کاربرد رایانش ابری در داده‌های حجیم

الهام عنایتی[1]

[1]مربی، گروه علوم کامپیوتر، دانشگاه بجنورد ، بجنورد
e.enayati@ub.ac.ir

چکیده

داده‌های حجیم یک اصطلاح برای مجموعه‌های داده خیلی بزرگ است که از نظر ساختار، پیچیدگی و منابع تولید بسیار متنوع هستند و ذخیره و آنالیز آنها کار پیچیده‌ای است. رایانش ابری یک تکنولوژی قدرتمند برای اجرای محاسبات پیچیده و سنگین است. رایانش ابری نیاز به استفاده از سخت‌افزارهای گران را حذف نموده و فضای محاسباتی و نرم‌افزار مورد نیاز را در اختیار کاربر قرار می‌دهد. رشد روزافزون حجم داده و ایجاد داده‌های حجیم از طریق رایانش ابری در سال‌های اخیر در بسیاری از کاربردها دیده شده است. داده‌های حجیم چالش مهمی است که احتیاج به زیرساختی قوی برای اطمینان از انجام موفق پردازش‌ها و آنالیزهای مورد نیاز دارد. موضوع حایز اهمیت این است که چگونه می‌توان از زیرساخت رایانش ابری برای دسترسی، پردازش و آنالیز داده‌های حجیم استفاده نمود. در این مطالعه به تعریف، خصوصیات و دسته‌بندی داده‌های حجیم در چارچوب رایانش ابری پرداخته شده است و کاربرد زیرساخت رایانش ابری برای آنالیزهای داده‌های حجیم مورد بررسی قرار گرفته‌است.

کلمات کلیدی

داده‌های حجیم، رایانش ابری، دسته‌بندی داده‌های حجیم، کاربرد داده‌های حجیم، Hadoop.

1- مقدمه

امروزه داده‌های حجیم[1] در مرکز توجه علوم مدرن و کسب و کار است. این داده‌ها از تراکنش‌های آنلاین، ایمیل‌ها، ویدیوها، صدا، تصاویر، جریان‌های کلیک، گزارش خطاها، پست‌ها، گزارشات جستجو، رکوردهای اطلاعات سلامت، عملیات متقابل در شبکه‌های اجتماعی، داده های علمی، حسگرها، تلفن‌های همراه و نرم‌افزارهای روی تلفن‌های همراه تولید می‌شوند[2-1]. دیتابیس‌های حاوی این داده‌ها به سرعت رشد می‌کنند و نظارت، فرم‌دهی، ذخیره، مدیریت، اشتراک گذاری، آنالیز و مجازی‌سازی آنها از طریق ابزارهای نرم‌افزاری معمول دشوار است. یکی از چالش‌های مهم محققین این است که با رشد سریع داده‌های حجیم ، نیاز به طراحی پلتفرم‌های رایانش ابری مناسب جهت آنالیز و بروز رسانی داده به سرعت افزایش می‌یابد.

رایانش ابری، یکی از بخش‌های مدرن تکنولوژی‌های ارتباطی فعلی است و سرویسی برای کاربردهای سازمانی است که با معماری قدرتمند قادر به اجرای محاسبات پیچیده در ابعاد بزرگ است. از مزایای رایانش ابری ایجاد منابع مجازی، پردازش‌های موازی، امنیت و تجمیع سرویس در انباره‌های داده است[3]. بخشی از اولین آداپتورهای اولیه داده‌های حجیم در رایانش ابری، کاربران هستند. این محیط‌ها توسط فراهم کنندگان خدمات ابری نظیر IBM، ویندوز Azure و AWS[2]آمازون[4]تهیه می‌شوند.

هدف از این مطالعه پیاده سازی یک تحقیق اختصاصی درباره وضعیت داده-های حجیم در محیط‌های رایانش ابری و تعریف، بیان ویژگی‌ها و دسته‌بندی داده‌های حجیم براساس مباحث رایانش ابری است. در این مطالعه، ارتباط بین داده‌های حجیم ، رایانش ابری و سیستم‌های ذخیره داده‌های حجیم مورد بررسی قرار گرفته است.

[1] Big Data

[2]Amazon Web Services

در ادامه مطالب به این صورت ارایه شده‌اند؛ بخش دوم در مورد تعریف، ویژگی‌ها و منابع تولید و دسته‌بندی‌های مختلف داده‌های حجیم است. بخش‌های سوم و چهارم در خصوص ارتباط و کاربرد داده‌های حجیم و رایانش ابری است. در بخش پنجم به بررسی چند نمونه عملی و آکادمیک از کاربردهای رایانش ابری در داده‌های حجیم می‌پردازیم و سرانجام خلاصه‌ای از مطالب بیان شده و مباحث آتی در بخش ششم آمده است.

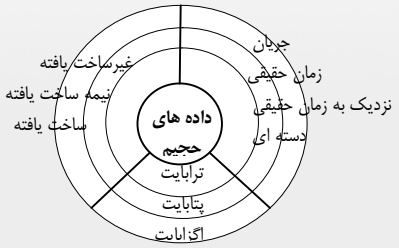
2- داده‌های حجیم چیست و چگونه تولید می شود؟

حجم اطلاعاتی که تا سال 2003 توسط انسان ایجاد شد تنها 5 اگزابایت (10^{18}بایت) است اما امروزه این حجم از اطلاعات تنها در عرض دو روز ایجاد می‌شود[5]. IBM [6] در تحقیقی نشان داد که هر روز 2/5 اگزابایت داده تولید می‌شود و حدود 90% داده‌های موجود تنها در دو سال اخیر تولید شده‌است[7]. هر کامپیوتر شخصی حدود 500 گیگابایت اطلاعات در خود نگهداری می‌کند و در دنیا حدود 20 میلیون کامپیوتر شخصی وجود دارد. در گذشته فرآیند توصیف ژن انسان حدود 10 سال طول می‌کشید در حالی که امروز در کمتر از یک هفته انجام می‌شود[8]. شرکتی مثل گوگل بیلیون‌ها سرور در سطح جهان دارد. حدود 6 بیلیون مشترک تلفن همراه در جهان همه روزه 10 میلیون پیام متنی ارسال و دریافت می‌کنند و تا سال 2020 حدود 50 بیلیون وسیله متصل به اینترنت و شبکه وجود خواهد داشت[9].

از سال 2012، داده‌های حجیم به عنوان یک پروژه مهم و جهانی مطرح شد. پروژه‌ای که به جمع‌آوری، بصری‌سازی[1] و آنالیز مقدار زیادی داده می‌پردازد. در راستای این پروژه اطلاعات آماری زیادی ارایه گردید. فیس‌بوک[2] ماهانه حدود 955 میلیون کاربر فعال به 70 زبان زنده دنیا دارد و حدود 140 بیلیون عکس در آن بارگذاری می شود و 125 میلیون ارتباط دوستی برقرار می‌گردد. هر روزه 30 بیلیون نوشته و 2/7 بیلیون لایک و توضیحات ارسال می‌گردد. در یوتیوب[3] هر دقیقه 48 ساعت ویدیو بارگذاری و هر روزه 4 بیلیون فیلم اجرا می‌گردد. گوگل نیز از سرویس‌های زیادی پشتیبانی می‌کند از جمله نظارت بر 7/2 بیلیون صفحه در هر روز و 20 پتابایت (10^{15}بایت) فرآیند روزانه و ترجمه به 66 زبان؛ یک بیلیون توییت[4] در هر 72 ساعت، بیشتر از 140 میلیون فعالیت کاربران توییتر[5] است. 571 وب سایت جدید در هر دقیقه از روز ایجاد می شود[10]. پیش‌بینی می‌شود در طی دهه آینده حجم اطلاعات 50 بار افزایش یابد، البته همزمان تعداد تکنولوژی های خاص اطلاعاتی که برای نگهداری این داده‌ها ایجاد می شود نیز 1/5 برابر می‌گردد[11].

2-1- تعریف داده‌های حجیم و ویژگی‌های آن

برای داده‌های حجیم تعاریف مختلفی ارایه شده است. داده‌های حجیم را می‌توان داده‌هایی که پردازش آنها خارج از حد توان سیستم‌های کنونی است تعریف نمود[12] و یا داده‌های حجیم را افزایش حجم داده دانست به نحوی که ذخیره، پردازش و آنالیز آن از طریق تکنولوژی‌های قدیمی دیتابیس‌ها به سختی ممکن باشد[13]. داده‌های حجیم سه ویژگی(3V) اساسی دارد: تنوع[6]، حجم[7] و سرعت[8][2, 7-6, 14] این ویژگی‌ها را در شکل (1) می‌بینید.



شکل (1): ویژگی های داده‌های حجیم(3V)

تنوع: داده‌های حجیم را واقعا حجیم می‌کند. داده‌های حجیم از منابع بسیار متنوع می‌آید و معمولا در سه نوع می باشد: ساخت‌یافته، نیمه‌ساخت‌یافته و غیرساخت‌یافته[9]. داده‌های ساخت‌یافته در انباره‌های داده تگ خورده و به آسانی ذخیره می‌شوند اما داده‌های غیرساخت‌یافته به صورت تصادفی بوده و آنالیز آنها دشوار است. داده‌های نیمه‌ساخت‌یافته از فیلدهای ثابتی تشکیل نشده‌اند اما تگ‌هایی برای جداسازی عناصر داده دارند.[1, 7]

حجم: حجم و اندازه داده‌های امروزه بزرگتر از ترابایت و پتابایت می‌باشد. با افزایش حجم داده، روش‌های ذخیره و تکنیک‌های آنالیز سنتی داده غیرقابل استفاده می شوند[1, 15].

سرعت: نه تنها برای داده‌های حجیم بلکه برای همه فرآیندها لازم است. برای فرآیندهای محدود به زمان در داده‌های حجیم می‌بایست جریان داده به داده سازمان یافته تبدیل گردد تا ارزش آن به حداکثر برسد[1, 15].

امروزه V های دیگری نیز به ویژگی‌های اساسی(3V) داده‌های حجیم اضافه شده است نظیر صحت[10]، اعتبار[11]، فراری[12]، ارزش[13][16]. در منبع [13] داده‌های حجیم به صورت 4V تعریف شده که ویژگی ارزش، به ویژگی‌های اساسی داده‌های حجیم اضافه شده‌است.

2-2- دسته بندی داده‌های حجیم

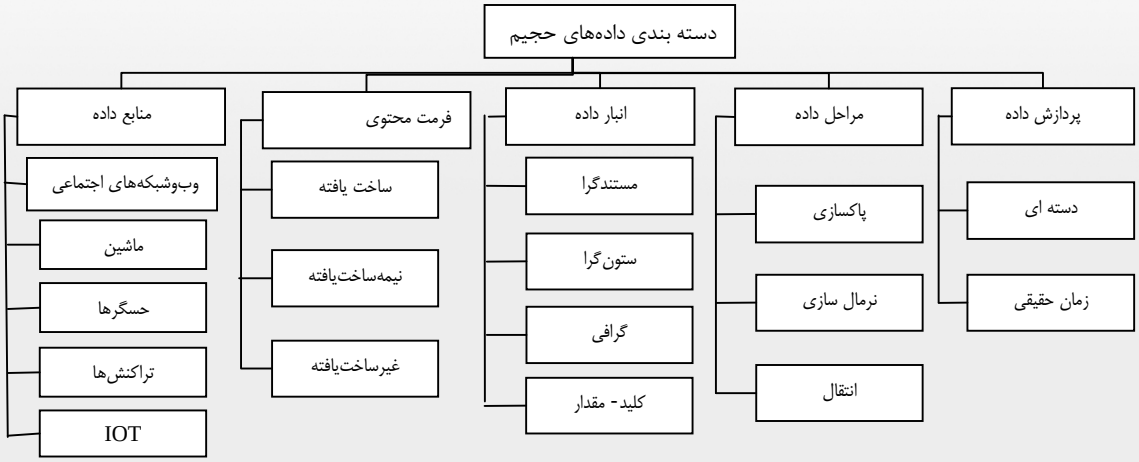
برای فهم بهتر داده‌های حجیم آنرا براساس مشخصه‌هایشان به دسته‌های مختلفی تقسیم‌بندی می‌نمایند. در شکل (2) دسته‌های مختلف داده‌های حجیم نشان داده‌شده‌است. دسته‌بندی براساس پنج جنبه انجام می‌شود: منبع داده، فرمت محتوی، انبار داده، مراحل داده و پردازش داده. هرکدام از این جنبه‌ها، ویژگی‌ها و پیچیدگی‌های خاص خود را دارند که در ادامه آمده است.

منابع داده‌های حجیم:

منابع تولید داده‌های حجیم شامل داده‌های اینترنتی، داده‌های کسب شده از حسگرها و اطلاعات ذخیره شده از تراکنش‌هاست که کلیه داده‌های غیرساخت یافته تا داده‌های ساخت یافته در فرمت‌های مختلف را در برمی‌گیرد[17]. در

[1] visualize
[2] Facebook
[3] YouTube
[4] Tweets
[5] Twitter

[6] Varity
[7] Volume
[8] Velocity
[9] Structed, Semi-Structed,Unstructed
[10] Veracity
[11] Validity
[12] Volacity
[13] variability or Value



شکل (2) دسته بندی داده‌های حجیم

نتیجه تنوع گسترده منابع داده، داده به دست آمده از نظر افزونگی، ثبات و نویز متفاوت است. انواع مختلف منابع تولید داده‌های حجیم به شرح زیر است:

الف- مدیای اجتماعی[1]: مدیای اجتماعی، اطلاعاتی است که از طریق به به اشتراک گذاری و یا تبادل اطلاعات از طریق آدرس های اینترنتی و یا از طریق ارتباطات مجازی و شبکه‌های مجازی به دست می‌آیند، نظیر اطلاعاتی که در پروژه‌های اشتراکی، بلاگ‌ها، میکروبلاگ‌ها، فیس‌بوک و توییتر تولید می‌شوند[13].

ب- داده‌های ماشین: داده ماشینی، اطلاعاتی است که به صورت اتوماتیک توسط سخت‌افزار و نرم‌افزارهای ابزارهایی نظیر کامپیوترها، وسایل پزشکی یا دیگر ماشین‌ها بدون دخالت انسان تولید می‌شود[13].

ج- حسگرها: وسایل حسگر مختلفی برای اندازه‌گیری کمیت‌های فیزیکی و تبدیل آنها به سیگنال وجود دارد[13] که بخشی از داده‌های حجیم را تولید می‌نماید.

د- تراکنش‌های اینترنتی(IoT)[2]: IoT یک مجموعه از اشیایی است که که به صورت یکتا قابل تعریف هستند و به عنوان بخشی از اینترنت می‌باشند. این اشیاء شامل تلفن‌های کوچک، دوربین‌های دیجیتال و تبلت‌ها هستند. وقتی این وسایل از طریق اینترنت به یکدیگر متصل می‌شوند قادرند بیشتر پردازش‌های کوچک و سرویس‌های پشتیبانی پایه‌ی اقتصادی، محیطی و سلامت مورد نیاز را فراهم آورند. تعداد زیاد وسایل متصل به اینترنت، انواع مختلفی از سرویس‌ها را فراهم می‌آورند و مقادیر زیادی داده و اطلاعات تولید می‌نمایند[18].

فرمت محتوی داده‌های حجیم:

داده‌های حجیم از نظر فرمت به سه دسته کلی تقسیم‌بندی می‌شود:

الف- داده ساخت‌یافته: داده‌های ساخت‌یافته اغلب در بانک‌های SQL مدیریت می‌شوند. برای مدیریت و پرس‌وجوی داده‌ها در RDBMS[3] یک زبان برنامه نویسی ویژه به وجود آمده است. ورود، پرس‌وجو، ذخیره و آنالیز داده‌های ساخت‌یافته به آسانی انجام می‌شود. نمونه‌ای از داده ساخت‌یافته شامل اعداد، کلمات و تاریخ‌ها می‌باشند.

ب- داده نیمه‌ساخت‌یافته: داده‌ای است که یک سیستم دیتابیسی معمول را دنبال نمی‌کند. داده نیمه‌ساخت‌یافته ممکن است به شکل یک داده ساخت-یافته باشد که در یک مدل دیتابیس رابطه‌ای نظیر جداول، سازمان‌دهی نشده است. به دست آوردن آنالیز داده نیمه‌ساخت‌یافته نسبت به آنالیز داده با فرمت ثابت، متفاوت است بنابراین کسب و کشف اطلاعات از داده‌های نیمه‌ساخت-یافته نیازمند استفاده از قوانین پیچیده‌ای است که بتواند به صورت دینامیک قادر به تصمیم در مورد پردازش‌های بعدی باشد[19].

ج- داده غیرساخت‌یافته: داده‌های غیرساخت‌یافته نظیر پیام‌های متنی، اطلاعات مکانی، ویدیوها و داده‌های حاصل از مدیای اجتماعی هستند که هیچ فرمت مشخصی را دنبال نمی‌کنند.

انباره‌های داده‌های حجیم:

داده‌های حجیم، در انباره‌های داده مختلفی ذخیره می‌شوند که از نظر ساختار و تکنولوژی دسترسی متفاوتند که در زیر به آن‌ها اشاره شده است.

الف- انبار داده مستندگرا: انباره‌های داده مستندگرا به صورت پایه برای ذخیره و بازیابی مجموعه‌های مستندات، اطلاعات و پشتیبانی داده‌های پیچیده شکل گرفته‌اند که در چندین فرمت استاندارد نظیر xml,json و باینری (Msword,pdf,...) قرار دارند. یک انبار داده‌ی مستندگرا شبیه یک رکورد یا یک سطر از یک دیتابیس رابطه‌ای است اما با انعطاف بیشتر و قابلیت بازیابی بهتر مستندات برپایه محتوی آنها (نظیر MongoDB, CouchDB,...)

ب- انبار داده ستون‌گرا: محتوی یک انبار داده ستون‌گرا، در ستون‌هایی از سطرها نگهداری می‌شود و مقادیر صفات متعلق به یک ستون به صورت پشت سرهم ذخیره می‌شوند. سیستم‌های دیتابیس ستون‌گرا نسبت به دیتابیس‌های کلاسیک که محتوایشان به صورت سطرهای پشت سرهم قرار دارند، متفاوت است[20] نظیر جداول بزرگ[21].

ج- انبار داده‌گرافی: یک دیتابیس گرافی نظیر Neo4j برای ذخیره و نمایش داده‌ها از یک مدل گرافی شامل نودها و لبه‌ها استفاده می‌کند که در آن خصوصیات داده‌ها از طریق روابط به یکدیگر مرتبط می‌شوند[22].

د- انبار داده کلید- مقدار: دیتابیس‌های کلید- مقدار، دیتابیس‌های ارتباطی متناوبی هستند که برای ذخیره و دسترسی به داده‌های در اندازه خیلی بزرگ طراحی شده اند[23]، Dynamo یک نمونه خوب برای سیستم‌های ذخیره کلید- مقدار با دسترسی بالاست[24] که توسط amazon.com در

[1] Social Media
[2] Internet of Things
[3] Relational Data Base Management System

برخی از سرویس‌ها استفاده می‌شود. به طور مشابه در [25] یک روش برای انبار داده کلید- مقدار مقیاس‌پذیر پیشنهاد شده است که در آن از تراکنش‌های چند کلیدی، تنها با دسترسی به یک کلید استفاده می‌شود؛ این روش برپایه روش کلید مقدار عمل می‌کند و برای استفاده در G-store طراحی شده است. در [26] یک روش خوشه‌ای مقیاس‌پذیر برای اجرای یک کار در مقیاس بزرگ از مجموعه داده ارایه شده است. نمونه‌های دیگر از انبار داده کلید-مقدار در ApachiHbase[27] و [28]ApachiCassandra و voldemort می‌باشد. Hbase از HDFS[1] استفاده می‌کند که یک نسخه کد باز از جدول بزرگ گوگل است که در Cassandra ساخته شده‌است. Hbase داده را در جداول، سطرها و سلول‌ها ذخیره می‌کند. هر سطر به وسیله کلید سطر ذخیره می‌شود. هر سلول در جدول به وسیله یک کلید سطر، یک کلید ستون و یک نسخه مشخص می‌شود و محتوای آن شامل آرایه‌ای از بایت‌های تفسیر نشده است.

مراحل داده‌های حجیم:

بر روی داده‌های حجیم سه مرحله فرآیند داده‌ای انجام می‌شود:

الف- پاکسازی داده: فرآیند شناسایی داده‌های ناکامل و غیرمعتبر[29]

ب- انتقال داده: فرآیند انتقال داده به فرم مناسب برای آنالیز[13]

ج- نرمال‌سازی داده: روشی در طرح دیتابیس ساخت‌یافته برای کاهش افزونگی داده[30]

پردازش داده‌های حجیم:

روش‌های مختلف پردازش داده‌های حجیم به شرح زیر است:

الف- پردازش داده دسته‌ای: در سال‌های اخیر از سیستم‌های برپایه MapReduce برای انجام کارهای دسته‌ای با زمان اجرای طولانی استفاده می‌شود[31]. چنین سیستم‌هایی به کاربردها اجازه می‌دهند تا از طریق خوشه‌های بزرگ ماشین‌ها، هزاران نود را با هم مقایسه کنند.

ب- پردازش داده زمان حقیقی: یکی از معروفترین و قدرتمندترین ابزارهای داده‌های حجیم برپایه پردازش زمان حقیقی، S4[2] است[32]. S4 پلتفرم محاسباتی توزیع شده‌ای است که به برنامه‌نویس اجازه می‌دهد کاربردها را برای انجام پردازش جریان‌های نامحدود داده در زمان حقیقی توسعه دهند.

3- ارتباط داده‌های حجیم و رایانش ابری[3]

رایانش ابری و داده‌های حجیم موضوعاتی به هم پیوسته اند. داده‌های حجیم برای کاربران امکان استفاده از محاسبات در فرآیندهای پرس‌وجوی توزیع شده را فراهم می‌آورد که این پرس‌وجو در مجموعه‌های داده مختلف است و امکان بازگشت نتیجه در زمان مناسب را دارد. نحوه استفاده از رایانش ابری در داده-های حجیم در شکل (3) نشان داده شده است. منابع داده‌های حجیم ابری و وبی در دیتابیس‌های توزیع شده با قابلیت تحمل‌پذیری خطا ذخیره شده‌اند و پردازش آنها از طریق یک مدل برنامه سازی برای مجموعه‌های داده‌های حجیم انجام می‌شود که از الگوریتم‌های توزیع شده موازی خوشه‌ای استفاده می‌کند. همان‌گونه که در شکل (3) می‌بینید هدف اصلی از ویژوال‌سازی داده‌ها استفاده از نتایج آنالیز به صورت تصویری برای اخذ تصمیمات مناسب است[13].

4- کاربرد رایانش ابری در داده‌های حجیم

داده‌های حجیم از تکنولوژی ذخیره توزیع شده، برپایه رایانش ابری استفاده می‌کند که نقطه مقابل ذخیره‌سازی محلی برروی وسایل الکترونیکی یا کامپیوترهاست. همزمان با رشد سریع کاربردهای ابری که از تکنولوژی‌های ویژوال‌سازی استفاده می‌کنند، داده‌های حجیم نیز توسعه می‌یابد. بنابراین رایانش ابری نه تنها امکان استفاده از محاسبات و پردازش‌های داده‌های حجیم را فراهم می‌آورد بلکه به عنوان یک مدل سرویس نیز می باشد. جدول(1) مقایسه چندین فراهم کننده ابر برای داده‌های حجیم را نشان می دهد.

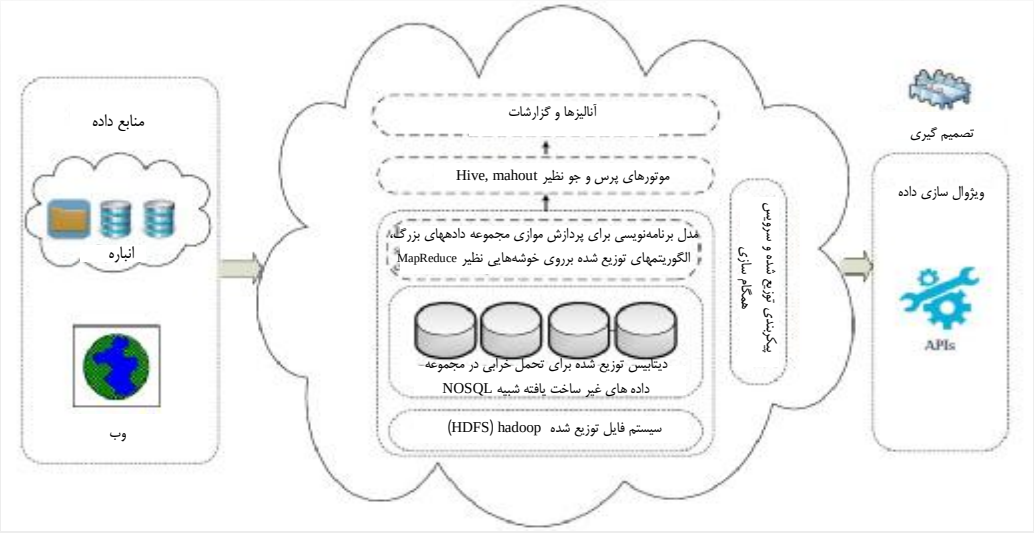
تالیا در منبع [33] در خصوص پیچیدگی و تنوع نوع داده‌ها و قدرت اجرای آنالیز بر روی مجموعه داده‌های حجیم بحث می‌کند. نویسنده می‌گوید زیرساخت رایانش ابری می‌تواند به صورت یک پلتفرم موثر برای آنالیز داده-های حجیم عمل نماید. رایانش ابری یک الگوی جدید زیرساخت محاسباتی است که روشی مناسب برای پردازش داده‌های حجیم در ابر فراهم می‌آورد و توسط همه انواع منابع در دسترس قابل استفاده است. به دلیل پیچیدگی زیادِ برخورد با داده‌های حجیم در پردازش‌های همزمان، چندین تکنولوژی مبتنی بر ابر، جهت مواجه با این محیطها به وجود آمده‌است[34]. MapReduce [35] یک نمونه خوب از فرآیندهای داده‌های حجیم در محیط ابری است که حجم زیادی از داده‌ها را در خوشه‌ها به صورت موازی ذخیره می‌کند.

همچنین بویلر و فایر استون[36] به توانایی محاسبات خوشه‌ای برای فراهم آوردن زمینه میزبانی رشد داده‌ها تاکید می‌کنند. میلر[37] به تاثیرکاهش میزان دسترس‌پذیری داده‌ها و هزینه گزاف این عدم دسترسی به دلیل انتقال تصمیمات در روش‌های تجربی از طریق کامپیوترها می‌پردازد. استفاده غلط از متدها یا ضعف ذاتی یک روش باعث تولید تصمیمات غلط و یا تحمیل هزینه می‌شود. DBMS[4]ها به عنوان بخشی از معماری فعلی رایانش ابری هستند و نقش مهمی در اطمینان از انجام آسان تراکنش‌های کاربردها از زیرساخت‌های قدیمی به معماری‌های زیرساخت ابری ایفا می‌کنند. از سوی دیگر رویکرد سازمان‌ها به پیاده‌سازی تکنولوژی‌هایی نظیر رایانش ابری، چالش‌های داده‌های حجیم و پردازش‌های مورد تقاضا در آن را با ریسک‌های پیش‌بینی نشده‌ای همراه می‌کند.

5- بررسی موردی ارتباط داده‌های حجیم و رایانش ابری

در اینجا با بررسی چند مطالعه موردی به بیان نحوه ارتباط بین داده‌های حجیم و رایانش ابری می‌پردازیم. این بخش به دو قسمت تقسیم می‌شود، قسمت اول بررسی نمونه‌هایی است که فروشندگان تکنولوژی‌های بزرگ در محیط ابری از آن استفاده می‌کنند و قسمت دوم بررسی تعدادی از مطالعات آکادمیک و مقالات علمی مرتبط با موضوع است.

[1] Hadoop Distributed File System
[2] Simple Scalable Streaming System
[3] Cloud Computing
[4] Data Base Management Systems



شکل (3) کاربرد رایانش ابری در داده‌های حجیم[13]

جدول(1) مقایسه چند پلتفرم ابری داده‌های حجیم

	گوگل	مایکروسافت	آمازون	کلودرا
انبار داده‌های حجیم	سرویسهای ابر گوگل	Azure	S3	
MapReduce	APPEngine	Hadoop on Azure	Elastic MapReduce(Hadoop)	YARN
آنالیز داده‌های حجیم	BigQuery	Hadoop on Azure	Elastic MapReduce(Hadoop)	Elastic MapReduce(Hadoop)
دیتابیس ارتباطی	ابرSQL	SQL Azure	اراکل یا MySQL	اراکل , MySQL , postgreSQL
دیتابیس NoSQL	انبار داده AppEngine	ذخیره جدولی	DynamoDB	AppachiAccumulo
پردازش جریان	جستجوی API	Streaminsight	هیچ جریان از پیش بسته بندی شده ای ندارد	Appachi Spark
یادگیری ماشین	پیش بینیAPI	Hadoop+Mahout	Hadoop+Mahout	Hadoop+Oryx
ورود اطلاعات	شبکه	شبکه	شبکه	شبکه
منابع داده	مجموعه داده‌نمونه‌کوچک	بازار ویندور Azure	دیتابیس های عمومی	دیتابیس های عمومی
دسترس پذیری	تعدادی از سرویسها در بتای خصوصی	تعدادی از سرویسها در بتای خصوصی	تولیدات عمومی	صنایع

5-1- بررسی موردی سازمان‌های استفاده کننده از تکنولوژی رایانش ابری در داده‌های حجیم

سازمان‌های مورد مطالعه از بین مشتریان شرکت‌هایی نظیر گوگل، آمازون و مایکروسافت انتخاب شده‌اند. بررسی نشان می‌دهد استفاده از تکنولوژی‌های رایانش ابری در آنالیز داده‌های حجیم و در مدیریت افزایش حجم، تنوع و سرعت اطلاعات دیجیتال کاربرد دارد. در اینجا یک مجموعه پنج موردی برای نمایش تنوع گسترده تحقیقات مرتبط با رایانش ابری انتخاب شده‌است.

swiftkey :

swiftkey یک تکنولوژی زبانی است که در سال 2008 در لندن به وجود آمد. این تکنولوژی به تایپ کردن در صفحات لمسی با استفاده از ایجاد پیش-نویس‌های شخصی‌سازی شده و انجام یکسری تصحیحات کمک می‌کند. کمپانی طراح، چندین ترابایت داده را جمع‌آوری و آنالیز نمود تا مدل‌های زبانی لازم برای کاربران فعال را به وجود آورد. بدین منظور کمپانی، به یک سیستم چند لایه با مقیاس‌پذیری بالا نیاز داشت تا به صورت مداوم همگام با افزایش تقاضا پیش رود و نیز یک موتور قوی برای پردازش تکنولوژی هوش مصنوعی نیاز داشت تا پیش‌گویی‌های لازم را انجام دهد. برای رسیدن به این اهداف از ApachiHadoop بر روی سرویس ذخیره‌سازی آمازون و رایانش ابری قابل انعطاف آمازون برای مدیریت پردازش استفاده شد تا پردازش چندگانه‌ای بر روی چندین ترابایت داده انجام شود. با استفاده از این راه حل جدید، swiftkey قادر به توزین سرویس در زمان‌های اوج تقاضا شد.

استودیو 343

هالو یک بازی داستانی تخیلی است که به یک پدیده عمومی تبدیل شده است. تاکنون بیشتر از 50 میلیون کوپی از بازی‌های ویدئویی هالو در جهان به فروش رسیده است. قبل از گسترش هالو4 توسعه‌دهندگان، داده‌ها را برای کسب آگاهی از نظرات بازیکنان و مسابقات آنلاین بررسی نمودند. برای کامل کردن این کار، تیم از سرویس تحت windows Azure HDInsight که در چارچوب داده‌های حجیم Apache Hadoop کار می‌کند، استفاده نمود. با این طراحی، تیم قادر به تهیه آمارهای بازی برای اپراتورهای مسابقات، جهت رده‌بندی بازیکنان شد؛ این آمارها با استفاده از سرویس HDInsight و پردازش و آنالیز داده‌های خام از طریق ویندوز Azure تهیه شد. تیم همچنین

از سرویس HDInsight برای بروزرسانی هفتگی هالو4 و پشتیبانی روزانه ایمیل‌های مسابقاتی جهت حفظ مشتریان خود استفاده نمود. دیگر سازمان‌ها نیز می‌توانند برای اخذ تصمیمات حرفه‌ای از چنین داده‌هایی استفاده نمایند.

RedBus

آژانس‌های مسافرتی آنلاین RedBus، برای اولین بار بلیط اینترنتی را در سال 2006 در هند عرضه نمودند. بدین ترتیب برای ده‌ها هزار اتوبوس عمل رزرو بلیط به صورت متمرکز انجام می‌شد. کمپانی به یک ابزار قدرتمند برای آنالیز داده‌های فروش و رزرو بلیط نیاز داشت. این داده‌ها از طریق اپراتورهای صدها اتوبوس که ده‌ها هزار مسیر را پشتیبانی کردند ارسال می‌شد. در ابتدا آنها از سرورهای خوشه‌های Hadoop برای پردازش داده استفاده نمودند اما این سیستم‌ها به زمان و منابع زیادی نیاز داشتند بنابراین با استفاده از خوشه‌های سرورهای Hadoop نمی‌توانستند آنالیز سریع و روشنی که مورد نیاز شرکت بود فراهم آورند. بنابراین RedBus از googleQuery برای آنالیز مجموعه‌های داده‌های حجیم که در زیرساخت پردازش داده گوگل وجود دارد استفاده نمود. درنتیجه، کسب دانش با استفاده از BigQuery سبب شد تا RedBus به یک کمپانی موفق تبدیل شود.

نوکیا

نوکیا یک کمپانی ارتباطی موبایل است که تولیدات آن بخش مهمی از زندگی مردم را فرا گرفته است. تعداد زیادی از مردم از موبایل‌های نوکیا برای ارتباطات، تصویربرداری و به اشتراک گذاری اطلاعات استفاده می‌کنند؛ بنابراین نوکیا به جمع‌آوری و آنالیز مقدار زیادی اطلاعات از طریق موبایل‌ها پرداخت. نوکیا برای استفاده وسیع از داده‌های حجیم به یک اکوسیستم از تکنولوژی‌ها نیاز دارد که شامل یک تراداتا انبار داده، تعداد زیادی مراکز داده اوراکل و MySQL و تکنولوژی‌های ویژوال سازی و Hadoop است. نوکیا بالغ بر 100ترابایت داده ساخت‌یافته بر روی چندین ترادادە دارد و همچنین چندین پتابایت داده چندساختاری بر روی انبار داده HDFS. HDFS انبار داده‌ای است که داده‌های چندساختاری و نیمه ساخت یافته را در خود ذخیره می‌کند[13].

Alacer

در خرده‌فروشی‌های آنلاین غالبا کسری درآمد وجود دارد که به علت غیرقابل اعتماد بودن هشدارهای زمان واقعی سرویس در پلتفرم ابری تجارت الکترونیک است. Alacer از الگوریتم‌های داده‌های حجیم برای تولید یک سیستم نظارتی استفاده می‌کند که هشدارهای فعال و غیرفعال را تحویل می‌دهد. با استفاده از پلتفرم نظارتی Alacer در رایانش ابری، زمان پاسخ از یک ساعت به چند ثانیه تقلیل یافت و درنتیجه افزایش رضایتمندی مشتریان و کاهش جریمه‌های توافقنامه سطح خدمات را به همراه داشت[13].

5-2- بررسی موردی کاربرد رایانش ابری و داده‌های حجیم در منابع آکادمیک و مقالات

بررسی موارد مطالعاتی که در ادامه آمده است نشان می‌دهد که محققان چگونه از تکنولوژی رایانش ابری در پروژه‌های داده‌های حجیم استفاده می‌کنند.

کاوش توییتر در ابر

نورد هوس و دیگران[38] از رایانش ابری برای آنالیز مقادیر زیاد داده در توییتر استفاده کردند. نویسنده برای به دست آوردن رتبه بندی کاربران توییتر الگوریتم PageRank را به کاربرده است. از زیرساخت ابری آمازون برای میزبانی همه محاسبات وابسته استفاده شد. محاسبات در دو گام انجام گرفت: 1- فاز خزیدن، همه داده‌ها از توییتر بازیابی شدند. 2- فاز پردازش، که از الگوریتم PageRank برای محاسبه داده به دست آمده استفاده شد. در طی فاز خزیدن یک گراف با 50 میلیون نود و 1/8 بیلیون لبه ایجاد شد که به صورت تخمینی به اندازه دو سوم کاربران توییتر بود. بنابراین یک راه حل ارزان برای دستیابی به داده و آنالیزهای وابسته به آن با استفاده از زیرساخت ابر آمازون به دست آمد.

پردازش داده‌های علمی

ژانگ و دیگران[39] یک کاربرد رایانش ابری برپایه Hadoop را توسعه دادند که دنباله‌ای از تصاویر میکروسکوپی از سلول‌های زنده را پردازش می‌نمود این پردازش با استفاده از متلب انجام می‌شد. پروژه، کاری مشترک بین گروه‌هایی در دانشگاه‌های Quebec/McGill در مونترال و دانشگاهی در واترلو بود. هدف از این پروژه، مطالعه واکنش‌های مولکولی پیچیده‌ای است که سیستم‌های بیولوژیکی را تنظیم می‌کند. کاربرد برپایه Hadoop ساخته شد به نحوی که به کاربران اجازه ارسال داده‌های کاری را در ابر می‌داد. نویسندگان از یک خوشه مشابه برای هدایت توسعه‌های اولیه سیستم و اثبات مفاهیم آزمایشی استفاده نمودند.

6- نتیجه

در حال حاضر، داده از نظر اندازه در حال بزرگ شدن است و این روند رو به رشد با افزایش تنوع داده تولید شده بیشتر می‌شود. سرعت تولید داده به دلیل استفاده زیاد از وسایل همراه و حسگرهای متصل به اینترنت در حال افزایش است. داده‌های تولید شده فرصتی مناسب برای همه صنایع و حرفه‌ها ایجاد می‌کنند تا با آنالیز داده‌های حجیم به آگاهی بهتر نسبت به کسب و کار خود دست یابند. امروزه سرویس‌های ابری برای ذخیره، پردازش و آنالیز داده‌های محیطی مناسب هستند. این سرویس‌ها چهره تکنولوژی‌های ارتباطی را تغییر داده‌اند. در این مطالعه به مرور داده‌های حجیم در محاسبات ابری پرداختیم و مفهوم داده‌های حجیم، دسته‌بندی داده‌های حجیم و مدل سرویس ابری را بیان نمودیم. این مدل با چندین پلتفرم ابری داده‌های حجیم مقایسه شد. در مطالعات آتی می‌بایست چالش‌های مهم و مباحث بیشتر در منابع آکادمیک و در صنایع بررسی شود و تحقیق، پژوهش و بررسی بیشتر جهت حصول اطمینان از مدیریت درست داده‌های حجیم با استفاده از رایانش ابری انجام گیرد. همچنین وقوع داده نامتعادل در دسته‌بندی داده‌های حجیم و محیط ابری موضوع حایز اهمیت دیگری است که می‌تواند عنوان تحقیقات آتی در زمینه داده‌های حجیم در رایانش ابری باشد.

مراجع

[1] C. Eaton, *et al.*, *Understanding Big Data: Analytic for Enterprise Class Hadoop and Streaming Data*: Mc Graw-Hill companies, 2012.

[2] R. D. Schnieder, *Hadoop for Dummies Special Edition*: John Widly&Sons Canada, 2012.

[3] L. Chih-Wei, *et al.*, "An Improvement to Data Service in Cloud Computing with Content SensitiveTransaction Analysis and Adaptation," in *2013 IEEE 37th Annual*, 2013, pp. 463–468.

[4] L. Chang, *et al.*, "Public Auditing for Big Data Storage in Cloud Computing – a Survey," in *2013 IEEE16thInternational Conference* 2013, pp. 1128–1135.

[5] S. Sagirogolu and D. Sinanc, "Big Data: A Review," *IEEE*, 2013.

[6] I. I. Center, "Planning Guide: Getting Started with Hadoop," in *Steps IT Managers Can Take to Move Forward with Big Data Analytics*, ed, 20.۱۲

[7] S. Singh and N. Singh, "Big Data Analytics," presented at the International Conference on Communication, Information & Computing Tecnology, Mumbai india, 2012.

[8] H. Rathod and T. Chauhan, "A Survey on Big Data Analysis Techniques," IJSRD - International Journal for Scientific Research & Development, vol. 1, pp. 1806-1808, 2013.

[9] A. vailaya, "What's All the Buzz Around "Big Data?"," *IEEE Women in Engineering Magazine*, pp. 24-31, December 2012.

[10] 11.03.2013). An introductory session on Big Data. Available: http://www.humanfaceofbigdata.com/

[11] C. Tankard, "Big Data Security," *Network Security Newsletter*, July 2012.

[12] M. Minelli, *et al.*, *Data, Big Data Analytics:Emerging Business Intelligence and Analytic Trend for Today's Businesses*: John Wiley & Sons, 2013.

[13] I. A. T. Hashem, *et al.*, "The Rise of "Big Data" on Cloud Computing:Review and Open Research Issues," *Information System*, vol. 47, pp ,۱۱۵-۹۸ . Agust 2014.

[14] B. Gerhardt, *et al.*, "Unlocking Value in the Fragmented World of Big Data Analytics," *Cisco Internet Business Solutions Group*, June 2012.

[15] S. Madden, "From Databases to Big Data," *IEEE Internet Computing*, vol. 16, pp. 4-6, June 2012.

[16] S. Del Rio, *et al.*, "On the Use of MapReduce for imbalanced big data using Random Forest," *Information Sciences*, vol. 285, pp. 112-137, 2014.

[17] J. Hurwitz, *et al.*, "Big data for dummies," *For Dummies*, 2013.

[18] B. P. Rao, *et al.*, "Cloudcomputing for Internet of Things&sensing based applications," in *Sensing Technology(ICST) 2012 Sixth International Conference on IEEE*, 2012, pp. 374–380.

[19] B. Franks, "Taming the Big DataTidal Wave:Finding Opportunities in Huge Data Streams with Advanced Analytics," *Wiley.com John Wiley Sons Inc*, 2012.

[20] D. J. Abadi, *et al.*, "Harizopoulos,Column-oriented database systems," *Processing VLDB Endow*, vol. 2, pp. 1664-1665, 2009.

[21] F. Chang, *et al.*, "Bigtable:a distributed storage system for structured data," *ACM Transaction COmputer System (TOCS)*, vol. 26, p. 4, 2008.

[22] P. Neubauer, "Graph databases, NOSQL and Neo4j," 2010.

[23] S. M. Seeger, Comput.Sci.Media, "Ultra-Large-Sites,Key-Value stores:a practical overview," *computer sciences Media*, 2009.

[24] G. DeCandia, *et al.*, "Dynamo:amazon's highly available key-value store ",.۲۲۰–۲۰۵*SOS*, vol. 41, 2007.

[25] S. Das, *et al.*, "G-store:a scalable datastore for transactional multi key access in the cloud," in *1st ACM symposiumon Cloud computing*, 2010 ,pp. 163–174.

[26] F. Lin and W. W. Cohen, "Powe riteration clustering," in *27th International Conference on Machine Learning (ICML-10)*, 2010, pp. 655-662.

[27] R. C. Taylor, "An overview of the Hadoop/MapReduce/Hbase frame-work and its current applications in bioinformatics," *BMC Bioinformatic*, vol. 11, 2010.

[28] A. Lakshman and P. Malik, "The Apache cassandra project," 2011.

[29] H. H. D. E. Rahm, , and .Bull.23(2000)3–13., "Data cleaning: problems and current approaches," *IEEE Data Engineering*, vol. 2 ,۳pp. 3-13, 2000.

[30] J. Quackenbush, "Micro array data normalization and transformation," *Nature Genetics*, vol. 32, pp. 496–501, 2002.

[31] S. A. Y.Chen, R.Katz, and P. VLDBEndow. 5(2012)1802–1813.,"Interactive analytical processingin big data systems :across-industry study of MapReduce workloads," *VLDB Endow*, vol. 5, pp. 1802-1813, 2012.

[32] B. R. L.Neumeyer, A.Nair,A.Kesari,2010,2010, pp.170–177, "S4:Distributed Stream Computing Platform," in *Data Mining Workshops(ICDMW) IEEE International Conferenceon*, 2010, pp. 170-177.

[33] Z. Talia, "Cloudsforscalablebigdataanalytics," *computer*, vol. 46, pp. 98-101, 2013.

[34] C. Ji, *et al.*, "Big Data Processing in Cloud Computing Environments, Pervasive Systems, Algorithms and Networks," in *Proceedings of the12th International Symposium on* 2012.

[35] J. Dean and S. Ghemawat, "MapReduce: Simplified Data Processingon Large Clusters," *ACM*, vol. 51, pp. 107-113, 2008.

[36] D. Boillier, *et al.*, "The Promise and Peril of Big Data," in *Communications and Society Program Washington*, A. Institiue, Ed., ed. DC,USA, 2010.

[37] H. Miller and E, "Big-Data in Cloud Computing:a Taxonomy of Risks," *Inf Res*, vol. 18, p. 571, 2013.

[38] P. Noordhuis, *et al.*, "Mining twitter in the cloud: A case study Cloud Computing(CLOUD ",(in *Proceedings of IEEE 3rd International Conference* Miami, FL, 2010, pp. 107-114.

[39] C. Zhang, *et al.*, "Case study of scientific data processing on a cloud using hadoop," *High Performance Computing Systems and Applications*, pp. 400-415, 2010.