داده های عظیم؛ تعاریف و چالشها

شیرین عباسی
دانشجوی کارشناسی ارشد، دانشگاه آزاد اسلامی-واحد تهران مرکز، تهران
Shirin.abbasi67@gmail.com

چکیده
در سالهای اخیر، با توجه به فراگیر شدن استفاده از خدمات الکترونیکی و همچنین استفاده از شبکه های اجتماعی، حجم زیادی از اطلاعات تولید می شود که این اطلاعات علاوه بر حجم زیاد، از انواع گوناگونی از قبیل فیلم، عکس، متن و .... تشکیل شده اند . به دلیل حجم بالا و عدم ساخت یافتگی این اطلاعات ، پوشش آنها از طریق پایگاه داده های سنتی و روشهای رابطه ای امکان پذیر نیست و باید از راهکارهای نوین برای پردازش آنها استفاده شود، به گونه ای که سرعت پردازش نیز تحت پوشش قرار گیرد.
ذخیره سازی اطلاعات برای پردازش و نحوه ی دسترسی به آنها در حافظه ، ارتباطات شبکه ای ، پوشش ویژگی های مورد نیاز برای سیستم توزیع شده در راهکارهای مورد استفاده در ذخیره سازی داده های بزرگ، از جمله مواردی است که باید مورد پوشش قرار گیرد . در این مقاله مجموعه ای از مزیتها و چالشها در داده های بزرگ، ویژگی هاو خصوصیات خاص آنها فراهم شده است و با معرفی تکنولوژیهای مورد استفاده ، راهکارهای ذخیره سازی مورد بررسی قرار می گیرند و فرصتهای تحقیقاتی برای ادامه راه ، معرفی خواهند شد.

کلمات کلیدی: داده های بزرگ،رایانش ابری، هادوپ،تجزیه و تحلیل داده های بزرگ، **Big Data ،Hadoop**

۱- مقدمه

فرمونت رایدر ، در مقاله ای در مورد آینده کتابخانه ی دانشگاه یل ، پیش بینی ارائه کرده بود که بر اساس افزایش سالیانه ی منابع تحقیقاتی، در سال ۲۰۴۰، دویست میلیون جلد کتاب موجود خواهد بود که اگر قرار باشد به صورت کاغذی نگه داری شود ، قفسه های آن ، مسافتی در حدود ششصد مایل را پوشش خواهد داد. مقادیر داده ای که تولید می شوند و مورد پردازش قرار می گیرند، روز به روز در حال افزایش است. داده‌های بزرگ به مجموعه‌ای از داده‌ها گفته می‌شود که به‌صورت ساخت‌یافته یا غیرساخت یافته، ذخیره می‌شوند و داده‌های پیچیده‌ای هستند که از ابعاد گوناگون تشکیل‌شده‌اند.اولین خصوصیت داده های

بزرگ، حجم آنهاست که به مقدار و کمیت آنها برمی گردد و به دلیل حجم بالا، مدیریت، تجزیه و تحلیل آنها متفاوت است و به واسطه ی پایگاه داده های سنتی انجام نمی شود. اگر از تعداد کمی گره های پردازشی استفاده شود، با توجه به این حجم بالا، پردازش با سرعت کمتری صورت می پذیرد. برای افزایش سرعت پردازش، گره های بیشتر و همچنین قدرت پردازش بیشتری مورد نیاز است که هزینه ی بالاتری را طلب می کند. یکی از راههای اولیه که در این زمینه پیشنهاد می شود ، فشرده سازی داده هاست. این امر در داده های بزرگ چندان کارساز نیست، زیرا یکی دیگر از خصوصیات داده های بزرگ، تنوع آنهاست . این داده ها از انواع مختلفی از قبیل فیلم، عکس، متن و ..... تشکیل شده اند که این غیر ساخت یافتگی، فشرده سازی آنها را دشوار کرده و در بعضی از شرایط به گونه ای است که همان زمانی که برای پردازش آنها به واسطه ی روشهای سنتی صرف می شود، برای فشرده سازی هدر می رود و از طرفی به دلیل این گوناگونی نوع، پیچیدگی خاص خود را دارد.به همین دلیل فشرده سازی کاربردی در پردازش داده های بزرگ ندارد. مورد دیگری که باید در پردازش داده های بزرگ در نظر گرفته شود، این موضوع است که این داده ها در برنامه های کاربردی به کار می روند که به صورت آنلاین اطلاعات را رد و بدل می کنند و یا باید در حالتی مورد تجزیه و تحلیل قرار گیرند که پاسخ افراد را در زمان معینی بدهند. بنابراین در امر پردازش ، زمان بندی به گونه ای است که پاسخگویی به صورت بلادرنگ انجام شود. شیوه‌های سنتی مدیریت داده، برای مدیریت داده‌های بزرگ پاسخگو نیستند در مدیریت داده‌های بزرگ، باید همه‌ی موارد ازجمله ساختارهای داده‌ای گوناگون، ابعاد مختلف داده‌ای و عدم ساختار آن‌ها در نظر گرفته شود[۳]

در مدیریت داده های بزرگ، باید مواردی خاص در نظر گرفته شود:

۱- اگر از پردازش موازی استفاده می شود ، باید گره ها در خوشه ها به گونه ای عمل کنند که در صورت رخداد شکست در هر گره ی خاص، اختلالی در پردازش داده های بزرگ پیش نیاید.

۲- فایل سیستمهایی که برای داده های بزرگ به کار می روند، باید به گونه ای باشند که حجم بالای داده های بزرگ را تحت پوشش قرار دهند و ظرفیت آنها در حد گیگابایت یا بالاتر باشد.

۳- عملیات خواندن و نوشتن در بسیاری از برنامه های کاربردی به صورت پی در پی اجرا می شوند و برای پاسخگویی به حالت بلادرنگ و سرعت پاسخگویی بهتر، باید بهینه شوند .

۴- برخی از عملیات برای پاسخگویی بهتر باید به سمت برنامه های کاربردی انتقال داده شوند و برای رهایی از سازگاری امکانات سخت افزاری و نیازهای ارتباطی، بهتر است از محیط های ابری برای پوشش این مورد استفاده شود.

زمانی که بحث داده های بزرگ مطرح می شود، تا حد زیادی رایانش ابری نیز مطرح می شود.چرا که نگرانی های مشترکی در بین هر دو مسئله وجود دارد. در هر دو، منابع سخت افزاری و نرم افزاری در یک شبکه قرار گرفته اند و زیر ساختهای فیزیکی برای کاربران، شبیه سازی شده اند تا بر اساس تقاضای افراد، فضای به ظاهر نامحدود، در اختیار آنها قرار گیرد.پایگاههای داده ی ابری، برای تخصیص صحیح منابع در داده های بزرگ به کار گرفته می شوند و از طرف پارامترهایی مانند مقیاس پذیری و در دسترس پذیری، برای محیط های ابری نیز به عنوان یک چالش مطرح می شوند.

در این مقاله، ابتدا کارهای پیشین در زمینه ی داده های عظیم ذکر خواهند شد و سپس به خصوصیات اصل در این داده ها خواهیم پرداخت و چالشهای پیش رو استخراج می شود و در نهایت ارزیابی از آینده ذکر خواهد شد.

۲- کارهای پیشین

در سالهای دهه ی ۱۹۷۰، مفاهیم اولیه ی پایگاه داده، مطرح شد. در ابتدا برای ذخیره سازی داده ها از آنها استفاده می شد و سپس با گسترش مفاهیم پایگاه داده های رابطه ای، برای تجزیه و تحلیل داده ها نیز، روش های جستجو و پرس وجو در داده ها، گسترش یافتند که امکان پردازش داده ها نیز فراهم گردید.هنگامی که به مرور زمان، حجم داده ها افزایش یافت ، امکان استفاده از یک کامپیوتر برای پردازش وجود نداشت.در اوایل دهه ی ۸۰، امکان استفاد از چندین سیستم و به اشتراک گذاری داده ها فراهم گردید و هر سیستم، پردازشگر و حافظه ی خاص خود را دارد و به صورت مستقل، پردازش را انجام می دهد. با توجه به افزایش حجم داده های به هم مرتبط و نیاز به پردازش همزمان این داده ها ، راهکارهای متفاوتی برای داده های بزرگ در نظر گرفته شده است. زمانی که گوگل، با مشکل کمبود حافظه و مشکلات تحلیل مقادیر زیادی از صفحات وب روبه رو شد، فایل سیستم های مورد استفاده را توسعه داد و از مدل تجزیه و تجلیل و پردازش توزیع شده، بهره مند شد. همراه با این پردازش موازی، گوگل، پایگاه داده ای با مقیاس پذیری بالا، طراحی کرد که BigTable نامیده می شد. این نوع پایگاه داده، قادر بود که اطلاعات را نمایه سازی و برچسب گذاری کند، به همین دلیل دسترسی به اطلاعات، با سرعت بیشتری صورت می گرفت و با توجه به استفاده در محصولات و خدمات گوگل به عنوان یک نقطه ی شروع برای پردازش داده های بزرگ مطرح شد و پس از گسترش محیطهای ابری، به عنوان یک استاندارد بالقوه در زمینه ی پردازش و پرس و جوی داده های بزرگ در محیط های ابری به کار گرفته شد.تکنولوژی گوگل، تکنولوژی متن باز نبود و به همین دلیل، یاهو برای گسترش این پایگاههای داده بر اساس چاچوبهای متن باز مانند هادوپ اقدام کرد که پایگاههایداده ای مانند HBase و Hive بر اساس این امر مطرح شده است.

در مدیریت داده‌های عظیم، رویکردهای مختلفی از قبیل استفاده از BigTable ها، استفاده از روش‌های مدیریت داده به‌صورت غیر رابطه‌ای [۸] و یا مدیریت داده به‌صورت موجودیت[۷] و ایجاد صفت بر مبنای ابعاد مختلف داده[۹]، وجود دارد.

همچنین در سرویس‌های ذخیره‌سازی ابری فعلی، مانند Amazon از روش‌هایی بهره می‌برند که بر اساس آن‌ها قابلیت اطمینان در مورد داده‌های ذخیره‌شده، بسیار بالا می‌رود و این سرویس‌ها برای سیستم‌های توزیع‌شده مورداستفاده قرار می‌گیرند. ازجمله این سرویس‌ها، Dyamo است [۱۰].. اگر سه خصوصیت اصلی انسجام داده‌ها، در دسترس‌پذیری و انعطاف‌پذیری قسمت‌بندی اطلاعات که در محیط‌های ابری، مهم هستند را در نظر بگیریم و بر اساس قضیه‌ی CAP که معتقد است، سیستم‌ها نرم‌افزاری، در هر حالتی فقط دو مورد از این سه مورد را تحت پوشش قرار می‌دهند، آن‌ها را موردبررسی قرار دهیم، مدل‌های رابطه‌ای و مقایسه‌ای فقط یکپارچگی و در دسترس‌پذیری را پشتیبانی می‌کنند. مدل‌های مبتنی بر کلید، فقط در دسترس‌پذیری و قسمت‌پذیری را تحت پوشش قرار می‌دهند و مدل‌های مبتنی بر Big Table ها، یکپارچگی و قسمت‌پذیری را تحت پوشش قرار می‌دهند و روش‌های مبتنی بر سند، به‌گونه‌ای عمل می‌کنند که در دسترس‌پذیری و قسمت‌پذیری را تحت پوشش قرار می‌دهند.

۳- تعریف داده های عظیم

داده های عظیم داده هایی هستند با حجم بالای داده ای ، که ترکیبی از داده های ساخت یافته و غیر ساخت یافته را تحت پوشش قرار می دهند و پردازش آنها با روشهای سنتی پایگاه داده های رابطه ای امکان پذیر نیست و به همین دلیل برای مدیریت آنها از تکنیک های خاصی استفاده می شود. سه ویژگی خاص داده های عظیم هستند که به عنوان یک معیار برای شناسایی داده های عظیم، استفاده می شوند. این سه ویژگی به 3V معروف هستند و شامل مقدار[۱]، نوع[۲] و سرعت پردازش[۳] می شوند که در ویژگی های داده های عظیم آنها را تعریف خواهیم کرد.[3]

۴- ویژگی های داده های عظیم

- حجم داده ها : مقدار داده ها در مجموعه های داده ای داده های عظیم، بالاست. این حجم یکی از ویژگی هایی است که برای داده های عظیم ، به عنوان یک خصیصه اصلی شناسایی می شود. همان طور که در بخش های پیشین ذکر شد، حجم داده ها در جهان امروز رو به افزایش است و در پردازش داده ها ، باید در نظر گرفته شود. چرا که در بسیاری از موارد نیاز به پالایش و فیلتر اطلاعات است و همچنین باید طرق دسترسی و ذخیره سازی اطلاعات نیز، بر مبنای این حجم، شخصی سازی شود.[3]
- سرعت پردازش : سرعت خلق، جریان ، پردازش و تجمیع اطلاعات باید به گونه ای باشد که متناسب با ویژگی های گروههای داده ای امروزی عمل کند. با توجه به سرعت تولید اطلاعات در دنیای امروز و نیاز به پاسخگویی بلادرنگ در بسیاری از برنامه های کاربردی و شبکه های اجتماعی، سرعت عمل و پردازش روی داده ها باید به گونه ای باشد که متناسب با این ویژگی ها انجام شود. از طرفی ، چون داده های عظیم معمولا به صورت توزیع شده، نگه داری می شوند، برقراری ارتباطات و نحوه دسترسی به حافظه نیز باید مورد توجه قرار گیرد.

---

[1] Volume
[2] variety
[3] velocity

- انواع داده ای گوناگون: داده هایی که در گروههای داده ای داده های عظیم قرار می گیرند،شامل انواع مختلف داده ای از قبیل عکس، متن، ویدئو و .... هستند که از منابع گوناگونی به دست آمده اند. فرمتهای مختلف دارند و دسته بندی آنها بسیار مشکل است و نمی توان قالب یا ساختار خاصی برای آنها تعریف کرد و داده های عظیم از این رو ، غیر ساختارمند نامیده می شوند.[4]
- ارزش داده ها : به دلیل حجم بالای داده های عظیم،مثالهای گوناگونی برای ارزیابی در اختیار قرار می گیرد که با توجه به این حجم بالا اگر در زبرگروهی از این داده ها، مشکل یا ناقصی مشاهده شود، می توان داده ها را مجددا پالایش و انتخاب کرد و این امر در زمانی که در نتایج به دست آمده از ارزشیابی نیز مشکل داشته باشد ، کاربرد دارد. همچنین باید در نظر گرفت که مقدار داده ای در مجموعه های داده ای تا چه زمانی، مورد اطمینان است و نتایج آن درست است. این امر در داده های عظیم بسیار اهمیت دارد ، چرا که بسیاری از تصمیمات و برنامه ریزی ها ، در حوزه ی صنایع مختلف؛ بر اساس پردازش این داده ها صورت می پذیرد.
- صحت و یکپارچگی داده ها، به دلیل توزیع پذیر بودن داده های حجیم اهمیت زیادی پیدا می کند ، چرا که باید بخش های مختلف از یک مجموعه داده ای روی سرویس دهنده ای مختلف قرار گرفته اند، یکپارچه باشند و آخرین نسخه ی به روز شده ی آن نیز در قسمتهایی که ممکن است این مجموعه داده ای روی سرویس دهنده های مختلف ، کپی شده اند ، وجود داشته باشد.[6]

۵- ابزارهای مفید در تحلیل داده های عظیم

ابزاری که برای تجزیه و تحلیل داده ها به کار می روند ، روز به روز در حال گسترش و توسعه هستند تا بتوانند بر مبنای انواع داده ای گوناگون، به جمع آوری و تجزیه و تحلیل داده ها ، در هر نوعی کمک رسانی کنند و در داده های عظیم به دلیل ویژگی های خاص آن ، ابزارهایی ویژه به کار می روند که با ابزارهای سنتی مبتنی بر رابطه ی اجزا، متفاوت هستند.

۵-۱- ابزار موازی سازی هادوپ[1]

تکنولوژی های بسیاری هستند که در زمینه ی پردازش داده های عظیم مطرح شده اند ، اما هادوپ یکی از معروفترین آنها است. هادوپ یک سکو[2]ی کد باز [3]است که برای پردازش و ذخیره سازی اطلاعات از انواع مختلف به کار می رود که صنایع مبتنی بر داده را در دسترسی سریع به ارزشهای نهان در داده ها و پردازش و کاوش آنها، یاری می کند. ویژگی های اصلی این ابزار به شرح زیر است :

- این ابزار به صورت متن باز است و به همین دلیل منابع آن و کتابخانه ها و توابع به راحتی در دسترس است.

[1] Hadoop
[2] Platform
[3] Opensource

- لایه ها و اجزای آن به صورت مستقل عمل می کنند و یکپارچه نیستند
- از دسترسی به فایلهای خارجی پشتیبانی می شود.
- در هنگام بار زیاد سیستم ، هادوپ، عملیات انجام یک دستور را به چندین گروه وظیفه می شکند. و به همین دلیل برنامه ریزی برای کارهایی که نیاز چندین گروه عملیات دارنئ، ساده تر صورت می گیرد
- برقراری تعادل خودکار باردر هرکدام از گره های سیستم توزیع شده، در هنگامی که ترافیک داده افزایش می یابد
- پشتیبانی از جایگزینی ماشین ها و گره ها در هنگام خرابی

این ابزار، یک معماری لایه ای دارد. درپایین ترین سطح یک لایه ی حافظه ای مبتنی بر رکورد قرار دارد و این مجموعه ی داده ای به صورت سطری و ستونی مدیریت می شود و در هر ماشین موجود در خوشه های توزیع شده، یک مدیر حافظه وجود دارد که حافظه ی موجود در سیستم را مدیریت می کند. لایه ی وسط، یک لایه ی اجرای جریان کاری است که در آن عملگرهای رابطه ای برای انجام عملیات بر روی مجموعه ی داده ها ، وجود دارند. در پایین ترین سطح نرم افزاری هادوپ، یک فایل سیستم توزیع شده وجود دارد که به اختصار HDFS[1] نامیده می شود. هر فایل در اینجا بخش بندی شده و در دنباله ای از مکانهای حافظه آدرس پذیر و ادامه دار قرار می گیرد و به صورت دسته ای پردازش می شود. در لایه ی وسطی نرم افزاری فایلها، تقسیم بندی و بخش بندی شده و هر قسمتی از پردازش در اختیار یک گره قرار می گیرد و در نهایت نتایج نیز از گره ها جمع آوری شده و تبدیل به خروجی نهایی خواهد شد. این تقسیم بندی و جمع بندی، بر اساس تابع Map-Reduce یا نگاشت- کاهش انجام می شود .[12]

۵-۱-۱- فایل سیستم توزیع شده هادوپ

همان طور که در قسمت معماری هادوپ اشاره ی کوچکی به این فایل سیستم شد ، به اختصار آن را HDFS می نامیم و این قسمت سیستم اصلی حافظه را در هادوپ در بر می گیرد. زمانی که اطلاعات وارد می شود ، این فایل سیستم، آنها را به قسمتهای کوچکتری تقسیم می کند و این بخشها را بین سرویس دهنده های مختلف که به عنوان یک گره در سیستم قرار گرفته اند، توزیع می کند هر سرویس دهنده ، فقط قسمت کوچکی از اطلاعات اصلی را نگه می دارد و بر روی آن پردازش انجام می دهد و هر کدام از این بخش ها برای تحمل پذیری در مقابل خطا، بر روی چندین سرویس دهنده ، کپی شده اند. گره ی مرکزی که وظیفه ی نظارت و تقسیم بندی و جمع آوری نتایج را بر عهده دارد Name نامیده می شود و گره های مختلف که اطلاعات را نگه داری می کنند، گره ی اطلاعاتی یا Data نامیده می شوند و لایه ی میانی که به آن در معماری اشاره شد برای بهینه سازی عملکرد HDFS به کار می رود که برای ارتباط بهتر داده ای به کار می رود.

۵-۱-۲- تابع نگاشت-کاهش

[1] Hadoop Distributed File System

یک مدل معماری برای پردازش موازی در سیستمهای محاسباتی توزیع شده است. در این مدل، داده ها به قسمتهای کوچکتر تقسیم شده و هر فرایند نیز به دستورات کوچکتر شکسته می شود و گره های مختلف در سیستمهای توزیع شده، بخشی از عملیات را بر مبنای این قسمتها مدیریت می کنند. این امر، باعث می شود که قدرت پردازش بالا برود .

بخش اولیه این الگوریتم، یعنی نگاشت، برای تقسیم داده ها استفاده می شود و زیر مجموعه ای از دستورات و وظایف را به هرکدام از گره های محاسباتی اختصاص می دهد. در این مرحله از الگوریتم، ابتدا اطلاعات ورودی خوانده می شود و مجموعه ای از رکوردهای میانی برای محاسبات تولید می شود و برچسب گذاری می شود و این رکوردها، در میان گره های محاسباتی، بر اساس استفاده از توابع Hash ، توزیع می شود و گره ها هرکدام فرآیند مربوط به خودشان را به صورت جداگانه و مستقل آغاز می کنندو هر کدام نتیجه ی تولید شده ی خود را به عنوان خروجی به گره ی مرکزی انتقال می دهند در این زمان، گام دوم الگوریتم یعنی کاهش آغاز خواهد شد. این تابع نتایج را جمع آوری کرده و نتیجه ی اصلی را بر مبنای فرمت درست خروجی تولید خواهد کرد.[12]

۵-۲- پایگاه داده NOSQL

یکی از سیستمهای مدیریت داده است که می تواند در محیط های ابری و برای داده های غیر رابطه ای استفاده شود. در این سیستم، داده ها بر اساس روابطشان شناسایی نمی شوند. انواع پایگاه داده ای آن می توانند می توانند متفاوت باشند ولی نقطه ی اشتراک آنها این است که غیر رابطه ای هستند و این ساختارهای متفاوت می تواند در ذخیره سازی داده های عظیم، تاثیر زیادی داشته باشد. همچنین در این سیستمها، راهکاری اندیشیده شده است که در صورت افزونگی در مورد بعضی از داده ها، داده ها فقط یک جا ذخیره شوند و در حالت دیگر با دسترسی به آنها، بتوان بر روی آنها، پردازش انجام داد.**[۱۱]**

۶- چالش ها در داده های عظیم

- به دلیل عدم ساختار داده ها، فیلتر کردن داده ها و تعیین سطوح دسترسی مختلف برای کاربران بسیار مشکل و هزینه بر است. برای حل این مشکل ، معمولا از برچسب گذاری اطلاعات استفاده می شود
- تولید اطلاعات مربوط به داده ها ، در داده های عظیم نیازمند استفاده از تکنیکهای خاصی است ، چرا که داده ها باید مورد تفسیر قرار گرفته و به دلیل حجم داده ای بالا و همچنین ساختارهای گوناگون و منابع اطلاعاتی مختلف، تعیین معیار، ارزش و نحوه به دست آوردن اطلاعات مشکل است.[6]
- ناهمگنی در منابع سخت افزاری: از آنجایی که داده های عظیم به صورت توزیع شده نگه داری می شوند ، چارچوبهای سخت افزاری متفاوت در سمت سرویس گیرنده ها ، می تواند مشکل ساز شود که با استفاده از منابع مشترک تامین نرم افزاری ، این سازگاری تا حد زیادی برقرار می شود.
- به دلیل اینکه این داده ها ساختار مند نیستند ، تعیین معیار برای سنجش آنها، و استخراج مناسب دسته های مختلف داده ای که در تصمیم گیری های مربووط به کسب و کار به کار آید ، مشکل است و بر اساس دیدگاههای مختلف می توان معیارهای گوناگونی را در نظر گرفت.

- مدیریت داده های عظیم، به شکلهای گوناگونی انجام می شود که مسائلی مانند تعیین سطح دسترسی و مدیریت منابع به دلیل متفاوت بودن منابع و توزیع شدگی سیستم ، باید درنظر گرفته شود. الگوریتمهای مختلفی در این زمینه به کار گرفته می شوند . اطلاعات بلوک بندی شده و روی سیستمهای گوناگون قرار می گیرند نا منابع سخت افزاری و نرم افزاری بیشتری در دسترس باشد.
- به دلیل حجم بالای داده ها، پالایش آنها در یک موضوع خاص ، زمان بر است . باید این امر سنجیده شود که آیا داده ها، همان داده هایی است که در تصمیم گیری ، مورد استفاده قرار خواهد گرفت یا خیر و همچنین باید اطمینان حاصل شود که مقدار داده ای آنها ، هنوز درست است و قابلیت اطمینان برای این مقدار وجود دارد . همچنین این امر که چه مقدار داده لازم است که بتوان پیش بینی های لازم را برای احتمالات پیش رو انجام داد، باید در پالایش داده ها، مورد توجه قرار گیرد.[9]
- سرعت پردازش داده ها نیز در داده های عظیم باید مورد توجه قرار گیرد . چرا که همان طور که در ویژگیهای داده های عظیم ، ذکر شده است ، سرعت پردازشی ، یکی از عواملی است که داده های عظیم ، به واسطه ی آن شناخته و سنجیده می شوند.
- به دلیل تجمیع داده های عظیم از چندین منبع مختلف، تعیین این مسائل که چه کسی مسئول دقت و صحت اطلاعات است و همچنین تعیین اینکه سطوح دسترسی چگونه باید تعیین شود و برای تنظیمات امنیتی، چه کسی صاحب اطلاعات محسوب می شود، چالش برانگیز است.
- کاهش افزونگی و فشرده سازی: به طور کلی، سطح بالایی از افزونگی در مجموعه های داده ای عظیم وجود دارد و برای کاهش هزینه ها به صورت غیر مستقیم لازم است ، این افزونگی، بدون آسیب به اطلاعات اصلی، به صورت نسبی کاهش یابد.[7]
- مکانیزمهای تجزیه وتحلیل در داده های عظیم، باید به گونه ای باشند که داده های غیر ساخت یافته را مورد تجزیه و تحلیل قرار دهند و همچنین بتوانند در زمان محدود پاسخگو باشند.راهکارهایی که در پایگاه داده های سنتی و رابطه ای به کار می روند ، در اینجا قابل توسعه نیستند، از طرفی در پایگاه داده های غیر رابطه ای نیز بحث عملکرد مطرح می شود. به همین دلیل باید راهکارهایی ترکیبی از این دو مسئله به کار گرفته شود.

۷- نتیجه گیری

با توجه به افزایش حجم داده ها به صورت روزافزون و همچنین اهمیت استفاده از آنها در حوزه های مختلف، مدیریت داده های عظیم، مسائل خاص خود را می طلبد. تنها چالش مدیریت داده های عظیم حجم آنها نیست، بلکه انواع مختلف داده ای و سرعت پردازش در داده ها ، مسائل بسیار مهمی هستند که حتی در تعریف داده های عظیم، مورد توجه قرار گرفته اند . در این مقاله؛ چالشهای دیگر مطرح در مدیریت داده های عظیم، مورد توجه قرار گرفت و مسائل ارتباطی، شبکه ای ، ذخیره سازی مورد بررسی قرار گرفت و خصوصیات؛معایب و مزایای داده های عظیم، به عنوان خروجی مقاله استخراج گردید.

مدیریت داده های عظیم، امر مهمی است که به واسطه عملکرد درست در آن، می توان دانش مورد نیاز در حوزه های مختلف را استخراج کرد.این مقاله می تواند نقطه شروعی برای تحقیقات گسترده، در زمینه داده های حجیم باشد.

مراجع

[1] V. Hristidis, S. Chen, T. Li, S. Luis, and Y. Deng, "Survey of Data Management and Analysis in Disaster Situations", the Journal of Systems and Software, Vol. 83, No. 10, PP. 1701-1714, 2010.

[2] K.Grolinger, "Disaster Data Management in Cloud Environments", PHD Thesis, the University of Western Ontario, PP21-27, December2013.

[3]D. Kossmann and T. Kraska, "Data Management in the Cloud: Promises, State-of the Art, and Open Questions," Datenbank-Spektrum, Vol. 10, No. 3, PP. 121-129,
2010.

[4]H.Dewan, R.Hansdah,"A Survey of Cloud Storage Facilities", IEEE World Congress, PP 224-231, 2011.

[5]W. Halton, "Security Issues and Solutions in Cloud Computing", Wolf Halton, June 2010.[Online], (http://wolfhalton.info/2010/06/25/security-issues-and- solutions-in- cloud-computing).

[6]F.Chung, J.Dean, "BigTable: a Distributed Storage System For Structured Data", ACM, 2008.

[7]W.Guo, C.Qiao, Y.Jin, "Demonstration of Joint Resource Scheduling in an Optical Network Integrated Computing Environment", IEEE Communications Magazine, Vol. 48, No.5, PP.76-83, 2010.

[8] W.Vogels, "Eventually Consistent", ACM, Vol 52, No.1, PP.40-44, 2009.

[9]P. Zadrozny and R. Kodali," Big Data Analytics using Splunk, Berkeley", CA, USA: Apress, 2013.

[10]G.Decandia, D.Hastorun, M.Jampani, "Dynamo: Amazon highly Available Ley-value Store", SOSP, PP.205-220, 2007.

[11] K.Grolinger, W.Higashino, A.Tiwari1,"Data management in cloud environments: NoSQL and NewSQL data stores", Jornal of Cloud Computing, 2013.

[12] *M.Minelli, M.Chambers , A.Dhiraj," Big Data, Big Analytics: Emerging Business Intelligence and Analytic Trends for Today's Businesses.", John Wiley & Sons, Inc.,2013.*