر را باید مد نظر قرار داشت:

- سازگاری[1]: یک سیستم توزیع شده نیازمند این است که چندین سرور بطور همکار داده‌ها را ذخیره کنند، هرچه تعداد سرورها بیشتر باشند، احتمال شکست سرور بیشتر می‌شود. معمولاً داده‌ها به چندین قطعه تقسیم می‌شوند تا در سرورهای مختلف ذخیره شوند تا قابلیت دسترسی را در صورت خطای سرور تضمین کنند. به هر حال شکست-های سرور و ذخیره موازی ممکن است باعث ناسازگاری در بین کپی‌های مختلف داده‌های مشابه شود. سازگاری یعنی اینکه اطمینان بدهیم کپی‌های چندگانه از داده‌های مشابه، مشابه باشند.
- قابل دسترس بودن[2]: یک سیستم ذخیره توزیع شده در چندین مجموعه سرور عمل می‌کند. هر چه سرورهای بیشتری استفاده شوند، شکست‌های سرور اجتناب ناپذیر خواهد بود. انتظار داریم کل سیستم بطور جدی در مقابل خواندن/نوشتن ترمینال‌های مشتریان متاثر نشود. یا به عبارتی دسترس پذیری به داده‌ها در صورت وجود خرابی امکان پذیر باشد.
- تحمل پذیری پارتیشن[3]: چندین سرور در یک سیستم ذخیره سازی توزیع شده توسط یک شبکه متصل می‌شوند. شبکه می‌تواند شکست‌های لینک/گره یا تراکم‌های موقت داشته باشد. سیستم توزیع شده باید سطح خاصی از تحمل‌پذیری در مقابل شکست‌های شبکه داشته باشد. مطلوب است که ذخیره توزیع شده هنوز کار کند اگر شبکه به چندین پارتیشن شکسته شود.

Eric Brewer در سال ۲۰۰۰ نظریه CAP را مطرح کرد که مشخص می‌کند یک سیستم توزیع شده بطور همزمان نمی-تواند نیازهای سازگاری، قابل دسترس بودن، و تحمل پارتیشن را داشته باشد. در سال ۲۰۰۲، اثبات شد که تئوری CAP درست می‌باشد. چون سازگاری C، در دسترس بودن A، و تحمل پذیری پارتیشن P بطور همزمان بدست نمی‌آید یا سیستم CA داریم با نادیده گرفتن تحمل پارتیشن، یا یک سیستم CP داریم با نادیده گرفتن قابل دسترس بودن، و یا یک سیستم AP با نادیده گرفتن سازگاری. براساس هدف طراحی یکی از اینها انتخاب می‌شود. این سه سیستم را در ادامه بحث می‌کنیم:

- سیستم‌های CA: تحمل‌پذیری پارتیشن را ندارند یعنی نمی‌توانند خطاهای شبکه را مدیریت کنند. بنابراین سیستم‌های CA معمولاً به عنوان سیستم‌های ذخیره سازی با یک سرور به نظر می‌رسند، مثل پایگاه داده‌های رابطه‌ای در مقیاس کوچک سنتی. چنین سیستمی یک کپی از داده‌ها را دارد بنابراین سازگاری به سادگی بدست می‌آید. قابل دسترس بودن با طراحی خوب پایگاه‌های داده رابطه‌ای بدست می‌آید. به هر حال چون سیستم‌های CA نمی‌توانند شکست‌های شبکه را مدیریت کنند، آنها نمی‌توانند برای استفاده جندین سرور توسعه داده شوند. به همین خاطر اکثر سیستم‌های ذخیره مقیاس-بزرگ، سیستم‌های CP یا سیستم‌های AP هستند.
- سیستم‌های CP: در مقایسه با سیستم‌های CA، تحمل‌پذیری پارتیشن را تضمین می‌کنند. بنابراین سیستم CP می‌تواند توسعه داده شود یا یک سیستم توزیع شده شود، سیستم‌های CP معمولاً چندین کپی از داده‌های مشابه را نگه می‌دارند تا سطحی از تحمل‌پذیری خطا را تضمین کنند. سیستم‌های CP همچنین سازگاری را تضمین می-

[1] Consistency
[2] Availability
[3] Partition Tolerance

کنند، مثلاً چندین کپی از داده‌های مشابه تضمین می‌شود که کاملاً مشابه باشند. به هر حال CP نمی‌تواند قابل دسترس بودن را تضمین دهد به خاطر هزینه بالای تضمین سازگاری. بنابراین سیستم‌های CP مناسب برای سناریوهای با بار متوسط هستند اما نیازهای شدید به درستی داده دارند (مثل تجارت داده). BigTable و Hbase دو سیستم CP رایج هستند. BigTable شناخته شده است زیرا برای مدیریت داده‌های زمینه موتور جستجوی گوگل موفق بود. چون اکثر داده‌های گوگل ساختار یافته هستند، BigTable عمدتاً داده‌ها را با جدول ذخیره می‌کند. وقتی مقدار زیادی از اطلاعات در یک جدول قرار داده می‌شوند، اندازه جدول رشد می‌کند. چنین اطلاعاتی باید پارتیشن شوند و بصورت مجزا ذخیره شوند. جدول معمولاً خیلی خلوت می‌باشد بنابراین BigTable ستون‌ها را به چندین خانواده ستون مختلف تقسیم می‌کند که هر خانواده ستون داده‌های از نوع مشابهی را ذخیره می‌کند. بنابراین داده‌ها مشابه با هم ذخیره می‌شوند و اطلاعات با نوع مشابه به روش مشابهی پردازش می‌شوند که آن را برای کاربران سیستم آسان می‌کند. در یک خانواده مشابه، ستون‌های جدید بطور دلخواه و اتفاقی درج می‌شوند بنابراین محدودیت توسعه BigTable را کاهش می‌دهند. BigTable مشابه GFS طراحی شده است که یک رئیس و چند سرورر Tablet بصورت هم‌بندی ستاره در سیستم هستند. ساختار ستاره باعث ایجاد SPOF[1] می‌باشد. بار سرور رئیس به منظور حداقل کردن خطاهای رئیس باید کاهش داده شود. در BigTable، انتقال داده و آدرس‌دهی داده در سیستم رئیس انجام نمی‌شود. بنابراین بار رئیس زیاد نیست. به منظور حل مشکل SPOF، BigTable یک مکانیزمی برای انتخاب رئیس دارد.

- سیستم AP: نیز امکان تحمل پارتیشن را تضمین می‌دهند. اما سیستم‌های AP از سیستم‌های CP متفاوتند زیرا سیستم‌های AP، قابل دسترس بودن را نیز اطمینان می‌دهند. به هر حال سیستم‌های AP فقط سازگاری مشروط را تضمین می‌کنند نه سازگاری قوی دو سیستم قبل را. بنابراین سیستم‌های AP به سناریوهایی که درخواست‌های متناوب وجود دارد اما نه با نیاز به دقت بالا مناسب هستند. به عنوان مثال در سیستم‌های SNS[2] (سرویس‌های شبکه اجتماعی)، ملاقات‌های همروند زیادی برای داده‌ها وجود دارد اما مقدار خاصی از خطاهای داده قابل تحمل است. بنابراین چون سیستم‌های AP سازگاری مشروط را تضمین می‌کنند داده‌های دقیق می‌توانند بعد از مقداری تأخیر بدست آیند. Dynamo, Cassandra دو سیستم AP رایج هستند. Cassandra با قابلیت توسعه صدا، برای ذخیره داده‌های متنی حجیم در شرکت‌های SNS برخط تجاری مثل فیس‌بوک و توییتر استفاده می‌شود.

۳-۴ روش‌های ذخیره سازی کلان داده

تحقیقات قابل توجهی روی کلان داده، روش‌های ذخیره‌سازی کلان داده را توسعه داده است. روش‌های ذخیره موجود برای کلان داده در سه سطح پایین به بالا کلاس‌بندی شده‌اند. الف: سیستم‌های فایل ب- پایگاه‌های داده ج-مدل‌های برنامه نویسی.

سیستم‌های فایل پایه کاربردهای سطح بالا می‌باشد. GFS گوگل یک سیستم فایل توزیع شده توسعه یافته برای پشتیبانی کاربردهای مقیاس بزرگ، توزیع شده، حساس به داده می‌باشد. GFS سرورهای ارزان را استفاده می‌کند تا به تحمل‌پذیری خطا برسد و برای مشتریان سرویس‌های با کارایی بالا تهیه کند. GFS کاربردهای فایل مقیاس-بزرگ را با

[1] Single Point Of Failure
[2] Social Network Services

فرکانس خواندن بیشتر از نوشتن پشتیبانی می‌کند. به هر حال GFS چندین محدودیت دارد، مثل SPOF و کارایی کم برای فایل‌های کوچک.

به علاوه سایر شرکت‌ها و محققین راه‌حل‌های خودشان را دارند تا تقاضای ذخیره‌سازی برای کلان داده را برآورده کنند. به عنوان مثال، HDFS و Kosmosfs مشتقات کدباز GFS هستند. میکروسافت Cosmos را برای پشتیبانی تحقیقاتش و کارهای تجاری‌اش توسعه داد. فیس بوک HayStack را بهره‌برداری می‌کند تا مقدار بزرگی از تصاویر با اندازه کوچک را ذخیره کند. Toobao همچنین TFS و FastDFS را توسعه داد. بطور خلاصه سیستم‌های فایل توزیع شده تقریباً بعد از سال‌ها توسعه و تحقیق تقریباً بالغ شده‌اند. بنابراین ما روی دو سطح دیگر در این بخش تمرکز می-کنیم.

۱-۳-۴ فن‌آوری پایگاه داده

فن‌آوری پایگاه داده برای بیشتر از ۳۰ سال تکامل یافته است. سیستم‌های پایگاه داده مختلف برای مدیریت مجموعه داده‌های با مقیاس‌های مختلف و پشتیبانی از کاربردهای مختلف توسعه داده شده‌اند. واضح است که پایگاه داده‌های رابطه ای سنتی نمی‌توانند چالش‌های مرتبط با مقیاس کلان داده را برآورده کنند. پایگاه داده‌های NoSQL (پایگاه داده‌های غیررابطه‌ای) برای ذخیره کلان داده رایج شدند. پایگاه داده‌های NoSQL ویژگی مدهای منعطف، پشتیبانی از کپی ساده، سازگاری مشروط، و پشتیبانی حجم زیاد داده‌ها را دارند. پایگاه داده‌های NoSQL فن‌آوری ساده برای کلان داده می‌باشند. ما سه پایگاه داده NoSQL عمده را در این بخش دنبال می‌کنیم: پایگاه داده‌های کلید-مقدار[1]، پایگاه داده‌های ستون گرا[2]، و پایگاه داده‌های سند گرا[3]. که هر کدام براساس مدل داده خاصی هستند.

۱-۱-۳-۴ پایگاه داده‌های مقدار-کلید

پایگاه داده‌های کلید-مقدار از یک مدل ساده تشکیل شده‌اند و داده‌ها براساس کلید-مقدار مرتبط ذخیره می‌شوند. هر کلید منحصر بفرد است و مشتریان ممکن است مقادیر پرس و جو شده را براساس کلیدها وارد کنند. چنین پایگاه داده‌های مقدار-کلید مدرن توسعه پذیری بالا و زمان پاسخ پرس و جوی کمتری نسبت به پایگاه داده‌های رابطه ای دارند. در سال‌های اخیر پایگاه داده‌های کلید-مقدار زیادی ظاهر شدند. مثل سیستم Dynamo آمازون. در ادامه Dynamo و چند پایگاه داده کلید-مقدار مهم را معرفی می‌کنیم.

Dynamo: یک سیستم ذخیره داده کلید-مقدار توزیع شده قابل توسعه است که بطور گسترده قابل دسترس است. آن برای مدیریت وضعیت ذخیره بعضی سرویس‌های هسته پلت فرم تجارت الکترونیک آمازون استفاده می‌شود. پلت فرم تجارت الکترونیک آمازون چندین سرویس ذخیره را فراهم می‌کنند که می‌تواند با دسترسی کلید برآورده شود. مد عمومی پایگاه داده‌های رابطه ای ممکن است داده‌های نامعتبر و مقیاس داده محدود و قابل دسترس ایجاد کند. Dynamo می‌تواند نیازهای چنین کاربردهایی را با رابط کلید-شیء ساده فراهم کند. رابط Dynamo با خواندن نوشتن ساده عناصر داده

[1] Key-Value
[2] Column-oriented
[3] Document-Oriented

تشکیل می‌شود. Dynamo به کشسانی و قابل دسترس بودن از طریق پارتیشن بندی داده‌ها، کپی داده‌هاه و مکانیزم‌های ویرایش شیء می‌رسد.

Voldemort: یک سیستم ذخیره کلید-مقدار است که ابتدا توسط Linkedin توسعه داده شد. کلمات کلیدی و مقادیر در Voldemort اشیاء مرکب هستند که جداول و تصاویر را تشکیل می‌دهند. Voldemort یک رابط به سه عمل دارد: خواندن، نوشتن، حذف که همه آنها با کلید انجام می‌شود.

۲-۱-۳-۴ پایگاه داده‌های ستون گرا

در پایگاه داده‌های ستون گرا، داده‌ها بر اساس ستون‌ها به جای سطرها پردازش می‌شوند. ستون‌ها و سطرها به چندین گره تقسیم می‌شوند تا توسعه پذیری برآورده شود. پایگاه داده‌های ستون-گرا عمدتاً توسط BigTable گوگل بررسی می‌شود. در این بخش ابتدا BigTable را بحث می‌کنیم و سپس چندین ابزار مشتق را معرفی می‌کنیم.

BigTable: یک سیستم حافظه توزیع شده، با داده ساختار یافته است که برای پردازش داده‌های با مقیاس بزرگ (PB) در هزاران سرور تجاری طراحی می‌شود. ساختار داده پایه BigTable نگاشتی چندبعدی با ذخیره خلوت، توزیع شده و دائمی است. ایندکس‌های نگاشت کلمات کلیدی، سطرها، کلیدهای ستون‌ها، و مهرهای زمانی می‌باشد و هر مقدار نگاشت یک آرایه بایت تجزیه تحلیل شده است. کلمات کلید سطرهای BigTable رشته کاراکترهای 64KB هستند، که سطرها براساس ترتیب ذخیره می‌شوند.

Cassandra: یک سیستم ذخیره سازی توزیع شده برای مدیریت مقدار حجیمی از داده‌های ساختار یافته توزیع یافته در چندین سرور تجاری می‌باشد. سیستم توسط فیس بوک توسعه داده شد و یک ابزار کد باز در ۲۰۰۸ شد.

ابزارهای مشتق شده از BigTable: Hbase, HyperTable از آن مشتق شده‌اند. Hbase یک محیط برنامه‌نویسی شده BigTable با جاوا می‌باشد و بخشی از فریم ورک MR، هدوپ آپاچی است.

۲-۱-۳-۴ پایگاه داده‌های سندی[1]

در مقایسه با ذخیره سازی بر اساس کلید-مقدار، ذخیره سازی بر اساس سند می‌تواند شکل‌های پیچیده‌تری از داده را پشتیبانی کند چون اسناد از مدهای strict تبعیت نمی‌کنند، نیازی به مهاجرت مد اتصال نیست. به علاوه زوج‌های کلید-مقدار هنوز می‌توانند ذخیره شوند. مثل MongoDB, SimpleDB, CouchDB.

۲-۳-۴ فاکتورهای طراحی

از انواع سیستم‌های پایگاه داده، یک سیستم تک که می‌توانند به کارایی بهینه تحت همه شرایط بارکاری برسد وجود ندارد. در هر سیستم پایگاه داده، بعضی اهداف کارایی باید برای رسیدن به عملیات بهینه برای کاربردهای خاص مورد مصالحه قرار گیرند.

---

[1] **Document Databases**

Cooper، مصالحه که در سیستم‌های مدیریت داده براساس رایانش ابری با آن مواجه می‌شویم را بحث کرده است، شامل کارایی خواندن و کارایی نوشتن، تأخیر و دوام، کپی‌های سنکرون و آسنکرون و بخش‌بندی داده و بقیه می‌باشد. بعضی محققین نیز سایر فاکتورهای طراحی را تجزیه تحلیل کرده‌اند. در زیر ما چندین ویژگی برحسته سیستم‌های پایگاه داده موجود را مقایسه کرده‌ایم.

- مدل داده‌ها: این بخش سه مدل داده مرکزی را بررسی کرده است. مثل مقدار-کلید، ستون گرا و سندگرا.
- مخزن داده‌ها: در بعضی سیستم‌ها، داده‌ها طوری طراحی شده‌اند تا در RAM ذخیره شوند و Snapshot آنها در دیسک ذخیره می‌شوند. سایر سیستم‌ها داده‌ها را در دیسک ذخیره می‌کنند.
- کنترل همروندی: سه مکانیزم کنترل همروند در سیستم‌های موجود دارند: Lock، MVCC و کنترل غیرهمروند.
- سازگاری:براساس نظریه CAP، سازگاری سخت نمی‌توانند بطور همزمان با دسترس‌پذیری و تحمل پارتیشن بدست آید. مدل‌های سازگاری ضعیف، مشروط و براساس زمان باید برای هرکدام در نظر گرفته شود.
- انتخاب CAP: نظریه CAP مشخص می‌کند که یک سیستم داده مشترک ممکن است به دو ویژگی برسد، از بین سازگاری، قابل دسترس بودن و تحمل پارتیشن باید نوع CAP را مشخص کرد.

۴-۳-۳ مدل برنامه نویسی پایگاه داده

مجموعه داده‌های انبوه کلان داده معمولاً در صدها و حتی هزاران سرور تجاری ذخیره می‌شوند. واضح است که مدل‌های موازی سنتی (مثل رابط عبور پیام MPI) و چندپردازشی بار (OpenMP) ممکن است برای پشتیبانی چنین برنامه‌های موازی مقیاس بزرگ مناسب نباشند.

بعضی مدهای برنامه نویسی موازی برای رشته‌های خاص پیشنهاد شده‌اند. این مدل‌ها بطور موثری کارایی NOSQL را بهبود می‌دهند و شکاف کارایی بین پایگاه داده رابطه‌ای را کاهش می‌دهند. بنابراین این مدل‌ها ستون تجزیه تحلیل داده‌های انبوه می‌باشند.

۴-۳-۳-۱ MapReduce

MR یک مدل برنامه‌نویسی ساده ولی قدرتمند برای محاسبات مقیاس-بزرگ با استفاده ار تعدا زیادی خوشه PCهای تجاری می‌باشد تا به پردازش موازی و توزیع خودکار برسد. در MR، وظایف[1] به کارها[2] تقسیم شده و بطور موازی به چندین سرور تقسیم شده و با هم انجام می‌شوند و خروجی کارها با انجام عمل Reduce با هم ترکیب شده و خروجی نهایی استخراج می-شود.

۴-۳-۳-۲ Dryad

[1] job
[2] task

یک موتور اجرایی توزیع شده همه‌منظوره است برای کاربردهای پردازش موازی داده‌های درشت دانه. ساختار عملیاتی Dryad یک گراف بدون دور جهت دار است که رأس‌ها نشان دهنده برنامه‌ها و یال‌ها نشان دهنده کانال‌های داده می‌باشند. Dryad عملیات را روی رأس‌های خوشه‌های کامپیوتر اجرا می‌کند و داده‌ها را از طریق کانال داده منتقل می‌کند، شامل اسناد، اتصالات TCP، و FIFO حافظه-مشترک. در طول عملیات، منابع در گراف عملیاتی منطقی بطور خودکار به منابع فیزیکی نگاشت داده می‌شوند.

ساختار عملیاتی Dryad توسط یک برنامه مرکزی به نام مدیر وظیفه هماهنگ می‌شود که می‌تواند در خوشه‌ها یا ایستگاه-های کاری کاربران اجرا شود. ایستگاه‌های کاری می‌تواند از طریق شبکه به خوشه دسترسی داشته باشند.

All-Pairs ۴-۳-۳-۳

سیستمی است که بطور خاص برای کارهای پزشکی و زیستی و کاربردهای داده کاوی طراحی شده است.

Pregel ۴-۳-۳-۴

برای تجزیه تحلیل گراف‌های شبکه و سرویس‌های شبکه استفاده می‌شود.

فصل پنجم تجزیه تحلیل کلان داده

چکیده: در این فصل، روش‌ها، معماری‌ها و ابزارهای تجزیه تحلیل کلان داده‌ها را معرفی می‌کنیم. تجزیه تحلیل کلان داده عمدتاً شامل روش‌های تجزیه تحلیل برای داده‌های سنتی و کلان داده، معماری تجزیه تحلیلی برای کلان داده، و نرم‌افزار استفاده شده برای داده‌کاوی و تجزیه تحلیل کلان داده است. تجزیه تحلیل داده فاز نهایی و مهمترین فاز در زنجیره کلان داده می‌باشد با هدف استخراج ارزش‌های مفید، تهیه پیشنهاد و تصمیم‌گیری می‌باشد. سطوح مختلف مقادیر می‌توانند از طریق تجزیه تحلیل مجموعه داده‌ها در فیلدهای مختلف تولید شوند.

۵-۱ تجزیه تحلیل داده‌های سنتی

تجزیه تحلیل داده‌های سنتی به معنی استفاده درست از روش‌های آماری می‌باشد تا داده‌های درجه-یک انبوه و داده‌های سطح دو به منظور تمرکز، استخراج، و تصحیح داده‌های مفید مخفی در یک دسته از داده‌های بی‌نظم، و حداکثر کردن ارزش داده تجزیه تحلیل شوند. تجزیه تحلیل داده‌ها نقش راهنمای بزرگ در تولید نقشه‌های توسعه یک کشور بازی می‌کند، مثل فهمیدن تقاضاهای مشتری، و پیش‌بینی جهت و سمت و سوی سرمایه گذاران بازار.

ممکن است به نظر برسد تجزیه تحلیل کلان داده مشابه تجزیه تحلیل نوع خاصی از داده‌ها می‌باشد. بنابراین بسیاری از روش‌های تجزیه تحلیل داده‌های سنتی ممکن هست هنوز هم برای تجزیه تحلیل کلان داده‌ها استفاده شوند. چند روش تجزیه تحلیل داده‌های سنتی مهم در زیر بررسی می‌شوند، که بسیاری از آنها از آمار و علوم کامپیوتر بدست می‌آیند:

- تجزیه تحلیل خوشه (کلاستر): تجزیه تحلیل خوشه یک روش آماری برای اشیاء گروهی می‌باشد، و بالاخص کلاس-بندی اشیاء براساس بعضی ویژگی‌ها. تجزیه تحلیل خوشه برای اشیاء مختلف با ویژگی خاص استفاده می‌شود و آنها را براساس این ویژگی‌ها، به گروه‌ها (خوشه) تقسیم می‌کند بطوریکه اشیاء در یک گروه تجانس زیادی دارند و گروه-های مختلف عدم تجانس بالایی دارند.
- تجزیه تحلیل فاکتور: تجزیه تحلیل فاکتور برای توصیف روابط بین شناسه‌ها و عناصر زیادی براساس چند فاکتور می‌باشد، و چند متغیر مرتبط را گروه می‌کند.
- تجزیه تحلیل همبستگی: تجزیه تحلیل همبستگی روشی برای مشخص کردن قانون همبستگی براساس چند پدیده مشاهده شده می‌باشد و بر این اساس پیش‌بینی و کنترل انجام می‌شود.
- تجزیه تحلیل رگرسیون: این روش یک ابزار ریاضی برای آشکار سازی روابط بین یک متغیر و سایر متغیرها می‌باشد.
- تست A/B: که به آن تست سطل نیز می‌گویند. یک فن‌آوری است برای تعیین نقشه‌ها تا متغیرهای هدف با مقایسه گروه تست شده بهبود یابند. کلان داده نیازمند تعداد زیادی تست است تا اجر و تجزیه تحلیل شود.
- تجزیه تحلیل آماری:تست آماری براساس نظریه آمار است، یک شاخه از applied mathematics. در نظریه آماری، احتمالی بودن و عدم قطعیت با نظریه آمار مدل می‌شوند.
- داده‌کاوی: یک پروسه برای استخراج اطلاعات مفید ناشناخته و مخفی می‌باشد و دانش از داده‌های تصادفی، فازی، نویزی، ناکامل، و انبوه بدست می‌آید. عبارات دیگری نیز مشابه داده‌کاوی هستند، مثل کشف دانش از پایگاه داده‌ها، و پشتیبانی تصمیم‌گیری.

داده‌کاوی عمدتاً برای این استفاده می‌شود تا شش کار متفاوت زیر را با روش‌های تجزیه تحلیلی مرتبط کامل کند: کلاس‌بندی، تخمین، پیش‌بینی، گروه بندی سببی یا قوانین انجمنی، خوشه بندی، و توصیف و بصری سازی. به نظر می‌رسد داده‌های اصلی به عنوان منبع دانش می‌باشند و داده‌کاوی پردازه‌ای از کشف دانش از داده‌های اصلی است. داده‌های اصلی، ممکن است داده‌های ساختار یافته باشند: به عنوان مثال داده‌ها در پایگاه داده رابطه‌ای، یا داده‌های شبه ساختار یافته باشند. مثلاً متن، گرافیک، داده‌های تصویر ، یا حتی داده‌های نامتجانس توزیع شده در شبکه. روش‌های کشف دانش ممکن است ریاضی یا غیر ریاضی، استقرایی یا استنتاجی باشند. دانش کشف شده ممکن است برای مدیریت اطلاعات، بهینه سازی پرس و جوها، پشتیبانی تصمیم‌گیری و پردازه کنترل و نگهداری داده استفاده شود.

روش‌های کاوش معمولاً به روش‌های یادگیری ماشین، روش‌های شبکه عصبی، و روش‌های پایگاه داده تقسیم شوند. یادگیری ماشین ممکن است به آموزش استنتاجی، آموزش براساس مثال، و الگوریتم‌های ژنتیک و غیره تقسیم شود. روش‌های شبکه عصبی ممکن است به شبکه عصبی feed forward و شبکه‌های عصبی خودسازمان دهی شده و ... تقسیم شود. روش‌های پایگاه داده عمدتاً شامل تجزیه تحلیل داده‌های چند بعدی یا OLAP و روش‌های استنتاجی ویژگی‌گرا می‌شود.

الگوریتم‌های داده‌کاوی مختلفی توسعه داده شده‌اند. شامل هوش مصنوعی، یادگیری ماشین، شناسایی روش، اجتماع پایگاه داده و آمار و .... در سال ۲۰۰۶، کنفرانس بین المللی IEEE روی داده‌کاوی ده الگوریتم داده‌کاوی موثر را مشخص کرد و شامل Naive Bayes, and Cart, C4.5, k-means, SVM, Apriori, EM, و غیره می‌باشند. این ده الگوریتم کلاس‌بندی، خوشه‌بندی، رگراسیون، آموزش آماری، تجزیه تحلیل انجمنی، و کاوش لینک را پوشش می‌دهد که همه مهمترین مسائل در تحقیقات داده‌کاوی می‌باشند. به علاوه سایر الگوریتم‌های پیشرفته مثل شبکه‌های عصبی، الگوریتم‌های ژنتیک می‌توانند به داده‌کاوی در کاربردهای مختلف اعمال شوند.

۲-۵ روش‌های تجزیه تحلیل کلان داده

در بدو تولد کلان داده، مردم نگران چگونگی استخراج سریع اطلاعات کلیدی از داده‌های انبوه بودند به طوریکه ارزش را به افراد و سرمایه گذاران بدهند. در حال حاضر، روش‌های پردازش کلان داده در زیر نشان داده شده است:

- Bloom Filter: یک آرایه بیتی و یک سری از توابع هش می‌باشد، مفهوم Bloom Filter ذخیره مقادیر هش برای فشرده سازی ذخیره داده‌ها می‌باشد.
- Hashing: روشی است که داده‌ها را به مقادیر عددی با طول کوچکتر یا مقادیر ایندکس تبدیل می‌کند. هش مزایایی دارد مثل خواند سریع، و سرعت پرس و جوی بالا.
- Index: ایندکس همیشه یک روش مفید برای کاهش هزینه خواندن و نوشتن دیسک و بهبود درج، حذف، اصلاح و سرعت پرس و جو هم در پایگاه داده‌های سنتی داده‌های ساختار یافته را مدیریت می‌کنند و هم فن‌آوری‌هایی که داده‌های نیمه ساختار یافته یا ساختار یافته را مدیریت می‌کنند می‌باشد. به هر حال، ایندکس یک عیب دارد که هزینه اضافی ذخیره فایل‌های ایندکس را دارد و فایل‌های ایندکس باید بطور پویا براساس بروز رسانی داده نگه داری شوند.
- Triel: که به درخت Trie نیز مشهور است که یک درخت هش متغیر می‌باشد. آن عمدتاً برای بازیابی سریع و وضعیت‌های فرکانس کلمه اعمال می‌شود تا مقایسه روی رشته‌های کاراکتری در حوزه سریعتری را کاهش دهد تا بازدهی پرس و جو بهبود یابد.
- Parallel computing: در مقایسه با محاسبات سری سنتی، محاسبات موازی به معنی بهره‌وری از چندین منبع محاسباتی برای کامل کردن یک کار محاسباتی می‌باشد. ایده اصلی تجزیه یک مساله است و انتساب آنها به چندین

پروسه مستقل تا بطور مستقل تکمیل شوند تا به موازات پردازشی برسیم. بعضی مدل‌های محاسباتی سنتی شامل MPI, MapReduce, Dryad می‌باشد.

۵-۳ معماری تجزیه تحلیل کلان داده

به خاطر منابع گسترده، ساختارهای مختلف، و کاربردهای گسترده رشته کلان داده، معماری‌های تجزیه تحلیلی مختلف می‌توانند برای کلان داده با نیازهای کاربردی مختلف مورد توجه قرار گیرد.

۵-۳-۱ تجزیه تحلیل زمان-قطعی در مقایسه با آف لاین

تجزیه تحلیل کلان داده می‌تواند براساس نیازهای زمان-قطعی به دو دسته تجزیه تحلیل زمان-قطعی و تجزیه تحلیل آف لاین کلاس‌بندی شود. تجزیه تحلیل زمان-قطعی عمدتاً در تجارت الکترونیک و مالی استفاده می‌شود. چون داده‌ها بطور ثابت تغییر می‌کنند، تجزیه تحلیل داده سریع مورد نیاز است. نتایج تجزیه تحلیلی باید در مدت زمان کمی برگردانده شود. مهمترین معماری‌های تجزیه تحلیل زمان-قطعی شامل (الف) خوشه‌های پردازش موازی با استفاده از پایگاه داده‌های سنتی و (ب) پلت فرم‌های محاسباتی براساس حافظه به عنوان مثال Greenplum از EMC و HANA از SAP معماری‌های تجزیه تحلیل زمان قطعی هستند.

تجزیه تحلیل آف لاین عمدتاً برای کاربردهایی بدون نیازهای زمان پاسخ بالا استفاده می‌شوند مثل یادگیری ماشین، تجزیه تحلیل آماری، الگوریتم-های توصیه نامه. تجزیه تحلیل آف لاین با ورود لاگ‌های کلان داده در یک پلت فرم خاص با استفاده از ابزارهای جمع‌آوری داده انجام می‌شود. تحت تنظیمات کلان داده، بسیاری از سرمایه گذاران اینترنت معماری تجزیه تحلیل آف لاین را براساس هدوپ به منظور کاهش قیمت تبدیل فرمت داده‌ها و بهبود کارایی گردآوری داده‌ها استفاده می‌کنند. مثل Scribe ابزار کدباز فیس بوک، ابزار Kafka کد باز لینکد این، ابزار کد باز Timetunnel از Taobo و غیره.

۵-۳-۲ تجزیه تحلیل در سطوح مختلف

تجزیه تحلیل کلان داده می‌تواند به تجزیه تحلیل سطح حافظه، تجزیه تحلیل سطح هوش تجاری (BI)، و تجزیه تحلیل سطح انبوه کلاس‌بندی شود که در زیر توضیح داده می‌شوند.

- سطح حافظه[1]: تجزیه تحلیل سطح حافظه برای حالتی است که کل حجم داده در داخل حداکثر سطح حافظه یک خوشه باشد. حافظه خوشه سرور جاری از صدها GB می‌گذرد در حالیکه سطح TB رایج است. بنابراین، یک تکنولوژی پایگاه داده داخلی ممکن است استفاده شود و داده‌های داغ باید در حافظه قرار گیرند بطوریکه بازدهی تجزیه تحلیل بهبود یابد. تجزیه تحلیل سطح حافظه برای تجزیه تحلیل زمان-قطعی خیلی مناسب است. MangoDB یک معماری تجزیه تحلیل سطح-حافظه مشهور می‌باشد. با توسعه SSD (درایو حالت جامد)، ظرفیت و کارایی تجزیه تحلیل داده سطح-حافظه بیشتر بهبود داده می‌شود و بطور وسیعی اعمال می‌شود.
- BI: تجزیه تحلیل BI برای حالتی است که مقیاس داده از سطح حافظه عبور کند، اما ممکن است به محیط تجزیه تحلیل BI وارد شود. در حال حاضر محصولات BI اصلی با تجزیه تحلیل داده‌ها که بیشتر از TB را پشتیبانی می‌کند تهیه شده است.
- انبوه[2]: تجزیه تحلیل انبوه برای حالتی است که مقیاس داده از ظرفیت محصولات BI و پایگاه داده‌های رابطه‌ای عبور کند. در حال حاضر مهمترین تجزیه تحلیل انبوه Hadoop HDFS می‌باشد برای ذخیره داده‌ها و MR برای تجزیه تحلیل داده‌ها، اکثر تجزیه تحلیل انبوه در گروه تجزیه تحلیل آف لاین قرار می‌گیرند.

۵-۳-۳ تجزیه تحلیل با پیچیدگی مختلف

پیچیدگی زمانی و فضایی الگوریتم‌های تجزیه تحلیل داده عمدتاً براساس انواع مختلف داده‌ها و نیازهای کاربردی با هم فرق می‌کنند. به عنوان مثال، برای کاربردهایی که متمایل به پردازش موازی هستند، یک الگوریتم توزیع شده ممکن است طراحی شود و یک مدل پردازش موازی ممکن است برای تجزیه تحلیل داده‌ها استفاده شود.

---

[1] Memory level
[2] Massive

۴-۵ ابزارهای کاوش و تجزیه تحلیل کلان داده

ابزارهای زیادی برای کاوش کلان داده و تجزیه تحلیل قابل دسترس هستند، شامل نرم‌افزار حرفه‌ای و آماتور، نرم‌افزار تجاری گران، و نرم افزار کد باز رایگان. در این بخش بطور مختصر پنج نرم‌افزار استفاده شده بطور گسترده براساس تحقیق "چه نرم-افزار کلان داده، داده‌کاوی شما در ۱۲ ماه گذشته روی یک پروژه واقعی استفاده کرده‌اید" که از ۷۹۸ حرفه‌ای که در سال ۲۰۱۲ توسط KBNuggets انجام شده است را بیان می‌کنیم:

- R (۳۷٪): R یک زبان برنامه‌نویسی کد باز و محیط نرم‌افزاری است که برای تجزیه تحلیل و کاوش داده و بصری سازی طراحی شده است.
- Excel (۲۹,۸٪): اکسل از مجموعه آفیس قابلیت پردازش داده و قابلیت تجزیه تحلیل آماری قوی دارد، و به تصمیم گیری کمک می‌کند. وقتی اکسل نصب می‌شود، چند افزونه پیشرفته مثل Analysis ToolPak و Solver Add-in به آن اضافه می‌شود با توابع قدرتمند برای تجزیه تحلیل داده‌ها که برای استفاده از آنها باید آنها را فعال کنید. اکسل تقریباً یکی از ۵ نرم افزار تجاری اصلی است.
- Rapid-I Rapidminer (۲۶,۷٪): Rapidminer یک نرم‌افزار کدباز است که برای داده‌کاوی، یادگیری ماشین، و تجزیه تحلیل استفاده می‌شود. برنامه‌های یادگیری ماشین و داده‌کاوی تولید شده توسط Rapidminer شامل استخراج، تبدیل و بارگذاری (ETL) بین پردازش داده‌ها و بصری سازی، مدل‌سازی، ارزیابی بکارگیری می‌شوند. جریان داده‌کاوی در XML توصیف می‌شود و در یک رابط گرافیکی قوی نشان داده می‌شود. Rapidminer در جاوا نوشته شده است. آن روش ارزیابی و آموزش Weka را ترکیب می‌کند و با R کار می‌کند. توابع Rapidminer با اتصال پردازه‌های پیاده‌سازی شده و عملگرها کار می‌کنند. می‌توان جریان کلی را به عنوان خط تولید کارخانه در نظر گرفت و داده‌های اصلی را به عنوان ورودی و نتایج مدل را خروجی.
- KNIME(۲۱,۸٪): KNIME یک پلت فرم داده-کاوی سودمند و کاربرپسند است. آن به کاربر اجازه می‌دهد جریان-ها داده یا کانال-های داده را به روشی تصویری ایجاد کند، تا بطور انتخابی بعضی یا همه روال‌های تجزیه تحلیلی را اجرا کند و نتایج تجزیه تحلیل را ایجاد کند، مدل‌ها و نماهای محاوره‌ای KNIME در جاوا نوشته شده است و براساس Eclipse است.
- Weka/pentaho: Weka[1] یک نرم افزار دادکاوی و یادگیری ماشین کد باز است که به زبان جاوا نوشته شده است. Weka توابعی به عنوان پردازش داده، انتخاب ویژگی، کلاس‌بندی، رگراسیون، خوشه‌بندی، قانون نصب و بصری سازی دارد.

[1] Waikato environment for knowledge analysis

فصل ششم کاربردهای کلان داده

در فصل قبل، تجزیه تحلیل کلان داده را بررسی کردیم، که مهمترین فاز و فاز پایانی زنجیره کلان داده می‌باشد. تجزیه تحلیل کلان داده می‌تواند مقادیر مهمی ایجاد کند از طریق قضاوت، توصیه‌ها، یا تصمیم‌ها. به هر حال تجزیه تحلیل کلان داده دامنه وسیعی از داده‌ها را دربرمی‌گیرد با تغییر متناوب و خیلی پیچیده. در این فصل، تکامل منابع داده بررسی می‌شود. سپس شش حوزه تجزیه تحلیل داده مهم بررسی می‌شوند شامل تجزیه تحلیل داده ساختار یافته، تجزیه تحلیل متن، تجزیه تحلیل وب سایت، تجزیه تحلیل چندرسانه‌ای، تجزیه تحلیل شبکه و تجزیه تحلیل موبایل. این فصل با بحث راجع به چند رشته کاربردی کلیدی کلان داده پایان می‌یابد.

۶-۱ تکامل کاربرد

اخیراً، کلان داده، و تجزیه تحلیل کلان داده برای توصیف مجموعه داده‌ها به عنوان فن‌آوری‌های تجزیه تحلیلی در برنامه‌های پیچیده با مقیاس-بزرگ پیشنهاد شده است، که نیازمند تجزیه تحلیل با روش‌های تجزیه تحلیلی پیچیده می‌باشد. به عنوان یک حقیقت، کاربردهای براساس داده در دهه‌های گذشته پدیدار شده‌اند. به عنوان مثال در سال ۱۹۹۰، هوش تجاری به عنوان یک فن‌آوری غالب برای کاربردهای شغلی ، موتورهای جستجوی شبکه براساس پردازش داده‌کاوی انبوه در اوایل قرن بیست و یک پدیدار شد. بعضی کاربردهای تأثیرگذار از فیلدهای مختلف و داده‌های آنها و ویژگی‌های تجزیه تحلیلی در زیر بحث می‌شوند.

- تکامل کاربردهای تجاری: اولین داده‌های تجاری غالباً داده‌های ساختار یافته بودند که توسط شرکت‌ها از سیستم-های قدیمی جمع‌آوری شدند و سپس در RDBMS ذخیره می‌شدند. فن‌آوری‌های تجزیه تحلیلی استفاده شده در چنین سیستم‌هایی در سال ۱۹۹۰ رایج شدند مثل گزارش‌ها، پنل‌های تجهیزات، پرس و جوهای خاص، هوش تجاری براساس جستجو، پردازش تراکنش برخط، بصری سازی محاوره‌ای، کارت‌های score، مدل‌سازی پیش‌بینی کننده، و داده‌کاوی. از ابتدای قرن بیست و یک، شبکه‌ها و وب‌سایت‌ها، یک روش منحصر به فرد برای سازمان‌هایی که نمایش برخط داشتند و مستقیماً با مشتری محاوره داشتند تهیه کردند و محصولات فراوان و اطلاعات مشتری شامل کلیک روی رشته‌های لاگ‌های click stream data logs و داده و رفتار کاربر و ... می‌توانند از وب‌سایت‌ها بدست آیند. در سال ۲۰۱۱ تعداد تلفن‌های موبایل و Pcهای تبلت از تعداد لپ‌تاپ‌ها و Pcها پیشی گرفتند. گوشی‌های موبایل و اینترنت اشیاء براساس سنسورها یک نسل جدید از کاربردهای ابداعی را باز کردند.
- تکامل کاربردهای شبکه: اینترنت اولیه عمدتاً ایمیل و سرویس‌های صفحه وب را فراهم می‌کرد. تجزیه تحلیل متن، داده‌کاوی، و فن‌آوری‌های تجزیه تحلیل صفحه وب به کاوش محتوای ایمیل و ساخت موتورهای جستجو اعمال می-شده است. امروزه اکثر کاربردها براساس وب هستند، بدون توجه به رشته کاربردی آنها و اهداف طراحی. داده‌های شبکه مسئول درصد عمده‌ای از حجم داده عمومی می‌باشد. وب یک پلت‌فرم رایج برای صفحات به هم متصل پر از انواع مختلف داده، مثل متن، تصویر، ویدیو، عکس و محتوای محاوره‌ای و غیره می‌باشد. بنابراین فن‌آوری‌های توسعه یافته کافی برای داده‌های ساختارنیافته یا نیمه ساختار یافته پدیدار شده در آن لحظه استفاده می‌شد. به عنوان مثال، فن‌آوری تجزیه تحلیل تصویر، ممکن است اطلاعات مفیدی از تصاویر استخراج کنند مثل تشخیص صورت . از ۲۰۰۴، رسانه اجتماعی برخط، مثل گروه‌های اینترنت، اجتماعات برخط، بلاگ‌ها، سرویس‌های شبکه اجتماعی، و

وب سایت‌های اجتماعی چندرسانه‌ای و غیره برای کاربر روش‌های متفاوتی تهیه کرد تا محتوای تولید شده توسط کاربران را برای جستجوی اخبار روزانه و اخبار افراد مشهور، انتشار عقاید سیاسی و اجتماعی آنها، و تهیه کاربردهای مختلف با فیدبک زمان‌دار استفاده شوند.

- تکامل کاربردهای علمی: تحقیقات علمی در رشته‌های زیادی در حال کسب داده‌های حجیم با سنسورهای با کارایی-بالا و تجهیزات شامل فیزیک نجومی، اقیانوس شناسی، ژن شناسی، و تحقیقات محیطی می‌باشد. بنیاد علوم ملی (NFS) آمریکا اخیراً اطلاعیه‌ای راجع به تحقیقات ابتکاری کلان داده دارد تا تلاش‌های تحقیقی برای استخراج دانش و بینش از مجموعه‌های بزرگ و پیچیده داده‌های دیجیتال را توسعه دهد. بعضی تحقیقات علمی پلت‌فرم‌های داده‌های انبوه و استخراج خروجی‌های مفید را توسعه داده‌اند. به عنوان مثال در بیولوژی iPlant زیرساخت شبکه، منابع محاسباتی فیزیکی، محیط هماهنگی، منابع ماشین مجازی و تجزیه تحلیل عملیاتی نرم‌افزاری و سرویس داده-ها را اعمال کرده است تا به محققین، دانش پژوهان، و دانش‌آموزان در توانا کردن همه علوم زمین کمک کند.

۶-۲ رشته‌های تجزیه تحلیل کلان داده‌ها

تحقیقات تجزیه تحلیل داده را می‌توان به شش رشته تکنیکی تقسیم کرد: تجزیه تحلیل داده‌های ساختار یافته، تجزیه تحلیل داده‌های متن، تجزیه تحلیل داده‌های وب سایت، تجزیه تحلیل داده‌های چندرسانه‌ای، تجزیه تحلیل داده‌های شبکه، و تجزیه تحلیل داده‌های موبایل. هدف چنین کلاس‌بندی تاکید روی ویژگی‌های داده دارد، اما بعضی رشته‌ها ممکن است از فن‌آوری-های مشابه بهره ببرند. چون تجزیه تحلیل داده‌ها دامنه وسیعی دارد و ساده نیست تا همگرایی جامعی داشته باشند، ما روی مسائل کلیدی و فن‌آوری‌ها در تجزیه تحلیل داده‌ها در بحث زیر تمرکز خواهیم کرد.

۶-۲-۱ تجزیه تحلیل داده‌های ساختار یافته

کاربردهای شغلی و تحقیقات علمی ممکن است داده‌های ساختار یافته انبوهی تولید کنند، که مدیریت و تجزیه تحلیل روی فن‌آوری‌های تجاری بالغ تکیه دارد. مثل RDBMS، ورهاوس داده، OLAP، BPM. تجزیه تحلیل داده عمدتاً براساس داده کاوی و تجزیه تحلیل آماری است که هر دو به خوبی در حدود ۳۰ سال گذشته مطالعه شده‌اند.

تجزیه تحلیل داده هنوز هم یک رشته تحقیقاتی خیلی فعال می‌باشد و کاربردهای جدید متقاضی روش‌های جدیدی می‌باشند. یادگیری ماشین آماری براساس مدل‌های ریاضی دقیق و الگوریتم‌های قدرتمند به تشخیص آنومالی و کنترل انرژی اعمال شده‌اند. استخراج ویژگی‌ها داده‌ها، کاوش زمان و فضا ممکن است ساختارهای دانش مخفی در جریان‌های داده سرعت بالا و مدل‌ها و روش‌های سنسورهای داده را استخراج کنند. براساس حفظ امنیت در تجارت الکترونیک، دولت الکترونیک، و کاربردهای حفاظت سلامت، حفظ امنیت داده‌کاوی یک رشته تحقیقاتی نوظهور است.

۶-۲-۲ تجزیه تحلیل داده‌های متنی

رایجترین قالب اطلاعات ذخیره سازی متن، مثل ارتباطات ایمیل، اسناد شغلی، صفحات وب، و رسانه اجتماعی می‌باشد. بنابراین به نظر می‌رسد تجزیه تحلیل متن پتانسیل شغلی بیشتری از داده‌کاوی ساختار یافته داشته باشد. عموماً، تجزیه تحلیل مالیات، نیز که متن کاوی نامیده می‌شود پردازه استخراج اطلاعات و دانش مفید از متن غیر ساختار یافته است. متن کاوی یک مساله داخلی است، شامل بازیابی اطلاعات، یادگیری ماشین، آمار، زبان شناسی کامپیوتری و داده‌کاوی بطور خاص.

اکثر سیستم‌های متن‌کاوی براساس عبارات متنی و پردازش زبان طبیعی (NLP) می‌باشند، با تمرکز بیشتر روی حرف معرفی سند و پردازش پرس و جو پایه مدل توسعه فضای برداری، مدل بازیابی بولین، و مدل بازیابی احتمال می‌باشد که پایه موتور جستجو می‌باشند. از سال ۱۹۹۰ موتورهای جستجو به یک سیستم شغلی بالغ تکامل یافته است، که معمولاً شامل خیزش توزیع شده سریع، ایندکس‌های معکوس شده، و تجزیه تحلیل Log جستجو می‌باشد. NLP کامپیوترها را قادر به تجزیه تحلیل، تفسیر و حتی تولید متن می‌کند. بعضی روش‌های NLP رایج عبارتند از: بدست آوردن کلمه‌ای (لغوی)، رفع ابهام احساس کلمه، تگ گذاری بخشی از گفتار، و گرامر آزاد از متن احتمالی.

۳-۲-۶ تجزیه تحلیل داده‌های وب

در دهه گذشته ما با رشد انفجاری اطلاعات اینترنت مواجه شدیم. تجزیه تحلیل وب به عنوان حوزه جستجوی فعال ظهور پیدا کرده است. هدف تجزیه تحلیل وب بازیابی خودکار، استخراج، و ارزیابی اطلاعات از اسناد وب و سرویس‌هایی است که برای کشف دانش مفید می‌باشد. تجزیه تحلیل وب مرتبط با چندین رشته تحقیقاتی مرتبط، شامل پایگاه داده، بازیابی اطلاعات، NLP و کاوش متن می‌باشد. براساس بخش‌های مختلف وب که باید کاوش شوند، ما تجزیه تحلیل وب را به سه حوزه کلاس-بندی می‌کنیم: محتوا کاوی وب، کاوش ساختار وب و کاوش کارایی وب.

- کاوش محتوای وب: پردازه کشف دانش مفید در صفحات وب است که معمولاً شامل انواع داده مختلف است مثل متن، تصویر، صدا، ویدئو، کد، متا دیتا، و ابر پیوند.

به تحقیقات روی صدا و ویدئو کاوی اخیراً تجزیه تحلیل چندرسانه‌ای می‌گویند، که در بخش ۶-۲-۴ بحث می‌شود. چون اکثر محتوای وب داده‌های متنی ساختارنیافته هستند، تحقیقات روی تجزیه تحلیل داده وب عمدتاً حول متن و ابر متن می‌چرخد. آموزش مدیریت و کلاس‌بندی یک نقش عمده در ابرمتن کاوی بازی می‌کند، مثل ایمیل، مدیریت گروه خبری، و نگهداری کاتالوگ وب. کاوش محتوای وب می‌تواند با دو روش انجام شود: روش بازیابی اطلاعات، و پایگاه داده. بازیابی اطلاعات عمدتاً به جستجوی اطلاعات کمک می‌کند یا آن را بهبود می‌دهد، یا اطلاعات کاربر را براساس پیکربندی اسناد یا استنتاج فیلتر می‌کند. روش پایگاه داده در ارزیابی و تجمع داده‌ها در وب کمک می‌کند، بطوریکه پرس و جوهای پیچیده‌تری از جستجوهای براساس کلمات کلیدی را شامل می‌شود.

- کاوش ساختار وب: شامل مدل‌های کشف ساختار لینک‌های وب می‌باشد. در اینجا ساختار به دیاگرام شماتیکی اشاره می‌کند که در یک وب سایت پیوند دارد یا در بین چندین وب سایت. مدل‌ها براساس ساختارهای هم‌بندی تهیه شده با ابرپیوندها یا بدون توصیف پیوند می‌باشد. چنین مدل‌هایی شباهت‌ها و همبستگی‌های بین چندین سایت مختلف را آشکار می‌کنند و برای کلاس‌بندی وب سایت‌ها استفاده می‌شوند. PageRank, Clever مدل-های کامل را استفاده می‌کنند تا صفحات وب مرتبط را جستجو کنند. خزنده‌های براساس موضوع، حالت موفق دیگری با بهره‌وری مدل‌ها می‌باشند.
- کاوش کارایی و بهره‌وری وب: هدف آن کاوش داده‌های جانبی تولید شده توسط رفتار یا دیالوگ‌های وب می‌باشد. کاوش محتوای وب و کاوش ساختار وب داده‌های اصلی را استفاده می‌کنند. کاوش بهره‌وری وب شامل دسترسی logها در وب سرورها، logها در سرورهای پراکسی، رکوردهای تاریخچه مرورگر، پروفایل‌های کاربر، داده‌های ثبت، جلسات یا تجارت‌های کاربر، کش، پرس و جوهای داده، داده‌های bookmark، کلیک ماوس و Scroll، و هر نوع دیگری از داده‌ها که در محاوره با وب به دست می‌آید. هر چه سرویس‌های وب و web 2.0 بالغتر و رایجتر شوند،

داده‌های کاربردی در وب باید بطور افزایشی تنوع زیادی داشته باشند. کاوش بهره‌وری وب نقش کلیدی در فضای شخصی، تجارت الکترونیک، امنیت/شخصی سازی شبکه، و سایر فیلدهای پدیدار شده بازی می‌کنند.

۶-۲-۴ تجزیه تحلیل چندرسانه‌ای

داده‌های چندرسانه‌ای (عمدتاً عکس‌ها، صداها، ویدئوها) با سرعتی متحیر کننده رشد کرده‌اند. اشتراک محتوای چندرسانه‌ای استخراج دانش مرتبط و فهم معنای داده‌های چندرسانه‌ای می‌باشد. چون داده‌های چندرسانه‌ای نامتجانس هستند و اکثر چنین داده‌هایی شامل اطلاعات غنی‌تری از داده‌های ساختار یافته ساده هستند استخراج اطلاعات با چالش بزرگی از تفاوت-های معنایی داده‌های چندرسانه‌ای مواجه می‌شود. تحقیقات روی تجزیه تحلیل چندرسانه‌ای، جنبه‌های زیادی را پوشش می-دهد. بعضی از اولویت‌های تحقیقاتی اخیر شامل خلاصه سازی چندرسانه‌ای، حاشیه نویسی چندرسانه‌ای، ایندکس چندرسانه-ای و بازیابی آن، پیشنهاد چندرسانه‌ای و تشخیص رخداد چندرسانه‌ای و غیره می‌باشد.

خلاصه سازی صدا: می‌تواند با استخراج کلمات یا عبارت برجسته از متا دیتا یا سنتز نمایش جدید بدست آید. خلاصه‌سازی ویدیو تفسیر مهمترین یا برجسته‌ترین رشته محتوای ویدیو می‌باشد و می‌تواند ایستا یا پویا باشد. روش‌های خلاصه‌سازی ویدیوی ایستا یک رشته فریم کلیدی یا فریم‌های کلیدی حساس به محتوا را بهره‌برداری می‌کند تا ویدیو را نشان دهد. چنین روش‌هایی خیلی ساده هستند و قابل اعمال به بسیاری از کاربردهای شغلی هستند (مثل yahoo, AltaVista, google) اما کارایی playback ضعیف می‌باشد. روش‌های خلاصه سازی پویا یک سری کلیپ‌های ویدئو را استفاده می‌کند تا یک ویدئو را نشان دهد، توابع ویدئوی سطح پایین را پیکربندی می‌کند، و واحدهای سنجش نرم دیگری را برای بدست آوردن خلاصه-سازی نهایی بطور طبیعی استفاده می‌کند. سیستم خلاصه‌سازی چندرسانه‌ای موضوع-گرا (TOMS) می‌تواند بطور خودکار اطلاعات مهم یک ویدئو را خلاصه سازی کند ویدئویی که در یک موضوع خاص قرار دارد براساس مجموعه داده شده از ویژگی‌های استخراج شده از ویدئو.

حاشیه نویسی چندرسانه‌ای: برچسب‌هایی را برای توصیف محتوای تصاویر و ویدئو در هر دو سطح لغوی و نحوی شامل می-شود. با کمک چنین برچسب‌هایی، مدیریت، خلاصه‌سازی و بازیابی داده‌های چندرسانه‌ای می‌تواند به سادگی پیاده‌سازی شود. چون حاشیه نویسی هم حساس به زمان و کوشش است، حاشیه نویسی خودکار چندرسانه‌ای بدون مداخله انسان می‌تواند خیلی جذاب شود. چالش‌های اصلی حاشیه نویسی خودکار چندرسانه‌ای تفاوت نحوی می‌باشد، به عنوان مثال تفاوت بین ویژگی‌های سطح پایین و حاشیه نویسی. هرچند که پیشرفت زیادی ایجاد شده است، کارایی روش‌های حاشیه نویسی خودکار موجود هنوز نیازمند بهبود است. در حال حاضر، تلاش‌های زیادی ایجاد شده است تا حاشیه نویسی چندرسانه‌ای خودکار و دستی بطور همزمان بررسی شود.

ایندکس بازیابی چندرسانه‌ای: شامل توصیف، ذخیره‌سازی و سازمان‌دهی اطلاعات چندرسانه‌ای و کمک به کاربران می‌باشد تا به آسانی و به سرعت منابع چندرسانه‌ای را جستجو کنند. معمولاً بازیابی و ایندکس چندرسانه‌ای شامل ۵ روال است: تجزیه تحلیل ساختاری، استخراج ویژگی، داده‌کاوی، کلاس‌بندی و حاشیه نویسی، پرس و جو و بازیابی. هدف تجزیه تحلیل ساختاری تقسیم کردن یک ویدئو به چندین عنصر ساختاری معنایی می‌باشد، شامل تشخیص مرز لنز، استخراج فریم کلیدی، و تقسیم‌بندی صحنه و غیره. براساس نتایج تجزیه تحلیل ساختاری، روال دوم استخراج ویژگی است که شامل کاوش ویژگی-های فریم‌های کلید مورد نیاز، اشیاء، متن‌ها، حرکات می‌باشد که پایه ایندکس و بازیابی ویدئو است. داده‌کاوی، کلاس‌بندی، و

حاشیه نویسی برای بهره‌برداری ویژگی‌های استخراج شده و پیدا کردن مدهای محتوای ویدئو و قرار دادن ویدئو در دسته-های زمانبندی شده برای تولید ایندکس ویدئو می‌باشد. براساس دریافت یک پرس و جو، سیستم یک روش اندازه‌گیری مشابه استفاده خواهد کرد تا یک ویدئوی کاندیدا را جستجو کند. نتیجه بازیابی فیدبک مربوطه را بهینه می‌کند.

هدف توصیه نامه چندرسانه‌ای: توصیه کردن یک محتوای چندرسانه‌ای خاص براساس ترجیح کاربر می‌باشد. آن به عنوان یک روش موثر برای تهیه سرویس‌های مشخص سازی کیفی اثبات شده است. مهمترین سیستم‌های توصیه موجود می‌تواند به سیستم‌های براساس محتوا، سیستم‌های براساس-فیلتر-هماهنگ کننده، تقسیم بندی شوند. روش‌های براساس محتوا، کاربران یا ویژگی‌های عمومی که کاربر به آن علاقه‌مند است را شناسایی می‌کند و محتوایی با ویژگی‌های مشابه را برای کاربر توصیه می‌کند. این روش‌ها بیشتر روی سنجش محتوای مشابه تکیه دارد. اما اکثر آنها با تجزیه تحلیل محتوا و توصیف زیاد محدود می‌شود. روش‌های براساس-فیلتر-هماهنگ کننده[1] گروه‌های با علایق مشابه را تشخیص می‌دهد و محتوا را برای اعضای گروه براساس رفتارشان توصیه می‌کند. البته در حال حاضر یک روش ترکیبی نیز معرفی شده است که مزایای دو نوع روش توضیح داده شده را ترکیب می‌کند تا کیفیت توصیه نامه را بهبود دهد.

US NIST یک روش تشخیص ارزیابی ویدئو TREC ایجاد کرده که ظهور یک رخداد در کلیپ-ویدئو را براساس Event kit تشخیص می‌دهد که شامل بعضی توصیفات متنی مرتبط به مفاهیم و مثال‌های ویدئو است. تحقیقات موجود روی تشخیص رخداد عمدتاً روی ورزش و اخبار، اجرای رخداد غیرعادی (ناهنجار) در مانیتورینگ ویدئو، و سایر رخدادهای مشابه با الگوهای تکراری تمرکز می‌کند. تحقیقات روی تشخیص رخداد ویدئو هنوز بالغ نشده است.

۵-۲-۶ تجزیه تحلیل داده شبکه

تجزیه تحلیل شبکه از تجزیه تحلیل کمی اولیه و تجزیه تحلیل شبکه جامعه شناسی شروع شد و به تجزیه تحلیل شبکه اجتماعی در ابتدای قرن بیست و یک رسید. بسیاری از سرویس‌های شبکه اجتماعی برخط مثل توییتر، فیس بوک و لینکد این و غیره در ده سال خیلی رایج شده‌اند. چنین سرویس‌های شبکه اجتماعی برخط معمولاً شامل داده‌های پیوندی و محتوایی انبوه هستند. داده‌های پیوندی عمدتاً به شکل ساختارهای گرافیکی هستند که ارتباطات بین دو ورودی را توصیف می‌کنند. داده‌های محتوا شامل محتوای متن، تصویر، و سایر داده‌های چندرسانه‌ای شبکه می‌باشد. محتوای غنی شبکه‌ها چالش‌هایی در مورد تجزیه تحلیل داده‌ها ایجاد می‌کند. براساس جلوه با مرکزیت داده، تحقیقات موجود روی محتوای سرویس شبکه اجتماعی می‌تواند به دو طبقه کلاس‌بندی شود: تجزیه تحلیل ساختاری و تجزیه تحلیل براساس محتوا.

تحقیقات روی تجزیه تحلیل ساختاری براساس پیوند همیشه به پیش‌بینی پیوند، کشف اجتماعات، تکامل شبکه اجتماعی و تجزیه تحلیل تأخیر اجتماعی منجر شده است. SNS می‌تواند به عنوان گراف‌هایی بصری شود، که هر رأس مرتبط با یک کاربر و یال‌ها مرتبط با وابستگی‌های بین کاربران می‌باشد. چون SNS شبکه‌های پویا هستند، رأس‌های جدید و یال‌های جدید می‌توانند بطور پیوسته به گراف‌ها اضافه شوند. پیش‌بینی یال، یعنی پیش‌بینی اینکه احتمال اضافه شدن اتصالات آینده بین دو رأس می‌باشد. فن‌آوری‌های زیادی می‌توانند برای پیش‌بینی یال استفاده شوند، مثلاً کلاس‌بندی‌های براساس ویژگی، روش‌های احتمالات، و جبر خطی. کلاس‌بندی براساس ویژگی انتخاب یک گروه از ویژگی‌ها برای یک راس بهره‌وری از اطلاعات یال موجود می‌باشد تا یال‌های آینده پیش‌بینی شود. هدف روش‌های احتمالاتی ساخت مدل‌هایی برای احتمال اتصال

[1] Collaborate-filtering-based methods

بین راس‌های SNS می‌باشد، جبر خطی تشابه بین دو راس را براساس یک ماتریس تشابه تک محاسبه می‌کند، اجتماع با یک ماتریس زیرگراف نشان داده می‌شود که یال‌ها راس‌ها را متصل می‌کند و چگالی در داخل یک اجتماع از چگالی خارج اجتماع بیشتر است.

روش‌های زیادی در مورد تشخیص اجتماع پیشنهاد و مطالعه شده اند که اکثر آنها براساس همبندی هستند که روی مفهوم تشخیص ساختار اجتماع تکیه دارند. هدف تحقیقات SNS جستجو برای یک مدل قانونی و استنتاجی برای تغییر تکامل شبکه می‌باشد.

تأثیر اجتماعی به معنی حالتی است که افراد رفتارشان را تحت تأثیر دیگران تغییر می‌دهند. قدرت تأثیر اجتماعی وابسته به رابطه بین افراد، فواصل شبکه، اثر زمان، و ویژگی‌های شبکه‌ها و افراد دارد. بازیابی، آگهی، توصیه نامه، و سایر کاربردها از تأثیر اجتماعی سود می‌برند با سنجش میزان تأثیر کمی و کیفی بر دیگران. به طور کلی اگر تکثیر محتوای بین SNS مورد توجه قرار گیرد، کارایی تجزیه تحلیل ساختاری براساس لینک می‌تواند بیشتر بهبود داده شود.

براساس پیشرفت انقلابی Web 2.0، استفاده از محتوای تولید شده بطور انفجاری در SNS رشد می‌کند، SNS برای تولید محتوا براساس فن‌آوری‌های مختلف، شامل بلاگ‌ها، ریزبلاگ‌ها، عقیده کاوی، تصاویر، اشتراک ویدیو، نشان‌گذاری اجتماعی، سایت‌های شبکه اجتماعی، اخبار اجتماعی و Wiki و .. استفاده می‌شود. تجزیه تحلیل براساس محتوا در SNS نیز به عنوان تجزیه تحلیل رسانه اجتماعی نیز شناخته می‌شود. رسانه اجتماعی شامل، متن، چندرسانه‌ای، مکان‌یابی، و توضیحات می‌باشد. تقریباً همه موضوعات تحقیقاتی مرتبط با تجزیه تحلیل ساختاری، تجزیه تحلیل متن، و تجزیه تحلیل چندرسانه‌ای می‌تواند به عنوان تجزیه تحلیل رسانه اجتماعی تفسیر شود، اما تجزیه تحلیل رسانه اجتماعی با چالش‌های بی‌سابقه‌ای مواجه می‌شود. اولاً، داده‌های رسانه اجتماعی شامل نویز بیشتری است، مثلاً blogSphere شامل مجموعه بزرگی از بلاگ‌های اسپم، است و بنابراین توییت‌های ناچیزی در توییتر دارد. ثالثاً، SNS شبکه‌های پویا هستند، که بطور متناوب و به سرعت تغییر داده شده و بروز می‌شوند.

چون رسانه اجتماعی نزدیکSNS است، تجزیه تحلیل رسانه اجتماعی ناگزیر توسط تجزیه تحلیل SNS تحت تأثیر قرار می‌گیرد. تجزیه تحلیل SNS به تجزیه تحلیل متن محتوای SNS و ویژگی‌های شبکه و اجتماع و تجزیه تحلیل چندرسانه‌ای اشاره می‌کند. تحقیقات موجود روی تجزیه تحلیل رسانه اجتماعی هنوز در ابتدای راه هستند. کاربردهای تجزیه تحلیل متن SNS شامل آموزش انتقال و جستجوی کلمات کلیدی، کلاس‌بندی، خوشه‌بندی، و شبکه‌های نامتجانس می‌باشند. جستجوی کلمه کلیدی سعی دارد تا بطور سنکرون محتویات و رفتارهای لینک جستجو را استفاده کند. محرک چنین کاربردهایی، فایل-های متنی می‌باشد که شامل کلمات کلیدی مشابه می‌باشند که معمولاً به یکدیگر متصل شده‌اند. در کلاس‌بندی فرض کنید همه گره‌های SNS برچسب‌هایی دارند، گره‌های اضافه شده با برچسب‌ها کلاس‌بندی می‌شوند. در خوشه‌بندی هدف محققین این است که مجموعه گره‌های با محتوای مشابه را مشخص کنند و براساس آن، آنها را گروه‌بندی کنند. فرض کنید SNS شامل اطلاعات انبوه اشیاء به هم پیوسته مختلف باشند، به عنوان مثال مقالات، برچسب‌ها، تصاویر، و ویدیوها، آموزش در شبکه‌های نامتجانس را انتقال می‌دهند، هدف انتقال دانش اطلاعات در طول یال‌های مختلف است.

مجموعه داده‌های چندرسانه‌ای در SNS به یک شکل ساختاریافته سازمان‌دهی می‌شوند، که اطلاعات غنی دارند، مثلاً، محاورات اجتماعی، رسانه اجتماعی، نقشه‌های جغرافیایی، و عقاید چندرسانه‌ای. تجزیه تحلیل چندرسانه‌ای ساختاری در

SNS شبکه‌های اطلاعاتی چندرسانه‌ای نامیده می‌شوند. ساختار یال شبکه‌های اطلاعاتی چندرسانه‌ای غالباً یک ساختار منطقی است، که اهمیت بالایی در شبکه‌های چندرسانه‌ای دارند. ساختارهای اتصال منطقی شبکه‌های اطلاعاتی چندرسانه‌ای می‌توانند به چهار نوع تقسیم شوند: هستی شناسی معنایی، رسانه اجتماعی، آلبوم‌های تصویر فردی، و مکان‌های جغرافیایی.

۶-۲-۲ تجزیه تحلیل ترافیک موبایل

با رشد سریع محاسبات موبایل، ترمینال‌های موبایل و کاربردهای آن در جهان به سرعت رشد می‌کند. در آپریل ۲۰۱۳، کاربردهای اندروید به بیش از ۶۵۰۰۰۰ کاربرد شدند، که تقریباً هرچیزی را می‌پوشانند. در آخر سال ۲۰۱۲، جریان داده موبایل ماهانه به بیشتر از 885PB رسیده است. داده‌های انبوه و کاربردهای زیاد یک رشته تحقیقاتی بزرگ برای تجزیه تحلیل موبایل بوجود آورد. اما با تعدادی چالش روبرو شد. در کل داده‌های موبایل ویژگی‌های منحصر بفردی دارند، مثل حس موبایل، انعطاف حرکت، نویز و میزان زیادی افزونگی. اخیراً تحقیقات جدیدی در تجزیه تحلیل موبایل در رشته‌های مختلف شروع شده‌اند. به خاطر عدم بلوغ تحقیقات در تجزیه تحلیل موبایل، فقط چند کاربرد تجزیه تحلیل مشهور را در این بخش معرفی می‌کنیم.

با رشد تعداد کاربران موبایل و بهبود کارایی، گوشی‌های موبایل هم اکنون برای ساخت و حفظ اجتماعات مهم هستند، مثل اجتماعات براساس مکان‌های جغرافیایی و اجتماعات براساس خوشه‌ها در علایق مختلف مثل Wechat. گوشی‌های موبایل می‌توانند محاورات غنی را در هر زمان و هر جایی پشتیبانی کنند. Wechat ارتباطات یک به یک و یک به چند و چند به چند را پشتیبانی می‌کند. اجتماعات موبایل به عنوان گروهی از افراد تعریف می‌شوند که علایق مشابهی دارند (سلامت، امنیت، سرگرمی و غیره) در کنار هم در یک شبکه جمع شده‌اند به هدف ایجاد یک هدف مشترک و حرکت برای رسیدن به هدف.

برچسب های RFID برای تشخیص، مکان یابی، ردیابی، و مدیریت اشیاء فیزیکی به یک روش سودمند-هزینه استفاده می-شوند. RFID بطور گسترده به مدیریت ابداعات و لوجیستیک اعمال می‌شود. به هر حال RFID چالش‌های بسیاری راجع به تجزیه تحلیل داده‌ها آورد. (الف) داده‌های RFID نویزی و افزونه هستند (ب) داده‌های RFID داده‌های جریان دار و فوری و حجم بالا هستند که نیاز به پردازش با محدودیت زمان دارند.

اخیراً پیشرفت سنسورهای بی‌سیم، فن‌آوری ارتباط موبایل، و پردازش جریان محققین را قادر می‌کند تا یک حوزه در مورد سلامت و ردیابی سلامتی مردم ایجاد کنند.

۶-۳ کاربردهای کلیدی

۶-۳-۱ کاربرد کلان داده در سرمایه‌گذاری

در حال حاضر، کلان داده عمدتاً از سرمایه‌گذاری‌ها نشأت می‌گیرد مثلاً BIو OLAP کاربردهای اولیه کلان داده بودند. کاربرد کلان داده در سرمایه‌گذاری می‌تواند کیفیت محصول و رقابت آنها را در جنبه‌های زیادی توسعه دهد. مثل پیش‌بینی رفتار مشتری‌ها، و کاوش حالت‌های شغلی جدید. در امور مالی کاربرد کلان داده بسیار است در تجارت الکترونیک نیز کاربرد آن گسترده است.

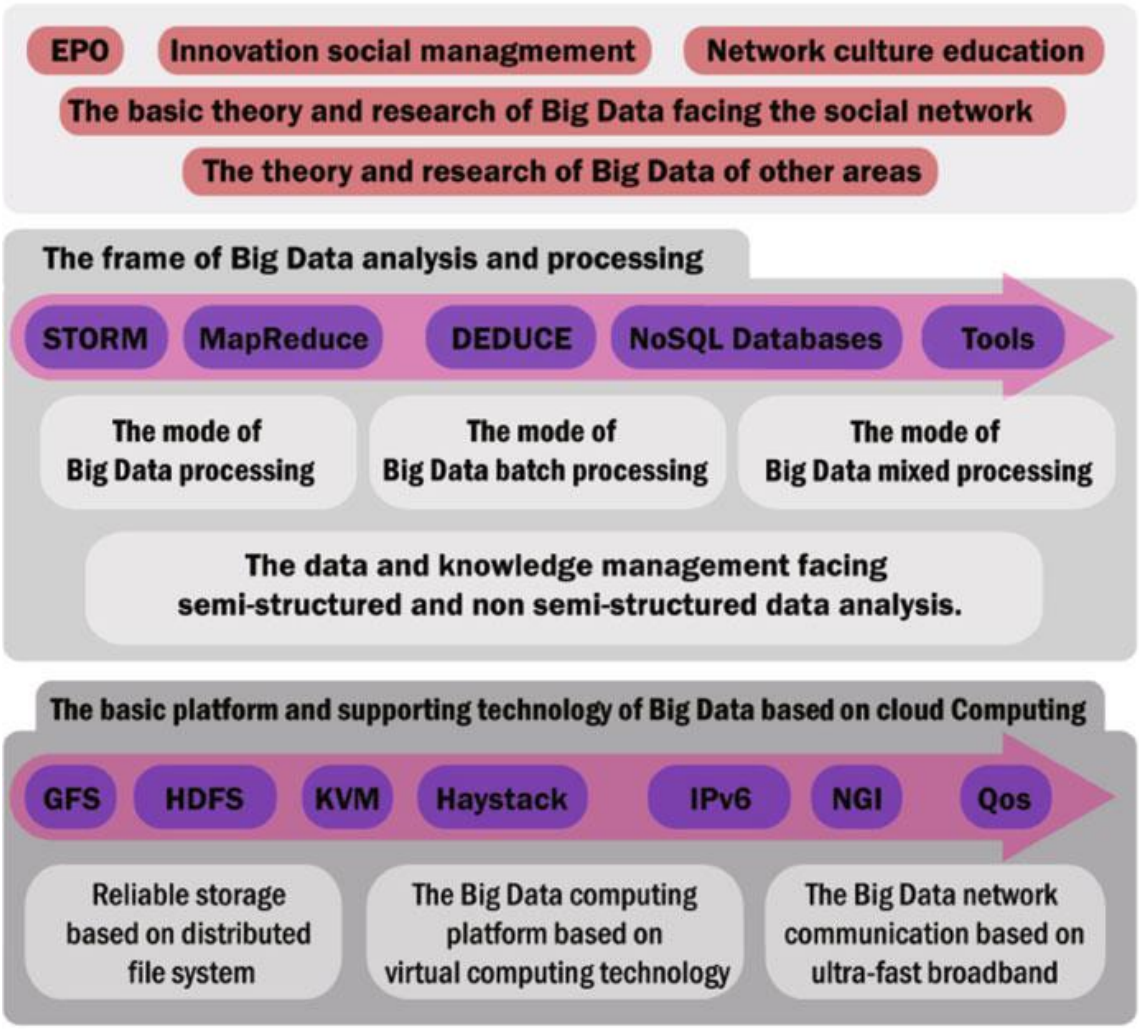
۶-۳-۲ کاربرد کلان داده در IOT

اینترنت اشیاء نه تنها منبع مهمی از کلان داده است، بلکه مهمترین بازار کاربردی کلان داده است. در اینترنت اشیاء، هر شیء در دنیای واقعی ممکن است هم تولید کننده و هم مصرف کننده داده‌ها باشد، زیرا به خاطر تنوع زیاد اشیاء، کاربرد IOT به طور نامتناهی تکامل می‌یابد. شهرهای هوشمند یک موضوع تحقیقاتی داغ براساس کاربرد داده‌های IOT می‌باشد. که به تصمیم گیری بهتر دولت در مصرف آب، منابع آب، مدیریت آن، کاهش ترافیک و بهبود سلامت عمومی کمک می‌کند.

۶-۳-۳ کاربرد کلان داده شبکه-گرای اجتماعی آن‌لاین

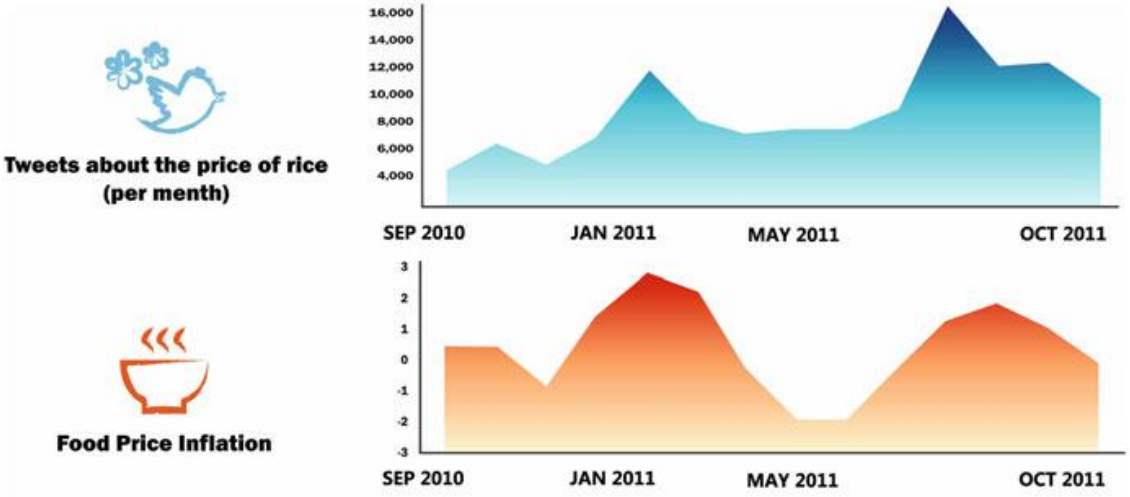
SNS برخط یک ساختار اجتماعی است که توسط افراد اجتماعی و اتصالات بین افراد براساس یک شبکه اطلاعاتی ایجاد شده است. کلان داده در SNS عمدتاً از پیام‌های مستقیم، اجتماع برخط، ریزبلاگ‌ها و فضای مشترک و ... می‌آید. چون کلان داده از SNS برخط نشان دهنده فعالیت‌های کاربردی مختلف است، تجزیه تحلیل چنین داده‌هایی توجه بیشتری را جلب می‌کند. تجزیه تحلیل داده‌های کلان SNS برخط روش تجزیه تحلیل محاسباتی را استفاده می‌کند که برای فهم روابط جامعه انسانی استفاده از تئوری‌ها و روش‌هایی مثل ریاضیات، اطلاعات انفورماتیک، جامعه شناسی و علم مدیریت و غیره استفاده می‌کند. کاربرد کلان داده در SNS برخط شامل تجزیه تحلیل عقیده عمومی، تجزیه تحلیل و جمع‌آوری هوش شبکه، بازارشناسی اجتماعی، پشتیبانی از تصمیم‌گیری دولتی، و آموزش برخط و غیره می‌باشد. شکل ۶-۱ فریم ورک فنی کاربردهای کلان داده SNS برخط را نشان می‌دهد. کاربردهای کلاسیک کلان داده SNS برخط در زیر معرفی می‌شوند، که عمدتاً کاوش و تجزیه تحلیل محتوای اطلاعات و اطلاعات ساختاری است یا ارزش از آن بدست می‌آید.

- کاربردهای براساس محتوا: زبان‌ها و متن دو تا از مهمترین شکل اطلاعاتی SNS می‌باشند. از طریق تجزیه تحلیل زبان و متن، رجوع‌های کاربر، احساسات، علایق، و تقاضاها و غیره ممکن است به دست آید.
- کاربردهای براساس ساختار: در SNSها که کاربران به عنوان گره هستند، روابط اجتماعی، علایق، و غیره روابط بین کاربران را در یک ساختار خوشه‌بندی شده جمع می‌کنند. چنین ساختاری با روابط بسته در بین افراد، اما روابط خارجی ضعیف نیز اجتماع نامیده می‌شوند. تجزیه تحلیل براساس-اجتماع از اهمیت زیادی برخوردار است تا انتشار اطلاعات بهبود یابد. مثل کشف جرم توسط پلیس و حتی نرخ جرایم در نواحی عمده. در آوریل ۲۰۱۳، Wolfarm Alphw یک موتور جستجوی و اجتماعی US، قانون رفتار اجتماعی کاربران را با تجزیه تحلیل داده‌های اجتماعی بیش از یک میلیون کاربر آمریکایی در فیس بوک مطالعه کرد. براساس تجزیه تحلیل، مشخص شد که اکثر کاربران فیس بوک در سن ۲۰ سالگی عاشق می‌شوند، در سن ۲۷ سالگی نامزد می‌کنند و در سن ۳۰ سالگی ازدواج می-کنند و تغییرات کمی در روابط ازدواج بین ۳۰ و ۶۰ سالگی دارند. این نتایج تحقیقاتی با داده‌های دفاتر ازدواج (کلیساها) نیز همخوانی دارد.



شکل ۶-۱ فن‌آوری‌های توانا در کلان داده شبکه گرای اجتماعی برخط

پروژه دیگری نیز در توییتر در سال ۲۰۱۱ انجام شد نتایج مرتبط با غذا، سوخت، خانه‌داری، و وام را بررسی کند که نتایج آن در شکل ۶-۲ نشان داده شده است.



[URL] http://www.unglobalpulse.org/projects/twitter-and-perceptions-crisis-related-stress

شکل ۶-۲ روابط بین توییت‌ها راجع به قیمت برنج و غذا

کاربرد کلان داده در SNS برخط ممکن است به فهم بهتر رفتار مردم کمک کند تا قوانین فعالیت‌های اجتماعی و اقتصادی را در سه جنبه زیر کارفرما کند.

- اخطار Earlywarning: بررسی جرائم از طریق تشخیص ناهنجاری‌ها در استفاده از دستگاه‌ها و سرویس‌های الکترونیکی
- مانیتور کردن زمان قطعی: تهیه اطلاعات دقیق برای قاعده سازی پلیس-ها و نقشه‌های آن‌ها براساس مانیتور رفتار جاری، احساسات، و رجوع‌های کاربران.
- بازخورد زمان قطعی: بدست آوردن فیدبک‌های گروهی در مقابل بعضی فعالیت‌های اجتماعی براساس مانیتور کردن زمان قطعی.

کاربرد کلان داده در SNS برخط شامل سه مساله تکنیکی عمده می‌باشد:

- مدل داده‌ها: سنتی ترین مدل‌های داده SNS براساس مد ایستا و الگوریتم تجزیه تحلیلی خاصی هستند و برای محاسبات داده‌های با مقیاس PB و بیشتر جوابگو نیستند. به عبارت دیگر تجزیه تحلیل SNS معمولاً تجزیه تحلیل نسبتاً پیچیده چندبعدی را روی داده‌های پویا پیاده‌سازی می‌کنند. بررسی تئوری‌ها و مدل‌های جدید برای پوشاندن این شکاف مورد نیاز می‌باشد.
- مدیریت و ذخیره سازی داده: روش‌های مدیریت ذخیره‌سازی اینترنتی موجود عمدتاً ذخیره سازی کلان داده و پرس و جوی سریع را پشتیبانی می‌کنند. به هر حال، روش‌های موجود بطور مفیدی محاسبات تجزیه تحلیلی کلان داده روی SNS برخط را که ویژگی همبستگی بالا، متغیر پویا، و تکامل چندبعدی و غیره دارند را پشتیبانی نمی‌کند. بنابراین نیازمند روش‌های مدیریت و ذخیره سازی جدید هستیم.
- تجزیه تحلیل داده: روش‌های بررسی تجزیه تحلیل موجود روی کلان داده SNS غالباً براساس ویژگی یک بعدی با دقت ناکافی است. به عبارت دیگر، تجزیه تحلیل SNS مثل تکامل موضوع، محاوره گروه، و احساسات عمومی و غیره نیازمند تجزیه تحلیل همبستگی پیچیده از دید ساختار، گروه و اطلاعات هستند. وجود یک تئوری پایه و روشی که همبستگی پیچیده داده‌های چند بعدی، داده پویا را پشتیبانی کند نیاز می‌باشد.

۶-۳-۴ کاربردهای کلان داده پزشکی و سلامت

داده‌های پزشکی بطور مداوم و سریعاً در حال رشد هستند و دربرگیرنده ارزش‌های اطلاعاتی مختلفی هستند. کلان داده پتانسیل نامحدودی برای ذخیره، پردازش، پرس و جو و تجزیه تحلیل داده‌های پزشکی دارد. کاربرد کلان داده پزشکی برسلامت افراد و بشریت تأثیر بسزایی دارد.

به عنوان مثال، شرکت بیمه Atenalife، ۱۰۲ مریض را از یک مجموعه هزار مریض انتخاب کرد تا یک آزمایش را به منظور کمک به پیش‌بینی بازگشت مریض‌ها با سندروم‌های متابولیک انجام دهد. در یک آزمایش مستقل، ۶۰۰۰۰۰ تست کتابخانه‌ای و ۱۸۰۰۰۰ ادعا از طریق یک سری نتیجه تست تشخیص سندروم متابولیک مریض‌ها در سه سال متوالی بدست آمد. در این پیش‌بینی دکترها می‌توانند ۵۰٪ مریضی‌های آینده را پیش‌بینی و جلوگیری کنند.

سایرکاربردها در شبکه قدرت و برق نیز می‌باشد برای جلوگیری از خاموشی‌ها، کنترل کیفیت و سایر موارد.

فصل ۷ چشم انداز و آینده کلان داده

چکیده: در فصل‌های قبل، زمینه‌ها و مطالب عمده کلان داده را بررسی کردیم. شکل ۷-۱ تمامی فن‌آوری‌های کلیدی معرفی شده در کتاب را خلاصه می‌کند. در این فصل، نقاط داغ تحقیقاتی را خلاصه می‌کنیم و جهت های تحقیقاتی ممکن روی کلان داده را پیشنهاد می‌کنیم. ما همچنین اهداف توسعه این موضوع تحقیقاتی گسترده را بحث می‌کنیم.

۷-۱ مطالب باز: تجزیه تحلیل کلان داده با چالش‌های زیادی مواجه است، اما تحقیقات جاری هنوز در ابتدای راه هستند. تلاش تحقیقاتی قابل توجهی مورد نیاز است تا بهره‌وری بهبود یافته نمایش داده، ذخیره داده و تجزیه تحلیل داده را داشته باشیم.

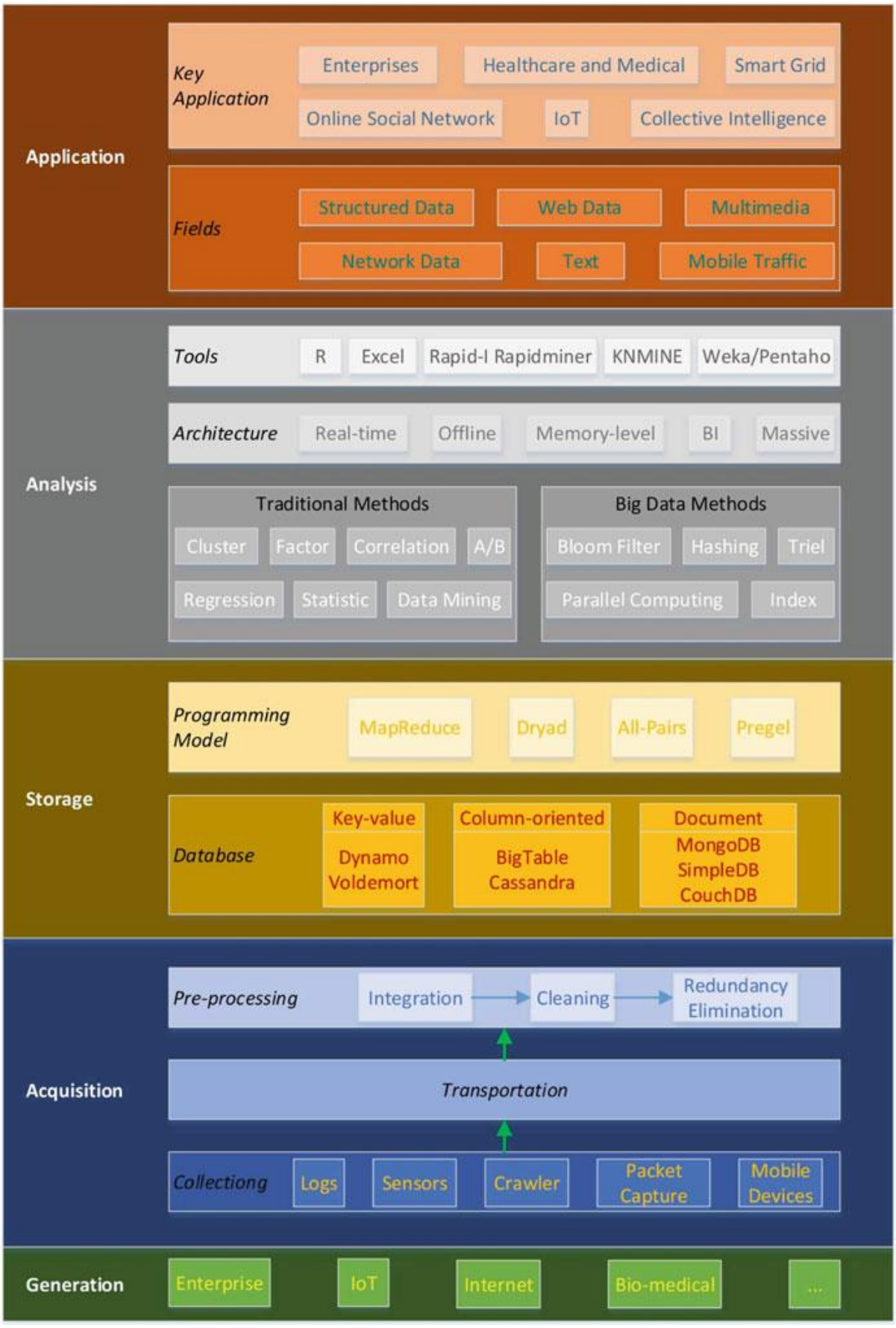
۷-۱-۱ تحقیقات تئوری

هرچند که کلان داده یک حوزه تحقیقاتی داغ هم از دید تجاری و آکادمیک است، مسائل مهمی وجود دارند که باید حل شوند که در زیر بحث می‌شوند.

- مسائل پایه: نیاز ضروری برای تعریف محکم کلان داده، یک مدل ساختاری از کلان داده، توصیف شکلی کلان داده، و یک سیستم تئوری از علم داده و غیره می‌باشد. در حال حاضر، بحث‌های زیادی از کلان داده مشابه تعمق تجاری و تحقیقات علمی وجود دارد. زیرا کلان داده کاملاً بطور ساختار یافته تعریف نشده است و به درستی بررسی نشده است.
- استاندارد سازی: یک سیستم ارزیابی از کیفیت داده و یک استاندارد ارزیابی از بهره‌وری محاسباتی داده باید توسعه داده شود. راه‌حل‌های زیادی از کاربردهای کلان داده ادعا می‌کنند که می‌توانند پردازش داده و ظرفیت‌های تجزیه تحلیل را در همه جنبه‌ها بهبود دهند، اما هنوز یک استاندارد ارزیابی یکنواخت و محک برای متعادل کردن بهره‌وری محاسباتی کلان داده با روش‌های ریاضی قوی وجود ندارد. چون کیفیت داده یک پایه مهم برای پیش‌پردازش داده، ساده‌سازی نمایش می‌باشد، آن یک مسأله فوری است که بطور موثر کیفیت داده را ارزیابی کنیم.
- تکامل مدهای محاسباتی کلان داده: این شامل مد ذخیره سازی خارجی، مد جریان داده، مد PRAM، و مد MR و غیره می‌باشد. ظهور کلان داده باعث، باعث توسعه طراحی الگوریتم شده است که ارزش حساس به محاسبه به روش حساس-به داده تبدیل می‌شود. انتقال داده مهمترین گلوگاه محاسبه کلان داده است.

۷-۱-۲ توسعه فن‌آوری

فن‌آوری کلان داده در ابتدای طفولیت خود است. بسیاری از مسائل فن‌آوری کلیدی، مثل رایانش ابری، محاسبات گرید، محاسبات جریان دار، محاسبات موازی، معماری کلان داده، مدل کلان داده، و سیستم‌های نرم‌افزاری که از کلان داده پشتیبانی می‌کند و غیره باید کاملاً بررسی شوند.



شکل ۷-۱ فن‌آوری‌های کلیدی در موضوع کلان داده

- تبدیل قالب: به خاطر منابع داده مختلف و وسیع، عدم تجانس همیشه یک ویژگی کلان داده است، مثل یک فاکتور کلیدی که بهره‌وری تبدیل فرمت داده‌ها را محدود کند. اگر تبدیل قالب بطور مفیدتری انجام شود کلان داده ممکن است ارزش بیشتری تولید کند.
- انتقال کلان داده: انتقال کلان داده شامل تولید کلان داده، بدست آوردن، انتقال، و سایر تبدیلات داده در دامنه فضایی می‌باشد. همانطور که بحث شد، انتقال کلان داده معمولاً با هزینه بالایی اتفاق می‌افتد که گلوگاه محاسبات کلان داده می‌باشد. به هر حال انتقال داده در کاربردهای کلان داده ضروری است. بهبود، بهره‌وری انتقال کلان داده یک فاکتور کلیدی برای بهبود محاسبه کلان داده است.
- کارایی زمان-قطعی: کارایی زمان قطعی کلان داده یک مساله کلیدی در سناریوهای کاربردی مختلف زیادی است. روش‌های تعریف چرخه عمر داده‌ها، محاسبه نرخ توصیف داده‌ها، و ساخت مدل‌های محاسباتی کاربردهای زمان قطعی و کاربردهای برخط بر ارزش و روش‌های تجزیه تحلیل و نتایج فیدبک کلان داده تاثیر دارد.

۳-۱-۷ استلزام عملی: هرچند که کاربردهای کلان داده موفق زیادی در حال حاضر وجود دارد، مسائل عملی زیادی باید حل شوند.

- مدیریت کلان داده: مدیریت مدل‌های حافظه و پایگاه داده سخت افزار جدید، تجمع داده ساختار یافته و نامتجانس و SNS و مدیریت داده توزیع شده.
- جستجو، کاوش و تجزیه تحلیل کلان داده: پردازش داده همیشه یک نقطه داغ در تحقیقات کلان داده است. مثل جستجو و کاوش مدل‌های SNS، الگوریتم‌های جستجوی کلان داده، جستجوی توزیع شده، جستجوی P2P، تجزیه تحلیل تصویری کلان داده، سیستم‌های توصیه انبوه، سیستم‌های رسانه اجتماعی، کاوش داده‌های کلان زمان قطعی، کاوش تصویر، متن کاوی، معناکاوی، داده‌کاوی چندساختاره، و یادگیری ماشین و غیره.
- تجمع و منشاء کلان داده:
- کاربرد کلان داده:

۴-۱-۷ امنیت داده

در IT امنیت و حفاظت دو مفهوم کلیدی هستند. در حوزه کلان داده هر چه حجم داده به سرعت رشد کند، ریسک‌های امنیتی بیشتری وجود دارد در حالیکه روش‌های محافظت داده‌های سنتی برای کلان داده عملی نیستند. موارد زیرچالش‌های کلان داده می‌باشند.

- اختفاء کلان داده
- کیفیت داده
- مکانیزم امنیت کلان داده
- کاربرد کلان داده در امنیت اطلاعات

۲-۷ چشم انداز

کلان داده روش اقتصادی و اجتماعی و حتی زندگی هر فردی را تغییر خواهد داد که هم اکنون شروع شده است، ما نمی‌توانیم آینده را پیش‌بینی کنیم اما چند رخداد ممکن در آینده عبارتند از:

- داده‌های با مقیاس بزرگتر، تنوع بیشتر، و ساختارهای پیچیده‌تر
- کارایی منابع داده‌ها
- کلان داده پیشرفت علمی را ارتقاء خواهد داد.
- کلان داده روش تفکر را منقلب خواهد کرد.
- مدیریت جداول مقیاس بزرگ برای شبکه‌های تعریف شده نرم‌افزاری با تکنیک‌های کلان داده
- شبکه-های بی‌سیم 5G: پشتیبانی کلان داده برای کلان داده موبایل.