بسمه تعالی

عنوان مستند

امنیت حجیم‌داده

بخش ۱ : معماری امن

فهرست مطالب

چکیده

در سال‌های اخیر، رشد سریع اینترنت و فراگیر شدن فن‌آوری‌های جدیدی نظیر اینترنت اشیاء، محاسبات ابری و شبکه‌های اجتماعی باعث رشد انفجاری تولید و جمع‌آوری داده‌ها در حوزه‌های مختلف حوزه فن‌آوری اطلاعات شده است. در کنار امکانات جدید سخت‌افزاری و روش‌های کلاسیک علوم داده، دانش یا فن‌آوری جدیدی به نام "حجیم‌داده" پایه‌گذاری شده است که به چالش‌های جدید این حوزه می‌پردازد. اصطلاح حجیم‌داده، یک واژه برای توصیف مجموعه داده‌هایی است که دارای حجم بزرگ، سرعت تولید زیاد و ساختار متنوع و پیچیده نسبت به پایگاه‌داده‌های معمولی هستند. از اینرو، برای ذخیره‌سازی، بازیابی، پردازش، تحلیل و همچنین بصری‌سازی آنها نیازمند ساختارها، الگوریتم‌ها و ابزارهایی متفاوت از گذشته هستیم. با پیشرفت‌های صورت گرفته در این چند سال در حوزه حجیم‌داده، کاربردهای این فن‌آوری روزبه‌روز بیشتر گسترش یافته و میزان داده‌هایی که در بسترهای مبتنی بر حجیم‌داده، ذخیره‌سازی و پردازش می‌شوند افزایش چشم‌گیری داشته است. یکی از چالش‌های اساسی در این زمینه، نحوه تامین امنیت داده در بستر حجیم‌داده است. یک بستر حجیم‌داده کارآمد نباید تنها روی حجم، سرعت یا تنوع داده‌ها تمرکز کند، بلکه با توجه به انبوده داده‌های مهم موجود در آن، باید حفاظت آنرا نیز تضمین نماید. تنوع در ساختار داده‌ها و امکان دسترسی گسترده به آن‌ها توسط کاربران متعدد، موجب شده تا روش‌های سنتی حفاظت در حجیم‌داده کارآمد نباشند و امنیت داده را با چالش‌های جدیدی روبرو سازند. از این روی شاهد پیدایش و رشد ابزارها و چارچوب‌های متنوع در بخش‌های مختلف حجیم‌داده در حوزه امنیت هستیم، که مدیران و صاحبان صنایع، کارشناسان حوزه علوم داده و کاربران علاقه‌مند و یا مجبور به رعایت و استفاده از این ابزارها و چارچوب‌ها در راستای حفاظت از داده می‌باشند. در این سلسله مستندات، ما قصد داریم مهمترین چالش‌ها و مباحث را در حوزه امنیت حجیم‌داده تشریح کرده و برای هرکدام از این چالش‌ها، ابزارهای کاربردی را معرفی نماییم. در نهایت، برخی از این ابزارها را که بیشتر مورد استفاده قرار می‌گیرند، به صورت عملی مورد نصب و بررسی اجمالی قرار خواهیم داد.

در این مستند، پس از بیان مفاهیم و کلیات حجیم‌داده، موارد مربوط به معماری حجیم‌داده مورد بررسی قرار می‌گیرد و راه‌کارهایی که به منظور توسعه معماری امن در این حوزه پیشنهاد شده، معرفی و بررسی می‌گردند.

۱. مقدمه

داده‌های حجیم یا حجیم‌داده ترجمه اصطلاح Big Data می‌باشد که معمولا به مجموعه‌ای از داده‌ها اطلاق می‌شود که اندازه آنها فراتر از حدی است که با روش‌ها و نرم‌افزارهای معمول بتوان آنها را در یک زمان معقول جمع‌آوری، ذخیره، بازیابی و پردازش نمود. مفهوم اندازه در حجیم‌داده به طور مستمر در حال تغییر است و به مرور بزرگتر می‌شود. در کنار حجم، فاکتورهای دیگری نظیر سرعت تغییرات داده، تنوع داده‌ها و یا دقت آنها نیز در این بحث مطرح می‌باشد. فن‌آوری حجیم‌داده مجموعه‌ای از دانش‌ها، فنون و ابزارهایی است که می‌توانند ارزش افزوده‌ای را که در مجموعه‌های حجیم و متنوع داده پنهان شده‌اند، آشکار سازند. با رشد روزافزون داده‌ها و نیاز به بهره‌برداری و تحلیل این داده‌ها، به کارگیری این فن‌آوری از اهمیت ویژه‌ای برخوردار شده است. این مبحث به این دلیل هر روز جذابیت و مقبولیت بیشتری پیدا می‌کند که با استفاده از پردازش حجم بیشتری از داده‌ها، می‌توان تحلیل‌های بهتر و پیشرفته‌تری را برای مقاصد مختلف، از جمله مقاصد تجاری، پزشکی و امنیتی، انجام داد و نتایج مناسب‌تری دریافت نمود. البته یکی از دلایل اصلی برای فراگیر شدن این فن‌آوری، رشد قابل ملاحظه قدرت پردازشی و فضای ذخیره‌سازی سامانه‌های کامپیوتری در کنار کاهش قیمت آنها در سال‌های اخیر بوده است. همینطور، ابزارهای مناسبی برای کار با داده‌های حجیم بوجود آمده که کار محققان، تحلیل‌گران و محققان را ساده ساخته‌اند.

طبق گزارش‌های مختلف، تا سال ۲۰۰۳ چیزی در حدود پنج اگزابایت داده توسط بشر ایجاد شده است. امروزه این مقدار از اطلاعات در طی دو روز ایجاد می‌شود. فیسبوک به صورت ماهانه ۹۵۵ میلیون حساب فعال با استفاده از ۷۰ زبان و ۱۴۰ میلیارد عکس آپلود شده و همینطور ۲.۷ میلیارد ابراز علاقمندی و نظر ارسال شده دارد. هر دقیقه ۴۸ ساعت ویدئو آپلود می‌شود و هر روز ۴ میلیارد نمایش ویدئو روی یوتیوب اجرا می‌شود. گوگل نیز سرویس‌های زیادی را پشتیبانی می‌کند: ۷.۲ میلیارد صفحه را در هر روز خزش می‌کند و همچنین ۲۰ پتابایت داده را روزانه به ۶۶ زبان ترجمه می‌کند. یک میلیارد توییت در هر ۷۲ ساعت از بیش از ۱۴۰ میلیون کاربر فعال روی توییتر فرستاده می‌شود. ۵۷۱ وب سایت جدید هر دقیقه از روز ایجاد می‌شوند. شواهد حالی از آنست که تا دهه آینده حجم اطلاعات تا صدها برابر افزایش خواهد یافت[۲]. امروزه، انفجار داده‌ها به وضوح در زمینه‌های اجتماعی و ارتباطی مختلف مشاهده می‌شود.

فن‌آوری‌های جدید کار با داده‌های حجیم مانند رایانش ابری، یک زیرساخت برای اتوماسیون تمام فرآیندها در جمع‌آوری داده‌ها، ذخیره‌سازی، پردازش و بصری‌سازی فراهم می‌کنند. در اثر استفاده از این فن‌آوری‌ها، چالش‌های تازه‌ای برای فن‌آوری‌های امنیتی سنتی بوجود می‌آید، که ممکن است نیاز به شناخت و طراحی مجدد ابزارها و حتی تجدیدنظر در مدل‌های امنیتی کنونی داشته باشد [۵]. در این مطالعه عمدتاً روی مسائل امنیتی مربوط به حجیم‌داده تمرکز می‌کنیم و فن‌آوری‌های حجیم‌داده در زمینه امنیت و حریم خصوصی را مورد بحث قرار می‌دهیم.

حجیم‌داده دارای مزایای متعدد و پیامدهای مثبت زیادی می‌باشد. داده‌ها می‌توانند به افزایش بهره‌وری اقتصادی، بهبود دسترسی به سرویس-های اجتماعی، تقویت امنیت، شخصی‌سازی خدمات و افزایش دسترسی به اطلاعات و پلت‌فرم‌های نوآورانه برای ارتباطات کمک کنند. به عنوان مثال، برنامه‌های مسیریابی برای راننده‌ها، اطلاعات بلادرنگی در مورد ازدحام جاده‌ها فراهم می‌کنند که به آن‌ها اجازه انتخاب مسیرهای کارآمد را

می‌دهد. حجیم‌داده می‌تواند با بهبود عملیات، تسهیل نوآوری و انطباق پذیری و بهینه‌سازی تخصیص منابع، سازمان‌های کارآمدتری ایجاد کند. داده‌های بزرگ و فن‌آوری تحلیل داده، زمان تجزیه و تحلیل برخی از انواع داده‌ها را از چند ماه تا چند روز کوتاه کرده‌اند. به ویژه، پیشرفت‌های اخیر در زمینه یادگیری ماشین، برای مثال یادگیری عمیق، امکان کشف به موقع الگوها و ناهنجاری‌ها را در حجم انبوهی از داده‌ها امکان‌پذیر می‌کنند [۷].

۲. مشکلات و چالش‌های حجیم‌داده

سه ویژگی کلیدی حجیم‌داده، یعنی حجم زیاد، تولید سریع و تنوع داده، روش‌های محاسباتی سنتی را برای پشتیبانی موثر از پردازش، ذخیره‌سازی و تجزیه و تحلیل داده سخت می‌کنند. محاسباتی از این دست به سادگی بر روی ابزارها و الگوریتم‌های قبلی قابل اجرا نیستند. ویژگی‌های جدید در پردازش حجیم‌داده، مانند نمونه‌های ناکافی، باز بودن و نامشخص بودن روابط داده‌ها، نامتعادل بودن توزیع چگالی مقدار، نه تنها فرصت‌های بزرگی فراهم می‌کنند، بلکه چالش‌های بزرگی را برای محاسبه‌پذیری حجیم‌داده و توسعه روش‌های محاسباتی جدید مطرح می‌کنند [۱۰]. اگرچه حجیم‌داده یک زمینه بزرگ تحقیقاتی با پتانسیل بالا، هم در مجامع علمی و هم در صنعت محسوب می‌شود، هنوز مشکلات حل نشده‌ی مهمی در مورد آن وجود دارد که برخی در زیر آمده است:

الف) چارچوب نظری: نیاز فوری به یک تعریف دقیق و جامع از حجیم‌داده، و از آن مهمتر، یک مدل ساختاری و نظری از علوم داده‌ها وجود دارد. در حال حاضر، بیشتر بحث‌های حجیم‌داده از نقطه نظر تجاری به جای تحقیق علمی به آن می‌پردازند.

ب) استانداردسازی: یک سامانه ارزیابی از کیفیت داده‌ها و یک استاندارد/معیار ارزیابی از محاسبات داده‌ها باید معرفی شود. بسیاری از راه حل‌ها و برنامه‌های کاربردی حجیم‌داده ادعا می‌کنند که می‌توانند پردازش داده‌ها و ظرفیت‌های تجزیه و تحلیل را در تمام جنبه‌ها بهبود دهند، اما هنوز یک استاندارد و معیار ارزیابی یکپارچه برای برآورد کارایی محاسبات حجیم‌داده با روش‌های ریاضیاتی دقیق وجود ندارد. در حال حاضر، عملکرد حجیم‌داده تنها در زمان پیاده‌سازی و استقرار سامانه قابل ارزیابی است.

ج) آموزش: ظهور حجیم‌داده باعث پیشرفت‌های طراحی الگوریتم شد که از یک رویکرد مبتنی بر محاسبات فشرده و متمرکز[1] به یک رویکرد داده‌ی فشرده و توزیع شده[2] تبدیل شده است. انتقال داده به گلوگاه اصلی محاسبات حجیم‌داده تبدیل شده است. از همین رو، مدل‌های محاسباتی جدیدی برای حجیم‌داده پدید آمده‌اند که شیوه کارکرد آنها با مدل‌های محاسبات قبلی تا حد زیادی متفاوت می‌باشد و به همین دلیل، بایستی نیروی متخصص و توانمند در این زمینه آموزش داده شود [۹].

[1] Centralized
[2] Distributed

در کنار مشکلات فوق، راه‌اندازی هر سامانه حجیم‌داده چالش‌هایی به همراه دارد که باید حتما به آنها توجه شده و راهکاری برای آن اندیشیده شود. البته باید توجه نمود که اصولا یک راهکار منحصربفرد برای این چالش‌ها وجود ندارد و متناسب با نوع مسئله و ماهیت داده و کاربردها، باید این چالش‌ها مرتفع شوند. مهمترین این چالش‌ها به شرح زیر می‌باشند:

الف) ذخیره‌سازی: اولین چالش اساسی از ماهیت اصلی خود حجیم‌داده سرچشمه می‌گیرد که حجم می‌باشد. چگونگی ایجاد یک سامانه ذخیره‌سازی پایدار برای نگهداری، بروزرسانی و افزایش نمایی داده‌ها یک چالش است. اصولا در یک محیط مبتنی بر حجیم‌داده از سرورهای متعدد برای ذخیره‌سازی داده به صورت توزیع شده استفاده می‌شود. ثانیاً، شناسایی مجموعه داده‌های با ارزش، ذخیره‌سازی ایمن و امن داده‌ها، افزایش کارائی و سرعت بازیابی داده‌ها و امکان درج و بروزرسانی لحظه‌ای و آنلاین داده‌ها همچنان جزو چالش‌های اصلی در ذخیره‌سازی حجیم‌داده است. برای مثال، معمولا در سامانه‌های حجیم‌داده، از هر داده نسخه‌های پشتیبان متعددی ساخته و نگهداری می‌شود.

ب) پردازش: عامل حجم و تنوع حجیم‌داده که ناشی از جمع‌آوری مجموعه داده‌های بزرگ از انواع مختلف (ساخت‌یافته، نیمه ساخت‌یافته و غیر ساخت‌یافته) و همینطور از منابع مختلف است همچنان چالش بزرگی در بحث پردازش محسوب می‌شود. بسته به کاربرد و نوع مسئله، پردازش این داده‌ها می‌تواند به صورت توده‌ای (بچ[1]) و یا لحظه‌ای (استریمی[2]) باشد. پردازش چنین مجموعه داده‌هایی نیازمند فن‌آوری‌های جدیدی بر روی بسترهای توزیع‌شده و خوشه‌ای[3] می‌باشد که می‌تواند الگوریتم‌ها و ساختمان داده‌های مختلفی را تولید و اجرا کند.

ج) امنیت و حریم خصوصی: نگران کننده‌ترین چالش کنونی حجیم‌داده، حفظ امنیت اطلاعات و حریم خصوصی می‌باشد. با توجه به ابعاد و ارزش بالای داده، تامین امنیت مجموعه داده‌ها یک چالش اصلی محسوب می‌شود. جلوگیری از نشت و یا دستکاری داده‌ها در هنگام ذخیره‌سازی، بازیابی، پردازش و دفاع از حملات بیرونی نیازمند یک مدل امنیت داده محور قابل اعتماد است. این فن‌آوری همچنین باید از جلوگیری تهدیدات امنیتی که ممکن است در طول چرخه از تولید تا نمایش چنین داده‌های بزرگی رخ دهد، مراقبت کند [۶]، به ویژه آنکه یکی از مزیت‌های سرویس‌های حجیم‌داده، توانایی اشتراک و انتشار داده‌ها بر روی شبکه است. از طرف دیگر، امروزه حریم خصوصی برای افراد و سازمان‌ها بسیار حیاتی شده است و به همین دلیل، این موضوع تبدیل به یک چالش عمده برای حجیم‌داده شده است. مشخص است که حفظ حریم خصوصی صرفا با از بین بردن ساده هویت صاحبان داده‌ها تامین نمی‌شود. در بسیاری از موارد از طریق تجزیه و تحلیل داده‌های منتشر شده، کشف مقدار قابل توجهی از اطلاعات خصوصی حتی پس از حذف هویت صاحبان داده‌ها ممکن است. حتی اگر داده‌ها شامل هیچ شناسه صحیحی نباشند، تجمیع اطلاعات می‌تواند یک هویت منحصر به فرد از شخص یا سازمان را مانند اثر انگشت شکل دهد [۴]. برای تضمین حفظ حریم خصوصی، داده‌ها باید با تکنیک‌های گمنام‌سازی پیشرفته‌تر پردازش شوند. با این حال، این تکنیک‌ها برای داده‌های حجیم خیلی مناسب نمی‌باشند[۱].

[1] Batch
[2] Streaming
[3] Cluster

۳. امنیت حجیم‌داده

علاوه بر حجم، سرعت تولید و تنوع داده که معمولاً به عنوان سه ویژگی بارز حجیم‌داده در نظر گرفته می‌شوند، ویژگی‌های مهم دیگری مانند صحت، اعتبارسنجی و تغییرات در حجیم‌داده وجود دارند که بر امنیت و حفظ حریم خصوصی تأثیر می گذارند. این ویژگی‌ها در زیر با توجه به تأثیر آن‌ها بر امنیت داده‌ها و حریم خصوصی حجیم‌داده بحث شده‌اند:

- **حجم:** حجم در واقع بیان کننده اندازه مجموعه داده‌ها است که در حجیم‌داده از گیگابایت تا اگزابایت یا حتی فراتر پیش می‌رود. طبیعتا حجم داده‌های زیاد نیاز به ذخیره‌سازی در سامانه‌های ذخیره‌سازی چندلایه و توزیع‌شده دارد. جابجایی داده‌ها بین لایه‌های مختلف منجر به شکل‌گیری مدل‌های تهدید جدید و پیدایش تکنیک‌های جدید برای رسیدگی به آنها شده است. مدل تهدید برای سامانه‌های مبتنی بر شبکه، توزیع شده و خودکار شامل سناریوهای مهم زیر است: محرمانه بودن و جامعیت، امنیت منبع داده و انتقال آن، دسترس‌پذیری، سازگاری، حملات تبانی، احراز هویت و حق دسترسی، ثبت وقایع و نگهداری عملیات.
- **سرعت:** سرعت نرخ جریان داده را توصیف می‌کند. داده‌ها معمولاً به صورت تکی و یا دسته‌ای وارد می‌شوند و به طور مداوم جاری و پخش می‌شوند. مانند سایر بانک‌های اطلاعاتی غیررابطه‌ای، چارچوب‌های برنامه‌نویسی توزیع شده لزوما با هدف حفظ امنیت و حریم خصوصی توسعه نیافته‌اند. در این فضا، گره‌های محاسباتی غیرامن ممکن است اطلاعات محرمانه را انتشار دهند. حملات زیرساختی به دلیل سطح بالای اتصال و وابستگی اجزاء می‌توانند بخش قابل توجهی از سامانه را به خطر اندازد. اگر سامانه نتواند احراز هویت قوی بین گره‌های توزیع شده برقرار کند، گره‌های مخرب می‌توانند در داده‌های محرمانه استراق سمع کنند.
- **تنوع:** ساختار داده‌ها را توصیف می‌کند، به این معنی که آیا داده‌ها ساخت‌یافته، نیمه ساخت‌یافته یا غیر ساخت‌یافته هستند. همواره مبحث امنیت پایگاه داده سنتی رابطه‌ای با داده‌های غیررابطه‌ای یک چالش بوده است. این سامانه‌ها اغلب بدون ملاحظات امنیتی و حفظ حریم خصوصی طراحی شده‌اند و این عملکرد منفی معمولاً به میان‌افزارها نیز منتقل می‌شود. همچنین رمزنگاری، سازماندهی داده‌ها بر اساس معنای آن‌ها را تا حدودی مختل می‌کند. پدیده جدیدی که توسط تنوع داده ارائه شده است و اهمیت قابل توجهی کسب کرده است، توانایی استنباط هویت از مجموعه داده‌های ناشناس با استفاده از داده‌های بانک‌های اطلاعاتی عمومیِ ظاهراً غیرمرتبط است. روند این استنتاج هویتی ظاهرا به دلیل حجم داده زیاد است، اما تنوع منابع داده نیز به عنوان یکی از علل این رخداد لحاظ می‌شود.
- **صحت:** صحت و اعتبار داده‌های بزرگ شامل چندین ویژگی فرعی است که در زیر شرح داده شده‌اند. صحت داده خود به صورت مستقیم یک بعد از امنیت حجیم‌داده را تشکیل می‌دهد.
    - **منبع داده:** تشخیص منبع اصلی داده‌ها به وسیله پایش فراداده یا با ابزارهایی فراتر انجام می‌پذیرد.
    - **جمع‌آوری:** تضمین درستی روش‌های بکار رفته برای جمع‌آوری اطلاعات را شامل می‌شود.

- **اصلاح:** یک مفهوم یکپارچه است که صحت را به کیفیت داده‌ها مرتبط می‌کند. به عنوان مثال، ممکن است داده‌های خام را با رفع خطاها، پر کردن مقادیر خالی، مدل سازی، کالیبراسیون مقادیر و سفارشی‌سازی جمع‌آوری داده‌ها بهبود بخشد.
- **اعتبار:** به دقت و درستی داده‌ها برای کاربرد آن اشاره دارد که به آن کیفیت داده نیز گفته می شود.

- **نوسان و تغییرات:** چگونگی تغییر ساختار داده‌ها با گذشت زمان را مشخص می‌کند که بر اصل داده و کیفیت آن تأثیر مستقیم می‌گذارد. داده‌های حجیم نیز تا حد زیادی در حال تغییر و دگرگونی هستند، زیرا داده‌های حجیم وابسته به ابزارهای تولید و جمع‌آوری داده هستند.

حجیم‌داده مسائل و چالش‌های بسیار مهمی در خصوص امنیت اطلاعات ایجاد می‌کند. در عین حال، تجزیه و تحلیل حجیم‌داده، فرصت‌های قابل توجهی را برای پیشگیری و تشخیص حملات سایبری پیشرفته وعده می‌دهد که با استفاده از داده‌های حجیم و تحلیل آنها قابل دستیابی می‌باشد. اگرچه امنیت اطلاعات برای حجیم‌داده از ابتدا وجود داشته است، این مسائل تا به حال توجه کمی را به خود جلب کرده‌اند. برخی از محققان اشاره داشته‌اند که به علت حجم زیاد داده‌ها و پیچیدگی کار با آنها، در حال حاضر حجیم‌داده برای حمله کنندگان جذاب نیست. اتحاد امنیت ابر (CSA)، یک گروه کاری که مسائل امنیتی حجیم‌داده را مطالعه می‌کند، به تازگی یک سند ارائه نموده است که ابزارهای محافظت از سامانه‌های حجیم‌داده را فهرست می‌کند: [۱۴].

۱. محاسبات امن در چارچوب‌های برنامه‌نویسی توزیع شده: استفاده از ابزارهای امنیتی در یک معماری توزیع‌شده که به صورت گسترده و همزمان به اجرا در می‌آیند، برای نمونه راهکارهای رمزنگاری توزیعی با مخازن رمز توزیع‌شده.
۲. روش‌های امنیتی برای ذخیره‌سازی داده‌های رابطه‌ای: در سامانه حجیم‌داده، داده‌های غیرساخت‌یافته که شامل انواع فایل‌ها و قالب‌های متنوع مانند تصویر، متن و غیره می‌باشند، حضور دارند که از ارتباط دادن آنها، اطلاعات مهمی استخراج می‌شود. بی‌نام کردن داده‌ها و مخدوش کردن هویت آنها به همراه درهم‌سازی محتوایی می‌توانند برای حفظ امنیت و حریم خصوصی گزینه‌های مناسبی باشند.
۳. ذخیره‌سازی امن و پایدار داده‌ها و نتایج تحلیلی: داده‌ها و به ویژه، نتایج تحلیل باید به صورت محرمانه و رمزشده در بستر داده ذخیره و بازیابی شود. همینطور، پایداری داده (برای مثال، از طریق تکرارسازی) در قبال بروز خرابی نیز اهمیت ویژه‌ای دارد.
٤. اعتبارسنجی/ فیلتر نقطه ورودی و خروجی داده: در مورد اعتبارسنجی داده‌ها که از منابع مختلف جمع‌آوری می‌شوند نیز باید دقت نظر لازم صورت گیرد، تا داده‌ها از مبدا فرآیند حجیم‌داده (جمع‌آوری) تا استنتاج و نمایش از کیفیت و اهمیت لازم برخوردار باشند و داده بی‌کیفیت و مخرب اجازه ورود پیدا نکند.

۵. امنیت بلادرنگ: وجود مکانیزم‌های امنیتی که بتوانند تهدیدات امنیتی، نفوذها، خرابکاری‌ها، و سایر تهدیدات را در پردازش‌های آنلاین به صورت لحظه‌ای تشخیص دهند، امری ضروری می‌باشد.

٦. داده‌کاوی و تجزیه و تحلیل مقیاس‌پذیر: استفاده از ابزارهای داده‌کاوی در بالا بردن سطح امنیت و حفظ حریم خصوصی به صورت تشخیص و پیش‌بینی انواع مشکلات امنیتی و رخنه‌های مختلف ضروری می‌باشد.

۷. اجرای داده امنیت‌محور به طور رمزی: قابلیت رمزنگاری یکی از مهم‌ترین راهکارهای مورد استفاده برای داده‌های مهم و بحرانی است. بدین معنی که استفاده از این داده‌ها چه در حین ذخیره‌سازی و چه درحین پردازش باید مشروط به استفاده از پروتکل‌های رمزنگاری باشد.

۸. کنترل دسترسی: این مورد یکی از مهم‌ترین چالش‌های بحث حجیم‌داده است، زیرا به دلیل ماهیت توزیع شده بسترها و زیرساخت-های آن حفظ یکپارچگی کنترل و مدیریت دسترسی کاربران به داده‌ها با سختی‌های زیادی همراه است.

۹. منبع اطلاعات (داده‌ها): تعیین منابع داده ارزشمند و مورد تایید که بستر حجیم‌داده از داده‌های آنها استفاده نماید نیز با چالش‌های مهمی روبرو است. تعیین منابع داده، اولویت داده‌ها، تعیین کیفیت، تعیین تمیز بودن داده از موارد قابل بحث در این مورد می‌باشند.

CSA این ابزارها را به چهار گروه تقسیم می‌کند: ۱) امنیت زیرساخت، ۲) حفاظت داده‌ها، ۳) مدیریت داده‌ها و ۴) امنیت واکنشی و بازیابی. در گروه نخست برای تأمین امنیت زیرساخت‌های حجیم‌داده، ذخیره‌سازی و پردازش‌های توزیع‌شده باید ایمن باشند، یعنی تا حد ممکن، هر یک از سرورها، ابزارها و روش‌های مورد استفاده به صورت مستقل، امنیت را تامین نمایند. برای تأمین حفاظت داده‌ها، انتشار اطلاعات باید حافظ حریم خصوصی باشد و از داده‌های حساس باید با استفاده از رمزنگاری و کنترل‌دسترسی جزیی و در سطح رکورد محافظت شود. مدیریت داده‌ها، راه‌حل‌های مقیاس‌پذیر و توزیع‌شده را برای تأمین مدیریت و کنترل دسترسی داده‌ها و همچنین ممیزی و نظارت پیشرفته بر روی آنها ایجاد می‌کند. در انتها، در گروه امنیت واکشی و بازیابی، داده‌های منتشرشده از منابع متنوع باید برای یکپارچگی بررسی شوند، سپس برای تشخیص رخدادهای امنیتی از الگوهای خواندن و نوشتن داده، مورد تجزیه و تحلیل آنلاین قرار گیرند و جهت حصول اطمینان از امنیت زیرساخت‌ها استفاده شوند.

۴. حریم خصوصی حجیم‌داده

در سال ۲۰۱۲ مرکز فن‌آوری اطلاعات اینتل، ۲۰۰ مدیر فن‌آوری اطلاعات را در شرکت‌های بزرگ مورد مطالعه قرار داد تا بفهمد چگونه به آنالیز داده می‌پردازند. آنها از مدیران فن‌آوری اطلاعات خواستند تا استانداردهایی را که برای تجزیه و تحلیل حجیم‌داده نیاز است بیان کنند. جواب‌ها اینگونه بود: امنیت داده‌ها، فن‌آوری برای نگه داشتن حریم خصوصی داده‌های مشتریان، شفافیت داده، تعیین معیار عملکرد، قابلیت همکاری سامانه و داده‌ها [۹].

معنای کلی حفظ حریم خصوصی جلوگیری از فاش شدن اطلاعات حساس است. در مورد حجیم‌داده، حجم بالا و انواع مختلف داده‌ها جمع‌آوری شده‌اند که ممکن است حاوی اطلاعات شخصی افراد یا سازمان‌ها باشند. برای جلوگیری از آشکارسازی تمام این اطلاعات شخصی و حساس، اصطلاح حریم خصوصی حجیم‌داده به کار می رود[۶].

حریم خصوصی حجیم‌داده شامل دو جنبه است:

- حفاظت از حریم خصوصی شخصی در طی جمع‌آوری داده‌ها: منافع شخصی، علاقمندی‌های شخصی، عادت‌ها و ویژگی‌های ظاهری و یا باطنی در مورد کاربران و یا شرکت‌ها ممکن است به آسانی بدست آورده شود و کاربران از این موضوع آگاه نباشند.
- حفاظت از حریم خصوصی شخصی در طی ذخیره‌سازی و انتقال داده‌ها: داده‌های شخصی حریم خصوصی ممکن است در طی ذخیره‌سازی، به هنگام انتقال و یا استفاده، حتی اگر با اجازه کاربران به دست آورده شود، به بیرون درز کند. برای مثال، در حال حاضر، فیس‌بوک به عنوان یک شرکت حجیم‌داده، با بیشترین داده‌های اجتماعی فرض می‌شود. بر اساس گزارشی که توسط یک محقق (Ron Bowes) انجام شد، وی موفق شد از طریق جمع‌آوری صفحات عمومی کاربران فیسبوک از کاربرانی که برای تغییر تنظیمات حریم خصوصی‌شان دچار مشکل شده بودند، اطلاعات مهمی را بدست آورد. این محقق، داده‌ها را در یک بسته ۲.۸ گیگا بایتی بسته‌بندی کرد و از طریق یک فایل بیت‌تورنت آنرا منتشر کرد. [۱۵]

ظرفیت تجزیه و تحلیل حجیم‌داده ممکن است منجر به استخراج حریم خصوصی از اطلاعات به ظاهر ساده شود. بنابراین، حفاظت حریم خصوصی یک مسئله جدید و چالش برانگیز خواهد بود [۹].

۱-٤ مشکلات و چالش‌های حریم خصوصی

حفظ حریم خصوصی چالش بزرگی در حجیم‌داده است. عوامل مختلفی که نقش مهمی در حریم خصوصی حجیم‌داده ایفا می‌کنند و آن را چالش برانگیزتر می‌کنند به شرح زیر هستند:

- حریم خصوصی مبتنی بر متن: اساسی‌ترین چالش حریم خصوصی تعریف حریم خصوصی مبتنی بر متن است که در آن هر مجموعه داده معنای متفاوتی در متن‌های متفاوت دارد و تصمیم‌گیری در مورد اینکه کدام مجموعه داده‌ها در آن محتوا حساس هستند، بسیار سخت است. برای مجموعه داده‌های مختلف در متن‌های متفاوت، پیدا کردن اینکه چقدر حریم خصوصی مورد نیاز است، دشوارتر خواهد بود و بنابراین رعایت حریم خصوصی در آن چالش‌برانگیز می‌شود.
- مجموعه داده‌های متراکم و مربوط به هم: مجموعه‌های داده در حجیم‌داده به یکدیگر مرتبط هستند. بنابراین، اگر آشکار کننده‌ای در حریم خصوصی یک مجموعه داده وجود داشته باشد ممکن است منجر به آشکار کننده حریم خصوصی در سایر مجموعه داده‌ها شود. تهدید حریم خصوصی همچنین ممکن است در زمان پردازش داده‌ها رخ دهد، از آنجا که این احتمال وجود دارد که یک

اطلاعات از یک مجموعه داده هنگام پردازش مجموعه دیگری مورد نیاز باشد. چنین مجموعه داده‌های یکپارچه و مرتبط به هم به عنوان شبه شناسه تصور می‌شود، که یک تهدید بزرگ حریم خصوصی است.

- مدل‌سازی تهدید: مدل‌سازی تهدید یک تکنیک ساخت‌یافته برای طراحی یک راه‌حل حفظ حریم خصوصی توسط شناسایی قبلی اهداف حریم خصوصی و حملات است که ممکن است رخ دهد. بنابر این، چنین رویکردی نیازمند یک ایده و مدل شفاف در مورد نوع تهدید حریم خصوصی و چگونگی رسیدگی به آنها است. البته چنین شناسایی از تهدید و ارائه راه حل طراحی شده برای آن تهدید کار آسانی در حجیم‌داده نیست، زیرا در حال حاضر چنین فن‌آوری که بتواند به این نوع از مجموعه داده‌های متنوع و بزرگ رسیدگی کند، وجود ندارد.
- بودجه بندی حریم خصوصی: هزینه ای که باید صرف حفظ حریم خصوصی حجیم‌داده شود، موضوع بسیار چالش برانگیز دیگر است. برآورد هزینه از نظر نیازهای محاسباتی انجام می‌شود، بدین معنی که امکان انتخاب تکنیکی که از نظر محاسباتی بسیار گران باشد وجود ندارد. البته هنوز در رویکردهای محاسباتی برای حفظ حریم خصوصی حجیم‌داده مشکلات دانشی و فنی وجود دارد.
- سیاست‌ها و مجوزهای قانونی: امروزه داده به عنوان یک دارایی ارزشمند در نظر گرفته می‌شود و به علت اهمیت در حال افزایش آن تاکنون، سیاستها و قوانین نقش بسیار مهمی ایفا می‌کند. کشورهای مختلف قوانین مختلفی برای داده و حریم خصوصی آن دارند. یک تکنیک جدید مورد نیاز است که بتواند تمام جنبه‌های قانونی را دنبال کند. حفظ حریم خصوصی با تحقق تمام این محدودیت-های قانونی یک مانع بزرگ است[۶].

۲-٤ نقش کاربران و رویکردهای حریم خصوصی

کاربران نقش‌های مختلفی در دسترسی به داده‌ها و استفاده از آن دارند و نوع نگرانی‌های حریم خصوصی کاربران نیز وابسته به نقش آنها می‌باشد. به عبارت دیگر، تعاریف و رویکردهای حفظ حریم خصوصی متناسب با نقش کاربران متفاوت می‌باشد [۱۲]

- برای ارائه دهنده داده‌ها، هدف حفظ حریم خصوصی، کنترل موثر مقدار داده حساسی است که به دیگران نشان داده شده است. برای رسیدن به این هدف، ارائه دهنده داده‌ها می‌تواند از ابزارهای امنیتی برای محدود کردن دسترسی دیگران به داده‌هایش استفاده کند یا داده‌های خود را در مزایده بفروشد تا حریم خصوصی از دست رفته را جبران کند (مثلا، برای حریم خصوصی از دست رفته به میزان کافی غرامت دریافت کند) و یا داده‌های خود را برای پنهان کردن هویت واقعی خودش تعریف کند.
- برای جمع‌آوری کننده داده‌ها، هدف حفظ حریم خصوصی، انتشار داده‌های مفید برای کاوشگران داده است، بدون اینکه هویت ارائه کنندگان داده‌ها و اطلاعات حساس در مورد آنها افشاء شود. برای رسیدن به این هدف، جمع‌آوری کننده داده‌ها نیاز به توسعه مناسب مدل‌های حریم شخصی دارد تا امکان از دست دادن حریم خصوصی را تحت حملات مختلف محدود نماید و تکنیک‌های گمنام‌سازی را روی داده‌ها اجرا نماید.

- برای کاوشگران داده‌ها، هدف حفظ حریم خصوصی، رسیدن به نتایج صحیح داده کاوی است به نحوی که اطلاعات حساس را در فرآیند داده کاوی یا در نتایج استخراج (کاوش) به صورت پنهان نگه دارد. برای رسیدن به این هدف، کاوشگر داده می‌تواند، قبل از اینکه الگوریتم‌های کاوش خاص روی داده اعمال شود، یک روش مناسب برای تغییر داده‌ها انتخاب کند یا از پروتکل‌های محاسباتی امن برای اطمینان از امنیت داده‌های خصوصی و اطلاعات حساس موجود در مدل یاد گرفته شده استفاده کند.
- برای تصمیم‌گیرنده، هدف حفظ حریم خصوصی، قضاوت صحیح در مورد اعتبار نتایج داده‌کاوی می‌باشد که به آن دست پیدا کرده است. برای رسیدن به این هدف می تواند از تکنیک‌هایی نظیر ردیابی برگشت به عقب استفاده کند، یا دسته‌کننده‌ای برای تفاوت قائل شدن بین اطلاعات درست از اطلاعات نادرست بسازد.

۵. معماری‌های حجیم‌داده

با بررسی معماری‌های حجیم‌داده که توسط شرکت‌های مختلف پیشنهاد شده‌اند، یک سری ویژگی‌های مشترک در ساختار آنها قابل مشاهده می‌باشد. هر یک از پلتفرم‌های مورد بررسی از مولفه‌های مشترکی (نظیر مدیریت و ذخیره‌سازی داده، تجزیه و تحلیل داده و غیره) در معماری خود استفاده کرده‌اند که در ادامه به طور خلاصه به بیان ویژگی‌های معماری‌های مرجع ارائه شده می‌پردازیم.

۵-۱ معماری مارکوس

آقای مارکوس، یکی از مشارکت‌کنندگان و محققین پیشرو در حوزه حجیم‌داده است. ایشان یک مدل معماری لایه‌ای را برای حجیم‌داده ارائه داده است [۱۶]. این مدل شش مؤلفه زیر را در بر می‌گیرد:

- منابع داده: این مؤلفه ورودی داده‌های خارج از سامانه حجیم‌داده را برای مؤلفه‌های داخلی حجیم‌داده فراهم می‌کند.
- رابط برنامه و کاربری: این مؤلفه برنامه‌های کاربردی و همینطور رابط کاربری (به عنوان مثال بصری‌سازی) حجیم‌داده را فراهم می‌کند.
- پایگاه داده‌های تحلیلی و رابط‌ها: این چارچوب ادغام بانک‌های اطلاعاتی را در معماری حجیم‌داده پیشنهاد می‌کند. این‌داده‌ها می‌توانند دارای پایگاه داده‌های مقیاس‌پذیر با داده‌های استخراج شده از داده‌های ذخیره شده باشند. این چارچوب پایگاه داده‌ها و رابط‌های متنوعی را مانند پایگاه داده تحلیلی، پایگاه داده عملیاتی به همراه رابط‌های دسته‌ای و تعاملی دسترسی به داده‌ها فراهم می‌کند.
- جریان مقیاس‌پذیر و پردازش داده‌ها: این مؤلفه فیلترها و مکانیزم‌های تبدیل بین جریان داده‌ها از منابع خارجی و ساختمان داده‌های مورد استفاده در سامانه‌های حجیم‌داده داخلی را فراهم می‌کند.

- زیرساخت مقیاس‌پذیر: این چارچوب زیرساخت‌های مقیاس‌پذیر را تعیین می‌کند که می‌توانند افزودن منابع جدید را (مثلا سرور و یا فضای ذخیره‌سازی) به سامانه تسهیل نماید. چارچوب‌های ممکن شامل ابرهای پردازشی عمومی و یا خصوصی و همچنین ذخیره‌سازهای مقیاس‌پذیر و تزویع شده داده می‌باشند.
- خدمات پشتیبانی: این چارچوب خدمات مورد نیاز برای پیاده‌سازی و مدیریت سامانه‌های داده‌های حجیم را مشخص می‌کند. اجزاء فرعی مشخص شده در خدمات پشتیبانی در زیر نام برده شده‌اند:
  - طراحی، توسعه و استقرار ابزارها: این چارچوب ابزارها و کتابخانه‌های آماده سطح بالا برای پیاده‌سازی و اجرای برنامه‌های حجیم‌داده ارائه می‌دهد. سطح مهارت لازم برای توسعه‌دهندگان بنگاه‌های اقتصادی و دولتی باید متناسب با کاربردهای حجیم‌داده ارتقا یابد.
  - امنیت: در این قسمت به عدم پشتیبانی از استانداردی مناسب برای پرداختن به امنیت داده‌ها و حریم خصوصی اشاره شده است. این چارچوب عنوان می‌کند که تنها احراز هویت Kerberos برای Hadoop و Knox وجود دارد.
  - مدیریت فرآیندها: این چارچوب یادآور می‌شود که توسعه‌دهندگان ابزارهای مدیریت فرآیند را برای تقویت پیاده‌سازی اولیه منبع‌باز عرضه می‌کنند.
  - مدیریت منابع و سامانه: این چارچوب خاطرنشان می‌کند که ابزارهای مدیریت سامانه منبع‌باز نابالغ هستند. با این حال، ابزارهای مقیاس‌پذیر تجاری در دسترس هستند.

۲-۵ معماری مایکروسافت

مایکروسافت یک معماری مرجع حجیم‌داده‌ای را تعریف می‌کند که بتواند چهار قابلیت عملکردی کلیدی را تامین کند [۱۶]. در ادامه به طور خلاصه، مؤلفه‌های این معماری بیان می‌شود:

- منابع داده : طبق تعریف مایکروسافت، "داده‌های موجود در حجیم‌داده" برای یک هدف خاص جمع‌آوری می‌شوند و از اینرو، اشیاء داده را به شکلی ایجاد می‌کند که از کاربرد شناخته شده در زمان جمع‌آوری داده‌ها پشتیبانی کند. پس از جمع‌آوری داده‌ها، می‌توان برای اهداف مختلفی از آن استفاده کرد. البته بعضی از این اهداف و کاربردها در زمان جمع‌آوری نامشخص است. مایکروسافت همچنین توضیح می‌دهد که منابع داده را می‌توان با چهار ویژگی که همه آنها مستقل از محتوا و مقادیر داده‌ها است، طبقه بندی کرد: حجم، سرعت، تغییرپذیری و تنوع.
- تبدیل داده: دومین مؤلفه معماری مرجع حجیم‌داده که مایکروسافت شرح داده است، تبدیل و انتقال داده است. مایکروسافت این مرحله را مرحله‌ای برای پردازش و تبدیل داده‌ها به روش‌های مختلف برای استخراج ارزش و دانش از اطلاعات خام تعریف می‌کند. هر تابع تبدیل ممکن است مرحله پیش‌پردازش خاص خود را داشته باشد. ممکن است از زیرساخت‌های داده‌های تخصصی مختلفی

استفاده کند که برای شرایط موردنیاز مناسب باشد. همچنین ممکن است حریم خصوصی و سایر ملاحظات مربوط به سیاست گذاری و قابلیت توسعه و همکاری نیز وجود داشته باشد. این عملیات در چهار مرحله جمع‌آوری داده، تجمیع داده، تطبیق داده و فراداده و داده‌کاوی انجام می‌گیرد.

- زیرساخت داده: مایکروسافت زیرساخت‌های حجیم‌داده را به عنوان مجموعه‌ای از نرم‌افزارهای ذخیره‌سازی داده یا بانک اطلاعاتی، سرورها، ذخیره‌سازی و شبکه‌سازی تعریف می‌کند که برای پشتیبانی از توابع تبدیل داده‌ها و ذخیره اطلاعات در صورت لزوم استفاده می‌شود. علاوه بر این، مایکروسافت برای دستیابی به راندمان‌های بالاتر، زیرساخت‌ها را به عنوان واسطه‌ای تعریف می‌کند. داده‌ها از حجم، تنوع، تغییرات و سرعت‌های مختلف معمولاً با استفاده از فن‌آوری‌های محاسبات و ذخیره‌سازی متناسب با خصوصیات داده ذخیره و پردازش می‌شوند.
- استفاده و کاربرد داده: آخرین مؤلفه چارچوب معماری حجیم‌داده استفاده از داده است. داده پس از طی کردن مسیر از طریق زیرساخت‌های معین، به صورت نتیجه نهایی می‌تواند در قالب‌های مختلف، دانه‌بندی مختلف و تحت ملاحظات امنیتی مختلف ارائه شود.

۳-۵ معماری آمستردام

دانشگاه آمستردام، یک چارچوب بزرگ حجیم‌داده را به عنوان بخشی از کل زیرساخت ابری و رایانش ابری در قالب پنج زیربخش ارائه می‌دهد [۱۶]. هر یک از پنج زیربخش در ادامه بررسی شده است.

- مدل داده: چارچوب معماری حجیم‌داده شامل مدل‌ها، ساختارها و انواع داده‌ها است که داده‌های گوناگون تولید شده توسط منابع داده‌های مختلف را پشتیبانی می‌کند.
- تجزیه و تحلیل حجیم‌داده: دراین معماری، تحلیل حجیم‌داده به عنوان مؤلفه‌ای است که از معماری‌ها و فن‌آوری‌های پردازش سریع یا HPC استفاده می‌کند. علاوه بر این، انتظار می رود که تجزیه و تحلیل حجیم‌داده به صورت عمودی و افقی مقیاس‌پذیر باشد. این امر می تواند به طور طبیعی هنگام استفاده از پلتفرم مبتنی بر رایانش ابری و ادغام و ارتباط بین ابرها محقق شود. این معماری قابلیت‌های تجزیه و تحلیل‌های متنوعی مانند پالایش داده، پردازش بلادرنگ، پردازش جریانی و دسته‌ای و غیره را پشتیبانی می‌کند که توسط معماری پردازش سریع پشتیبانی می‌شوند.
- مدیریت حجیم‌داده: این معماری، خدمات مدیریتی را در قالب اجزا زیر بیان می‌کند:
  - تهیه نسخه پشتیبان از داده‌ها، تکرار داده، پیشرو بودن
  - ثبت، نمایه سازی/ جستجو، فراداده، هستی شناسی، فضای نام

- زیرساخت حجیم‌داده: این معماری، زیرساخت‌های حجیم‌داده را تعریف می‌کند که نیاز به دسترسی گسترده به شبکه و زیرساخت-های محاسباتی و ذخیره‌سازی به جهت ارائه عملکرد قابل اطمینان دارد.
- امنیت حجیم‌داده: معماری مرجع، امنیت داده‌های حجیم را بدین شکل توصیف می‌کند که باید از داده‌ها در حالت غیرپردازشی و در حین پردازش محافظت کند، از محیط‌های پردازش قابل اعتماد و عملکرد و کاربرد اطمینان حاصل کند، کنترل دسترسی با جزییات را فراهم کند و از اطلاعات شخصی کاربران محافظت کند.

٤-٥ معماری آی بی ام

این مدل مرجع حجیم‌داده، یک چارچوب بزرگ داده را ارائه می‌دهد که می‌تواند در چهار بلوک عملکردی اصلی خلاصه شود، که در زیر شرح داده شده‌اند [۱۶].

- اکتشاف داده: این معماری تجزیه و تحلیل داده‌ها را به عنوان اولین مرحله از فهم و درک از منابع داده برمی‌شمرد و در ادامه کیفیت داده‌ها و ارتباط آن با سایر عناصر داده را توضیح می‌دهد. کشف داده همچنین از نمایه سازی، جستجو و ارتباطات داده‌ای پشتیبانی می‌کند. این کشف مستقل از منابع داده است که شامل بانک‌های اطلاعاتی رابطه‌ای، فایل‌های مسطح و سامانه‌های مدیریت محتوا است. ذخیره‌سازی داده از داده‌های ساخت‌یافته، نیمه‌ساخت‌یافته یا بدون ساختار پشتیبانی می‌کند.
- تحلیل داده: معماری حجیم‌داده، اجرای هر دو نوع عملیات ذخیره‌سازی و تجزیه و تحلیل را بر روی یک پلتفرم یکسان پیشنهاد می‌دهد که این رویکرد بر خلاف رویکرد سنتی که نرم افزار تجزیه و تحلیل بر روی زیرساخت‌های مجزا اجرا می‌کند، است. منطق این است که محیط‌های داده برای دسترسی سریع‌تر به داده‌ها بهینه‌سازی می‌شوند، اما لزوماً برای محاسبات ریاضی پیشرفته نیستند. بنابراین، با تجزیه و تحلیل به عنوان یک بار کاری مشخص و مجزا که باید در یک زیرساخت جداگانه اداره شود، برخورد می شود. لیکن در معماری حجیم‌داده، به دلیل توزیع‌شدگی حجم زیادی از داده و نیاز به انتقال داده‌ها، این دو عملیات بهتر است بر روی یک پلتفرم انجام گیرند. از جمله این کارکردهای تحلیلی می‌توان هوش تجاری و گزارشات، برنامه‌های کاربردی، برنامه‌های صنعتی و تحلیل‌های پیشگویانه نام برد.
- زیرساخت حجیم‌داده: یکی از اهداف اصلی یک بستر حجیم‌داده باید کاهش زمان چرخه تحلیلی باشد (یعنی مدت زمانی که برای کشف و تبدیل داده‌ها، توسعه و ارزش‌گذاری مدل‌ها و تجزیه و تحلیل و انتشار نتایج به طول می انجامد). یک بستر حجیم‌داده نیاز به تعامل با متداول‌ترین ابزارهای تحلیلی دارد. این نوع کارکرد نیاز به مجموعه‌ای غنی از الگوریتم‌های "موازی‌سازی" دارد که برای اجرا در حجیم‌داده ایجاد و آزمایش شده‌اند. همچنین باید قابلیت نمایش و انتشار نتایج را به روشی بصری و با کاربرد آسان فراهم آورد.

- ادغام اطلاعات و مدیریت: آخرین مؤلفه چارچوب حجیم‌داده، ادغام و اداره همه منابع داده است. این شامل ادغام داده‌ها، کیفیت داده‌ها، امنیت، مدیریت چرخه زندگی داده‌ها و مدیریت ارشد داده‌ها است.

۵-۵ معماری اوراکل

معماری ارائه شده توسط اوراکل دارای چهار زیربخش است که در زیر معرفی می‌گردند [۱۶].

- تحلیل اطلاعات: بخش تحلیل اطلاعات دارای دو فضای اصلی است. تحلیل توصیفی و تجزیه و تحلیل پیش‌گویانه که هر کدام به صورت جزیی‌تر دارای زیربخش‌هایی هستند. از زیربخش‌های تحلیل توصیفی می‌توان گزارش‌گیری و داشبورد را نام برد که بیشتر به بیان نتایج تحلیل‌ها می‌پردازند. در مورد تجزیه و تحلیل پیش‌گویانه نیز تحلیل‌های آماری و معنایی، داده‌کاوی و متن‌کاوی به همراه تحلیل‌های نگاشت/کاهش وجود دارند.
- ارائه و تامین اطلاعات: این مؤلفه عملیات کشف، تبدیل و پردازش داده‌های حجیم ساخت‌یافته، بدون ساختار و داده‌های جریانی را انجام می‌دهد. این امر توسط بانک اطلاعاتی عملیاتی و هم انباره داده پشتیبانی می‌شود.
- منابع داده: معماری اوراکل، از انواع ساختار داده‌ای پشتیبانی می‌کند. داده‌هایی همچون فایل‌های توزیع شده، داده‌های جریانی، داده‌های رابطه‌ای، داده‌های غیرساخت‌یافته و داده‌های فضایی/رابطه‌ای در این معماری مورد استفاده قرار می‌گیرند.
- خدمات زیرساختی: مباحثی همچون سخت‌افزار، سامانه عامل، ذخیره‌سازی، امنیت، شبکه و ارتباطات و مدیریت جامع در این دسته خدمات قرار می‌گیرند.

همانگونه که مشاهده می‌شود در برخی از معماری‌های مطرح شده، امنیت به عنوان یک بخش مجزا در سامانه‌های حجیم‌داده بیان و الزاماتی برای آن تعیین شده است، اما در برخی دیگر به صورت ادغام شده در بستر سامانه توزیع شده در نظر گرفته شده است و گاها حتی از ارائه قابلیت‌ها و یا استانداردهای امنیتی صرف‌نظر می‌شود. در ادامه به بیان معماری مرجع و اجزای آن به تفصیل پرداخته می‌شود.

۶-۵ معماری مرجع

معماری حجیم‌داده یا معماری مرجع NBDRA، توسط گروه کاری عمومی‌حجیم‌داده NBD-PWG در مؤسسه ملی فناوری و استانداردها (NIST)که یک گروه معتبر استانداردسازی در آمریکا است، ارائه شده است [۱۷]. این مدل بعد از دریافت و مقایسه ۹ معماری مرجع برای حجیم‌داده‌ها از سازمان‌ها، شرکت‌ها و دانشگاه‌های مختلف زیر پیشنهاد شده است:

- ET Strategies
- Microsoft
- University of Amsterdam
- IBM
- Oracle

- EMC/Pivotal
- SAP
- sight Consulting
- LexisNexis

نکته قابل توجه این است که معماری مرجع NBDRA وابسته به یک فناوری یا سازنده خاص یا یک زیرساخت خاص نیست. NBDRA از لحاظ منطقی از پنج جزء کارکردی اصلی تشکیل شده است که با یکدیگر در ارتباط هستند. همچنین دو نقش مهم مدیریت و امنیت با تمام اجزاء اصلی دیگر در این مدل عجین بوده و با آنها در ارتباط هستند. معماری مرجع NBDRA طوری طراحی شده است که افراد مختلف شامل مهندسان سیستم، دانشمندان داده، توسعه‌دهندگان نرم افزار، معماران داده و تصمیم گیران بتوانند راهکارهای مورد نظر خود را برای مواجهه با مسائلی که روش‌های بسیار متفاوتی برای حل آنها در زیست‌بوم حجیم‌داده وجود دارد، توسعه دهند. همچنین این معماری مرجع طوری طراحی شده که بتواند در حوزه‌های متنوع با مدل‌های مختلف کسب‌وکار و طرح های تجاری متفاوت مورد استفاده قرار گیرد.

معماری مرجع NBDRA در راستای دو محور ترسیم شده است: زنجیره ارزش اطلاعات (محور افقی) و زنجیره ارزش فناوری اطلاعات (محور عمودی). زنجیره ارزش در راستای محور افقی، با جمع‌آوری داده‌ها، یکپارچه‌سازی آنها، آنالیز و اعمال نتایج حاصل می‌شود. زنجیره ارزش در راستای محور عمودی، با فراهم کردن شبکه، زیرساخت‌ها، بسترها، ابزارهای کاربردی و سرویس‌های دیگر فناوری اطلاعات جهت میزبانی و عملیاتی کردن حجیم‌داده‌ها شکل می‌گیرد. در محل تلاقی دو محور، کاربردهای داده‌های عظیم است که نشان می‌دهد آنالیز داده و پیاده سازی آن برای متولیان حجیم‌داده ایجاد ارزش می‌کند. دلیل استفاده از واژه «فراهم کننده» در دو جزء فراهم کننده چارچوب و فراهم کننده کاربردها به این‌دلیل است که نشان دهد که این‌دو موجودیت، وظیفه فنی خاصی را در سیستم پیاده‌سازی یا از آن پشتیبانی می‌کنند.

بازیگران و متولیان حوزه حجیم‌داده‌ها در معماری مرجع NBDRA به هفت گروه تقسیم‌بندی شده‌اند، که عبارتند از:

- هماهنگ کننده سامانه
- فراهم کننده داده
- فراهم کننده کاربرد حجیم‌داده
- فراهم کننده چارچوب حجیم‌داده
- مصرف کننده داده
- موجودیت امنیت و شخصی‌سازی
- موجودیت مدیریت

در ادامه هر یک از گروه‌های فوق توضیح داده می‌شود.

۵-۶-۱ هماهنگ‌کننده سیستم

این موجودیت وظیفه فراهم کردن نیازمندی‌های فراگیری را که می‌بایست توسط سیستم برآورده شود بر عهده دارد. این نیازمندی‌ها عبارتند از: سیاست، حکومت، ساختار، منابع و نیازمندیهای تجاری و همچنین نیازمندی‌های حسابرسی و نظارت بر فعالیت‌ها. هماهنگ کننده سیستم، نیازمندی‌های سیستم و طراحی‌های سطح بالا و نظارت سیستم داده‌ای را فراهم می‌کند. اگرچه نقش هماهنگ‌کننده سیستم قبل از سیستم‌های حجیم‌داده نیز وجود داشته است، ولی برخی از فعالیت‌های طراحی مرتبط با آن، در حوزه حجیم‌داده تغییر یافته است. معمولا هماهنگ کننده سیستم، اختیار و نقش‌های ویژه بیشتری نسبت به دیگر بازیگران دارد و عملیات سیستم حجیم‌داده را مدیریت و تنظیم می‌کند. دیگر وظیفه هماهنگ کننده سیستم، پیکربندی و مدیریت اجزای دیگر معماری حجیم‌داده است، به نحوی که بتواند یک یا چند بار کاری، که معماری برای اجرای آن طراحی شده است را اجرا کند. علاوه بر آن، ممکن است که هماهنگ کننده سیستم از طریق موجودیت مدیریت، بارهای کاری و کل سیستم را پایش کند تا تضیمن کند که نیازمندی‌های کیفیت سرویس مشخص شده برای هر بار کاری ارائه می‌شود. در حقیقت، ممکن است هماهنگ کننده سیستم به صورت پویا منابع فیزیکی یا مجازی اضافی مورد نیاز را فراهم کرده و تخصیص دهد تا به نیازمندیهای بار کاری که ناشی از تغییرات یا افزایش‌های ناگهانی در داده‌ها یا تعداد تعاملات یا تعداد کاربران است، پاسخ دهد.

۵-۶-۲ فراهم کننده داده

موجودیت فراهم کننده داده، داده را در دسترس خود یا دیگران قرار می‌دهد. این موجودیت در جهت انجام رسالت خود، تجریدی از انواع مختلف منابع داده (نظیر داده خام یا داده‌هایی که توسط سیستم‌های دیگر آماده و تبدیل شده است) را ایجاد می‌کند و از طریق رابط‌های تعاملی مختلف، آنها را در دسترس قرار می‌دهد. موجودیتی که این نقش را ایفا می‌کند، می‌تواند بخشی از سیستم حجیم‌داده‌ها باشد، بخش داخلی یک سازمان‌در سیستم دیگری باشد و یا خارج از سازمانی باشد که سیستم را هماهنگ می‌کند.

معمولا تعامل میان فراهم کننده داده و فراهم کننده کاربردهای حجیم‌داده در سه مرحله صورت می‌پذیرد: ایجاد ارتباط، انتقال داده و قطع ارتباط. مرحله ایجاد ارتباط می‌تواند توسط هر یک از طرفین آغاز شود و معمولا شامل سطوحی از تشخیص هویت و اعتبارسنجی می‌باشد. البته این مراحل می‌تواند بسیار ساده بوده و به این صورت انجام شود که یک طرف ارتباط، سوکت خود را برای یک پورت شناخته شده در طرف مقابل باز کند و ارتباط برقرار شود و در نهایت آنرا ببندد.

۵-۶-۳ فراهم کننده کاربرد حجیم‌داده

موجودیت فراهم کننده کاربرد حجیم‌داده، تغییرات مربوط به چرخه عمر داده‌ها را اجرا می‌کند تا نیازمندی‌هایی که توسط هماهنگ کننده سیستم ایجاد شده است، پاسخ داده شود. در این موجودیت، توانمندی‌های عمومی موجود در چارچوب‌های حجیم‌داده با یکدیگر ترکیب می‌شود تا یک سیستم داده‌ای مشخص ایجاد شود.

فراهم کننده کاربرد حجیم‌داده، وظیفه اجرای مجموعه‌ای از عملیات در راستای چرخه عمر داده‌ها را به عهده دارد به نحوی که بتواند پاسخگوی نیازمندی‌های امنیت و شخصی‌سازی علاوه بر نیازمندیهای تعیین شده توسط هماهنگ کننده سیستم باشد. فراهم کننده کاربرد حجیم‌داده در واقع جزئی از معماری است که منطق تجاری و کارکردهایی که می‌بایست توسط معماری اجرا شود را در هم می‌آمیزد. فعالیت‌های اصلی فراهم کننده کاربرد عبارت است از:

- جمع‌آوری: این فعالیت شامل ایجاد ارتباط با فراهم کننده داده و برقراری رابط تعاملی با آن می‌باشد.
- آماده‌سازی: این فعالیت شامل اعتبارسنجی (بررسی فرمت، جمع‌های کنترلی یا درهم‌سازی)، تمیزسازی (حذف فیلدها یا نمونه‌های بد) حذف داده‌های خارج از محدوده، استانداردسازی، فرمت بندی مجدد و یا بسته‌بندی می‌باشد.
- تحلیل: پیاده‌سازی و اجرای منطق پردازش حجیم‌داده که توسط هماهنگ کننده سیستم مشخص می‌شود، می‌باشد.
- بصری‌سازی: هدف این فعالیت این است که داده را به نحوی فرمت دهی و ارائه کند که به صورت بهینه، دانش و معانی را منتقل کند. آماده سازی بصری سازی ممکن است شامل ایجاد یک گزارش متنی یا تفسیر نتایج تحلیل به صورت گرافیکی باشد.
- دسترسی: ارتباط یا تعامل با مصرف کننده داده را فراهم می‌کند.

اجرای فعالیت‌های فوق برای هر کاربرد متفاوت بوده و در نتیجه گزینه مناسبی برای استانداردسازی وجود ندارد. اگرچه بسیاری از فعالیت‌های مذکور در سیستم‌های پردازش داده سنتی نیز وجود دارد، ولی ویژگی‌های حجم، تنوع، سرعت و تغییرپذیری که در سیستم‌های حجیم‌داده وجود دارد باعث می‌شود که پیاده‌سازی این فعالیت‌ها نیاز به تغییرات اساسی داشته باشد. الگوریتم‌ها و مکانیزم‌های موجود در پیاده‌سازی‌های پردازش داده سنتی می‌بایست تنظیم شده و برای مواجهه با حجیم‌داده بهینه شود.

۵-۶-۴ فراهم کننده چارچوب حجیم‌داده

این موجودیت، سرویس‌ها یا منابع عمومی‌که قرار است توسط فراهم کننده کاربرد حجیم‌داده جهت ساختن یک کاربرد مشخص مورد استفاده قرار گیرد را فراهم می‌کند. تعداد زیادی از اجزای جدیدی وجود دارد که فراهم کننده کاربردهای حجیم‌داده‌ها می‌تواند از بین آنها جهت ساختن یک سیستم مشخص انتخاب کند که از بین آنها می‌توان به پنج بخش زیرساخت (شبکه و محیط و ...)، بستر داده، بستر پردازش، بستر پیام‌رسانی و بستر مدیریت منابع اشاره کرد.

۵-۶-۵ مصرف کننده داده

مصرف کننده داده، خروجی با ارزش سامانه حجیم‌داده را دریافت می‌کنند. این واحد در بسیاری از زمینه‌ها همان رابط‌های تعاملی کاربردی را دریافت می‌کند که فراهم کننده داده در اختیار فراهم‌کننده کاربردهای حجیم‌داده‌ها قرار می‌دهد. علاوه بر آن، مصرف کننده داده مشابه نقش فراهم کننده داده، می‌تواند یک کاربر انتهایی واقعی یا یک سامانه دیگر باشد. فعالیت‌هایی که برای نقش مصرف کننده داده تعریف شده است شامل موارد زیر می‌باشد:

- جستجو و بازیابی
- آنالیز کردن به صورت محلی
- گزارش‌گیری
- بصری‌سازی داده

مصرف کننده از رابط‌های تعاملی یا سرویس‌هایی که توسط فراهم کننده کاربرد حجیم‌داده فراهم می‌شود استفاده می‌کند تا به اطلاعات مورد علاقه خود دست پیدا کند. این تعاملات می‌تواند شامل گزارش‌گیری داده، بازیابی داده و تفسیر داده باشد. تعاملات می‌توانند مبتنی بر درخواست باشند که طی آن مصرف کننده داده، تعامل یا دستور را آغاز می‌کند و فراهم کننده کابرد حجیم‌داده پاسخ او را می‌دهد. این تعامل می‌تواند شامل ایجاد گزارش‌ها یا حرکت در مسیر بطن‌داده‌ها با استفاده از توابع هوش تجاری که توسط فراهم کننده کاربرد حجیم‌داده فراهم شده است، باشد. در تمامی موارد، موجودیت امنیت و شخصی سازی در معماری حجیم‌داده از محرمانگی، هویت سنجی و اعتبارسنجی میان مصرف کننده داده و معماری پشتیبانی می‌کند.

۵-۶-۶ موجودیت امنیت و شخصی‌سازی

مباحث امنیت و شخصی سازی، تمامی‌اجزای معماری مرجع NBDRA را تحت تأثیر خود قرار می‌دهد. این واحد با واحد هماهنگ کننده سیستم جهت دریافت سیاست، نیازمندی‌ها و نظارت‌ها تعامل می‌کند. در معماری مرجع NBDRA نقش امنیت و شخصی سازی به صورت بسیار کلی ارائه شده است و جزئیات آن ارائه نشده است و خود نیازمند یک معماری مرجع جداگانه است.

۵-۶-۷ موجودیت مدیریت

ویژگی‌های خاص حجیم‌داده نیازمند یک بستر مدیریت تطبیق‌پذیر جهت ذخیره‌سازی، پردازش و مدیریت داده در سامانه‌های پیچیده است که بتواند هر دو جنبه سیستمی و مباحث داده‌ای را مدیریت کند. به بیان‌دیگر، با توجه به ویژگی های حجیم‌داده‌ها نظیر حجم زیاد، سرعت تولید زیاد، تنوع داده‌ها و تغییرپذیری داده‌ها، ناگزیر نیازمند یک سیستم و بستر نرم افزار مدیریتی تطبیق‌پذیر هستیم تا پایش، تنظیم و مدیریت نرم افزارها و بسته‌های نرم‌افزاری و همچنین مدیریت منابع و پایش کارآیی را به صورت اتوماتیک انجام دهد. مدیریت حجیم‌داده شامل برخی ملاحظات در سیستم، داده، امنیت و شخصی‌سازی می‌باشد تا بتواند کیفیت بالای داده و دسترسی امن را نیز فراهم کند. موجودیت مدیریت در معماری مرج NBDRA، دو گروه از فعالیت‌های اصلی را شامل می‌شود: مدیریت سیستم و مدیریت چرخه عمر حجیم‌داده. مدیریت سیستم شامل فعالیت‌هایی نظیر فراهم کردن، پیکربندی، مدیریت بستهها، مدیریت نرم‌افزاری، مدیریت نسخه پشتیبان، مدیریت توانمندی، مدیریت منابع و مدیریت کارآیی می‌باشد. مدیریت چرخه عمر حجیم‌داده شامل فعالیت‌هایی مرتبط با چرخه عمر از جمله جمع‌آوری داده، آماده‌سازی داده، انتخاب داده، تحلیل داده، بصری‌سازی و دسترسی است. از آنجایی که معماری مرجع NBDRA بسیار عمومی است و مختص یک محصول یا یک مدل کسب‌وکار نیست، ملاحظات مختلف و راهکارهای متفاوتی می‌تواند مبتنی بر آن پیاده‌سازی شود.

۷-۵ امنیت معماری حجیم‌داده

با توجه به انبوه داده‌های جمع‌آوری شده در سامانه‌های حجیم‌داده، تامین امنیت آنها به یک نیاز و دغدغه اصلی تبدیل شده است. در ادامه معرفی معماری‌های موجود برای حجیم‌داده، یک معماری برای تامین امنیت در محیط‌های حجیم‌داده ارائه می‌شود. همان‌طور که شرکت‌های امنیتی بزرگی مانند سیمنتک و مک‌آفی گزارش داده‌اند، در طی چند سال گذشته، حملات سایبری به شدت افزایش یافته است این حملات عمدتاً در نتیجه افزایش سطح حمله به افراد و بنگاه‌های اقتصادی، سهولت دسترسی ابزارهای بهره‌برداری، افزایش قابلیت و توانایی مجرمان سایبری و عدم درک نحوه عملکرد مجرمان سایبری می‌باشد. سامانه‌های نظارت و تحلیل امنیتی سنتی، از جمله سامانه تایید هویت، سامانه تشخیص و پیشگیری از نفوذ (مبتنی بر کانال‌های داده HTTP، DHCP، DNS یا NetFlow)، ابزارهای اطلاع رسانی امنیت و مدیریت رویداد (SIEM)، و کشف تقلب، همگی ابزارهای مبتنی بر امضا بوده‌اند که فقط تهدیدهایی را که در گذشته مشاهده شده‌اند شناسایی می‌کنند. این ابزارها از منابع اطلاعاتی بسیار محدودی برای شناسایی و جلوگیری از حملات استفاده می‌کنند که عمدتاً بر داده‌های ایجاد شده در شبکه متمرکز شده و مقادیر زیادی از اطلاعات موجود را که می‌توانند برای ایجاد درک بهتر از وضعیت امنیتی یک شرکت بکار گیرند، نادیده می‌گیرند.

پیشرفت‌های سریع در زمینه فن‌آوری‌های مختلف اطلاعات و ارتباطات (از جمله محاسبات ابری، ارتباطات سیار، شبکه‌های حسگر بی‌سیم، سامانه‌های فیزیکی سایبری، اینترنت اشیاء) منجر به ایجاد سامانه‌ها و فرآیندهایی شده است که می‌توانند مقادیر عظیمی از داده‌ها را تولید و ذخیره کنند. این پیشرفت‌ها همراه با توسعه تکنیک‌های جدید تجزیه و تحلیل داده‌ها، باعث بروز این پتانسیل شده که تجزیه و تحلیل عمیق بر روی داده‌های جمع‌آوری شده انجام گیرد. این امر مدیران امنیتی را به صورت واقعی به چالش کشیده و شرکت‌ها را در معرض خطر قرار می‌دهد. از سوی دیگر، با درک عمیق از نحوه کار مجرمان سایبری و نحوه انتخاب اهدافشان، می‌توان به مدیران امنیتی کمک کرد تا بتوانند زیرساخت‌های خود را با مقاومت بیشتری ارائه دهند. این چالش‌ها عمدتا در انتخاب معماری نیز اثرگذار است که در ادامه به آن پرداخته می‌شود.

۱-۷-۵ چالش‌های مربوط به داده‌های امنیتی

برای آنکه نظارت امنیتی جامع و فراگیر باشد، باید مقادیر زیادی داده از منابع مختلف جمع‌آوری شود. با این حال، با افزایش حجم داده و منابع داده‌ها، چالش ذخیره داده‌های جمع‌آوری شده و انجام مؤثر محاسبات بر روی داده‌های چند وجهی (ترکیبی از داده‌های ساخت‌یافته و غیرساخت‌یافته) به طور فزاینده‌ای آشکار می‌شود.

هر شرکت به دلایل مختلف، به طور مداوم حجم زیادی از داده‌های امنیتی (به عنوان مثال، رویدادهای شبکه مانند ویژگی‌های ارتباطات عبوری از دروازه، رویدادهای نرم‌افزارهای کاربردی مثل وب‌سرور و اقدامات مربوط به افراد) را جمع‌آوری و ذخیره می‌کند. در استفاده از این داده‌ها برای تجزیه و تحلیل امنیتی، چالش‌های زیر وجود دارد:

۱. مقدار داده: با بزرگتر شدن مجموعه داده‌ها، فضای ذخیره‌سازی مورد نیاز برای ذخیره آنها نیز افزایش می‌یابد. تجزیه و تحلیل امنیتی در مقیاس بزرگ، نیازمند فرآیندهای ذخیره‌سازی کارآمد و دسترسی سریع به حجم زیادی از داده‌ها دارد، که فن‌آوری‌های ذخیره‌سازی سنتی قادر به تهیه آن‌ها با هزینه مالی پایین نیستند.

۲. ناسازگاری داده‌ها: داده‌های جمع‌آوری شده از منابع ناهمگن (به عنوان مثال لاگ‌های مربوط به شبکه، لاگ‌های مربوط به برنامه‌های نرم افزاری مانند وب‌سرور و وقایع مربوط به رویدادهای بیومتریک) در ساختار و قالب بندی متفاوت هستند. منابع داده ناهمگن ممکن است در نحوه قالب‌بندی زمانی (به عنوان مثال، یونیکس یا GMT)، اطلاعات مختلف موجود در لاگ‌ها (به عنوان مثال آدرس‌های IP در لاگ‌های مربوط به وب‌سرور در مقابل URL در درخواستهای وب) و نوع داده‌های ذخیره شده در لاگ‌ها متفاوت باشند. بدون لایه پیش‌پردازش، داده‌ها غالباً متناقض، ناقص و دارای نویز هستند که انجام تجزیه و تحلیل را دشوار می‌کند.

۳. تجزیه و تحلیل داده‌ها: با افزایش اندازه مجموعه داده، درک داده‌ها دشوارتر می‌شود. برنامه‌های کاربردی حجیم‌داده، به ویژه در تجزیه و تحلیل امنیتی، محاسبه در زمان واقعی را برای شناسایی و گزارش موثر تهدیدات و ناهنجاری‌ها را انجام می‌دهند. با این حال، ایجاد همبستگی و کشف رابطه بین آن‌ها، در زمان واقعی یا برخط بر روی داده‌هایی که از نظر حجم زیاد و اغلب متناقض هستند، کاری پیچیده است.

٤. نمایش داده‌ها: هدف از نمایش داده‌ها در تجزیه و تحلیل امنیت حجیم‌داده، ارائه اطلاعات پنهان در مجموعه داده‌های پیچیده و بزرگ، و اعلام وضعیت امنیتی یک زیرساخت در زمان واقعی، به افراد، با استفاده از مؤثرترین روش نمایش بصری است. این به مدیر یا تحلیلگر امنیتی کمک می‌کند تا داده‌ها و نتایج را به راحتی تفسیر کند. با این حال، هنگام تلاش برای نمایش داده‌های بزرگ به دلیل حجم و ماهیت چند جانبه داده‌ها، مشکلاتی ایجاد می‌شود. نمایش داده‌های بزرگ جاری، از عملکرد ضعیف، مقیاس پذیری و زمان پاسخ و تعامل پایینی برخوردار است. یکی دیگر از محدودیت‌های مهم در توسعه نمایش داده برای حجیم‌داده، کمبود استعداد انسانی است. تحلیلگران حجیم‌داده باید مهارت‌های پیچیده ریاضی را داشته باشند که آموزش آن دشوار است و مدت زمان طولانی برای آموزش طول می‌کشد.

علاوه بر ذخیره داده‌ها و چالش‌های محاسباتی، حریم خصوصی داده‌ها نیز موضوعاتی را مطرح می‌کند که قابل تامل است. مشکلات امنیتی قابل توجه، شامل سرقت از مالکیت معنوی، محافظت از حریم خصوصی اطلاعات شناسایی شخصی و محافظت از اسرار تجاری و اطلاعات مالی است. قوانین مربوط به حفظ حریم خصوصی و قوانین مربوط به محافظت از داده‌ها، تناقضاتی در مورد نوع تجزیه و تحلیل در مجموعه داده‌های حساس ایجاد می‌کند. به موازات آن، پیشرفت‌ها در تجزیه و تحلیل حجیم‌داده‌ها، فرآیندهای استخراج و کشف ارتباط بین داده‌های حجیم را ساده‌تر کرده است که این امر نقض حریم خصوصی را آسان‌تر می‌کند.

در ادامه دو راهکار برای افزایش امنیت سامانه حجیم‌داده بیان می‌شود، ابتدا یک معماری امن بر اساس معماری مرجع استاندارد که در بخش قبل توضیح داده شد، بیان می‌شود و در ادامه، یک چارچوب را که توسعه‌یافته هدوپ است و موارد امنیتی را تا حدود زیادی رعایت می‌کند، معرفی می‌کنیم.

۵-۸ معماری مرجع امنیتی حجیم‌داده

معماری مرجع استاندارد به عنوان یک معماری نرم‌افزاری که چندین حوزه مختلف از بحث حجیم‌داده را پوشش می‌دهد شناخته شده است [۱۷]. افزودن الگوهای امنیتی برای کنترل شناسایی تهدیدات امنیتی، آنرا به معماری مرجع امنیتی تبدیل کرده که در برگیرنده الزامات امنیتی و سیاست های آن، تهدیدات، آسیب‌پذیری‌ها و غیره است. مهمترین نگرانی و هدف این طرح، بهبود امنیت و صحت در محیط حجیم‌داده است. برای رسیدن به این هدف معماری مرجع را که قبلا معرفی شد به یک مرجع کامل‌تر و بهتر تغییر می‌دهد تا با استفاده از درک بهتر اکوسامانه حجیم‌داده به الزامات امنیتی در آن پرداخته شود.

همانطور که در بخش‌های قبل ملاحظه شد، معماری مرجع که توسط انستیتو استاندارد و فن‌آوری برای سامانه‌های حجیم‌داده ارائه شده است، دارای پنج بخش اصلی است که نکات امنیتی را به طور کامل در بر نمی‌گیرد. در ادامه، معماری مرجع امنیتی [۱۸] که بر اساس معماری مرجع ارائه شده است و علاوه بر مزایای معماری مرجع، امنیت و محرمانگی را نیز شامل می‌شود معرفی می‌گردد. در ادامه، بخش‌های مختلف آن به اختصار توضیح داده می‌شود.

۵-۸-۱ هماهنگ کننده

مهمترین هدف این بخش اجرای الزامات مختلف است که باید در یک سامانه حجیم‌داده انجام گیرد. همچنین نحوه ارتباط این الزامات با اجزای مختلف معماری حجیم‌داده را معین می‌کند. در معماری امنیتی ارائه شده برای این بخش به فعالیت‌های امن مرتبط با الزاماتی که باید برآورده شود پرداخته می‌شود. این فعالیت‌ها از نحوه چگونگی پیاده‌سازی تا مانیتورینگ در حین اجرا را پوشش می‌دهد. با استفاده از سازوکارهای کلی مانند مجوز و تأیید اعتبار کاربر، تشخیص کلاهبرداری، کنترل ریسک، رمزگذاری، کنترل دسترسی به شبکه، تشخیص نفوذ، یا تضمین کیفیت و امنیت داده‌ها در هنگام تهیه منابع داده‌ای مختلف می‌توان این موارد را تامین نمود. یکی از راه‌های دستیابی به الزام امنیتی، استفاده از روش‌های تأیید اعتبار است که با استفاده از الگوی امنیتی "کنترل دسترسی مبتنی بر نقش" می‌توان به اجرای این راه‌حل امنیتی دست یافت. الگوهای امنیتی باید برای موارد مختلف و با توجه به کارکرد سامانه تعریف شوند و قابلیت استفاده مجدد نیز داشته باشند. برای تعریف این الگوها می‌توان از موارد نقض امنیت اتفاق افتاده در گذشته استفاده کرد. از این روش به عنوان راهی برای درک هر حمله و استخراج الگوهای امنیتی مختلف برای شناسایی تهدید استفاده می‌شود. یکی دیگر از ملاحظاتی که باید در نظر گرفته شود، نوع داده‌ها می‌باشد. به طور مثال، رویکردهای امنیتی تصاویر پزشکی و سوابق فایل‌ها متفاوت است.

۵-۸-۲ فراهم‌کننده داده

این بخش از حجیم‌داده یک انتزاع از داده‌ها با استفاده از فراداده امنیتی منابع داده‌ای ایجاد می‌کند. این فراداده اجازه می‌دهد تا این بخش از معماری حجیم‌داده انواع دسترسی، تجزیه و تحلیل مجاز از منابع داده و الزامات امنیتی آن‌ها را شناسایی کند. این بخش معمولا دارای رابط کاربری‌های متفاوتی است که این رابط‌ها باید محدودیت‌های هر منبع داده و همچنین سیاست‌ها و الزامات امنیتی متفاوت مشخص شده توسط هماهنگ کننده را برای آنها در نظر بگیرند. در اینجا ممکن است گاهی بین الزامات امنیتی منابع داده و یکی از بخش‌های حجیم‌داده تضاد و تداخل پیش بیاید که باید راهکاری مشخص برای این گونه موارد اندیشیده شود. یکی از بزرگترین چالش ها در این بخش مشخص کردن مبدا داده‌ها و بحث اعتبارسنجی و صحت آنها می باشد. لذا در این بخش رابط کاربری‌ها و فراداده‌های امنیتی منابع داده به همراه الزامات و سیاست‌های امنیتی آنها وجود دارند که در طول استفاده داده‌ها به عنوان مرجع امنیت داده به صورت مداوم ارزیابی می‌شوند.

۵-۸-۳ فراهم کننده کاربرد حجیم‌داده

این مولفه مسئول تامین الزامات امنیتی تبیین شده توسط هماهنگ کننده می‌باشد. برای دستیابی به این هدف، این مولفه از خدمات یا فعالیت‌های مختلفی تشکیل شده است که می تواند به عنوان لایه سرویس در اکوسامانه حجیم‌داده در نظر گرفته شود. مرحله آماده‌سازی دارای اعتبارسنجی، تمیز کردن و ذخیره اطلاعات است، اما در یک سناریوی بلادرنگ و برخط که داده‌ها به محض ورود به سامانه مورد تجزیه و تحلیل قرار می‌گیرند، می‌توان از این فعالیت‌ها صرفنظر کرد. عملیات مشابهی در مرحله نمایش نتایج رخ می‌دهد، اگر مصرف‌کننده داده کاربر انسانی نباشد و سامانه هوشمند دیگری مانند انباره داده یا حتی اکوسامانه حجیم‌داده‌ی دیگری باشد، تمام فعالیت‌های امنیتی ممکن است لازم نباشد.

۵-۸-۴ فراهم کننده چارچوب حجیم‌داده

این بخش از مجموعه‌ای از خوشه‌ها تشکیل شده است که هریک مجموعه‌ای از گره‌های پردازشی می‌باشند. این گره‌ها می توانند با استفاده از ماشین‌های مجازی یا کانتینرهایی مانند داکر ایجاد شوند، که با سخت‌افزار و سیستم‌عامل تعامل دارند. در این بخش، فعالیت‌ها باید روی رمزنگاری و مدیریت کلیدی داده‌ها، جداسازی و محصورسازی اجرای فرآیند، مجوز، تأیید اعتبارسنجی، ورود به سامانه حسابرسی و نحوه تأمین امنیت ذخیره‌سازی و شبکه متمرکز شوند. این موارد امنیتی باید با استفاده از راه‌حل‌های امنیتی تعریف شده در هماهنگ‌کننده برطرف شوند. در این بخش، راه‌حل‌های امنیتی هماهنگ‌کننده برای محافظت از داده‌ها، از جمله استفاده از رمزنگاری و مکانیزم‌های مجوز دسترسی اختصاصی، اجرایی می شود.

۵-۸-۵ مصرف کننده داده

این بخش مشابه فراهم کننده داده است که توسط مجموعه‌ای از رابط‌ها تشکیل شده است. نحوه تعامل می‌تواند شامل نمایش تعاملی نتایج، ایجاد گزارش‌های متنوع با استفاده از تکنیک‌های داده‌کاوی و هوش تجاری باشد. نکته مهم این است که این رابط‌ها باید عملکرد مربوط به تأیید

اعتبار و صحت را به درستی انجام دهند تا به هدفی که با استفاده از فراداده مربوط به کاربران نهایی و الزامات امنیتی و سیاستهای اطلاعات مطابقت داشته باشد، برسند.

۹-۵ نمونه معماری امنیتی G-Hadoop

نگاشت-کاهش (MapReduce) به عنوان یک مدل برنامه نویسی مناسب وبرای برنامه‌های کاربردی حجیم‌داده مقیاس بزرگ در نظر گرفته شده است. چارچوب هدوپ (Hadoop) یکی از پیاده‌سازی‌های شناخته شده مدل نگاشت-کاهش است که توسط بنیاد آپاچی (Apache) پشتیبانی می‌شود و توسط دانشمندان به عنوان پایه کار تحقیقاتی استفاده شده است. این چارچوب وظایف نگاشت-کاهش را روی یک سامانه خوشه‌ای اجرا می‌کند چارچوب G-Hadoop یک گسترش از چارچوب نگاشت-کاهش هدوپ، با قابلیت امکان دادن به وظایف نگاشت-کاهش برای اجرا روی خوشه‌های متعدد در یک محیط شبکه‌ای است [19]. با این حال، G-Hadoop به سادگی از تایید هویت کاربر و مکانیزم ثبت کار هدوپ استفاده مجدد می‌کند که برای یک خوشه منفرد طراحی شده است؛ و از این رو، برای محیط شبکه مناسب نیست. یک مدل امنیتی جدید برای G-Hadoop پیشنهاد شده است [۳]. مدل امنیتی مبتنی بر چندین راه حل امنیتی مانند رمز نگاری کلید عمومی و پروتکل SSL است و به طور اختصاصی برای محیط‌های توزیع شده مانند شبکه طراحی شده است. این چارچوب امنیتی تایید هویت کاربران را ساده می‌کند و فرآیند تسلیم کار مربوط به پیاده سازی G-Hadoop را ساده می‌سازد. علاوه بر این، چارچوب امنیتی طراحی شده یک تعداد از مکانیزم‌های امنیتی مختلف را برای محافظت از سامانه G-Hadoop از حملات سنتی فراهم می‌کند.

معماری نگاشت-کاهش هدوپ براساس یک مدل ارتباطی ارباب/رعیت یا همان Master/Slave است. با یک دنبال کننده کار به عنوان ارباب و چندین دنبال کننده وظیفه که در نقش رعیت‌ها فعالیت می‌کنند. هدوپ از سامانه فایل خود، سامانه فایل توزیع شده هدوپ (HDFS)، برای مدیریت داده‌های ورودی/خروجی برنامه‌های نگاشت-کاهش استفاده می‌کند. G-Hadoop یک پیاده‌سازی نگاشت-کاهش است که روی یک سامانه توزیع شده با چندین خوشه، به عنوان یک زیر ساخت شبکه، یک ابر، ماشین های مجازی توزیع شده، یا یک زیر ساخت با چند مرکز داده قرار داده می‌شود. به منظور اشتراک داده‌ها در طی زمینه‌های اجرایی چند گانه، G-Hadoop سامانه فایل توزیع شده بومی هدوپ را با سامانه فایل شبکه Gfarm جایگزین می‌کند. برنامه‌های نگاشت-کاهش در G-Hadoop در طی خوشه‌های متعدد با استفاده از یک رویکرد زمان‌بندی سلسله مراتبی برنامه‌ریزی شده‌اند. ابتدا، وظایف نگاشت-کاهش در میان خوشه‌ها با استفاده از سیاست برنامه‌ریزی آگاه از داده هدوپ برنامه‌ریزی شده، سپس گره‌های محاسباتی با استفاده از برنامه‌ریز خوشه موجود روی خوشه‌های هدف برنامه‌ریزی می‌شوند. G-Hadoop مدل ارباب/رعیت مربوط به هدوپ را حفظ می‌کند، که در آن گره‌های رعیت کارگران ساده‌ای هستند. درحالی که، گره ارباب کارهای ارسال شده توسط کاربر را می‌پذیرد، آنها را به وظایف کوچک‌تر تقسیم می‌کند و در نهایت وظایف را بین گره‌های رعیت توزیع می‌کند. سامانه G-Hadoop فعلی از مکانیزم‌های هدوپ برای تایید اعتبار کاربر و واگذاری کار، استفاده مجدد می‌کند که در واقع برای محیط‌های تک خوشه‌ای طراحی شده است. مکانیزم گفته شده برای برقراری ارتباط امن بین کاربر و خوشه هدف، پروتکل پوسته امن را بکار می‌برد (SSH). این نوع مکانیسم مانند Grid که حاوی چندین خوشه مقیاس بزرگ

است، برای یک محیط توزیع شده مناسب نیست. در G-Hadoop برای مثال، یک ارتباط منحصر به فرد HSL مجبور است بین کاربر و هر خوشه تک ساخته شود. این رویکرد، سامانه را به عنوان یک سامانه کلی رسیدگی نمی‌کند؛ بلکه به طور جداگانه‌ای با اجزای آن کار می‌کند. علاوه بر این، با رویکرد امنیت هدوپ، یک کاربر باید به منظور اینکه تصدیق شود وارد سامانه شود، قبل از اینکه قادر باشد از منابعش برای اجرای وظایف نگاشت-کاهش استفاده کند. این بدون شک یک کار خسته‌کننده برای کاربران است. بنابراین، یک راه حل کلی‌تر برای پنهان کردن جزئیات معماری سامانه، و همچنین برای آزادسازی کاربر از بار مسئولیت موردنیاز است. در این کار، یک مدل امنیتی جدید برای چارچوب هدوپ برای رویارویی با چالش های فوق طراحی شده است. چارچوب امنیتی با مشخصات زیر طراحی شده است:

- **یک فرآیند ورود به سامانه فقط با یک نام کاربری و رمز عبور:** یک کاربر یک‌بار با نام کاربری و رمز عبور خود به سادگی وارد سامانه G-Hadoop می شود. در ادامه، همه منابع سامانه برای کاربر قابل دسترسی می‌باشند. روش تصدیق شدن توسط خوشه‌های مختلف سامانه اصلی به طور خودکار توسط چارچوب امنیت در پس زمینه انجام شده است.
- **حریم خصوصی اطلاعات کاربر:** اطلاعات کاربر، مانند اطلاعات تایید اعتبار، در سمت گره‌های رعیت پنهان است. گره‌های رعیت وظایفی را انجام می‌دهند که بدون آگاهی از اطلاعات کاربر، ازجمله نام کاربری و رمز عبور، توسط گره ارباب اختصاص داده شده است.
- **کنترل دسترسی:** چارچوب امنیتی از منابع کل سامانه در برابر سوء استفاده یا استفاده غلط کاربران غیرحرفه‌ای، محافظت می‌کند. کاربران سامانه تنها حق دسترسی به منبع یک خوشه را دارند که با یک ارتباط SSH می‌تواند به آن دسترسی داشته باشد.
- **مقیاس‌پذیری:** یک خوشه می‌تواند به راحتی از محیط اجرایی بدون هیچ تغییری روی کد گره‌های رعیت یا هر تغییری از چارچوب امنیتی حذف یا ادغام شود.
- **تغییر ناپذیری:** چارچوب امنیتی مکانیزم امنیتی موجود در داخل خوشه‌ها را تغییر نمی‌دهد. کاربران یک خوشه، برای دستیابی به خوشه‌ها تنها می‌توانند به مشخصات تایید اعتبار خودشان وابسته باشند.
- **حفاظت در برابر حملات:** چارچوب امنیتی از سامانه در برابر حملات رایج مختلف محافظت می‌کند و امنیت و حریم خصوصی را توسط تبادل اطلاعات حساس تضمین می‌کند، مثل اطلاعات تایید اعتبار و رمز گذاری. همچنین این چارچوب قادر به تشخیص جعلی بودن یک نهاد برای جلوگیری از سوء استفاده یا دسترسی غیر قانونی به منابع توسط یک حمله کننده است.

با این مکانیزم‌های امنیتی، چارچوب طراحی شده توانایی جلوگیری از شایع‌ترین حملات را مثل حمله MITM، حمله پخش و حمله تاخیر دارد و یک ارتباط امن از G-Hadoop را در شبکه‌های عمومی تضمین می‌کند. علاوه بر این، آن مکانیزم‌های مختلفی را برای محافظت از منابع G-Hadoop و جلوگیری از سوء استفاده یا استفاده غلط اتخاذ می‌کند.

۶. مراجع

[1] li, X., & Yang, T. (2015). Signal Processing Oriented Approach for Big Data Privacy. 16 th International Symposium on High Assurance Systems Engineering (pp. 275- 276). Daytona Beach Shores, FL: IEEE.

[2] Sagiroglu, S., & Sinanc, D. (2013). Big data: A review. Collaboration Technologies and Systems (CTS) (pp. 42-47). San Diego: IEEE.

[3] Zhao, J., Wang, L., Tao, J.,Chen, J., Sun, W., Ranjan, R., Kolodziej, J., Streit, A., Georgakopoulos, D.(2014). A security framework in G-Hadoop for big data computingacross distributed Cloud data centres. Journal of Computer and System Sciences 80 994–1007

[4] Rahmani, A., Amine, A., & Hamou, M. (2015). De-identification of Textual Data Using Immune System for Privacy Preserving in Big Data. Computational Intelligence & Communication Technology (CICT) (pp. 112-116). Ghaziabad: IEEE.

[5] Demchenko, Y., Membrey, P., Ngo, C., De Laat, C., & Gordijenko, D. (2013). Big Security for Big Data: Addressing Security Challenges for the Big Data Infrastructure. International Conference on Collaboration Technologies and Systems (CTS) (pp. 26-30). Trento: Springer.

[6] Shrivasta, K., Rizvi, M., & Singh, S. (2014). Big Data Privacy Based on Differential Privacy a Hope for Big Data. Computational Intelligence and Communication Networks (CICN) (pp. 776-781). Bhopal: IEEE.

[7] Kshetri, N. (2014). Big data′s impact on privacy, security and consumer welfare. Telecommunications Policy, 38 (11), 1134-1145.

[8] Matturde, B., Xianwei, Z., Shuai, L., & Fuhong, L. (2014). Big Data security and privacy: A review. Communications, 135-145.

[9] Chen, M., Mao, S., & Lio, Y. (2014). Big Data: A Survey. Mobile Networks and Applications, 19 (2), 171-209.

[10] Mishra, R., & Sharma, D. R. (2015). BIG DATA: OPPORTUNITIES AND CHALLENGES. Computer Science and Mobile Computing, 4 (6), 27-35.

[11] XU, L., JIANG, C., WANG, J., YUAN, J., & REN, Y. (2014). Information Security in Big Data: Privacy and Data Mining. IEEE Acces, 2 , 1149-1176.

[12] Alguliyev, R., & Imamverdiyev, Y. (2014). Big Data: Big Promises for Information Security. Application of Information and Communication Technologies. Astana.

[13] Liang, K., Susilo, W., & K. Liu, J. (2015). Privacy-Preserving Ciphertext Multi-Sharing Control for Big Data Storage. Information Forensics and Security,
10 (8), 1578-1589.

[14] Cloud Security Alliance (CSA): Expanded Top Ten Big Data Security and Privacy Challenges, April 2013

[15] Bowes, R. https://blog.skullsecurity.org/2010/followup-to-my-facebook-research

[16] NIST Big Data Public Working Group, NIST Big Data Interoperability Framework: Volume 5, Architectures White Paper Survey

[17] NIST Big Data Public Working Group, NIST Big Data Interoperability Framework: Volume 6, Reference Architecture

[18] Moreno, J., Serrano, M., Medina, E., Fernandez, E.(2018). Towards a Security Reference Architecture for Big Data. EDBT/ICDT Joint Conference (Vienna, Austria) 1613-0073

[19] Wang, L., Tao, J.,Ranjan, R.,Marten, H., Streit, A., Chen, J., Chen, D.(2013). G-Hadoop: MapReduce across distributed data centers for data-intensive computing. Future Generation Computer Systems.739-750