مروری بر مباحث مطرح در داده های بزرگ

مهدی قزوینی [۱]، محمد داودی [۲]، حامد داودی [۳]

[۱] مدرس دانشگاه علمی کاربردی جهاد دانشگاهی سمنان
[۲] دانشجوی کارشناسی دانشگاه علمی کاربردی جهاد دانشگاهی سمنان
[۳] دانشجوی کارشناسی ارشد دانشگاه رشد دانش سمنان

چکیده

امروزه با رشد نمایی حجم داده و اطلاعات، دستیابی به فناوری های مرتبط با حجم زیاد داده نیز به عرصه رقابت اضافه شده است (سرگلریزان جوان مرتضی، اکبری محمد کاظم،). تعریف کوتاهی می تواند این باشد که داده های بزرگ به مجموعه های داده ای اشاره دارند که اندازه آن فراتر از توانایی ابزار های نرم افزاری استاندارد پایگاه داده برای ضبط،ذخیره،مدیریت و تجزیه و تحلیل است (yen shin،kaynak okyay، ۲۰۱۵). هم اکنون مهمترین چالش در یک موتور جست و جو ، وجود حجم زیاد اطلاعات برای پردازش می باشد . برای مثال در حال حاضر بیش از پانصد میلیون سند وب در پارسی جو وجود دارد که می بایست در زمان مناسب پردازش و نمایه سازی شوند. همچنین تعداد واژه های موجود بیش از پانصد میلیون میباشد که مدیریت این داده ها نیازمند به یک سیستم بهینه برای پردازش داده ها می باشد (زارع بیدکی علی محمد، کاوه یزدی فاطمه،).داده های حجیم یک اصطلاح برای مجموعه داده های خیلی بزرگ است که از نظر ساختار، پیچیدگی و منابع تولید بسیار متنوع هستند و ذخیره و آنالیز آنها کار پیچیده ای است(عنایتی الهام،۱۳۹۵).

واژه های کلیدی

داده های حجیم ، دسته بندی داده های حجیم ، کاربرد داده های حجیم ، امنیت،

مقدمه

امروزه داده های حجیم در مرکز توجه علوم مدرن کسب و کار است. این داهد ها از تراکنش های آنلاین،ایمیل ها، ویدیو ها، صدا، تصاویر، جریان های کلیک، گزارش خطاها ، پست ها،گزارشات جستجو، رکورد های اطلاعات سلامت، عملیات متقابل در شبکه های اجتماعی، داده های علمی، حسگر ها، تلفن های همراه و نرم افزار های روی تلفن همراه تولید می شوند . دیتابیس های حاوی این داده ها به سرعت زشد می کنند و نظارت، فرم دهی، ذخیره، مدیریت، اشتراک گذاری، آنالیز و مجازی سازی آنها از طریق ابزار های نرم افزاری معمول دشوار است (عنایتی الهام،۱۳۹۵). داده های عظیم به مجموعه داده هایی اطلاق می شود که مدیریت،کنترل و پردازش آنها فراتر از توانایی نرم افزاری در زمان قابل تحمل و مورد انتظار باشد. سرعت تولید اطلاعات در سیستم های رایانه ای به سرعت در حال افزایش است، در سال ۲۰۱۰ سرعت تولید اطلاعات به حدی رسید که در هر دو روز، بیش از کل داده هایی که تا سال ۲۰۰۳ تولید شده بود داده تولید می شد. این در حالی است که بر اساس یکی از تحقیقات، اطلاعات تولید شده در سال ۲۰۲۰ پنجاه برابر داده های تولید شده در سال ۲۰۱۱ خواهد بود(تاجیکی محمد مهدی، اکبری بهزاد). داده های عظیم به مجموعه داده هایی اطلاق می شود که به قدری بزرگ و پیچیده باشند که به دست آوردن،نگهداری، جستجو، آنالیزو مشاهده آنها دشوار باشد(صادقی مهدی).

داده های حجیم چیست

حجم اطلاعاتی که تا سال ۲۰۰۳ توسط انسان ایجاد شد تنها ۵ اگزابایت است اما امروزه این حجم از اطلاعات تنها در عرض دو روز ایجاد می شود. در تحقیقی نشان داد که هر روز ۲/۵ اگزابایت داده تولید می شود و حدود ۹۰٪ داده های موجود تنها در دو سال اخیر تولید شده است. هر کامپیوتر شخصی حدود ۵۰۰ گیگابایت اطلاعات در خود نگهداری می کند و در دنیا حدود ۲۰ میلیون کامپیوتر شخصی وجود دارد . از سال ۲۰۱۲، داده های حجیم به عنوان یک پروژه مهم و جهانی مطرح شد. پروژه ای که به جمع آوری، بصری سازی و آنالیز مقدار زیادی داده می پردازد (عنایتی الهام). به طور کلی داده های عظیم سه ویژگی دارند . حجم داده ها ، سرعت پردازش داده ها، و تنوع منابع داده ها(صادقی مهدی). یکی از مهمترین مسائلی که در رابطه با مدیریت داده های عظیم وجود دارد انتقال این داده ها جهت ذخیره سازی یا پردازش توزیع شده می باشد. داده های عظیم به دلیل آنکه میزان بسیار زیادی از منابع شبکه را به خود اختصاص می دهند می توانند باعث بروز ازدحام و متعاقبا کاهش بهره وری شبکه و زمان تحویل بسته ها شوند(تاجیکی محمد مهدی، اکبری بهزاد).

کلان داده