کنگره بین المللی علوم و مهندسی

آلمان - هامبورگ

اسفند ماه ۱۳۹۶

تجزیه و تحلیل داده های کلان

محمد داودی [۱]*، عاصفه ماهیچیان [۲]،

۱- دانشجوی کارشناسی دانشگاه علمی کاربردی جهاد دانشگاهی سمنان ،mohamad.dawoodi۱۳۷۳@gmail.com

۲- مدرس دانشگاه علمی کاربردی جهاد دانشگاهی سمنان ،mohamad.dawoodi۱۳۷۳@gmail.com

چکیده

هدف از پژوهش : با توجه با اینکه روز به روز بر مقدار داده و اطلاعات در دنیا افزوده می شود ، نیاز به وجود فناوری هایی که بتواند این حجم انبوه داده و اطلاعات را تجزیه و تحلیل کند روز به روز بیشتر احساس می شود. در حال حاضر یکی از مهمترین چالش های روز دنیا رشد نمایی داده ها و اطلاعات است . سرعت تولید اطلاعات در سیستم های رایانه ای به طرز باورنکردنی در حال افزایش است و با توجه به اینکه سیستم های نرم افزاری استاندارد قادر به تجزیه و تحلیل و ذخیره و مدیریت این حجم از داده را نیستند ، نیاز به ایجاد سیستم هایی قوی برای تجزیه و تحلیل و به طور کلی مدیریت این داده ها بیش از پیش احساس می شود . روش پژوهش : این مقاله به شیوه مروری نوشته شده است و از روش کتابخانه ای برای پژوهش در آن استفاده شده است . نتایج این پژوهش ها این شده است که برای اینکه بتوانیم داده های حجیم را به درستی مدیریت کنیم تا بتوانیم از دانش درون آن ها بهره ببریم باید از روش های گفته شده درون این مقاله پیروی کنیم .

واژه‌های کلیدی: داده های حجیم ، دسته بندی داده های حجیم ،تجزیه و تحلیل داده های حجیم، HADOOP،

۱- 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،۱۳۹۳).امروزه در بسیاری از کاربرد های داده-محور،با حجم انبوهی از داده ها مواجه هستیم که پردازش و بهره برداری از آن ها الگو های گذشته مدیریت داده ها را نیازمند تغییر نموده است (صفائی علی اصغر،ظهیری مطلق سید علی، ۱۳۹۳). 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،۱۳۹۳). حجم داده های تجاری با نرخ تقریبی ۹ برابر در هر ۵ سال افزایش می یابد. یکی از مسادل در مواجهه با این حجم زیاد از داده ها تحلیل و استخراج دانش سودمند به عنوان