معرفی مختصری از کلان داده‌ها

سحر فخریه کاشان[1]، نگین بنای شاهانی[2]

[1] دانشجو، دانشگاه گیلان، رشت

[2] دانشجو، دانشگاه گیلان، رشت

چکیده

کلان داده مجموعه ای از دیتا ست‌ها و دیتا ولیوم‌های عظیم و پیچیده برحسب ترابایت یا پتابایت است که شامل حجم زیادی از داده‌ها، قابلیت‌های مدیریت داده، تجزیه و تحلیل شبکه‌های اجتماعی و داده‌های بی درنگ می‌باشد. در این مقاله، خصوصیات "5V" در کلان داده‌ها و همچنین تکنیک‌ها و تکنولوژی‌هایی که برای رسیدگی به کلان داده‌ها استفاده می‌شود ارائه کردیم.

چالش‌ها شامل ضبط، تجزیه و تحلیل، ذخیره‌سازی، جستجو، به اشتراک گذاری، تجسم و انتقال است. چالش‌هایی که وجود دارد این است که نه می‌توان از کوئری‌های سنتی SQL استفاده کرد و نه می‌توان از سیستم مدیریت پایگاه داده رابطه‌ای (RDBMS) برای ذخیره‌سازی استفاده کرد؛ اگرچه، طیف گسترده ای از ابزارها و تکنیک‌های پایگاه داده مقیاس‌پذیر تکامل یافته است. بهترین راه‌حل برای این مشکل استفاده از هدوپ می‌باشد. هدوپ یک فریم‌ورک متن باز و مبتنی بر جاوا است که برای ذخیره‌سازی و پردازش کلان داده‌ها استفاده می‌شود و سیستم فایل توزیع شده ی آن، پردازش موازی و تحمل خطا را امکان پذیر می‌کند.

کلمات کلیدی

کلان داده، هدوپ، 5V، آپاچی، پایگاه داده

۱- مقدمه

کلان داده‌ها به حدی حجیم و پیچیده هستند که برای سیستم‌ها و ابزارهای ذخیره‌سازی داده به صورت سنتی غیرممکن است که بتوانند داده‌هایی که توسط دستگاه‌ها، انسان‌ها و همچنین طبیعت تولید می‌شوند را پردازش کنند. با رشد تکنولوژی‌ها و سرویس‌ها، داده‌های حجیمی‌که ازمنابع مختلف تولید می‌شوند، می‌توانند به صورت ساختاری، نیمه ساختاری وهمچنین بدون ساختار باشند.

از آن جایی که نه می‌توان با استفاده از SQL سنتی مانند کوئری‌ها، روی کلان داده‌ها کار کرد و نه می‌توان از سیستم مدیریت پایگاه داده رابطه‌ای برای ذخیره‌سازی استفاده کرد، طیف گسترده ای از ابزارها و تکنیک‌های مقیاس‌پذیر پایگاه داده مجبور به تکامل شده اند. هدوپ یکی از راه‌حل‌های برجسته و شناخته شده است. پایگاه داده ی NOSQL هم با نمونه‌هایی مانند MongoDB و DynamoDB از سرور آپاچی، به عنوان یک پایگاه داده‌ی رابطه‌ای امتیاز خوبی به دست آورده است[1,2].

نیاز به کلان داده‌ها از شرکت‌های بزرگی مانند گوگل و فیس بوک و به منظور تجزیه و تحلیل حجم زیادی از داده‌ها شروع شد. پردازش این نوع از داده‌ها بسیار دشوار است چون شامل میلیاردها رکورد از اطلاعات میلیون‌ها نفر است. اطلاعاتی که شامل داده‌های تولید شده درشبکه‌های اجتماعی، تصاویر، داده‌های سمعی و غیره می‌باشد. در این مقاله ابتدا با معرفی شروع می‌کنیم و در مورد خصوصیات کلان داده‌ها یا همان "5V " صحبت می‌کنیم در ادامه اجزای مختلف کلان داده‌ها را بر اساس چارچوب هدوپ بررسی می‌کنیم. هدوپ یک فریم‌ورک متن باز است که برای ذخیره‌سازی و پردازش دیتا ستِ‌های موجود بر روی خوشه‌های سخت افزاری استفاده می‌شود. هدوپ توسط Doug Cutting و Mike Cafarella در سال ۲۰۰۵ ساخته شد[1].

۲- خصوصیات (5V)

کلان داده را می‌توان با ویژگی‌های زیر توصیف کرد:

- ظرفیت[1]: این ویژگی بدیهی‌ترین چالش را برای ساختار فناوری اطلاعات ارائه می‌دهد. اولین ویژگی که به ذهن اکثر مردم وقتی به کلان داده‌ها فکر می‌کنند می‌آید، ظرفیت است. بسیاری از شرکت‌ها در حال حاضر حجم زیادی از داده‌های بایگانی شده را به صورت فایل گزارش (لاگ) ذخیره می‌کنند، اما ظرفیت پردازش این داده‌ها را ندارند. مزیتی که از توانایی پردازش حجم زیادی از اطلاعات به دست می‌آید، جذابیت اصلی تجزیه و تحلیل کلان داده‌ها است.
- سرعت[2]: همان گونه که سرعت تولید داده در حال افزایش است، باید سرعت پردازش، ذخیره، تجزیه و تحلیل داده‌ها در پایگاه داده‌ای رابطه‌ای نیز افزایش می‌یافت. در واقع سرعت، تندی تولید داده ی جدید است و به محیط‌های مختلف منتقل می‌شود. به عنوان مثال در سال ۱۹۹۹ وقتی پیام‌های شبکه‌های اجتماعی بارها توسط کاربران به اشتراک گذاشته شدند حدود ۱۰۰۰ ترابایت (۱۰۰۰۰۰۰ گیگابایت) داده توسط پایگاه داده ی تحلیلی وال‌مارت ذخیره شد یا در سال ۲۰۱۲ توانست به ۲/۵ پتابایت (۲۵۰۰۰۰۰گیگابایت) داده دسترسی داشته باشد. کاربران روزانه به ازای هر یک دقیقه صدها ساعت فیلم در یوتیوب آپلود می‌کنند و از طریق جیمیل ۲۰۰ میلیون پیام ارسال می‌شود.
- تنوع[3]: جنبه ی بعدی کلان داده، تنوع آن است. کلان داده‌ها همیشه به صورت داده‌های ساخت یافته نیستند و قرار دادن این نوع داده‌ها در پایگاه داده‌های رابطه‌ای آسان نیست. مقوله‌ی کلان داده یک حقیقت حیاتی است که باید توسط تحلیلگران داده شناخته شود؛ تحلیلگرانی که با داده‌های ساخت یافته و ساخت نیافته (که به شدت پیچیدگی ذخیره‌سازی و تحلیل کلان داده‌ها را افزایش می‌دهد) دست و پنجه نرم می‌کنند. همچنین ۹۰ درصد داده‌هایی که تولید می‌شوند در مقوله‌ی داده‌های ساخت نیافته قرار می‌گیرند.
- صحت و صداقت[4]: هنگامی‌که ما با حجم، سرعت و تنوع بالایی از داده‌ها روبرو هستیم، ممکن نیست که همه ی داده‌ها صد در صد درست باشند، داده‌های سرکش نیز وجود دارد. کیفیت داده‌های موجود می‌تواند به شدت متفاوت باشد. در واقع دقت تحلیل داده‌ها به صحت و درستی داده‌های منبع بستگی دارد.
- مقدار[5]: مهم ترین جنبه در کلان داده، مقدار آن است. پتانسیل مقدارهایی که کلان داده‌ها دارند بسیار زیاد است. دسترسی داشتن به کلان داده‌ها بسیار خوب است اما اگر نتوانیم آن را به مقدار تبدیل کنیم کاملا بی فایده می‌شود. همچنین پیاده سازی زیرساخت سیستم‌های فناوری اطلاعات برای ذخیره‌سازی کلان داده‌ها بسیار هزینه بر است و بیزنس‌ها نیاز به بازگشت سرمایه دارند.

۳- تکنیک‌ها و تکنولوژی‌ها

کلان داده‌ها نه تنها بزرگ بلکه متنوع نیز هستند و رشد سریعی دارند، پس به منظور تلاش برای استخراج اطلاعات، به تکنولوژی‌ها و تکنیک‌های تحلیلی زیادی نیاز است و برای پردازش حجم زیادی از داده‌ها، به تکنولوژی‌های استثنایی نیاز خواهیم داشت. این تکنیک‌ها و تکنولوژی‌ها برای دستکاری، تجسم و تجزیه و تحلیل داده‌ها به کار برده می‌شوند. بنابراین برای مدیریت کلان داده‌ها، راه‌حل‌های زیادی وجود دارد، اما فناوری هدوپ یکی از پرکاربردترین فناوری‌ها است[3,4].

۳-۱- تکنیک‌ها

هنگام انجام یک پروژه ی کلان داده، انواع مختلفی از تکنیک‌ها وجود دارد که می‌توان از آنها استفاده کرد. این که کدام یک از آنها استفاده می‌شود به نوع داده‌ی درحال پردازش، تکنولوژی‌های قابل دسترس و همچنین به سوالات تحقیقاتی که تلاش در یافتن پاسخ آن دارید، بستگی دارد. در ادامه به بررسی برخی از ابزارهایی که مکرراً استفاده می‌شوند می‌پردازیم.

۳-۲- تکنولوژی

برای تسهیل تجزیه و تحلیل کلان داده‌ها، چندین محصول نرم افزاری و تکنولوژی‌های قابل استفاده وجود دارد. در ادامه به بررسی برخی از تکنولوژی‌هایی که معمولا مورد استفاده قرار می‌گیرند، می‌پردازیم.

هدوپ یک تکنولوژی کلیدی است که برای مدیریت کلان داده‌ها، تجزیه و تحلیل آنها و محاسبات جریانی استفاده می‌شود. درواقع یک پروژه ی نرم‌افزاری منبع باز است که پردازش توزیعی دیتا ست‌های بزرگ را در بین خوشه‌های سرورهای کالا امکان پذیر می‌کند.

۴- مولفه‌های کلان داده در فریم‌ورک هدوپ

آپاچی هدوپ یک فریم‌ورک منبع باز است که با محاسبات توزیعی امکان پردازش دیتا ستِ‌های بزرگ را در خوشه‌های رایانه ای فراهم می‌کند و از مدل‌های ساده ی برنامه‌نویسی کمک می‌گیرد. هدوپ به گونه ای طراحی شده است که مقیاس را از سرورهای واحد به هزاران ماشین افزایش بدهد، در حالی که هر کدام از ماشین‌ها مرکز محاسبه و ذخیره‌سازی محلی خود را دارند. پس این افزایش مقیاس یکی از مزایای ارائه شده توسط هدوپ است زیرا ما می‌توانیم از سخت افزارهای ارزان قیمت استفاده کنیم. این فریم‌ورک در سال ۲۰۰۵ توسط Doug Cutting و Mike Cafarella ساخته شد، اسم آن را از روی یک عروسک فیل انتخاب کرده اند[6,8,9].

هدوپ شامل این ماژول‌ها است:

۴-۱- سیستم فایل توزیعی هدوپ (HDFS)

یک سیستم فایل توزیعی است که دسترسی به داده‌های برنامه را با توان عملیاتی بالا فراهم می‌کند. این سیستم به ما کمک می‌کند که حجم زیادی از داده‌ها را با روشی قابل اطمینان ذخیره کنیم همچنین یک سیستم فایلی با تحمل خطای بالا برای ما ایجاد می‌کند.

گره‌ی اصلی گره‌ی name نامیده می‌شود و فراداده‌های خوشه را مدیریت می‌کند. از یک ساختار ارباب/برده پیروی می‌کند که در آن یک یا چند دستگاه به عنوان دستگاه‌های برده توسط یک دستگاه اصلی (ارباب) کنترل می‌شوند. گره‌ی برده، گره‌ی داده نامیده می‌شود و داده‌ها را در خود ذخیره می‌کند. این سیستم یک سیستم مبتنی بر جاوا می‌باشد[5].

Hadoop YARN/ Map Reduce -۴-۲

فریم‌ورکی برای برنامه ریزی شغلی و مدیریت منابع خوشه‌ای که به عنوان یک مدل برنامه‌نویسی در چارچوب هدوپ برای پردازش مقدار زیادی داده در یک محیط توزیع شده و موازی روی یک خوشه ساخته شده است.

Hbase -۴-۳

HBase یک پایگاه داده مقیاس‌پذیر و توزیع شده‌ی هدوپ است که از ذخیره‌سازی داده‌های ساختار یافته برای جداول بزرگ پشتیبانی می‌کند. همچنین با اجازه دادن به به‌روزرسانی‌ها،الحاق‌ها، حذف‌ها و غیره، قابلیت‌های معاملاتی را فراهم می‌کند. HBASE یک پایگاه داده‌ی غیر-رابطه‌ای (NOSQL) است که در بالای HDFS اجرا می‌شود. HBASE امکان خواندن/نوشتن به صورت تصادفی و بلادرنگ را برای کلان داده‌ها فراهم می‌کند. به صورت ستونی است و فضای ذخیره‌سازی تحمل پذیر و دسترسی سریعی را فراهم می‌کند[7].

Pig -۴-۴

یک چارچوب اجرای جریان داده و زبان سطح بالا برای محاسبه موازی است. Apache PIG یک زبان برنامه‌نویسی است که به کاربران امکان می‌دهد تا تغییرات پیچیده MapReduce از جمله خلاصه/جمع بندی، پیوست دادن، مرتب سازی و غیره را بنویسند. یکی از ویژگی‌های اصلی PIG پردازش موازی است که به آن امکان می‌دهد مجموعه دیتا سِت‌های بسیار بزرگ را مدیریت کند.

Hive -۴-۵

یک زیرساخت برای انبار داده است که برای سیستم ما خلاصه سازی داده‌ها و جست جوی موقت را فراهم می‌کند. Hive یک نرم افزار انبار داده است که برای مدیریت، استعلام، خلاصه‌سازی و تجزیه و تحلیل دیتا سِت‌های بزرگ استفاده می‌شود.

HiveQL یک زبان شبیه به SQL است که برای استعلام و یافتن پاسخ از میان پتابایت داده (۱۰۰۰۰۰۰ گیگابایت داده) استفاده می‌شود. این زبان برای تجزیه و تحلیل داده‌ها در HDFS استفاده شده و از مدل برنامه‌نویسی map/reduce کاملا پشتیبانی می‌کند.

مزیت‌هایی که Hive ارائه می‌دهد، خیلی شبیه زبان SQL سنتی است. در دیتا سِت‌های بزرگ سریع عمل می‌کند، مقیاس‌پذیر و قابل توسعه است و گزارش‌های مختلفی را فراهم می‌کند.

Sqoop -۴-۶

Sqoop ابزاری نرم افزاری است که برای انتقال داده‌های انبوه بین پایگاه داده‌های رابطه‌ای و هدوپ طراحی شده است. Sqoop برای انتقال اطلاعات از پایگاه داده‌های خارجی به HDFS یا HBASE یا HIVE ساخته شده است. این امکان را برای واردات و صادرات داده‌ها به پایگاه‌های ارتباطی خارجی و انتقال داده‌های موازی فراهم می‌کند.

ZooKeeper -۴-۷

یک سرویس هماهنگی با کارایی بالا برای برنامه‌های توزیع شده است. Zookeeper خدمات عملیاتی را در چارچوب هدوپ ارائه می‌دهد. Zookeeper یک سرویس متمرکز است که برای نگهداری اطلاعات، پیکربندی، همگام سازی داده‌ها و خدمات گروهی با استفاده از برنامه‌های توزیع شده استفاده می‌شود. معماری Zookeeper از دسترسی زیاد از طریق سرویس‌های اضافی (redundant services) پشتیبانی می‌کند، این فرایندهای توزیع مختلف را قادر می‌سازد تا از طریق فضای سلسله مراتبی مشترک ثبات‌ها به اسم znodes بین خودشان هماهنگ شوند.

Avro -۴-۸

Avro یک سیستم مرتب سازی داده‌هاست که داده‌ها را با استفاده از داده‌های باینری فشرده به صورت سریالی تنظیم می‌کند و ساختارهای داده‌ای غنی و یک پرونده کانتینر را برای ذخیره‌سازی داده‌های مداوم فراهم می‌کند، که به طرح‌های خواندن و نوشتن داده‌ها متکی است. برای تعریف انواع داده‌ها و پروتکل‌ها از(JSON (Java script open notation استفاده می‌کند. Avro از فرمت سیمی[۶] برای برقراری ارتباط بین گره‌های هدوپ و بین برنامه و سرویس‌های مشتری استفاده می‌کند.

Cassandra -۴-۹

یک پایگاه داده چند کاره، مقیاس‌پذیر و بدون هیچ نقطه خرابی است. Apache Cassandra یک سیستم مدیریت پایگاه داده توزیع شده منبع باز، بسیار مقیاس‌پذیر و با کارایی بالا است که توانایی مدیریت تعداد زیادی داده در چندین سرور را دارد. Cassandra باعث تحمل خطا می‌شود و غیرمتمرکز است.

Mahout -۴-۱۰

یک کتابخانه مقیاس‌پذیر برای یادگیری ماشین و داده کاوی است، Apache Mahout پروژه ای بر اساس نرم افزار آپاچی برای تولید و پیاده سازی رایگان الگوریتم‌های توزیع شده یا برای یادگیری ماشین مقیاس‌پذیر که عمدتاً در زمینه فیلتر مشترک، خوشه و طبقه بندی متمرکز است.

Tez -۴-۱۱

Tez یک فریم‌ورک برنامه ریزی شده برای جریان داده‌ها و یک موتور قدرتمند و انعطاف پذیر برای اجرای یک DAG (Directed Acyclic Graph) است، که وظایف دلخواه برای پردازش داده‌ها در موارد استفاده دسته ای و تعاملی فراهم می‌کند. Tez توسط Hive، Pig و فریم‌ورک‌های دیگر در اکوسیستم هدوپ و همچنین توسط سایر نرم افزارهای تجاری مانند ابزارهای ETL برای جایگزینی Hadoop MapReduce[۷] به عنوان موتور اصلی اجرا در حال اتخاذ است.

Spark -۴-۱۲

Spark یک مدل برنامه‌نویسی ساده و رسا است که از طیف گسترده ای از برنامه‌ها، یادگیری ماشین، فرآیند جریان و محاسبات عظیم نمودارها پشتیبانی می‌کند. Apache spark یک موتور الگوریتمی تجزیه و تحلیل سریع داده و یادگیری ماشین است که برای پردازش داده‌ها در مقیاس بزرگ استفاده می‌شود. Spark با هدوپ یکپارچه شده و دارای موتور تحلیلی پیشرفته ای است که با استفاده از آن در پردازش حافظه، سرعت آن را صد برابر Hadoop MapReduce می‌کند.

Flume -۴-۱۳

Flume یک سرویس توزیع شده قابل اعتماد برای جمع آوری و انتقال مقدار زیادی از اطلاعات لاگ است. این به کاربران کمک می‌کند تا از داده‌های ارزشمند لاگ استفاده کنند. Flume به جریان داده منابع مختلف اجازه می‌دهد که حجم زیادی از لاگ‌های مربوط به وب را در لحظه جمع‌آوری کند.

۵- نتیجه گیری

کلان داده‌ها فرصتی را برای تجزیه و تحلیل گسترده فراهم می‌کند که منجر به فرصت‌های بزرگ برای پیشرفت کیفیت زندگی یا حل رازهای جهان می‌شود. در این مقاله جزئیات مربوط به کلان داده با استفاده از چارچوب هدوپ به عنوان پایه مورد بحث قرار گرفته است. ما مشخصات و اطلاعات عمیقی راجع به اجزای مختلف کلان داده از نظر هدوپ را ارائه کردیم. امروزه با متمرکز کردن و تلفیق داده‌ها در ابر (cloud) ، بار زیاد اطلاعات را در همه جا میبینیم. روش‌های کلان داده بینش جدیدی در مورد مجموعه داده‌های موجود ارائه می‌دهد. آپاچی هدوپ یک چارچوب داده با رشدی سریع است[9,10].

آپاچی هدوپ یک پلت‌فرم منسجم و رایگان ارائه می‌دهد که یکپارچه سازی و پردازش داده‌ها، نظارت و برنامه‌ریزی گردش کار و غیره را دربرمی‌گیرد. کارهای آینده شامل یک مطالعه دقیق در مورد چالش‌ها و مسائل مربوط به کلان داده‌های صنایع مختلف است.

مراجع

[1] Apache Software Foundation. Official website https://hadoop.apache.org/

[2] University of Texas at Austin School of Information Big Data Analytics Dylan Maltby 1616 Guadeloupe, Austin, TX 78701 512-471-3821

[3] Bakshi, K, (2012) , "*Considerations for big data: Architecture and approach*"

[4] Chen, H. , Chiang, R. H. L. , & Storey, V. C. (2012) . Business Intelligence and Analytics: *From Big Data to Big Impact*. MIS Quarterly, 36(4) , 1165-1188.

[5] Picciano, A. G. (2012) . The Evolution of Big Data and Learning Analytics in American Higher Education. *Journal of Asynchronous Learning Networks*, 16(3) , 9-20.

[6] Big Data basics from oreilly: https://strata. oreilly. com/2012/01/what-is-big data. html

[7] White, Tom. *Hadoop the Definitive Guide 2nd Edi-tion. United States*: O'Reilly Media, lnc. , 2010.

[8] A. Vailaya "What's All the Buzz Around “Big Data? ”", IEEE Women in Engineering Magazine, December 2012, pp. 24-31.

[9] S. Madden, "From Databases to Big Data", IEE Inter-net Computing, June 2012, v. 16, pp. 4-6

[10] Katal, A Wazid, M. ؛ Goudar, R. H. (Aug,2013) , “Big data: Issues, challenges, tools and Good practices”.

پانویس‌ها

[1] Volume

[2] Velocity

[3] Variety

[4] Veracity

[5] Value

[6] Wire format: قالب‌های سیم فرمی را برای ارسال یا دریافت پیام توسط نقاط انتهایی تعریف می‌کند. ActiveEnterprise Message یک فرم XML پیام خارجی است که توسط TIBCO Adapter SDK پشتیبانی می‌شود.

[7] MapReduce: مجموعه ای از توابع کتابخانه را در دل خود دارد که جزئیات و پیچیدگی را از دید برنامه‌نویس پنهان می‌کند