****

**نگاشت کاهش: تجزیه و تحلیل ساده شده کلان داده**

**چکیده**

با توسعه تکنولوژی رایانه، افزایش قابل توجهی در رشد داد­ه­ها وجود دارد. دانشمندان با توجه به مقدار میزان افزایش نیازهای پردازش داده­ها که در حوزه علم ایجاد می­شوند همیشه غرق هستند. یک مسئله بزرگ در زمینه­های مختلف استفاده از داده­های با مقیاس بزرگ وجود دارد و این مسئله همیشه با پشتیبانی تصمیم­گیری مواجه است. داده کاوی تکنیکی است که می­تواند الگوهای جدیدی را از مجموعه کلان داده­ها پیدا کند. در طی این سال­ها تمام زمینه­های کاربردی مورد مطالعه و بررسی قرار گرفتند و بسیاری از روش­های داده کاوی را توسعه داده­اند. اما در سال­های مقدار زیادی از داده­ها، محاسبات و تجزیه و تحلیل آنها به طور قابل توجهی افزایش یافته­اند. در چنین موقعیتی، اکثر روش­های داده کاوی در عمل برای دسترسی به چنین کلان داده­هایی از دسترس خارج شدند. الگوریتم موازی/همزمان و تکنیک­های پیاده­سازی به طور موثر کلید ارزیابی مقیاس­پذیری و عملکرد مورد نیاز در تجزیه و تحلیل کلان داده­ها در مقیاس بزرگ می­باشند. تعدادی از الگوریتم­های موازی دارای تکنیک­های مختلف پیاده­سازی هستند و می­توانند: از نگاشت کاهش، MPI، بندها، و mash-up یا گردش کار که دارای ویژگی­های عملکردی و قابلیت­های متفاوتی هستند استفاده کنند. مدل MPI به طور موثر در محاسبه مسئله، به ویژه در شبیه­سازی به دست می­آید. در حقیقت استفاده از آن کار ساده­ای نیست. نگاشت کاهش از یک مدل تجزیه و تحلیل داده­ها در زمینه بازیابی داده­ها است و به صورت فناوری ابر توسعه پیدا کرده است. تاکنون، چندین معماری نگاشت کاهش برای دست زدن به کلان داده­ها توسعه پیدا کرده­اند. معروف­ترین آنها گوگل است. یکی دیگر از ویژگی­های هادوپ این است که محبوب­ترین نرم­افزارها، نرم افزار منبع باز نگاشت کاهش است و توسط بسیاری از شرکت­های فناوری اطلاعات مانند یاهو، فیس بوک، eBay و غیره مورد پذیرش قرار گرفته است. در این مقاله، ما به طور خاص روی هادوپ و اجرای موثر نگاشت کاهش به منظور تحلیل پردازش تمرکز می­کنیم.

**کلمات کلیدی:** کلان داده، داده کاوی، تکنیک­های موازی کردن، HDFS، نگاشت کاهش، هادوپ

**1. مقدمه**

سازمان­ها از چندین مقادیر که داده­های ساختاری بسیار دشواری دارند و از تکنولوژی DBMS برای پردازش و تجزیه و تحلیل داده­ها استفاده می­کنند. این نوع مسئله با شرکت­های مبتنی بر وب مانند فیس بوک، یاهو، گوگل و لینکدین همیشه مواجه است و نیاز به پردازش داده­های با حجم بسیار زیاد و هزینه کافی مستر (ارباب) دارند. تعداد زیادی از این سازمان­ها سیستم­های غیر رسمی خود را برای غلبه بر این موضوع توسعه داده­اند. به عنوان مثال، گوگل، نگاشت کاهش و سیستم فایل گوگل را توسعه داده است. همچنین یک سیستم DBMS به نام بیگ تیبل (جدول بزرگ) نیز ساخته شده است. امکان جستجو در میلیون­ها صفحه نیز وجود دارد و نتایج در آن به صورت میلی ثانیه یا کمتر به منظور کمک به الگوریتم­هایی که هر کدام از سرویس­های جستجو را در یک چارچوب نگاشت کاهش گوگل به ارمغان می­آورند برگردانده می­شوند ]1[. این یک مسئله چالش برانگیز در تحلیل داده­های بزرگ نیز است. کلان داده­ها برای کار کردن بسیار بزرگ هستند و بنابراین یک کار بزرگ برای تجزیه و تحلیل کلان داده­ها انجام می­شود. تکنولوژی­های موجود در تجزیه و تحلیل کلان داده­ها به سرعت در حال تکمیل شدن هستند و به طور قابل توجهی علاقه زیادی به رویکردهای تحلیلی مانند هادوپ، نگاشت کاهش و Hive و توسعه نگاشت کاهش در برابر ارتباط DBMS دارند ]2[.

استفاده از چارچوب نگاشت کاهش به طور گسترده در جهت مقابله با کلان داده­ها بسیار موثر بوده است. در چند سال گذشته، نگاشت کاهش به عنوان رایج­ترین نمونه محاسباتی موازی، تحلیل کلان داده­ها به نظر می­رسید ]3[.

نگاشت کاهش محبوبیت خود را زمانی که با موفقیت توسط گوگل مورد استفاده قرار گرفت به دست آورد. در واقع، این یک ابزار پردازش داده کاوی است و با خطا مقابله می­کند و قادر است پردازش داده­های با حجم بسیار زیاد را به موازات گره­های محاسباتی ارائه دهد ]4[. به لطف سادگی آن، مقیاس پذیری و تحمل خطا، و نگاشت کاهش در حال تبدیل شدن در همه جا هستند، و به طور قابل توجهی هر دو صنعت علمی دانشگاه را به دست آورده­اند. ما می­توانیم عملکرد بالا را با توقف پردازش واحدهای کوچک به پایان برسانیم و می­توانیم به صورت موازی چندین گره را در خوشه اجرا کنیم ]5[. در چارچوب نگاشت کاهش، سیستم فایل توزیع شده (DFS) ابتدا داده­ها را در چندین ماشین تقسیم­بندی کرده و سپس داده­ها را به صورت جفت شده (key,value) بیان می­کنند. چارچوب نگاشت کاهش توابع اصلی ماشین مستر (ارباب) را اجرا می­کند و ما ممکن است داده­های ورودی را قبل از نگاشت توابع که پس پردازنده نام دارد و خروجی عملکرد کاهش را پردازش کنند. توابع نگاشت و کاهش به صورت دوتایی ممکن هستند یکبار یا چندین بار اجرا شوند، به این دلیل که به ویژگی­های برنامه بستگی دارند ]6[. هادوپ یک برنامه محبوب منبع باز است که مجموعه­ای از نگاشت کاهش کلان داده­ها را تجزیه و تحلیل می­کند. این یک فایل سیستمی توزیع شده در سطح کاربر است و برای مدیریت منابع ذخیره­سازی در میان خوشه­ها مورد استفاده قرار می­گیرد ]7[. با این وجود، سیستم سرعت­های ناخواسته در مجموعه­ای از داده­ها که کمتر تولید می­شوند از بین می­برد. اما سرعت قابل قبولی را با مجموعه­ای از کلان داده­ها که تعداد گره­های محاسباتی را کامل می­کند تولید می­کند و زمان اجرای آن را 30 درصد کاهش می­دهد و آنها را با داده کاوی و سایر روش­های پردازش مقایسه می­کند ]8[.

به طور کلی بخش 2 تکامل نگاشت، کاهش و هادوپ را بیان می­کند. بخش 3 توضیحات مختصر کلان داده­ها و مدل برنامه­نویسی نگاشت کاهش را ارائه می­دهد. بخش 4 معماری هادوپ را تنظیم می­کند. بخش 5 روشی علمی از تکنولوژی نگاشت کاهش و هادوپ که ترکیبی از عملکرد نگاشت و کاهش هادوپ است را ارائه می­دهد.

**2. کارهای مرتبط**

کلان داده­ها به اشکال مختلف مجموعه­ای از داده­های بزرگ اشاره دارد و این کلان داده­ها نیاز به سیستم­های محاسباتی خاصی دارند تا تحلیل شوند. برای تجزیه و تحلیل کلان داده­ها کارهای زیادی مورد نیاز است. اما، امروزه برای تجزیه و تحلیل چنین کلان داده­هایی مسائل چالش برانگیز نیز وجود دارد. چارچوب نگاشت کاهش به تازگی توجه زیادی را برای چنین داده­های گسترده­ای را به کار می­برد. نگاشت کاهش یک مدل برنامه­نویسی و پیاده­سازی مرتبط با پردازش و تولید مجموعه کلان داده­ها می­باشد و به طیف گسترده­ای از وظایف در دنیای واقعی پاسخ می­دهد ]9[. نگاشت کاهش نمونه­ای از ویژگی برنامه­نویسی موازی را به سادگی ارائه می­دهد. در عین حال، متوازن کننده و ظرفیت تحمل­پذیری خطا به همراه این ویژگی­ها ارائه می­شود ]10[. سیستم فایل گوگل (GFS) معمولا تحت عنوان یک سیستم نگاشت کاهش داده­های توزیع شده را به صورت کارآمد و با قابلیت اطمینان ذخیره می­کند و برنامه­های کاربردی را در یک سیستم پایگاه داده بزرگ را که مورد نیاز است ارائه می­دهد ]11[. نگاشت کاهش از طریق عنصر اولیه نگاشت و کاهش در توابع زبان­های برنامه کاربردی انجام می­شود ]12[. در حال حاضر برخی از پیاده سازی­ها قابل دسترس هستند: اشتراک سیستم چند هسته­ای با حافظه ]13[، پردازنده­های چند هسته­ای نامتقارن، پردازنده­های گرافیکی، و خوشه­ای ماشین­های شبکه ]14[. تکنولوژی نگاشت کاهش گوگل امکان توسعه­ برنامه­های توزیع شده در مقیاس وسیع را به شیوه­ای ساده­تر و با هزینه کم را فراهم می­کند. ویژگی اصلی مدل نگاشت کاهش این است که قادر است کلان داده­ها را به صورت موازی که در میان گره­های مختلف توزیع شده است پردازش کند ]15[. نرم افزار نوین نگاشت کاهشی یک سیستم اختصاصی گوگل است و بنابراین برای استفاده از منابع باز قابل دسترس نیست. محاسبات توزیع شده نظریه عناصر اولیه نگاشت و کاهش را ساده می­کند، سپس به زیرساخت عملکرد مورد نظر که غیربدیهی است دسترسی پیدا می­کند ]16[. یک زیرساخت کلیدی دارای نگاشت کاهش گوگل، سیستم فایل توزیع شده است و با قابلیت اطمینان بالا به داده­ها دسترسی پیدا می­کند ]9[. با ترکیب روش زمانبندی نگاشت کاهش و سیستم فایل توزیع شده، می­توان به راحتی به محاسبات توزیع شده به صورت موازی که بیش از هزاران گره محاسباتی دارد دست یافت؛ و پردازش داده­ها را در مقیاس ترابایت و پتابایت و همچنین قابلیت اطمینان و بهینه­سازی سیستم توزیع شده را می­توان بهبود داد. ابزار نگاشت کاهش در بهینه­سازی داده­ها بسیار کارآیی دارد و دارای قابلیت اطمینان نیز است به این دلیل که زمان دسترسی به داده­ها یا بارگیری از آنها را 50 تا کاهش می­دهد ]16[. گوگل اولین روش تکنیک نگاشت کاهش را تعمیم می­دهد ]17[. تکنولوژی نگاشت کاهش که اخیرا معرفی شده است از جامعه علمی نشات می­گیرد و کلان داده­های بزرگ را تجزیه و تحلیل می­کند ]18[. هادوپ یک برنامه منبع باز از مدل برنامه­نویسی نگاشت کاهش است و به سیستم فایل توزیع شده هادوپ (HDFS) متکی است. اما سیستم فایل گوگل (GFS) وابسته نیست. HDFS بلوک­های داده­ای را با قابلیت اطمینان بالا در گره­های مختلف قرار می­دهد و آن­ها را کپی می­کند و سپس محاسبات را بعد از هادوپ در این گره­ها انجام می­دهد. HDFS شبیه به سیستم­های دیگر است اما طوری طراحی شده است که در برابر خطا بسیار مقاوم است. سیستم فایل توزیع شده (DFS) هیچ سخت افزار بالایی ندارد و می­تواند در رایانه­ها و نرم افزارها اجرا شود. همچنین مقیاس­پذیر نیز است و یکی از اهداف اصلی طراحی در اجرا است. همانطور که مشخص شد HDFS مستقل از هرگونه سیستم عامل سخت افزار و نرم افزار است، بنابراین در سیستم­های ناهمگن به راحتی قابل حمل هستند ]19[. دستاورد بزرگی که توسط نگاشت کاهش حاصل شده است باعث شبیه سازی هادوپ که یک برنامه منبع باز می­باشد شده است. هادوپ یک چارچوب منبع باز است که نگاشت کاهش را اجرا می­کند ]20[. این یک مدل برنامه­نویسی موازی است که از یک موتور نگاشت کاهش و یک سیستم فایل که سطح کاربر را مدیریت می­کند و در میان منابع ذخیره­سازی خوشه تشکیل شده است ]9[. حمل و نقل سراسری سیستم عامل­های مختلف-لینوکس، Mac OS/X، FreeBSD، سولاریس و ویندوز- هر دو در جاوا نوشته شده­اند و فقط نیاز به سخت افزار کالا دارند.

**3. اهمیت کلان داده­ها**

سازمان­ها باید سیستم عامل محاسباتی تحقیقاتی خود را برای بهبود بخشیدن مقادیر کامل کلان داده­ها ایجاد کنند. این کار کاربران را قادر می­سازد تا از ساختار تجزیه و تحلیل کلان داده­ها برای استخراج داده­های مفید که به راحتی قابل کشف هستند را استفاده کنند. اهمیت کلان داده­ها را می­توان به صورت زیر توصیف کرد ]21[:

1) کلان داده­ها باعث انگیزه در یک اصطلاح می­شوند.

2) این افزایش و مشهوری از هر دو کاربر تجارت و صنعت فناوری اطلاعات به دست می­آیند.

3) از دیدگاه تجزیه و تحلیل هنوز هم تراکم کاری و راه حل­های مدیریتی که قبلا نمی­توانستند از هزینه/یا محدودیت­ها پشتیبانی کنند نشان داده شده­اند.

4) راه حل­ها قادر هستند تصمیم­گیری هوشمندتری را که زمان بیشتری را برای تحلیل تکنولوژی و محصولات صرف کنند ارائه دهند.

5) تجزیه و تحلیل داده­ها در چندین ساختار تصمیم­گیری­های هوشمندانه­ای را می­توانند اتخاذ کنند. تا به امروز، این نوع داده­ها برای پردازش­های پیچیده از تجزیه و تحلیل سنتی تکنولوژی­های پردازش استفاده می­کرده است.

6) تصمیم­گیری­های سریع قابلیت فعال بودن را دارند به این دلیل که راه حل­های کلان داده­ها از تجزیه و تحلیل سریع داده­های دقیق با حجم بالا پشتیبانی می­کنند.

7) در نظر گرفتن زمان سریع امکان پذیر است به این دلیل که سازمان­ها می­توانند داده­های خارج از انبار داده­های سازمانی را پردازش و تجزیه و تحلیل کنند.

برنامه­نویسان از مدل برنامه­نویسی نگاشت کاهش برای بازیابی اطلاعات از کلان داده­ها استفاده می­کنند. ویژگی­های اصلی و مسائل مربوط به تحویل انواع مختلف مجموعه­ای از کلان داده­ها در جدول زیر خلاصه شده­ هستند. داده­هایی که درباره تکنولوژی کلان داده­ها هستند می­توانند به حل آنها کمک کنند ]22[.

جدول 1: خلاصه­ای از ویژگی­های اصلی، چالش­ها و پاسخ­های تکنولوژی مربوط به تحویل انواع مختلف کلان داده­ها در زیر نشان داده شده است

|  |  |  |
| --- | --- | --- |
| پاسخ به پرسش­ها و مهارت­ها | ویژگی­ها  | جنبه  |
| اینترنت باعث افزایش قابل ملاحظه­ای در تولید داده­های سراسری شده است. پاسخ به این وضعیت از طریق تعمیم راه حل­ها براساس ابر بوده است. رویکرد پایگاه داده noSQL یک پاسخ برای ذخیره کوئری و حجم زیادی از داده­ها به صورت توزیع شده در نظر گرفته است. | میزان داده­های تولید شده در سال­های گذشته به شدت افزایش یافته است. با این حال، در عمل جنبه چالش برانگیز کمتری دارد.  | حجم صدا  |
| میلیون­ها دستگاه روزانه به صورت متصل (گوشی­های هوشمند) اضافه می­شوند و نه تنها باعث افزایش حجم بلکه سرعت نیز می­شود. برای دستیابی به یک مزیت رقابتی، شرکت­های جهانی سیستم­های پردازش داده را به عنوان یک نیاز ضروری در نظر گرفتند.  | تولید داده­های با سرعت بالا در حال افزایش است و داده­های تولید شده باید در فریم­های کوتاه­تری جمع آوری شوند.  | سرعت  |
| روش­های جاری که برای جمع­آوری و تجزیه و تحلیل داده­های غیر ساخت یافته یا نیمه ساخت یافته به کار می­روند، کاملا مخالف شیوه مدل رابطه­ای کوئری هستند. این واقعیت منجر به تکامل انواع جدیدی از ذخیره داده شده است و قادر است از مدل­های انعطاف پذیر داده پشتیبانی کند.  | انفجار فرمت­های داده­ای که با داده­های ساختار متن منابع داده­ای بسیار متفاوت است وجود دارد.  | گونه |
| تکنولوژی­های کلان داده­ها باعث ایجاد، و استفاده کردن از حجم زیادی از داده­ها شده است. در عمل، زمانی که داده­های ناقص به داده­های حاوی مقادیر که دارای چالش­هایی هستند تبدیل می­شوند می­توانند در تصمیم­گیری­ها و یا سایر شرایط تجارت مورد استفاده قرار گیرند.  | تا همین اواخر، تمرکز بیشتر بر روی ثبت حجم زیادی از داده­ها بود با این حال نگرانی عمده­ای در نحوه تسخیر آنها وجود داشت.  | مقدار  |

**3.1 مدل برنامه­نویسی نگاشت کاهش**

نگاشت کاهش طوری طراحی شده است که برنامه­نویسان بتوانند از آن به جای کاربران تجاری استفاده کنند. این یک مدل برنامه­نویسی است، نه یک زبان برنامه­نویسی. این مدل برنامه­نویسی برای راحتی، عملکرد و کنترل کلان داده­ها طراحی شده است. مراحل مربوط به کار نگاشت کاهش را می­توان به صورت زیر نشان داد:



شکل : مراحل نگاشت کاهش به منظور پردازش پایگاه داده

برنامه­های کاربردی که شامل نمایه­سازی و جستجو، تجزیه و تحلیل نمودار، و متن، یادگیری ماشین، تبدیل داده­ها است با استفاده از SQL استاندارد DBMSهایی که آسان نیستند را ایجاد می­کند. در چنین مناطقی ماهیت رویه نگاشت کاهش به راحتی برای برنامه­نویسان ماهر قابل درک است. همچنین این مزیت را هم دارد که توسعه­دهندگان مجبور نیستند از اجرای محاسبات موازی استفاده کنند-و به صورت شفاف در سیستم بکار گرفته می­شوند. اگرچه نگاشت کاهش برای برنامه­نویسان طراحی شده است، با این حال برنامه­های غیربرنامه­نویسی می­توانند از برنامه­های نگاشت کاهش از پیش ساخته شده و توابع کتابخانه­ای بهره ببرند [3]. معماری نگاشت کاهش را می­توان به صورت زیر نشان داد:



شکل 2: نگاشت کاهش به همراه ترکیب کننده، و تقسیم کننده

جدول 2: توصیف نگاشت کننده­ها، کاهنده­ها، تقسیم کننده­ها و ترکیب کننده­ها

|  |  |
| --- | --- |
| * نیاز به تولید تعداد دلخواهی از جفت­های میانی
 | نگاشت کننده­ها  |
| * اعمال تمام مقادیر میانی که مرتبط با کلید میانبر است.
 | کاهنده­ها |
| * وظیفه اصلی آن تقسیم فضای کلیدی میانجی است، سپس برای جفت کردن مقادیر میانی به کاهنده­ها تخصیص داده می­شود.
 | تقسیم کننده­ها |
| * ترکیب کننده­ها یک مسئله بهینه­سازی هستند (اختیاری).
* قبل از انجام فاز بهم زدن و مرتب­سازی، اجازه می­دهد که داده­های محلی جمع­آوری شوند.
* اساسا، ترکیب کننده­ها برای ذخیره مورد استفاده قرار می­گیرند، به عنوان مثال برنامه شمارش لغات
 | ترکیب کننده­ها |

برنامه­های نگاشت کاهش معمولا در جاوا نوشته می­شوند. آنها همچنین می­توانند به زبان­های دیگر مانند C++، پایتون، روبی و غیره کدگذاری شوند. این برنامه­ها ممکن است داده­های ذخیره شده در فایل­ها و سیستم­های پایگاه داده را پردازش کنند. به عنوان مثال در گوگل، نگاشت کاهش در بالای سیستم فایل گوگل (GFS) اجرا می­شود.

**4. تنظیم مسائل**

هادوپ: یاهو! اولین عامل اصلی در سال 2006 شده است



شکل 3: عامل اصلی هادوپ

آپاچی هادوپ شامل چندین مولفه است. مواردی که در یک پایگاه داده و پردازش تحلیلی مورد توجه هستند عبارتند از [23]:

سیستم فایل توزیع شده هادوپ (HDFS)، نگاشت کاهش، Pig، Hive، Hbase، اسکوپ

HDFS می­تواند یک منبع یا سیستم فایل برای برنامه­های نگاشت کاهش باشد. این بهترین درخواست برای فایل­های بسیار بزرگ می­باشد. استفاده از داده­های تکراری برای دسترسی به داده­ها در HDFS امکان­پذیر است. اما این باعث افزایش ذخیره­سازی مورد نیاز برای مقابله با داده­ها می­شود. چارچوب نگاشت کاهش هادوپ به توزیع نگاشت فرآیند کمک می­کند بنابراین داده­های HDFS به برنامه محلی نیاز پیدا می­کنند. برای پردازش، تمام فایل­های خروجی ایجاد می­شوند و توسط فرآیند نگاشت و برنامه کاهش بیشتر حرکات و دسترسی به داده­های گره داخلی را انجام می­شوند. در زمان اجرا، هر دو برنامه نگاشت و کاهش ، داده­های انجام شده را به سیستم فایل محلی ارسال می­کنند تا بتوانند از سربارگذاری تکرار HDFS جلوگیری کنند. HDFS از خوانندگان متعدد و یک نویسنده (MROW) پشتیبانی می­کند. مکانیسم شاخص در HDFS قابل دسترس نیست، از این رو، برای خواندن برنامه­ها به منظور اسکن محتوای کامل یک فایل مناسب و بهتر است. در HDFS، محل واقعی داده­ها برای برنامه­ها و نرم افزار خارجی شفاف است.

**معماری HDFS**

معماری HDFS شامل گره­های ارباب/برده است و گره ارباب گره نام و گره برده گره داده نام دارد. HDFS فقط شامل تک گره نام ارباب و تعداد زیادی گره داده (برده) در میان خوشه و معمولا در هر گره است. HDFS یک فضای نامی (شبیه بسته­ای در جاوا) را برای ذخیره داده­های کاربران اختصاص می­دهد. یک فایل ممکن است به یک یا چند بلوک داده تقسیم شود و این بلوک­های داده ممکن است در مجموعه­ای از گره­های داده نگهداری شوند. گره نام اطلاعات ضروری فراداده را که در مورد نحوه اتصال بلوک­ها به یکدیگر است و نحوه ذخیره بلوک­ها در گره نام را نشان می­دهد. نیازهایی که توسط کلاینت برای خواندن و نوشتن سیستم فایل ساخته شده بودند به صورت مستقیم توسط گره داده پردازش شده بودند، در حالی که عملیات فضای نام مانند باز کردن، بستن و تغییر نام دایرکتوری­ها توسط گره­های نام انجام می­شود. مسئولیت­های گره نام و گره­های داده باید مربوط به فعالیت­های خاصی مانند ایجاد بلوک داده، تکرار و حذف باشد [20]. معماری HDFS (سیستم فایل توزیع شده هادوپ) در زیر نشان داده شده است [23]:



شکل 4: یک مدل ساده از چندین گره خوشه هادوپ

نمونه­ای از HDFS که دارای یک ماشین اختصاصی است فقط گره نام را اجرا می­کند. معمولا هر یک از ماشین­ها در خوشه یک نمونه از نرم افزار گره داده را اجرا می­کنند، و این معماری به شما اجازه می­دهد که چندین گره داده را در یک ماشین اجرا کنید. گره نام مربوط به محل ذخیره­سازی فراداده و کنترل است، در صورتی که گره نام مربوط به موارد ذکر شده نباشد اطلاعات کاربر هرگز مدیریت نمی­شود. گره نام از یک نوع خاص ورود به سیستم، به نام EditLog، و برای پایداری فراداده استفاده می­کند.

**توسعه هادوپ**

اگرچه هادوپ یک پیاده­سازی خالص در جاوا است، با این حال ما می­توانیم از آن به دو روش مختلف استفاده کنیم. ما می­توانیم از یک جریان موجود در API یا از لوله­های هادوپ استفاده کنیم. گزینه دوم این امکان را فراهم می­کند تا برنامه­های هادوپ با استفاده از C++ ایجاد شوند. از این رو، ما روی سوابق تمرکز خواهیم کرد. هدف اصلی طراحی هادوپ فراهم کردن ذخیره­سازی و ارتباطات موجود در بسیاری از ماشین­های همگن است. برنامه­نویسان لینوکس را به عنوان سیستم عامل اولیه خود برای توسعه و آزمایش انتخاب کردند؛ از این رو، اگر علاقمند به کار با هادوپ در ویندوز باشید، لازم است نرم افزار جداگانه­ای را برای تقلید محیط پوسته نصب کنید.

هادوپ می­تواند به سه روش مختلف بستگی به نحوه توزیع فرآیندها اجرا شود [24]:

* حالت مستقل: این حالت به صورت پیش فرض با هادوپ ارائه می­شود. همه این موارد در یک فایل جاوا اجرا می­شوند.
* حالت شبه کد توزیع شده: در اینجا، هادوپ برای اجرای بر روی یک ماشین واحد، با انواع دمون­های هادوپ به عنوان فرآیندهای مختلف جاوا پیکربندی می­شود.
* حالت خوشه­ای یا توزیع شده: در اینجا، یک دستگاه در خوشه به عنوان گره نام نامگذاری می­شود و دستگاه دیگری به عنوان Job Tracker تعیین می­شود. فقط یک گره نام در هر خوشه قرار می­گیرد و فضای نام، فایل سیستمی فراداده و کنترل دسترسی را مدیریت می­کند. دومین گره نام نیز می­تواند به قابلیت تحمل­پذیری خطا دستیابی پیدا کند. بقیه دستگاه­های موجود در خوشه هر دو به عنوان گره نام و Task Tracker عمل می­کنند. گره داده، داده­های سیستم را نگه می­دارد؛ و هر گره داده حافظه محلی را ذخیره می­کند یا هارددیسک محلی آن را مدیریت می­کند. Task Tracker عملیات نگاشت و کاهش را انجام می­دهد.

**5. آزمایشات**

نوشتن برنامه کاربردی نگاشت کاهش هادوپ

بهترین راه برای درک و کار کردن با هادوپ این است که از طریق فرآیند نوشتن برنامه، نگاشت کاهش هادوپ پیاده سازی شود. ما با یک برنامه ساده نگاشت کاهش کار می­کنیم و این برنامه می­تواند بسیاری از رشته­ها را معکوس کند. مثالی که در زیر ارائه شده است از طریق تعدادی از مراحل ابتدا تمام داده­ها را به گره­های مختلف تقسیم می­کند، عملیات را انجام می­دهد تا داده­ها معکوس شوند و نتیجه رشته­ها را باهم مرتبط می­سازد و سپس نتایج را تولید می­کند. این نرم افزار فرصتی را برای بررسی تمام مفاهیم اصلی هادوپ فراهم می­کند. ابتدا، ما در مراحل زیر نگاهی به اعلان و وارد کردن بسته می­اندازیم. بسته در کلاس رشته­ای com.javaworld.mapreduce قرار دارد. این را می­توان در دو مجموعه به صورت زیر نشان داد:



اولین مجموعه اعلان­ها برای کلاس­های استاندارد جاوا است و دومین مجموعه برای پیاده­سازی نگاشت کاهش است. بازخوانی کلاس با توسعه org.apache.hadoop.conf.Configured و اجرای رابط org.apache.hadoop.until.Tool آغاز می­شود.

**نگاشت و کاهش**

حالا شما می­توانید به پیاده­سازی نگاشت کاهش واقعی پرش پیدا کنید. دو کلاس داخلی عبارتند از: نگاشت: شامل عملکردی برای پردازش جفت­های کلیدی ورودی به منظور تولید جفت خروجی کلیدی است.



**6. مشارکت ما**

به تازگی، در برخی از آزمایشات کشف شده است که برنامه­های کاربردی­ای که از هادوپ استفاده می­کرده­اند در مقایسه با برنامه­های مشابه که از پایگاه داده­های موازی استفاده می­کرده­اند کارایی کمتری دارند. هدف اصلی ما این است که بهینه­سازی HDFS و به طور قابل توجهی عملکرد کلی چارچوب نگاشت کاهش را که باعث افزایش کارایی کل برنامه­های نگاشت کاهش در هادوپ می­شود را فراهم کنیم. ممکن است نتیجه نهایی نگاشت کاهش در مقابل پایگاه داده موازی هیچ تغییری نداشته باشد، با این حال رویکرد جدید هادوپ و نگاشت کاهش مطمئنا امکان مقایسه دقیق­تر مدل­های برنامه­نویسی را فراهم خواهد کرد. اگرچه هادوپ قابلیت­های داخلی را برای نمایش نگاشت و کاهش فراهم می­کند، با این حال هیچ ابزار ساخته شده­ای برای تنظیم چارچوب، که بتواند موانع عملکرد را بدون انعطاف نگه دارد وجود ندارد. این مقاله تعاملات بین هادوپ و ذخیره­سازی را بازیابی می­کند. در اینجا، ما بسیاری از موانع عملکرد را که به طور مستقیم به کد برنامه مربوط نمی­شود (یا سبک برنامه­نویسی نگاشت کاهش)، بلکه به واسط زمانبندی کار و سیستم­های توزیع شده تمام برنامه­های هادوپ مربوط می­شود را توضیح دادیم. HDFS همزمان می­تواند به طور قابل توجهی با استفاده از برنامه ورودی/خروجی زمانبندی شود و می­تواند حفظ قابلیت حمل و نقل را بهبود ببخشد. بهبودهای بیشتر می­توانند با کاهش پراکندگی و حافظه نهان و کاهش هزینه قابل حمل انجام شوند. هادوپ از قابلیت انتقال برای پشتیبانی کاربران و کاهش پیچیدگی نصب و راه اندازی استفاده می­کند. این نتایج نمونه­ای گسترده از محاسبات موازی است.

**7. نتیجه­گیری**

کلان داده­ها و فناوری­ها می­توانند مزایای قابل توجهی را برای کسب و کار به ارمغان بیاورند. اما استفاده­های فوق­العاده از این تکنولوژی­ها برای یک سازمان و برای کنترل مجموعه­های ناهمگن داده­ها به منظور بررسی بیشتر بسیار دشوار است. اثرات چندگانه استفاده از کلان داده­ها وجود دارند. برای روبرو شدن با رقابت­های شدید شرکت­های خصوصی، آنها از یک پتانسیل بزرگ پشتیبانی کردند. بعضی از جنبه­ها نیاز به پیگیری دارند تا ما بتوانیم نتایج را به موقع از کلان داده­ها بدست آوریم، به این دلیل که استفاده دقیق از کلان داده­ها می­تواند به گسترش، نوسازی و اثربخشی کل بخش­ها منجر شود. برای اینکه بتوانید مزایای کلان داده­ها را استخراج کنید، این بسیار مهم است که بدانید مدیریت و استفاده مجدد از منابع داده از جمله تاثیر داده­های کانتری و ایجاد برنامه­های کاربردی و مورد اطمینان چگونه حاصل می­شوند. مهم این است که بهترین روش را به منظور استفاده از فیلتر کردن/و یا تجزیه و تحلیل داده­ها ارزیابی کنید. برای پردازش تحلیلی بهینه­سازی، هادوپ با نگاشت کاهش مورد استفاده قرار می­گیرد. در این مقاله، ما اصول برنامه­نویسی نگاشت کاهش را با چارچوب هادوپ منبع باز ارائه کردیم. این یک چارچوب فوق­العاده از سرعت پردازش مقادیر زیادی از داده­های هادوپ است که از طریق پروسه­های توزیع شده و پاسخ­های بسیار سریع ارائه می­شود. این می­تواند برای رفع نیازهای مختلف توسعه مورد پذیرش قرار گیرد و می­تواند با افزایش تعداد گره­های موجود به منظور پردازش مقیاس­پذیری شود. قابلیت امتداد و سادگی چارچوب، و متمایزکننده­های کلیدی که یک ابزار امیدوار کننده است را برای پردازش داده­ها ایجاد می­کند.



