

هدوپ راهکاری برای داده‌های انبوه

صفاناز حیدری^{۱*}، محمدعلی افشار کاظمی^۲

۱- دانشجوی دکتری گروه مدیریت فناوری اطلاعات، واحد علوم تحقیقات تهران، دانشگاه آزاد اسلامی، تهران، ایران

safanaz.heidari@srbiau.ac.ir

۲- دانشیار دانشکده مدیریت دانشگاه آزاد اسلامی واحد تهران مرکز

dr.mafshar@gmail.com

چکیده

داده‌های انبوه بیانگر مجموعه‌های داده‌ای با حجم بسیار بزرگ، متنوع و پیچیده‌ای هستند که مدیریت آنها با استفاده از روش‌های معمول پایگاه داده میسر نیست و نمی‌توان دانش‌های نهان آنها را در زمان‌های پیش‌بینی شده استخراج نمود در این راستا دیتابیس‌هایی برپایه مفهوم داده‌های انبوه که NOSQL نامیده شده‌اند، می‌توانند به راحتی بر بستری از سرورهای کلاستر شده پیاده‌سازی و اجرا شوند که برخلاف روش‌های معمول پایگاه داده، که اطلاعات در جداول جداگانه ذخیره می‌شوند در اینجا اطلاعات به صورت مجتمع ذخیره می‌گردند. هدوپ از معروفترین پایگاه‌داده‌های NOSQL است. هدوپ یک اکوسیستم اطلاعاتی است که با ترکیب چندین مولفه، پلتفرمی مناسب را برای داده‌های انبوه در بستر ابر ایجاد می‌کند. پیاده‌سازی داده‌های انبوه در بستر سیستم‌های پردازش ابری از یک سو منجر به کاهش هزینه‌ها و از سوئی دیگر باعث بالا رفتن کارایی سیستم‌ها شده است، چرا که حجم انبوهی از اطلاعات از طریق سرورهای کلاستر شده، جمع‌آوری و پردازش می‌شوند. یکی از چالش‌های مهم این است که با رشد سریع داده‌های انبوه و متنوع نیاز به ساختاری مناسب جهت آنالیز و بروزرسانی داده‌های در حال رشد می‌باشد. سوال مطرح شده این است که چگونه می‌توان از هدوپ برای دسترسی، پردازش و آنالیز داده‌های انبوه استفاده نمود؟ در این مقاله سعی به پاسخ این سوال با بررسی نقش هدوپ در پردازش داده‌های انبوه داریم.

واژه‌های کلیدی: داده‌های انبوه، هدوپ، رایانش ابری، نگاشت-کاهش

۱- مقدمه

داده‌های انبوه به داده‌هایی اطلاق می‌شود که به لحاظ حجم و تنوع به اندازه‌ای پیچیده هستند که سیستم‌های سنتی و ابزارهای سنتی انبارداده قادر به پردازش و کار بر روی این داده‌ها نمی‌باشند (Ishwarappa & J, 2015). با رشد سریع و مستمر جریان داده‌ها که از طریق تراکنش‌های آنلاین، ایمیل‌ها و ویدئوها، صدا و تصویر، گزارش خطاها،... به صورت ساختاریافته و ساختار نیافته تولید می‌شوند، نظارت، فرم‌دهی، ذخیره، مدیریت، اشتراک‌گذاری و تحلیل آنها از طریق ابزارهای سنتی امکان‌پذیر نیست. لذا نیاز به رویکرد و ابزارهای جدید ذخیره‌سازی داده‌های انبوه مطرح می‌شود که برای کاربران توانایی پردازش جستجوهای توزیع شده از میان پایگاه داده‌های چندگانه و استخراج به موقع پاسخ با استفاده از یک ابزار رایانشی را ایجاد می‌کند. رایانش ابری، یکی از بخش‌های مدرن تکنولوژی‌های ارتباطی فعلی است که روش مناسب را برای پردازش داده‌های انبوه در محیط ابر فراهم می‌نماید.

هدوپ ابزاری مناسب برای این منظور می‌باشد. هدوپ یک ساختار برنامه نویسی مبتنی بر جاوا می‌باشد که از پردازش مجموعه‌های بزرگ داده‌ای در محیط رایانش توزیع شده پشتیبانی می‌کند. در این بستر امکان ذخیره‌سازی اطلاعات در چندین