

معرفی یک ترکیب کننده تطبیقی برای بهبود کارایی پردازش در کلاستر Hadoop

فرهنگ پدیداران مقدم، ندا صدر*

*Nsadr61@gmail.com

چکیده

زیرساختها و بستر نقل و انتقال نسل آینده برای کلان داده، با چالش ها و مساله های متفاوتی روبرو هستند. از این رو معماری ها و راه حلهای متفاوتی نیز در این زمینه توسط فراهم کنندگان سرویس، ارائه می شود که البته همگی آنها در اهداف مشترکی چون دستیابی به انعطاف پذیری، قابلیت گسترش و صرفه بیشتر، مشترک می باشند. داده های عظیم بدلیل آنکه میزان بسیار بالایی از منابع شبکه را بخود اختصاص می دهد می توانند باعث بروز ازدحام و متعاقبا کاهش بهره وری شبکه و زمان تحویل بسته ها شود. از این رو لازم است جریان های ترافیکی مربوط به داده های عظیم مورد شناسایی قرار گرفته و مسیریابی آنها در شبکه ها با در نظر گرفتن ماهیت این نوع ترافیک صورت گیرد. به همین منظور در این پژوهش روشی جهت بهبود حل ترافیک شبکه با استفاده از ترکیب کننده تطبیقی Combine در تابع Map Reduce معرفی می گردد.

واژگان کلیدی: ترکیب کننده تطبیقی، Map Reduce، Hadoop، Combine.

۱-مقدمه

داده های عظیم الگوی ترافیک درون مرکز داده را دچار تغییر و تحول زیادی می کند. در کاربردهای سنتی که مبتنی بر معماری client-server می باشند ترافیک شبکه مراکز داده بصورت شمالی-جنوبی است. این در حالی است که با توسعه کاربردهای داده های عظیم، الگوی ترافیک اصطلاحاً شرقی-غربی در شبکه مراکز داده بسیار زیاد شده است. علاوه بر این، بکارگیری فناوری رایانش ابری در زیرساختهای پردازشی و ذخیره سازی مراکز داده موجب پویایی ترافیک شبکه در این مراکز شده است.

حجم زیاد ترافیک شرقی-غربی، تغییر الگوی ترافیک بصورت پویا و ماهیت بی درنگ از ویژگی های مهم کاربردهای داده های عظیم در شبکه مراکز داده هستند. بنابراین، ایجاد تغییرات اساسی در بخش زیرساخت شبکه ای مراکز داده جهت اجرای کاربردهای متعدد داده های عظیم، بسیار ضروری است. تحلیل داده های عظیم نقش بسزایی در پیشرفت علم در حوزه های مختلف نظیر پزشکی، بهره وری و سود آوری اقتصادی، انرژی و پیش بینی بیماریها دارد. [9]

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 است.

در چهار چوبهای نرم افزاری که برای پردازش و ذخیره سازی داده های عظیم طراحی شده اند، از قبیل Hadoop و MapReduce داده ها بر روی تعداد بسیار زیادی سرور توزیع شده و نتایج حاصل توسط تعدادی سرور دیگر جمع می شود و در صورت نیاز مجدداً بین سرورهای دیگر توزیع می گردد.