علوم رایانش و فناوری اطلاعات
نشریه علمی انجمن کامپیوتر ایران
مجلد ۱۶، شماره ۲، ۱۳۹۷
صفحات ۳۱- ۳۸
مقاله عادی

افزایش کیفیت پاسخ[1] با استفاده از گوناگونی داده در پردازش داده‌های بزرگ[2]

حسین احمدوند مازیار گودرزی

دانشکده مهندسی کامپیوتر، دانشگاه صنعتی شریف، تهران، ایران

چکیده

تعداد زیادی از شرکت‌ها با پردازش داده‌های بزرگ برای تحلیل داده‌های مالی، داده‌های تجاری و سایر تحلیل‌ها روبرو هستند. با توجه به زیر ساخت بزرگ و گران قیمت برای پردازش داده‌های بزرگ، ممکن است نتوان تمام داده‌ها را مورد پردازش قرار داد. این موضوع بر روی کیفیت پاسخ تأثیر گذاشته و کیفیت پاسخ را کاهش می‌دهد. راه‌حل ارائه شده در این مقاله در مواقعی که با محدودیت بودجه و زمان اتمام پردازش روبرو هستیم می‌تواند مورد استفاده قرار بگیرد. در این مقاله ما برای افزایش کیفیت پاسخ داده‌ها با تأثیر[3] بیشتر را به منابع با توان پردازشی بیشتر اختصاص می‌دهیم. بعد از آن اگر بودجه‌ای برای استفاده در دسترس بود، سایر داده‌ها را نیز مورد پردازش قرار می‌دهیم. در این مقاله با استفاده از روش‌های آماری با سطح اطمینان قابل قبولی میزان تأثیر هر قسمت از داده را بر روی پاسخ نهایی مشخص می‌کنیم. با استفاده از این روش قادر خواهیم بود در صورت وجود محدودیت زمانی و بودجه‌ای کیفیت پاسخ را افزایش دهیم. در فاز ارزیابی داده‌هایی از حوزه‌های مختلف را مورد بررسی قرار داده‌ایم. بررسی نشان می‌دهد این روش دارای کارایی خوبی برای افزایش کیفیت پاسخ در صورت وجود محدودیت زمانی و بودجه است. در کاربردهای مورد ارزیابی در این مقاله موفق شده‌ایم تا ۳۳ درصد بهبود در کیفیت پاسخ ایجاد کنیم.

کلمات کلیدی: پردازش داده‌های بزرگ، تأثیر، کیفیت پاسخ، گوناگونی داده، محدودیت بودجه.

۱- مقدمه

یکی از راه‌های پردازش داده‌های بزرگ استفاده از پردازش مقیاس بزرگ[4] است. با توجه به زیر ساخت مورد نیاز برای این نوع از پردازش هزینه استفاده از آن میزان قابل توجهی خواهد شد. محدودیت بودجه ممکن است مانع پردازش تمام داده‌ها و کاهش کیفیت پاسخ شود. کیفیت پاسخ یکی از مسائل مهم در پردازش داده‌های بزرگ بوده که ما در این مقاله بر روی آن تمرکز کرده‌ایم.

این مقاله توسعه یافته کار پیشین در همین زمینه است. در این مقاله به تحلیل بیشتر در مورد راه‌کار ارائه شده، بررسی دقیق‌تر اختصاص منابع و بررسی حالت‌های بیشتر برای محدودیت‌های موعد زمانی و هزینه خواهیم پرداخت [۱]. در این مقاله علاوه بر توسعه مباحث مطرح شده در مقاله کنفرانس بررسی توضیحات تکمیلی در برخی قسمت‌های مقاله، اندازه‌های مختلف برای بخش‌های داده، در نظر گرفتن محدودیت‌های مختلف برای زمان پردازش و هزینه‌ی مورد نیاز و بررسی بیشتر نتایج خواهیم پرداخت. سربار نمونه‌برداری و دقت جداسازی قسمت‌های با اهمیت بیشتر از قسمت‌های کم اهمیت‌تر عوامل تأثیرگذار در این موضوع هستند.

راه‌حل ارائه شده در این مقاله بر روی بهبود کیفیت پاسخ برای کاربردهای جمع‌شونده[5] تمرکز کرده است. در این نوع از کاربردها، مقادیر مختلف با هم ترکیب می‌شوند تا مقدار خروجی خاصی تولید گردد. کاربردهای جمع‌شونده از جمله کاربردهای پر استفاده در تحلیل‌های تجاری و مالی است. به عنوان مثال محاسبات و آنالیز تراکنش‌های مالی، آنالیز مربوط به تاریخچه سیستم[6]، پردازش متن[7] و پردازش داده‌های مربوط به حسگرها[8] در دسته کاربردهای جمع‌شونده قرار دارند. در بین شرکت‌های استفاده‌کننده از داده‌های بزرگ، ۶۴٪ با تراکنش‌های مالی، ۵۹٪ با تاریخچه سیستم، ۳۰٪ با پردازش داده‌های حسگرها و ۳۰٪ با پردازش متن سروکار دارند. در این مقاله ما به کاربردهای جمع‌شونده مانند حاصل جمع و میانگین می‌پردازیم [۲].

یکی از ویژگی‌های داده‌های بزرگ، گوناگونی[۹] داده است. گوناگونی داده به دلیل این ایجاد می‌گردد که داده‌ها از منابع و انواع مختلف جمع‌آوری می‌شوند. گوناگونی سبب می‌گردد که قسمت‌های مختلف داده تأثیر متفاوتی بر روی پاسخ نهایی داشته باشد [۳].

در پژوهش پیشین [۴] ما نشان دادیم که قسمت‌های مختلف داده از منابع متفاوت تأثیر متفاوتی بر روی پاسخ نهایی می‌گذارند. در این مقاله به بررسی تأثیر گوناگونی داده در بهبود کیفیت پاسخ در صورت وجود محدودیت در زمان پردازش و بودجه می‌پردازیم. برای این منظور بخش‌های داده با تأثیر بیشتر را به سرورهایی با توان پردازشی بالاتر و بقیه بخش‌های داده را به منابع پردازشی موجود اختصاص داده‌ایم.

برای این موضوع ابتدا مفهوم تأثیر را برای هر کاربرد مشخص کرده‌ایم. سپس داده‌ی ورودی را به بخش‌های داده با اندازه یکسان تقسیم نموده و با استفاده از روش نمونه‌برداری تضمین شده میزان تأثیر آن‌ها تعیین نموده‌ایم. با استفاده از نمونه‌برداری زمان و هزینه پردازش بخش‌های داده‌ی مختلف را بر روی سرورهای گوناگون تخمین زده‌ایم. با توجه به محدودیت زمان اتمام پردازش و بودجه مورد نیاز برای پردازش و با هدف افزایش کیفیت پاسخ بخش‌های داده با تأثیر بیشتر را به سرورها با توان پردازشی بالاتر و بقیه بخش‌های داده را به سرورهای موجود اختصاص داده‌ایم.

در نهایت راه‌کار پیشنهادی را با استفاده از کاربردهایی از حوزه‌های مختلف مورد ارزیابی قرار داده‌ایم.

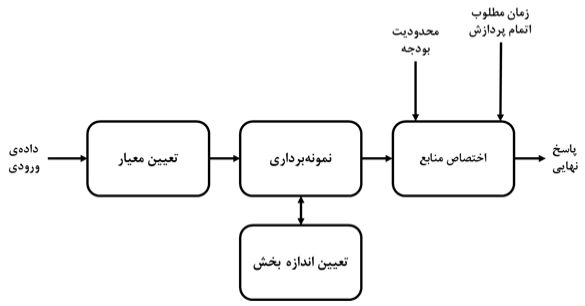
۱-۱- چالش‌ها

در این مقاله با چالش‌های زیر مواجه بوده‌ایم:

کاربردهای مورد نظر برای این راه‌کار. براساس ساختار داخلی کاربردها و وابستگی درونی داده‌ها راه‌کار مورد نظر ما دارای کارایی متفاوتی خواهد بود. راه‌کار ما برای کاربردهای جمع‌شونده و کاربردهای مشابه دارای کارایی مناسب است. کاربردهای جمع‌شونده مورد نظر این مقاله جمع و میانگین هستند.

تعیین معیار تأثیر. تعیین معیار تأثیر یکی از بخش‌های مهم و اصلی پژوهش حاضر است. میزان تأثیر هر بخش داده[۱۰] بعد از پردازش آن تعیین کننده میزان پیشرفت پردازش آن کاربرد است. معیار تأثیر باید به خوبی تعیین گردد تا بتواند به درستی میزان پیشرفت پردازش را نشان دهد.

روش نمونه‌برداری. برای تعیین میزان تأثیر هر بخش داده باید از روش نمونه‌برداری مناسبی استفاده کنیم. برای این منظور باید ابتدا اندازه هر بخش داده را به خوبی مشخص نموده و با توجه به میزان سربار نمونه‌برداری، نمونه‌برداری را انجام دهیم.



شکل ۱- مراحل انجام راه‌کار پیشنهادی

اختصاص آگاه از کیفیت پاسخ. برای اختصاص بخش‌های داده‌ی مختلف به سرورهای موجود باید هدف پژوهش که بیشینه کردن کیفیت پاسخ است را در نظر گرفت. بدین منظور ابتدا بخش‌های داده با تأثیر بیشتر را به سرورها با کارایی بالاتر از نظر زمان و هزینه اختصاص داده و سپس سایر بخش‌های داده را به منابع پردازشی موجود اختصاص می‌دهیم.

۱-۲- نوآوری

نوآوری ما در این مقاله به صورت زیر است:

- توجه به گوناگونی داده و تأثیر آن در کیفیت پاسخ
- بهبود کیفیت پاسخ در صورت وجود محدودیت در زمان پردازش و بودجه با استفاده از گوناگونی داده
- ارائه راه‌کار برای استفاده از گوناگونی داده در افزایش کیفیت پاسخ

۱-۳- ارزیابی

برای ارزیابی روش پیشنهادی از کاربردهای مختلف استفاده کرده‌ایم. این کاربردهای از حوزه‌های مختلف از جمله: متن، تاریخچه سیستم، داده‌های مالی و تجاری هستند. برای بررسی میزان زمان و هزینه پردازش بخش‌های داده بر روی سه نوع سرور پردازش کرده‌ایم. در فصل ۵ این ارزیابی‌ها نشان داده شده و مورد بررسی قرار گرفته‌اند.

در ادامه مقاله در فصل ۲ به بررسی کارهای پیشین می‌پردازیم. در فصل ۳ مثال انگیزشی برای پژوهش ارائه می‌کنیم. فصل ۴ راه‌کار مورد نظر را شرح می‌دهد. در فصل ۵ به ارائه نتایج آزمایش می‌پردازیم. در نهایت در فصل ۶ به جمع‌بندی، نتیجه‌گیری و کارهای آتی پرداخته خواهد شد.

۲- کارهای پیشین

کارهای پیشین در این زمینه به کیفیت پاسخ[۱۱]، پاسخ تقریبی[۱۲] و اختصاص منابع به صورت آگاه از کیفیت پاسخ پرداخته‌اند.

بعضی از کارهای پیشین بر روی بهبود کیفیت پاسخ تمرکز کرده‌اند. در کارهای [۵] و [۶] به موضوع محاسبات تقریبی پرداخته شده است. محاسبات تقریبی جواب قابل قبول را تولید می‌کند. در مرجع شماره [۵] ساختار نگاشت-کاهش[۱۳] تغییر یافته و برای تولید پاسخ تقریبی مورد استفاده قرار گرفته است. در روش ما نیز با اختصاص قسمت‌های مهم‌تر داده به سرور با توان پردازشی بالاتر پاسخ تقریبی در زمان مناسب تولید خواهد شد. همانند کار ما مرجع شماره [۶] نیز یک چهارچوب کاری[۱۴] برای استفاده کاربران برای تولید پاسخ تدریجی معرفی کرده است.

مرجع [۷] نیز به ارائه یک چهارچوب کاری برای افزایش کیفیت پاسخ در مدت زمان کمتر پرداخته است. در این پژوهش گوناگونی در کد[۱۵]، تسک[۱۶] و متغیر[۱۷] را در نظر گفته است. در این مرجع قسمت‌های با تأثیر بیشتر کد، تسک و متغیر در نظر گرفته شده و از آن‌ها برای بهبود کیفیت پاسخ استفاده شده است.

در مرجع [۸] کیفیت پاسخ مد نظر قرار گرفته است در این پژوهش وظایف با تأثیر بیشتر بر روی سخت‌افزار با قابلیت اطمینان بالاتر و وظایف با تأثیر کمتر بر روی سخت‌افزار با قابلیت اطمینان[۱۸] کمتر پردازش شده‌اند. ما نیز در این پژوهش قسمت‌های مهم داده را شناسایی کرده و آن‌ها را با منابع پردازشی با توان بیشتر مورد پردازش قرار داده‌ایم.

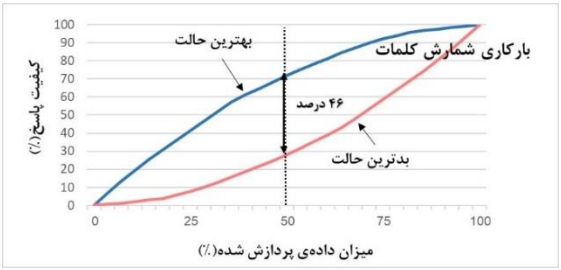
شناسایی و انتخاب قسمت‌هایی از داده برای بهبود کیفیت پاسخ، در مرجع شماره [۹] در نظر گرفته شده است. در این پژوهش ساختار هادوپ[19] تغییر به گونه‌ای تغییر یافته است که بتوان از خط لوله[20] در آن استفاده کرد. در این پژوهش قسمت‌های مختلف داده به گونه‌ای برای پردازش انتخاب شده‌اند که بتوانند کارایی خط لوله را افزایش دهند.

در این مقاله ما با استفاده از گوناگونی داده، کیفیت پاسخ را در شرایط کمبود بودجه و زمان بهبود می‌دهیم. برای این موضوع با توجه به ساختار و وابستگی‌های درونی کاربردها و داده‌های ورودی، کاربردهای مناسب و معیار تأثیر را تعیین می‌نماییم. در هیچ یک از پژوهش‌های پیشین از گوناگونی داده برای بهبود کیفیت پاسخ استفاده نشده است.

۳- مثال انگیزشی

برای نشان دادن انگیزه انجام این پژوهش مثالی را در نظر گرفته‌ایم. برای این منظور چند کاربرد از BigDataBench از مرجع [۱۰] و سایر منابع در نظر در نظر گرفته‌ایم. داده‌ی ورودی را به بخش‌هایی با اندازه ۰٫۵ گیگابایت تقسیم نموده و تمامی جایگشت‌های آن را در نظر گرفته‌ایم. شکل ۲ نشان‌دهنده‌ی تفاوت میان بهترین و بدترین ترتیب پردازش از نظر سرعت رسیدن به پاسخ نهایی است. محور عمودی این نمودار نشان دهنده «کیفیت پاسخ» و محور افقی این نمودار نشان دهنده «میزان داده‌ی پردازش شده» است. در شکل ۲ بارکاری شمارش کلمات[21] در نظر گرفته شده است. در این بار کاری، تعداد کلمات موجود در فایل به عنوان معیار تأثیر در نظر گرفته شده است.

با در نظر گرفتن دنباله‌های مختلف برای پردازش داده‌ها نتایج متفاوتی به دست می‌آید. بهترین حالت زمانی است که داده‌ها با تأثیر بیشتر در ابتدا و داده‌ها با تأثیر کمتر در انتها مورد پردازش قرار بگیرند. میزان این تفاوت به تفاوت بین بخش‌های داده بستگی دارد.



شکل ۲- تفاوت کیفیت پاسخ ایجاد شده در اثر پردازش بهترین/ بدترین ترتیب داده

همان‌طور که در شکل ۲ نشان داده شده است تأثیر انتخاب دنباله‌های متفاوت داده در سرعت رسیدن به پاسخ تأثیر زیادی داشته و تا ۴۶٪ در بارکاری شمارش کلمات تفاوت ایجاد می‌کند. در مورد سایر بارهای کاری نیز این موضوع مشاهده می‌شود. این موضوع انگیزه‌ی اصلی ما برای انجام این پژوهش است.

۴- راه‌حل پیشنهادی

راه حل ارائه شده در این مقاله برای استفاده از گوناگونی داده در بهبود کیفیت پاسخ از چند جزء تشکیل شده است. تعیین معیار برای تصمیم‌گیری در مورد میزان پیشرفت پردازش، نمونه‌برداری برای مشخص کردن میزان تأثیر هر بخش داده و اختصاص منابع به صورت آگاه از کیفیت پاسخ، اجزای این راه‌کار هستند.

۴-۱- تعیین معیار اهمیت برای تصمیم‌گیری در مورد میزان پیشرفت پردازش

برای تعیین میزان پیشرفت پردازش، ابتدا باید معیار مناسب برای کاربرد تعیین گردد. کاربردهای مورد نظر باید ویژگی‌های زیر را داشته باشند:

۱- داده‌ها در این کاربرد از هم مستقل هستند.

۲- نتیجه نهایی در این کاربردها از جمع نتایج جزئی به دست می‌آید.

۳- خروجی این کاربردها معمولاً یکی از پارامترهای آماری (مجموع، میانگین و ...) است.

۴- به‌وسیله نمونه‌برداری می‌توان میزان اهمیت هر بخش داده را تعیین نمود.

در پردازش کاربردهای جمع‌شونده معمولاً قسمت‌هایی از داده که دارای اهمیت بیشتری هستند منابع پردازشی (مانند پردازنده و حافظه) بیشتری نیز مصرف می‌کنند.

برای بعضی کاربردهایی که ویژگی‌های فوق را نداشته باشند نیز می‌توان از این راه‌کار استفاده کرد. به عنوان مثال کاربرد ایندکس معکوس[22] می‌توان حجم داده‌های میانی تولید شده را به عنوان معیار برای استفاده در این راه‌کار در نظر گرفت.

تعیین معیار. تعیین معیار برای میزان پیشرفت پردازش در کاربردهای مختلف با توجه به ساختار درونی آن‌ها صورت می‌گیرد. این معیار باید نشان دهنده میزان پیشرفت پردازش در آن کاربرد باشد. جدول ۱ معیار مناسب برای هر کاربرد را نشان می‌دهد. به توجه به این‌که معیار باید میزان پیشرفت در پردازش را مشخص نماید، معیار همان خروجی تابع و یا پارامتری مربوط به آن (مانند حجم فایل میانی) تعیین می‌شود.

جدول ۱- معیار مناسب برای هر کاربرد

کاربرد	معیار
شمارش کلمات[23]	تعداد کلمات موجود در فایل مورد نظر
گرپ[24]	تعداد خطوطی که در آن یک کلید خاص پیدا می‌شود.
ایندکس معکوس[25]	حجم فایل میانی تولید شده
متوسط طول کلمات[26]	متوسط طول کلمات موجود در فایل
سرمایه‌گذاری[27]	متوسط سرمایه‌گذاری در ایالات مختلف آمریکا
بورس[28]	متوسط معاملات انجام شده در بورس
سلامت[29]	متوسط ضربان قلب داوطلبان
شمارش آدرس اینترنتی[30]	شمارش تعداد یک آدرس خاص در لاگ مربوط به تجهیزات شبکه

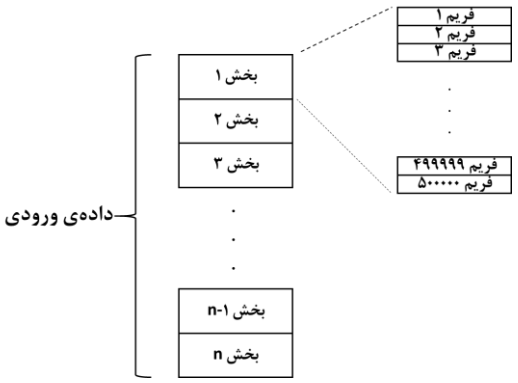
۴-۲- نمونه‌برداری

برای تعیین میزان تأثیر هر بخش داده نمونه‌برداری انجام داده‌ایم. در شکل ۳ نحوه‌ی نمونه‌برداری از داده‌ی ورودی نشان داده شده است. داده‌های ورودی در ابتدا به بخش[31]های هم‌اندازه تقسیم شده و اهمیت هرکدام از این بخش‌ها تعیین می‌شود. برای تعیین اهمیت باید میزان تأثیر هر بخش بر روی پاسخ نهایی تعیین گردد. هر بخش از تعدادی فریم تشکیل شده است. حجم این فریم‌ها ۱ کیلوبایت در نظر گرفته شده است. به عنوان مثال برای کاربرد شمارش کلمات، فریم می‌تواند یک پاراگراف باشد. برای نمونه‌برداری از بخش‌های داده ما از روش کوکران استفاده نموده‌ایم [۱۱].

فرمول کوکران برای روش ما به صورت زیر است.

$$N_0 = z^2.p.q / e^2 \qquad (۱)$$

که در فرمول (۱) z ضریب مربوط به توزیع نرمال است که برای سطح اطمینان ۹۵٪ برابر با ۱,۹۶ در نظر گرفته می‌شود. ضریب e مربوط به میزان خطای قابل قبول است. مقدار در نظر گرفته شده در این مقاله برای e برابر با ۵٪ است. ضریب p پیش بینی وقوع یک ویژگی خاص در مجموعه آماری مد نظر است. ضریب q برابر با "1-p" است. مقادیر ضریب p و q در این مقاله برابر با ۰,۵ است. براساس فرمول نمونه‌برداری کوکران برای داشتن سطح اطمینان ۹۵٪ و خطای ۵٪ باید ۳۸۵ عدد از این فریم‌ها از هر بخش داده به صورت تصادفی انتخاب و مورد بررسی قرار بگیرند. با توجه به این‌که از هر بخش داده که ۰,۵ گیگابایت است ۳۸۵ نمونه‌ی ۱ کیلوبایتی مد نظر قرار می‌گیرد، سربار نمونه‌برداری در این راه‌حل کمتر از ۰,۱٪ است.



شکل ۳- نحوه نمونه‌برداری از داده‌های ورودی

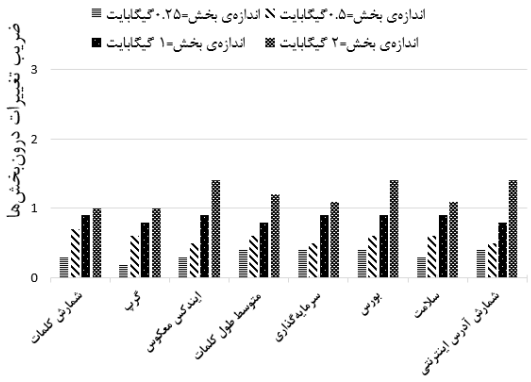
۴-۲-۱- تعیین اندازه بخش

برای استفاده از نمونه‌برداری جهت تعیین تأثیر قسمت‌های مختلف ابتدا باید اندازه مشخصی برای بخش‌ها در نظر گرفته شود. بدین منظور ما اندازه‌ی مختلف را برای بخش در نظر گرفته و آن‌ها را از نظر سربار نمونه‌برداری و میزان دقت جداسازی قسمت‌های مهم و کم اهمیت مورد بررسی قرار داده‌ایم.

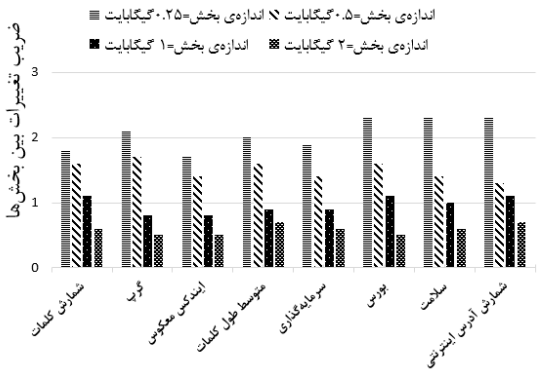
برای این موضوع از «ضریب تغییرات[۳۲]» استفاده کرده‌ایم. ضریب تغییرات به صورت نسبت انحراف معیار به میانگین تعریف می‌شود. این معیار بدون بعد است به همین دلیل برای مقایسه داده‌های آماری که دارای واحدهای مختلف هستند مورد استفاده قرار می‌گیرد.

شکل ۴- سربار نمونه‌برداری به ازای اندازه‌های مختلف برای بخش

شکل ۴ سربار نمونه‌برداری به ازای اندازه‌های مختلف بخش را نشان می‌دهد. شکل ۵ و شکل ۶ ضریب تغییرات درون و بین بخش‌ها را نشان می‌دهند هر چقدر اندازه‌ی بخش کوچک‌تر شود، ضریب تغییرات داخل بخش کمتر و بین بخش‌ها بیشتر می‌شود و به عبارت دیگر دقت جداسازی قسمت‌های مهم از قسمت‌های کم اهمیت بیشتر خواهد شد اما کوچک‌تر شدن اندازه‌ی بخش سبب می‌شود سربار نمونه‌برداری افزایش یابد. به همین دلیل اندازه‌ی هر بخش در این مقاله ۰,۵ گیگابایت در نظر گرفته شده است. در روش نمونه‌برداری کوکران در بدترین حالت ۳۸۴ نمونه باید در نظر گرفته شود. با انتخاب این اندازه علاوه بر دقت جداسازی مناسب داده‌ها سربار نمونه‌برداری دارای مقدار قابل قبولی خواهد بود. با در نظر گرفتن ۰,۵ گیگابایت برای اندازه‌ی بخش سربار نمونه‌برداری کمتر از ۰,۱٪ خواهد بود.



شکل ۵- ضریب تغییرات درون بخش‌ها



شکل ۶- ضریب تغییرات بین بخش‌ها

۴-۳- اختصاص به صورت آگاه از کیفیت پاسخ

در این مرحله باید بخش‌های داده‌ی مختلف با هدف افزایش کیفیت پاسخ به سرور مناسب اختصاص پیدا کند. این اختصاص باید به گونه‌ای انجام گیرد که بتواند در زمان مورد نظر برای اتمام پردازش و با بودجه موجود کیفیت پاسخ بهتری را ایجاد کند. برای این منظور ما از یک الگوریتم به صورت زیر استفاده کرده‌ایم. در جدول ۲ علائم مورد استفاده در الگوریتم معرفی شده‌اند.

در الگوریتم ۱ ما ابتدا همه بخش‌های داده را به ارزان قیمت‌ترین سرور اختصاص می‌دهیم (خط ۳). سپس بودجه و زمان مورد نیاز برای پردازش را محاسبه می‌کنیم. در صورتی که بودجه مورد نیاز پردازش از بودجه موجود بیشتر شد مجبور به حذف مقداری از داده هستیم. برای این منظور به همان مقدار که کمبود بودجه وجود دارد قسمتی از کم‌اهمیت‌ترین بخش داده[۳۳] را حذف می‌کنیم (خط ۷). این کار را تا زمانی ادامه می‌دهیم که هزینه مورد نیاز برای پردازش به اندازه بودجه‌ی موجود باشد. در قسمت بعدی الگوریتم زمان اتمام پردازش در نظر

گرفته شده است. برای کاهش زمان اتمام پردازش، پراهمیت‌ترین بخش داده‌ای که در مسیر بحرانی زمانی قرار دارد به سرور با توان پردازشی بالاتر اختصاص داده می‌شود (خط ۹ تا ۱۱). سرورهای در نظر گرفته شده در این مقاله مطابق مشخصات ارائه شده توسط آمازون هستند؛ بنابراین منظور از سرور با مشخصات بالاتر، سروری است که مطابق جدول ارائه سرویس آمازون نزدیک‌ترین سرور به سرور موجود با مشخصات بالاتر است. این کار ادامه پیدا می‌کند تا زمان اتمام پردازش از زمان مورد نظر برای اتمام پردازش کمتر گردد. مسیر بحرانی زمانی در نهایت تعیین‌کننده‌ی زمان اتمام پردازش است.

در بدترین حالت این الگوریتم ممکن است تمام بخش‌های داده را بر روی تمام سرورها را مورد بررسی قرار بدهد؛ بنابراین از نظر پیچیدگی دارای پیچیدگی O(N*M) است. در این عبارت N نشان‌دهنده‌ی تعداد بخش‌های داده و M نشان‌دهنده‌ی تعداد سرورهای موجود است.

جدول ۲- حروف اختصاری استفاده شده در شبه کد

متغیر	توضیح
NS[۳۴]	تعداد سرور (تعداد نوع ماشین پردازشی)
NP[۳۵]	تعداد بخش داده
PTF[۳۶]	زمان اتمام مطلوب برای پردازش
AB[۳۷]	بودجه در دسترس
FT[۳۸]	زمان اتمام پردازش
BRP[۳۹]	بودجه مورد نیاز برای پردازش
LES[۴۰]	ارزان قیمت‌ترین سرور
TCP[۴۱]	مسیر بحرانی از نظر زمانی
MSP[۴۲]	پراهمیت‌ترین بخش داده

الگوریتم ۱- الگوریتم اختصاص منابع

```
1: Input: NS, NP, PTF, AB
2: output: FT, BRP
3:  assign all portions to LES
4:   while (!process or remove all portions)
5:     estimate(QoR, FT, BRP)
6:  if (BRP > AB)
7:       remove (amount of) LSP (to meet the AB)
8:  end  if
9:     if(FT > PFT)
10:       detect TCP
11:       move MSP in TCP to higher server
12:    end if
13: end While
```

۵- نتایج پیاده‌سازی

در این قسمت به ارائه نتایج می‌پردازیم.

۵-۱- مشخصات در نظر گرفته شده برای پیاده‌سازی

برای نشان دادن زمان و هزینه متفاوت اجرای پردازش‌ها بر روی سرورهای متفاوت از سرورهایی با مشخصات ارائه شده در آمازون [۱۲] استفاده کرده‌ایم. جدول ۳ مشخصات سرورهای مورد استفاده ما در این آزمایش‌ها را نشان می‌دهد. همچنین در این مقاله برای پردازش بارهای کاری از Apache Spark استفاده شده است [۱۳].

جدول ۳- مشخصات سرورهای مورد استفاده

نام	حافظه (گیگابایت)	پردازنده (core)	قیمت به ازای واحد زمان
سرور شماره ۱ (ارزان)	۴	۴	۰,۲۳۹
سرور شماره ۲ (متوسط)	۸	۸	۰,۴۷۹
سرور شماره ۳ (گران)	۱۶	۱۶	۰,۹۵۹

کاربردهای مورد نظر برای این مقاله به صورت زیر است:

شمارش کلمات[۴۳]: این کاربرد تعداد کلمات موجود در یک فایل را مورد شمارش قرار می‌دهد.

گرپ[۴۴]: این کاربرد الگوهای مختلف را جست‌وجو و مورد شمارش قرار می‌دهد.

ایندکس معکوس: این کاربرد یک کاربرد ایندکسی است که عمل نگاشت را بین محتوا و محل قرارگیری در دیتا بیس را انجام می‌دهد.

متوسط طول کلمات[۴۵]: این کاربرد متوسط طول کلمات موجود در فایل را مورد محاسبه قرار می‌دهد.

سلامت[۴۶]: این کاربرد متوسط ضربان قلب داوطلبان را محاسبه می‌کند.

شمارش آدرس اینترنتی[۴۷]: این کاربرد تعداد یک URL[۴۸] خاص را مورد شمارش قرار می‌دهد.

بورس[۴۹]: این کاربرد متوسط معاملات انجام شده در بورس را محاسبه می‌کند.

سرمایه‌گذاری[۵۰]: این کاربرد متوسط سرمایه‌گذاری در ایالات مختلف را محاسبه می‌کند.

در مورد کاربردهای سلامت، شمارش آدرس اینترنتی، بورس و سرمایه‌گذاری برای تولید داده به اندازه ۲۰ گیگابایت با داشتن مجموعه داده[۵۱]ی اولیه، با استفاده از روش بوت استرپ[۵۲] داده‌ی مورد نیاز را تولید کرده‌ایم. اطلاعات مربوط به داده‌های مورد استفاده برای هر کاربرد در جدول ۴ توضیح داده شده است.

جدول ۴- مشخصات داده‌های مورد استفاده

کاربرد	داده‌های ورودی	حجم داده
شمارش کلمات، گرپ، ایندکس معکوس، متوسط طول کلمات	IMDB, Gutenberg, Quotes, Wikipedia	۲۰ گیگابایت
سلامت	ضربان قلب داوطلبان [۱۴]	۲۰ گیگابایت
شمارش URL	لاگ سرور	۲۰ گیگابایت
بورس	داده‌های بورس تهران [۱۵]	۲۰ گیگابایت
سرمایه‌گذاری	سرمایه‌گذاری در ایالات مختلف آمریکا [۱۶]	۲۰ گیگابایت

جدول ۵ بودجه موجود و زمان مورد نظر برای اتمام پردازش را برای کاربردهای در نظر گرفته شده در این مقاله نشان می‌دهد. در جدول زیر محدودیت‌های بودجه و زمان مطلوب پردازش در ۲ حالت در نظر گرفته شده‌اند.

محدودیت بودجه در حالت متعارف[۵۳] به صورت ۸۵٪ بودجه مورد نیاز پردازش بر روی گران قیمت‌ترین سرور در نظر گرفته شده است.

محدودیت بودجه در حالت سخت[۵۴] به صورت ۷۵٪ بودجه مورد نیاز پردازش بر روی گران قیمت‌ترین سرور در نظر گرفته شده است.

محدودیت زمان اتمام پردازش در حالت متعارف به صورت به صورت ۸۵٪ زمان اتمام پردازش بر روی سرور ارزان قیمت در نظر گرفته شده است.

محدودیت زمان اتمام پردازش در حالت سخت به صورت ۷۵٪ زمان اتمام پردازش بر روی سرور ارزان قیمت در نظر گرفته شده است.

جدول ۵- محدودیت بودجه و زمان اتمام پردازش برای کاربردها

کاربرد	بودجه موجود (واحد هزینه)		زمان مطلوب برای اتمام پردازش (ثانیه)	
	متعارف	سخت	متعارف	سخت
شمارش کلمات	۶۰۰	۵۳۰	۴۰۰	۳۵۰
گرپ	۳۵۰	۳۱۰	۲۵۰	۲۲۰
ایندکس معکوس	۱۳۰۰۰۰	۱۱۵۰۰۰	۹۰۰۰۰	۷۹۰۰۰
متوسط طول کلمات	۷۵۰	۶۶۰	۵۰۰	۴۴۰
سلامت	۵۰۰	۴۴۰	۳۰۰	۲۶۵
شمارش آدرس اینترنتی	۳۰۰	۲۶۵	۲۰۰	۱۷۵
بورس	۵۰۰	۴۴۰	۳۵۰	۳۱۰
سرمایه‌گذاری	۵۰۰	۴۴۰	۳۵۰	۳۱۰

سناریوهای پیاده‌سازی شده به شرح زیر هستند:

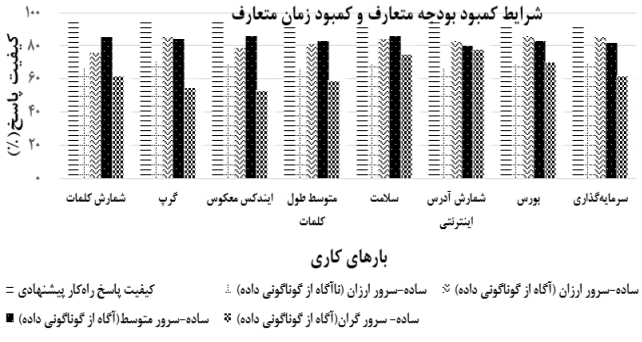
ساده[۵۵] - سرور ارزان - ناآگاه از گوناگونی داده: در این سناریو محاسبات بر روی تنها یک نوع سرور از نوع ارزان قیمت انجام می‌شود. در این سناریو گوناگونی داده در نظر گرفته نمی‌شود.

ساده - سرور ارزان - آگاه از گوناگونی داده: در این سناریو محاسبات بر روی تنها یک نوع سرور از نوع ارزان قیمت انجام می‌شود. در این سناریو گوناگونی داده در نظر گرفته می‌شود.

ساده - سرور متوسط - آگاه از گوناگونی داده: در این سناریو محاسبات بر روی تنها یک سرور از نوع متوسط انجام می‌شود. در این سناریو گوناگونی داده در نظر گرفته می‌شود.

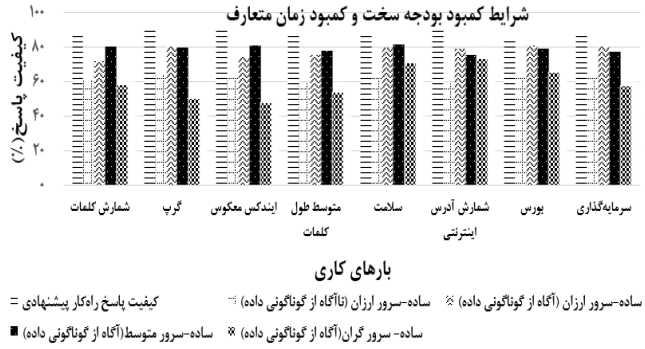
ساده - سرور گران قیمت - آگاه از گوناگونی داده: در این سناریو محاسبات بر روی تنها یک نوع سرور از نوع گران قیمت انجام می‌شود. در این سناریو گوناگونی داده در نظر گرفته می‌شود.

با توجه شکل ۷ مشخص می‌شود با استفاده از راه‌کار پیشنهادی موفق شده‌ایم در شرایط کمبود بودجه متعارف و کمبود زمان متعارف، در کاربرد شمارش کلمات تا ۲۷٪، در کاربرد گرپ تا ۲۴٪، در کاربرد ایندکس معکوس تا ۲۹٪، در کاربرد متوسط طول کلمات تا ۲۵٪، در کاربرد سلامت تا ۲۳٪، در کاربرد شمارش آدرس اینترنتی تا ۲۷٪، در کاربرد بورس تا ۲۱٪ و در کاربرد سرمایه‌گذاری تا ۲۳٪ کیفیت پاسخ را افزایش دهیم.



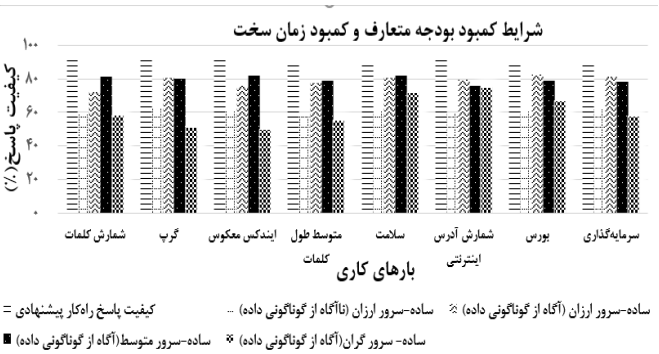
شکل ۷- نتایج در شرایط کمبود بودجه متعارف و کمبود زمان متعارف

با توجه شکل ۸ مشخص می‌شود با استفاده از راه‌کار پیشنهادی موفق شده‌ایم در شرایط کمبود بودجه سخت و کمبود زمان متعارف، در کاربرد شمارش کلمات تا ۳۰٪، در کاربرد گرپ تا ۲۷٪، در کاربرد ایندکس معکوس تا ۲۹٪، در کاربرد متوسط طول کلمات تا ۲۹٪، در کاربرد سلامت تا ۲۶٪، در کاربرد شمارش آدرس اینترنتی تا ۲۹٪، در کاربرد بورس تا ۲۴٪ و در کاربرد سرمایه‌گذاری تا ۲۷٪ کیفیت پاسخ را افزایش دهیم.



شکل ۸- نتایج در شرایط کمبود بودجه سخت و کمبود زمان متعارف

با توجه شکل ۹ مشخص می‌شود با استفاده از راه‌کار پیشنهادی موفق شده‌ایم در شرایط کمبود بودجه سخت در کاربرد شمارش کلمات تا ۳۱٪، در کاربرد گرپ تا ۲۹٪، در کاربرد ایندکس معکوس تا ۳۰٪، در کاربرد متوسط طول کلمات تا ۳۲٪، در کاربرد سلامت تا ۲۹٪، در کاربرد شمارش آدرس اینترنتی تا ۳۳٪، در کاربرد بورس تا ۲۸٪ و در کاربرد سرمایه‌گذاری تا ۲۸٪ کیفیت پاسخ را افزایش دهیم.



شکل ۹- نتایج در شرایط کمبود بودجه متعارف و کمبود زمان سخت

با استفاده از راه‌کار پیشنهادی با استفاده از ۲ راه‌حل ارائه شده برای غلبه بر کمبود بودجه و اتمام زمان مورد نظر برای پردازش، موفق شده‌ایم کیفیت پاسخ را افزایش دهیم. این ۲ راه‌کار عبارت‌اند از: ۱. حذف مقداری از کم تاثیرترین بخش داده، برای غلبه بر کمبود بودجه برای پردازش کل داده‌ها. ۲. انتقال مؤثرترین بخش داده در مسیر بحرانی به سرور با توان پردازشی بالاتر، برای غلبه بر اتمام زمان مورد نظر برای پردازش.

جدول ۶ دلیل کاهش کیفیت پاسخ را در سناریوهای مختلف نشان می‌دهد. در سناریوهایی که از سرور گران قیمت‌تر استفاده می‌کنند، کمبود بودجه و در سناریوهایی که از سرور ارزان قیمت‌تر استفاده می‌کنند، اتمام زمان مورد نظر برای پردازش سبب کاهش کیفیت پاسخ شده است. در راه‌کار پیشنهادی به دلیل آن‌که برای مقابله با کمبود بودجه و اتمام زمان پردازش راه‌کاری اندیشیده شده است، کیفیت پاسخ از دیگر راه‌کارها بیشتر شده است.

راه‌کار پیشنهادی موفق شده است تا با مدیریت آگاهانه از تغییرات داده‌ای میزان استفاده از زمان و هزینه پردازش را برای بخش‌های مختلف مدیریت کرده و کیفیت پاسخ را بهبود بخشد.

در حالت‌های مختلف کمبود بودجه و زمان راه‌کار پیشنهادی از سایر راه‌حل‌ها بهتر عمل کرده و کیفیت پاسخ بهتری را ایجاد می‌کند. دلیل این موضوع، توجه به گوناگونی داده و مدیریت منابع موجود در ساختار آن است.

جدول ۶- دلایل کاهش کیفیت پاسخ در شرایط مختلف محدودیت

کاربرد	پیشنهادی	ساده – ارزان – ناآگاه	ساده – ارزان – آگاه	ساده – متوسط – آگاه	ساده – گران – آگاه
شمارش کلمات	کمبود بودجه	اتمام زمان	اتمام زمان	کمبود بودجه	کمبود بودجه
گرپ	کمبود بودجه	اتمام زمان	اتمام زمان	کمبود بودجه	کمبود بودجه
ایندکس معوس	کمبود بودجه	اتمام زمان	اتمام زمان	کمبود بودجه	کمبود بودجه
متوسط طول	کمبود بودجه	اتمام زمان	اتمام زمان	کمبود بودجه	کمبود بودجه
سرمایه‌گذاری	کمبود بودجه	اتمام زمان	اتمام زمان	کمبود بودجه	کمبود بودجه
بورس	کمبود بودجه	اتمام زمان	اتمام زمان	کمبود بودجه	کمبود بودجه
سلامت	کمبود بودجه	اتمام زمان	اتمام زمان	کمبود بودجه	کمبود بودجه
شمارش آدرس	کمبود بودجه	اتمام زمان	اتمام زمان	کمبود بودجه	کمبود بودجه

۶- نتیجه‌گیری و کارهای آتی

در این مقاله به بررسی تأثیر گوناگونی داده در کاهش هزینه پردازش داده‌های بزرگ پرداختیم. از وجود گوناگونی داده استفاده نموده و با اختصاص آگاهانه بخش‌های مختلف داده سبب افزایش کیفیت پاسخ در پردازش داده‌های بزرگ شده‌ایم. برای این منظور ابتدا کاربردهای مورد نظر را مورد بررسی قرار داده و برای هرکدام معیار تأثیر تعیین نموده‌ایم. در مراحل بعد با استفاده از نمونه‌برداری تأثیر هر قسمت از داده را مشخص کرده‌ایم. در نهایت با اختصاص آگاهانه کیفیت پاسخ را افزایش داده‌ایم. راه‌کار پیشنهادی را در شرایط مختلف از نظر محدودیت بودجه‌ای متعارف و سخت همچنین محدودیت زمانی متعارف و سخت مورد بررسی قرار داده وکارایی این راه‌کار را نشان دادیم. گوناگونی داده در این مقاله بر اثر جمع‌آوری داده‌ها از منابع مختلف بود. جمع‌آوری داده از انواع مختلف (مانند متن، صوت، تصویر و ...) نیز می‌تواند موجب گوناگونی داده و مصرف منابع پردازشی به صورت متفاوت باشد. این موضوع در مقالات بعدی مورد بررسی قرار خواهد گرفت. در کارهای آتی بر روی الگوریتم‌های دیگر اختصاص منابع و بهبود الگوریتم ارائه شده متمرکز خواهیم شد.

مراجع

[1] H. Ahmadvand, and M Goudarzi, "Improving Quality of Results by Taking Advantage of Data Variety in Big Data Processing (in Persian)," The 23rd Computer Society of Iran Computer Conference (CSICC'96), Tehran, Iran, Mar. 2018.

[2] "Big Data Analysis of Practically All Data Types is on the Rise," 6 April 2017. [Online]. Available: https://bi-survey.com/data-types-big-data.

[3] J. Gantz, and R. David, "The digital universe in 2020: Big data, bigger digital shadows, and biggest growth in the far east," IDC iView: IDC Analyze the future, pp. 1-16, Dec. 2012.

[4] H. Ahmadvand, and M. Goudarzi, "Using Data Variety for Efficient Progressive Big Data Processing in Warehouse-Scale Computers," IEEE Computer Architecture Letters, vol. 16, no. 2, pp. 166-169, 2017.

[5] I. Goiri, B. Ricardo, N. Santosh, and D. N. Thu, "Approxhadoop: Bringing approximations to mapreduce frameworks," ACM SIGARCH Computer Architecture News, vol. 43, no.1, pp. 383-397, 2015.

[6] S. Mittal, "A survey of techniques for approximate computing," ACM Computing Surveys (CSUR), vol. 48, p. 62, 2016.

[7] V. Vassiliadis, R. Jan, D. Jens, P. Konstantinos, D. A. Christos, B. Nikolaos, L. Spyros, and N. Uwe, "Towards automatic significance analysis for approximate computing," In IEEE/ACM International Symposium on Code Generation and Optimization (CGO), 2016, pp. 182-193.

[8] J.-D. Fekete, and P. Romain, "Progressive analytics: A computation paradigm for exploratory data analysis," arXiv preprint arXiv: 1607.05162, 2016.

[9] T. Condie, C. Neil, A. Peter, M. H. Joseph, E. Khaled, and S. Russell, "MapReduce online," In Nsdi, vol. 10, no. 4, p. 20, 2010.

[10] "BigDataBench," [Online]. Available: http://prof.ict.ac.cn/. [Accessed 22 Dec. 2017].

[11] W. G. Cochran, "Sampling techniques," John Wiley & Sons, 2007.

[12] "Amazon EC2 Dedicated Instances," [Online]. Available: https://aws.amazon.com/ec2/purchasing-options/dedicated-instances/. [Accessed 22 Dec. 2017].

[13] "Apache Spark™ - Lightning-Fast Cluster Computing," [Online]. Available: http://www.spark-project.org/. [Accessed 22 Dec. 2017].

[14] "UCI Machine Learning Repository," [Online]. Available:https://archive.ics.uci.edu/ml/datasets/MHEALTH%20Dataset. [Accessed 22 Dec. 2017].

[15] Tse.ir. (2017). بورس اوراق بهادار تهران - آرشیو. [online] Available at: http://tse.ir/archive.html [Accessed 22 Dec. 2017].

[16] "Sample CSV Data," [Online]. Available: https://support.spatialkey.com/spatialkey-sample-csv-data/. [Accessed 22 Dec. 2017].

حسین احمدوند مدرک کارشناسی خود را از دانشگاه صنعتی خواجه نصیرالدین طوسی و مدرک کارشناسی‌ارشد را از دانشگاه صنعتی شریف در رشته‌ی کامپیوتر گرایش معماری کامپیوتر دریافت نموده است. وی هم‌اکنون دانشجوی دکتری دانشگاه صنعتی شریف است. زمینه‌های تحقیقاتی مورد علاقه ایشان پردازش داده‌های بزرگ، پردازش ابری و معماری کامپیوتر است

آدرس پست‌الکترونیکی ایشان عبارت است از:

ahmadvand@ce.sharif.edu

مازیـــار گـــودرزی مـدرک کارشناسـی، کارشناسی‌ارشـد و دکتری خود را از دانشـگاه صـنعتی شـریف دریافـت نمـوده است. ایشان هم‌اکنون عضو هیات علمی و دانشـیار دانشـکده کامپیوتر دانشگاه صنعتی شریف اسـت. زمینـه‌ی تحقیقـاتی ایشان در حال حاضر معماری سیستم‌های بزرگ کامپیوتری، رایانش سبز، و هم طراحی سخت‌افزار -نرم‌افزار است.
آدرس پست‌الکترونیکی ایشان عبارت است از:
goudarzi@sharif.edu

اطلاعات بررسی مقاله:

تاریخ ارسال: ۱۳۹۷/۰۲/۱۶
تاریخ اصلاح: ۱۳۹۷/۰۶/۰۹
تاریخ قبول شدن: ۱۳۹۷/۱۰/۰۹
نویسنده مرتبط: حسین احمدونـد، دانشـکده مهندسـی کـامپیوتر، دانشـگاه صنعتی شریف، تهران، ایران.

[1]Quality of Result
[2]Big Data
[3]Significance
[4]Large Scale Computing
[5]Aggregation Application
[6]System Logs
[7]Document
[8]Sensors
[9]Variety
[10]Data Portion
[11]Progressive Computing
[12]Approximate Computing
[13]Map Reduce
[14]Framework
[15]Code
[16]Task
[17]Variable
[18]Reliability
[19]Hadoop
[20]Pipeline
[21]Word Count
[22]Inverted Index
[23]Word Count
[24]Grep
[25]Inverted Index
[26]Average Length
[27]Investment
[28]Exchange
[29]Health
[30]URL Counting
[31]Portion
[32]Coefficient of Variation
[33]The Least Significance Portion (LSP)
[34]Number of Servers
[35]Number of Portions
[36]Preferred Finishing Time
[37]Available Budget
[38]Finishing Time
[39]Budget Required for Processing
[40]The Least Expensive Server
[41]Time Critical Path (TCP)
[42]The Most Significance Portion
[43]Word Count
[44]Grep
[45]Average Length
[46]Health
[47]URL Counting
[48]Uniform Resource Locator
[49]Exchange
[50]Investment
[51]Data Set
[52]Bootstrap
[53]Firm
[54]Hard
[55]Naïve