الگوریتم بهینه فشرده‌سازی جهت انتقال کلان داده

نویسندگان: بیتا ثابت- bta.sabet@gmail.com
الهام بختیاری اصل- elh.bakht@gmail.com
سیده نگین دادگر- negin20_67@yahoo.com

چکیده: امروزه گستره وسیعی از امور محاسباتی افراد و شرکت‌ها توسط سیستم‌های کامپیوتری صورت می‌گیرد و بسیاری از ارگان‌ها و سازمان‌ها مدام در حال بررسی و دست‌وپنجه نرم کردن با حجم زیادی از داده‌ها می‌باشند و تسریع، دقت و اطمینان از نتایج باعث توفیق روزافزون این سیستم‌ها شده است. در این میان اگر حجم داده‌ای برای محاسبات از حدی عبور کند، نه‌تنها سیستم‌های معمولی قادر به پاسخگویی نخواهند بود، بلکه ممکن است نتایج نادرست به‌عنوان خروجی به کاربر ارائه شود. اینجا نقطه ظهور کلان داده است، داده‌هایی حجیم و متنوع و با سرعت رشد نمایی و غیرقابل‌مهار. دراین‌بین روش‌هایی نیز جهت کنترل، مدیریت و تبدیل این تهدیدات به فرصت به وجود آمده است، که کاربردی‌ترین آن فشرده‌سازی بهینه است. در این مقاله ضمن ترسیم ابعاد کلان داده به بررسی راهکارهای موجود برای مدیریت آن‌ها خواهیم پرداخت و به‌طور خاص توانمندی‌های فشرده‌سازی در این حوزه را بیان کرده و نتایجی مقایسه‌ای از کاربرد ترکیبی آن را بررسی می‌کنیم.

کلیدواژه: الگوریتم فشرده‌سازی، بیگ‌دیتا، داده‌های حجیم، کلان داده، Bigdata

1 – مقدمه

با ورود به عصر اطلاعات و ارتباطات و آغاز استفاده از داده‌ها و اطلاعات به‌عنوان سرمایه‌های اصلی درحرکت علمی، اقتصادی، اجتماعی و فرهنگی جوامع، سازمان‌ها و شرکت‌های مختلف و از طرفی توسعه مشارکت افراد در اینترنت و ارتباطات شبکه‌ای در دنیا، دغدغه‌ای بُروز پیدا کرد که از جنس همین داده‌هایی بود که همه روز و با سرعت بسیار زیاد در دنیا و در عرصه‌های مختلفی که فناوری اطلاعات واردشده بود، تولید می‌شد و آن اینکه چگونه این حجم بزرگ و متنوع داده‌ها و اطلاعات را با توجه به ساختارهایی که در فضای فناوری اطلاعات وجود دارد، می‌توان مدیریت، کنترل، پردازش و انتقال داد و از آن در جهت بهبود ساختارها و سودآوری بیشتر بهره جست؟

از سال 2012 به بعد در هرروز هزار پتا بایت (Petabyte) داده تولید می‌شود که به دنبال خود مستلزم ذخیره‌سازی، تحلیل، جستجو، تمیزکاری داده‌ها، اشتراک‌ها و ... در داده‌هاست که باید در حوزه مختلف انجام شود.

به عبارت دیگر فشرده‌سازی روشی برای بیان منابع اطلاعاتی است که امکان استفاده کمتر از بیت‌ها را فراهم می‌نماید، از طرفی ذخیره‌سازی اطلاعات و انتقال داده‌ها هزینه‌بر است و این هزینه با افزایش مقادیر اطلاعات افزایش پیدا می‌کند. علاوه بر آن هزینه‌های حاصل از فشرده‌سازی را می‌توان با پردازش داده‌ها کاهش داد، به‌طوری‌که از حافظه کمتر استفاده شود. مباحث فوق باعث شده است که ما در پژوهشی تحت عنوان الگوریتم بهینه فشرده‌سازی برای انتقال داده‌های حجیم به بررسی بپردازیم.

به طور کلی در این مقاله، ابتدا مواردی درزمینه‌ی کلان داده و فشرده‌سازی ذکر خواهند شد و سپس به خصوصیات اصلی در این داده‌ها خواهیم پرداخت و چالش‌های پیش رو استخراج می‌شود و درنهایت ارزیابی از آینده و پیشینه پژوهشی ذکر خواهد شد.

2- بیان مسئله

در دنیای امروز با افزایش چشم‌گیر حجم داده‌ها و اطلاعات تمایل زیادی به توسعه روش‌های سریع‌تر در مسائل تحلیل داده و نیز رویکردهای داده‌کاوی به وجود آمده است. به‌طورکلی مقادیر داده ای که تولید می‌شوند و مورد پردازش قرار می‌گیرند، روزبه‌روز در حال افزایش است. کلان داده یا Big Data به مجموعه‌ای از داده ها گفته می‌شود که به‌صورت ساخت‌یافته یا غیر ساخت‌یافته، ذخیره می‌گردند و از ابعاد گوناگون تشکیل‌شده‌اند. اولین خصوصیت کلان داده، حجم آن‌هاست که به مقدار و کمیت آن‌ها برمی‌گردد و به دلیل حجم بالا، مدیریت، تجزیه‌وتحلیل آن‌ها

متفاوت است و به‌واسطه‌ی پایگاه داده‌های سنتی انجام نمی‌شوند و اگر از تعداد کمی گره‌های پردازشی استفاده شود، با توجه به این حجم بالا، پردازش با سرعت کمتری صورت می‌پذیرد. برای افزایش سرعت پردازش، گره‌های بیشتر و همچنین قدرت پردازش بیشتری موردنیاز است که هزینه‌ی بالاتری را طلب می‌کند. یکی از راه‌های اولیه که در این زمینه پیشنهاد می‌شود، فشرده‌سازی داده‌ها می‌باشد.

3- مبانی نظری پژوهش

در ابتدا به بررسی برخی اصطلاحات رایج در این مقاله می‌پردازیم.

3-1- کلان داده

کلان داده معمولاً به مجموعه از داده‌ها اطلاق می‌شود که اندازه آن‌ها فراتر از حدی است که با نرم‌افزارهای معمول بتوان آن‌ها را در یک‌زمان معقول اخذ، دقیق سازی، مدیریت و پردازش کرد. مفهوم «اندازه» در کلان داده به‌طور مستمر در حال تغییر است و به‌مرور بزرگ‌تر می‌گردد.

کلان داده مجموعه‌ای از تکنیک‌ها و تاکتیک‌هایی است که نیازمند شکل جدیدی از یکپارچگی هستند تا بتوانند ارزش‌های بزرگی را که در مجموعه‌های بزرگ، وسیع، پیچیده و متنوعِ داده پنهان‌شده‌اند، آشکار سازند.

در سال 2001 موسسه گارتنر (Gartner)، سه بُعد از چالش‌ها و فرصت‌های پیش رو در حوزه رشد داده‌ها را مطرح نمود که عبارت‌اند از افزایش حجم(Volume) ، سرعت و شتاب (Velocity) افزایش سرعت تولید داده‌های ورودی و خروجی، تنوع یا تفاوت (Variety)افزایش محدوده تنوع و منابع داده‌ها، ارزش داده‌ها (Value) و صحت و یکپارچگی داده‌ها یا Veracity.

3-2- بررسی چالش‌ها در کلان داده‌ها

به دلیل عدم ساختار داده‌ها، فیلتر کردن داده‌ها و تعیین سطوح دسترسی مختلف برای کاربران بسیار مشکل و هزینه‌بر است. برای حل این مشکل، معمولاً از برچسب‌گذاری اطلاعات استفاده می‌شود. همچنین تولید اطلاعات مربوط به داده‌ها، در کلان داده نیازمند استفاده از تکنیک‌های خاصی است، چراکه داده‌ها باید مورد تفسیر قرارگرفته و به دلیل حجم داده‌ای بالا و همچنین ساختارهای گوناگون و منابع اطلاعاتی مختلف، تعیین معیار، ارزش و نحوه به دست آوردن اطلاعات مشکل است.

ازآنجایی‌که کلان داده‌ها به‌صورت توزیع‌شده نگه‌داری می‌شوند، چارچوب‌های سخت‌افزاری متفاوت در سمت سرویس‌گیرنده‌ها، می‌تواند مشکل‌ساز شود که با استفاده از منابع مشترک تأمین نرم‌افزاری، این سازگاری تا حد زیادی برقرار می‌شود و این موضوع تحت عنوان ناهمگنی در منابع سخت‌افزاری بررسی می‌شود.

به دلیل اینکه این داده‌ها ساختارمند نیستند، تعیین معیار برای سنجش آن‌ها و استخراج مناسب دسته‌های مختلف داده‌ای که در تصمیم‌گیری‌های مربوط به کسب‌وکار به کار آید، مشکل است و بر اساس دیدگاه‌های مختلف می‌توان معیارهای گوناگونی را در نظر گرفت.

مدیریت کلان داده‌ها، به شکل‌های گوناگونی انجام می‌شود که مسائلی مانند تعیین سطح دسترسی و مدیریت منابع به دلیل متفاوت بودن منابع و توزیع‌شدگی سیستم، باید در نظر گرفته شود. الگوریتم‌های مختلفی در این زمینه به کار گرفته می‌شوند. اطلاعات بلوک‌بندی شده و روی سیستم‌های گوناگون قرار می‌گیرند تا منابع سخت‌افزاری و نرم افزاری بیشتری در دسترس باشد.

به دلیل حجم بالای داده‌ها، پالایش آن‌ها در یک موضوع خاص، زمان‌بر است. باید این امر سنجیده شود که آیا داده‌ها، همان داده‌هایی است که در تصمیم‌گیری، مورداستفاده قرار خواهد گرفت یا خیر و همچنین باید اطمینان حاصل شود که مقدار داده‌ای آن‌ها، هنوز درست است و قابلیت اطمینان برای این مقدار وجود دارد؛ علاوه بر این موارد، میزان داده‌ی لازم که بتوان پیش‌بینی‌های لازم را برای احتمالات پیش رو انجام داد، باید در پالایش داده‌ها، موردتوجه قرار گیرد.

سرعت پردازش داده‌ها نیز در کلان داده باید موردتوجه قرار گیرد. چراکه همان‌طور که در ویژگی‌های کلان داده، ذکرشده است، سرعت پردازشی، یکی از عواملی است که کلان داده، به‌واسطه‌ی آن شناخته و سنجیده می‌شوند.

به دلیل تجمیع کلان داده از چندین منبع مختلف، تعیین این مسائل که چه کسی مسئول دقت و صحت اطلاعات است و همچنین تعیین اینکه سطوح دسترسی چگونه باید تعیین شود و برای تنظیمات امنیتی، چه کسی صاحب اطلاعات محسوب می‌شود، چالش‌برانگیز است. به‌طورکلی، سطح بالایی از افزونگی در مجموعه‌های کلان داده وجود دارد و برای کاهش هزینه‌ها به‌صورت غیرمستقیم لازم است، این افزونگی، بدون آسیب به اطلاعات اصلی، به‌صورت نسبی کاهش یابد.

مکانیسم‌های تجزیه‌وتحلیل در کلان داده، باید به‌گونه‌ای باشند که داده‌های غیر ساخت‌یافته را مورد تجزیه‌وتحلیل قرار دهند و همچنین بتوانند در زمان محدود پاسخگو باشند. راهکارهایی که در پایگاه داده‌های سنتی و رابطه‌ای به کار می‌روند، در اینجا قابل توسعه نیستند، از طرفی در پایگاه داده‌های غیر رابطه‌ای نیز بحث عملکرد

مطرح می‌شود. به همین دلیل باید راهکارهایی ترکیبی از این دو مسئله به کار گرفته شود.

3-3- بررسی مفهوم فشرده‌سازی

در علوم کامپیوتر و نظریه اطلاعات (Information Theory) فشرده‌سازی به معنای کد نمودن اطلاعات به نحوی است که بیت کمتری را نسبت به نسخه‌ی اصلی آن داشته باشد.

3-3-1- انواع فشرده‌سازی

به‌طورکلی فشرده‌سازی به دودسته‌ی اصلی Lossless و Lossy تقسیم می‌شود که در نوع فشرده‌سازی بدون اتلاف یا Lossless، کاهش بیت بر اساس شناسایی و حذف افزونگی آماری انجام می‌گیرد و هیچ اطلاعاتی از بین نمی‌رود و در فشرده‌سازی با اتلاف (Lossy) کاهش بیت بر اساس شناسایی و حذف اطلاعات کم‌ارزش‌تر صورت می‌گیرد و این اطلاعات برگشت‌پذیر نخواهند بود.

3-3-2- مزایای استفاده از فشرده‌سازی

در ادامه به بیان چند مزیت مهم فشرده‌سازی اطلاعات می‌پردازیم:

- کاهش حجم لازم برای نگه‌داری اطلاعات.
- کاهش پهنای باند لازم برای ارسال اطلاعات.
- عملیات موردنظر به‌صورت بهینه و با ساده‌ترین شکل به انجام می‌رسد.

اطلاعات و آگاهی خوبی نسبت به اینکه کدام داده‌ها و اطلاعات قابلیت فشرده‌سازی را دارند حاصل می‌شود.

3-3-3- مشکلات فشرده‌سازی

علیرغم تمامی مزایایی که فرآیند فشرده‌سازی به همراه دارد برخی مشکلاتی را نیز در بر خواهد داشت.

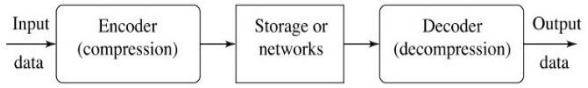
- درصورتی‌که از روش‌های فشرده‌سازی اتلافی استفاده شود بخشی از اطلاعات از بین می‌رود و امکان دسترسی به آن مقدور نمی‌گردد.
- پردازش‌های اضافی برای رمزگشایی اطلاعات گاهی نیازمند سخت‌افزار مخصوصی است که این مسئله متناسب با پیچیدگی روش پیاده‌سازی است (سربار محاسباتی).

3-3-4- بلوک دیاگرام فرایند فشرده‌سازی و رمزگشایی

میزان فشرده‌سازی بر اساس معیاری تحت عنوان نرخ فشرده‌سازی ارزیابی می‌شود که رابطه آن به‌صورت است:

Compression Ratio = B0/B1

که در آن B0 تعداد بیت‌ها قبل از فشرده‌سازی و B1 تعداد بیت‌ها بعد از فشرده‌سازی است.



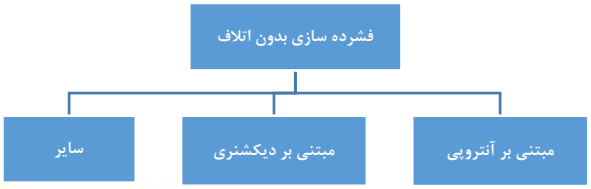
شکل 1: بلوک دیگرام فشرده‌سازی و رمزگشایی

3-3-5- انواع پسوندهای فشرده‌سازی

برخی از ابزارهای فشرده‌سازی با کاربری عمومی، عبارت‌اند از: WinRar, WinZip, pkzip, compress, gzip. همچنین پسوندهای مختلفی برای مقوله فشرده‌سازی وجود دارد که از آن جمله می‌توان در حوزه فشرده‌سازی فایل‌ها به.rar،.zip و .gz اشاره نمود. پسوندهای مطرح در حوزه فشرده‌سازی تصویر شامل JPEG, JPEG 2000, PCX, TGA, TIFF, WMP می‌باشند و برخی از پسوندهای مطرح در حوزه فشرده‌سازی ویدئو MPEG-(1,2,4), DivX, Quick time (MOV),Windows Media Video هستند. علاوه بر موارد فوق پسوندهایی نیز در حوزه فشرده‌سازی رسانه‌های شنیداری وجود دارند که ازجمله آن‌ها می‌توان به MP3, RealAudio (RA, RAM, RP), AAC, WMA اشاره نمود.

3-3-6- دسته‌بندی روش فشرده‌سازی LOSSLESS

در ادامه به بررسی چند روش پراستفاده از فشرده‌سازی Lossless می‌پردازیم.



شکل 2: انواع فشرده‌سازی بدون اتلاف

3-3-7- تکنیک انتقال داده فشرده‌شده

درروش پیشنهادی جهت فشرده‌سازی داده‌های غیر ساختاریافته ایجاد رفتار مؤثر و بهره‌وری در نظر گرفت می‌شود. در آغاز

www.kahrobaonline.ir

عملکرد، نود اسمی و نود داده در هنگام ایجاد ارتباطات TCP باهم تلاقی پیداکرده و ارتباطی برقرار می‌کنند. همان‌گونه می‌دانید ارتباط از نوع TCP یکی از پرفایده‌ترین نوع ارتباط برقرار کردن ازاین‌جهت است که این نوع ارتباطات قادر به بازگرداندن پاسخ به مبدأ جهت ایجاد برقراری ارتباط می‌باشند. لذا Client یک Ack را هنگام برقراری ارتباط از نود اسمی دریافت می‌کند. بدین شرح که نود داده می‌بایست در چارچوب ردیف‌های ایجادشده قرار گیرد. سپس Client بر اساس دستورالعملی که به‌وسیله نود اسمی دریافت کرده بود یک موقعیت و جایگاه مناسب و مؤثری را در بلاک‌هایی که برای نود داده در نظر گرفته‌شده است ثبت می‌کند.

کل مجموعه داده‌ها در ابتدای یک بلاک اولیه متصل می‌شوند (مرتبط می‌شوند) این بلاک‌های ابتدایی چیزی نیستند مگر دسته‌هایی متشکل از مجموعه داده‌های اصلی از این بلاک‌ها که چند بار در موقعیت‌های متفاوت کپی‌برداری می‌شود سپس این بلاک‌های اولیه به‌وسیله الگوریتم LZ فشرده‌سازی می‌شوند. پس قبل از انجام عمل فشرده‌سازی یک کوپلینگ اولیه از مجموعه‌ای از داده‌ها ساخته می‌شود.

3-3-8- فشرده‌سازی و Decode کردن

در ابتدا داده‌ها تا جایی که امکان‌پذیر باشد از هم تفکیک می‌شوند. سپس گروه‌های شبیه به هم از داده ایجاد می‌شود و دوباره تقسیم می‌شوند. زمانی که یک‌بار داده‌ای در یک گروه همسان خود بعد عمل جداسازی Sort شدند. بلاک‌های داده شکل می‌گیرند و سپس در این بلاک داده‌های تشکیل‌شده، عملیات فشرده‌سازی انجام می‌شود. با این نوع از فشرده‌سازی حجم قابل‌توجهی از داده کاهش می‌یابد به طبع زمان انتقال داده نیز کاهش پیدا می‌کند.

حال داده فشرده‌سازی شده به نود داده منتقل می‌شود و عملیاتی در Data Node بر روی داده صورت می‌گیرد. درواقع عملیات Mapping و Reducing الگوریتم اجرا می‌شود. با انجام عملیات Map داده مجدداً بر تکه‌ها مجزا تقسیم می‌شود و یک‌بار که یک قطعه و یا Sub تقسیم‌شده ایجاد گردید بلافاصله الگوریتم Map بر روی آن اجرا می‌شود؛ و یا برخی تکنیک‌های چندمنظوره برنامه‌ریزی‌شده به‌منظور انجام عملیات موردنظر بر روی‌داده صورت می‌گیرد.

یک‌بار کد دیتا و داده‌های فشرده‌شده به دست نود دیتا رسید در همان‌جا نیز داده فشرده‌شده Decode می‌شود. با این کار داده اصلی حاصل می‌شود.

سپس عملیات موردنظر بر روی‌داده اصلی انجام می‌گردد، بعد از انجام این مرحله داده خروجی بازیابی می‌شود؛ و این داده بازیابی شده مجدداً به سمت Client برمی‌گردد و Client این مجموعه پردازش‌ها را دریافت می‌کند و درواقع آنالیز اصلی، اساساً بر روی داده‌ای که در سمت Client ایجاد و ارسال‌شده است صورت می‌گیرد. البته لازم به ذکر است این مراحل بعد از انجام تمام مراحلی که برای پردازش صورت گرفته است، انجام می‌پذیرد.

3-3-9- الگوریتم فشرده‌سازی ILC

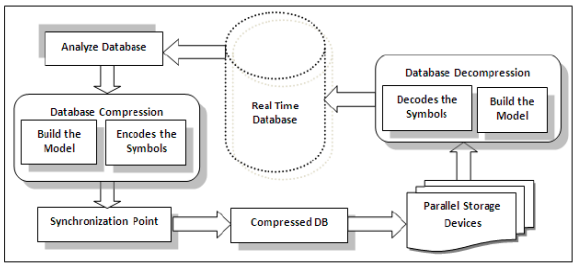
کلا الگوریتم ILC می تواند شرایطی را فراهم آورد که فشرده‌سازی

Backupها در چندین فضای ذخیره‌سازی که به‌موازات هم در حال انجام عملیات می‌باشند انجام گردد.

الگوریتم ILC با فضای ذخیره‌سازی موازی برای Backup گیری اطلاعات Real Time به‌طورکلی، شامل سه عملیات عمده می‌باشد.

1. شناسایی و آنالیز دیتابیس.
2. فشرده کردن داده‌های دیتا بیس به‌منظور برداشتن و دریافت اطلاعات به‌عنوان یک Backup.
3. ذخیره اطلاعات Backup فشرده‌شده در فضاهای ذخیره‌سازی که به‌موازات هم کار می‌کنند و ساختار ILC پیشنهادی تهیه‌شده از اطلاعات دیتابیس Real Time که به‌صورت بهینه فشرده‌شده‌اند.

در شکل زیر این شرایط نشان داده می‌شود.



شکل 3: روال کار الگوریتم ILC

فاز اول: درگام اول داده‌های پایگاه داده بر اساس شرایط محیطی که داده‌ها به‌واسطه آن تولیدشده و به وجود آمده‌اند آنالیز می‌شود. درواقع مشخصات و خصوصیات جاری که در سیستم پایگاه داده موجود است و به‌منظور هدفی خاص ایجادشده‌اند آنالیز و شناسایی می‌گردند. (بر اساس جداولی در پایگاه داده که خصوصیات را دسته‌بندی کرده که این مشخصات موردنظر قابل‌شناسایی می‌شوند).

در فاز دوم عملیات کد کردن و Decode نمودن انجام می‌شود. بر اساس الگوریتم ILC در این فاز رخ می‌دهد. خوبی این الگوریتم این است که تکنیک فشرده‌سازی را در خود دارد که بر اساس دسته‌بندی‌های روی نوع داده‌ها به‌منظور فشرده‌سازی انجام می‌شود. عملیات فشرده‌سازی و سرعت پردازش آن را به‌صورت فزاینده‌ای افزایش می‌یابد. ازاین‌رو ضرایبی حاصل از فشرده‌سازی به این شکل بسیار راحت و آسان به‌منظور Backup گیری قابل‌استفاده خواهد بود.

در فاز سوم: مرحله انجام عملیات ذخیره‌سازی Backupهای فشرده‌شده

www.kahrobaonline.ir

در سطوح مختلف و در فضاهای ذخیره‌سازی متفاوت است که همه به‌موازات هم قرار داشته و کار می‌کنند.
این موازی کاری باعث می‌شود ما یک کپی از Backupها همواره داشته باشیم لذا احتمال از بین رفتن داده‌های حجیم کد شده در طول مسیر انتقال کاهش می‌یابد و ضمناً امکان احیا و بازیابی اطلاعات در صورت از دست رفتن اطلاعات وجود خواهد داشت.

3-3-10- ارزیابی‌های عملی الگوریتم ILC

نرخی که برای نمونه‌های دیتابیس در الگوریتم ILC در نظر گرفته می‌شود 1GB رم است درصورتی‌که در مدل آزمایشگاهی که از تجربیات به‌دست‌آمده از Intel Core 2 DuoP-IV حاصل گردیده است مشاهده می‌شود که از نرخ به‌دست‌آمده در الگوریتم بهینه ILC به‌مراتب کمتر می‌باشد. ضمناً ضرایب و سرعت پردازش سیستمی که از این الگوریتم استفاده می‌کند نیز به‌مراتب بهتر خواهد بود.
حجم داده‌هایی که در این روش برای فشرده‌سازی در نظر گرفته می‌شود تنها منوط به بخش مهم اطلاعات نمی‌باشند که به‌ندرت ایجادشده‌اند بلکه شامل اطلاعات و داده‌های تکراری هم هستند. ضمناً با این شرایط ما قادر به ایجاد نمودارهایی برای ایجاد یکسری روابط و قاعده‌ها هم هستیم چراکه پیوستگی بین داده‌ها بسیار ضعیف است (به دلیل ایجاد بخش‌های مجزا جهت فشرده‌سازی) یعنی حتی دو محدوده اطلاعات که حتی در همسایگی هم قرار دارند به دلیل اینکه جدا مدل می‌شوند و کد می‌شوند کمترین تأثیری بر روی‌هم نداشته و اطلاعات را مخدوش نمی‌کنند بر اساس آزمایشگاه‌های به‌دست‌آمده میزان کیفیت اطلاعات بعد از دی کد کردن اطلاعات تا 85٪ گزارش‌شده است.
نحوه به دست آمدن ضریب فشرده‌سازی نیز به شرح زیر می‌باشد:

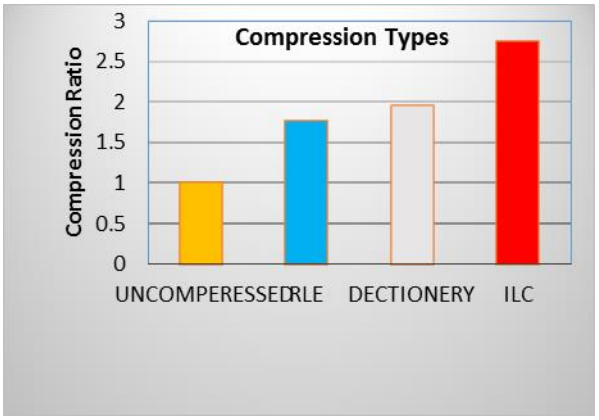
Compression Ratio=uncompressed size/compressed size (1)

زمانی که هرکدام از انواع فشرده‌سازی استفاده می‌شود معمولاً با مجموعه‌ای از فضاها برای ذخیره‌سازی مواجه می‌شویم که معمولاً یک سبک و سنگینی توسط سیستم صورت می‌گیرد که آیا در Disk و بافر Pool اطلاعات Backup ذخیره شود، یا از فضای I/O که مربوط به حافظه Cache سیستم است استفاده شود و گاهی از هر دو استفاده می شود و در شرایطی از یکی به‌جای دیگری استفاده می‌شود و معمولاً این میزان فضا به‌واسطه این فرمول و به‌صورت درصد به دست می‌آید.

Space saving=100*(1-compressed size/uncompressed size) (2)

همان‌طور که مطرح شد شرایط آزمایش انجام‌شده به شرح زیر در نظر گرفته‌شده است، یک سیستم کامپیوتری با 3G رم و یک CPU Intel Core 2Due که دیتابیسی با 1 G در آن استفاده گردید.

شرایط آزمایش را برای الگوریتم‌های مختلف بررسی شد و مشاهده گردید که نسبت فشرده‌سازی در الگوریتم ILC شرایط بهتری نسبت به الگوریتم‌های دیکشنری و RLE داشته است

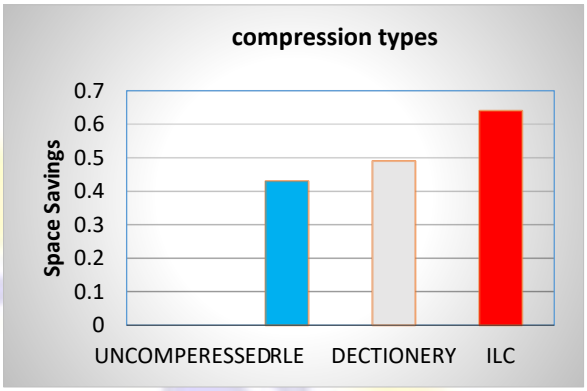


شکل 4: نسبت فشرده‌سازی در الگوریتم‌های مختلف

جدول 1: نسبت فشرده نسبت به‌صرفه جویی در فضا

Compression types	Compressed filesize (MB)	Compression Ratio	Space savings
Uncomperesed	1024	1.00	٪0
RLE	580	1.77	٪43
Dectionery	525	1.95	٪49
llc	370	2.76	٪64

پس از بررسی‌های به‌عمل‌آمده از آزمایش مطرح‌شده، مشخص گردید که در فشرده‌سازی ILC بیشترین میزان صرفه‌جویی در فضای دیسک پیش خواهد آمد.

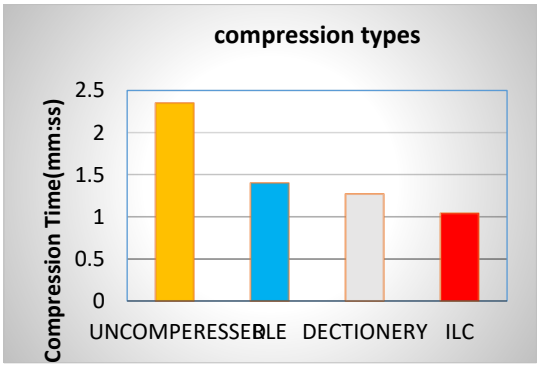


شکل 5: انواع فشرده‌سازی در مقابل صرفه‌جویی در فضا

همچنین علاوه بر میزان فشرده‌سازی و میزان فضای اشغال‌شده از مدت‌زمانی که صرف فرآیند فشرده‌سازی و همچنین Decompress نمودن فایل‌ها گردیده بود موردبررسی قرار گرفت که با توجه به نتایج نمایش داده‌شده در جدول 10، در این میان ILC کمترین زمان فشرده‌سازی و Decompression را داشته است.

جدول 2: تست عملکرد فشرده‌سازی داده‌ها

Compression Types	Compressed Filesize (MB)	Compression Time(mm:ss)	Decompression Time(mm:ss)
Uncomperesed	1.24	2.35	**1.22**
RLE	580	1.40	**1.06**
Dectionery	525	1.27	**0.58**
ILC	370	1.04	**0.39**



شکل 6: انواع فشرده‌سازی در میزان زمان فشرده‌سازی

4- نتیجه‌گیری

با توجه به مستندات ارائه‌شده جهت انتقال کلان داده و همچنین نوع داده‌ها و مشخصاتی که در این مقاله بیان گردید نیازمند استفاده و انجام عملیات مربوط به فشرده‌سازی می‌باشیم. با توجه به اینکه در هنگام فشرده‌سازی کامل داده، سایز آن کاهش می‌یابد لذا شرایط مناسب برای انتقال ایجاد می‌شود. انتقال حجم کمی از اطلاعات به طبع میزان و زمان انتقال را کاهش می‌دهد و درواقع نوعی صرفه‌جویی در زمان انتقال فراهم می‌گردد و درنهایت انجام پروسه انتقال به‌آسانی صورت می‌پذیرد. از طرفی سرعت اجرا عملیات هم به‌مراتب افزایش می‌یابد.

جهت دستیابی به این امر با توجه به بررسی‌های به‌عمل‌آمده و بنا بر آمار ارائه‌شده در این مستند، الگوریتم‌های گوناگونی با روش‌های متفاوت و متعدد جهت انجام فرآیند فشرده‌سازی موجود می‌باشد که از بین آن‌ها الگوریتم ILC جهت انتقال داده‌های حجیم و کلان داده، به‌صورت بهینه و Real Time به دلیل کمترین زمان انتقال پس از فشرده‌سازی و همچنین کاهش فضای ذخیره‌سازی پس از فشرده‌سازی به‌عنوان مناسب‌ترین الگوریتم فشرده‌سازی می‌تواند مورداستفاده قرار گیرد.

سپاسگزاری

نویسندگان این مقاله از هم‌فکری و همیاری استاد گرامی، جناب آقای دکتر محمدهادی زاهدی کمال سپاسگزاری را دارند.

مراجع

[1] عظیمی, مینا؛ میثم محمدی و علی خسرو زاده، ۱۳۹۵، بهینه‌سازی فاکتور زمان در فشرده‌سازی فراکتالی تصاویر با استفاده از الگوریتم مورچگان، سومین کنفرانس سراسری نوآوری‌های اخیر در مهندسی برق و کامپیوتر، تهران، موسسه آموزش عالی نیکان.

[2] ابوالفتحی, احد؛ احمد کشاورز و مازیار گنجو، ۱۳۹۵، ارایه روشی برای بهبود ترافیک کنترلی اوپن فلو در شبکه‌های نرم افزار محوربا فشرده‌سازی پیام گرای پیام چند بخشی، سومین کنفرانس بین المللی مهندسی دانش بنیان و نوآوری، تهران، دانشگاه پیام نور استان تهران- فرمانداری شهر قدس - انتشارات KBEI.

[3] میرزایی ثمرین, عباس و محمد رضا دلداده شیرین، ۱۳۹۵، فشرده‌سازی بدون اتلاف تصاویر رنگی با استفاده از الگوریتم رقابت استعماری، سومین کنفرانس سراسری نوآوری‌های اخیر در مهندسی برق و کامپیوتر، تهران، موسسه آموزش عالی نیکان.

[4] کوکبی, مهدی؛ امین محمدوزیری؛ علی درودی و سعید شعرباف تبریزی، ۱۳۹۵، چندی سازی برداری با استفاده از روش‌های آماری برای فشرده‌سازی تصویر، چهارمین کنفرانس ملی و دومین کنفرانس بین المللی پژوهش‌هایی کاربردی در مهندسی برق، مکانیک و مکاترونیک، تهران، دانشگاه صنعتی مالک اشتر.

[5] خانمیرزا, زهرا؛ فرهاد رمضانی موزیرجی و همایون موتمنی، ۱۳۹۴، بهینه سازی فشرده‌سازی فراکتالی تصاویر سه بعدی با استفاده از الگوریتم ژنتیک، کنفرانس بین المللی سیستمهای غیر خطی و بهینه سازی مهندسی برق و کامپیوتر، دبی، شرکت پندارانديش رهپو.

[6] احمدی, سمیه و مهدی افضلی، ۱۳۹۶، خوشه بندی داده‌های ابعادبالا با استفاده از آنتروپی و مدل MapReduce در Hadoop-Mahout، سومین کنفرانس ملی محاسبات توزیعی و پردازش داده‌های بزرگ، تبریز، دانشگاه شهید مدنی آذربایجان.

[7] خرم آبادی, زهرا، ۱۳۹۶، استخراج الگو از داده کلان مصرف انرژی وسایل برقی در خانه‌های هوشمند، کنفرانس بین المللی مهندسی و فن آوری اطلاعات، امارات- دبی، پژوهشکده فرهنگ و فناوری پژوهشگاه فرهنگ و هنر.

حمدی, سمیه و مهدی افضلی، ۱۳۹۶، رویکردمدل MapReduce در پردازش داده‌های عظیم دراکوسیستم Hadoop-Mahout، سومین کنفرانس ملی

محاسبات توزیعی و پردازش داده‌های بزرگ، تبریز، دانشگاه شهید مدنی آذربایجان.

[8] اردشیری, گلناز، ۱۳۹۵، مدیریت و پردازش کلان داده‌ها، کنفرانس بین المللی مهندسی کامپیوتر و فناوری اطلاعات، تهران، دبیرخانه دایمی کنفرانس.

صیادی, سعید؛ عبادالله زهره وندی و رسول روستایی، ۱۳۹۵، ارایه روشی جدید برای حفظ امنیت در بیگ دیتاهای شبکه اجتماعی با استفاده از خوشه بندی، اولین همایش ملی مهندسی کامپیوتر و فناوری اطلاعات، اهواز، دانشگاه آزاد اسلامی‌واحد سوسنگرد.

[9] محمدی جلکانی, میثم و مهدی نقوی، ۱۳۹۶، چالش‌های داده کاوی داده‌های بزرگ، سومین کنفرانس ملی رویکردهای نوین در مهندسی کامپیوتر و برق، رودسر، دانشگاه آزاد اسلامی‌واحد رودسر و املش- باشگاه پژوهشگران جوان و نخبگان واحد رودسر و املش.

[10] خلیلی, فاطمه و احمد فراهی، ۱۳۹۶، کاربردهای بیگ دیتا در شبکه‌های اجتماعی، چهارمین کنفرانس ملی فناوری اطلاعات، کامپیوتر و مخابرات، مشهد، دانشگاه تربت حیدریه.

[11] Snijders, C., Matzat, U., & Reips, U. D. (2012) . " Big Data": big gaps of knowledge in the field of internet science. *International Journal of Internet Science*, *7*(1) , 1-5.

chung – sheng, composable architecture for rack scale big data computing august 2016.

[12] Beyer, Mark A (2012) . "The importance of ‘big data’: a definition". Stamford, CT: Gartner.

[13] K.Grolinger, “Disaster Data Management in Cloud Environments”, PHD Thesis, the University of Western Ontario, PP21-27, December2013.

[14] P. Zadrozny and R. Kodali,” Big Data Analytics using Splunk, Berkeley”, CA, USA: Apress, 2013.

[15] K.Grolinger, W.Higashino, A.Tiwari1,”Data management in cloud environments: NoSQL and NewSQL data stores”, Jornal of Cloud Computing, 2013.

[16] M.Minelli, M.Chambers , A.Dhiraj,” Big Data, Big Analytics: Emerging Business Intelligence and Analytic Trends for Today's Businesses.”, John Wiley & Sons, Inc.,2013.

[17] Object Management Group. *Unified Modeling Language: Superstructure*, Version 2.0, ptc/03-07-06, July 2003, http://www.omg.org/cgi-bin/doc?ptc/2003-08-02.