مجله‌ی بررسی‌های آمار رسمی ایران
سال ۳۲، شماره‌ی ۱، بهار و تابستان ۱۴۰۰، صص ۹۷ –۱۲۳

تحلیل مه‌داده هزینه و درآمد خانوری کشور با بهره‌گیری از سیستم فایل توزیع‌شده هدوپ

رضا علی‌پور و رضا انتظاری ملکی*

دانشگاه علم و صنعت ایران

چکیده. مه‌داده از منابع مهم در دنیای امروز است، که با استفاده از تجزیه و تحلیل‌های گوناگونی که روی آن انجام می‌گیرد اطلاعات و دانش ارزشمندی از آن بدست می‌آید. طی دو دهه اخیر حجم این داده‌ها در حال گسترش بوده و رفته رفته بر حجم آن نیز افزوده می‌شود. چارچوب هدوپ برای توزیع و پردازش مه‌داده یکی از پرکاربردترین ابزاری است که با زبان برنامه‌نویسی جاوا نوشته شده است. هدوپ یک ابزار مناسب است که این امکان را می‌دهد تا پردازش بر روی مجموعه داده‌های بزرگ با خوشه‌بندی انجام پذیرد و مدیریت داده‌های نیمه‌ساختاریافته و ساختارنیافته را تسهیل کند.

در ایران نیز همچون کشورهای دیگر هر ساله در حوزه آمارهای رسمی کشور داده‌های خانواری جمع‌آوری می‌شود. این داده‌ها حاوی اطلاعات ارزشمندی است که نتایج آن فقط در سطح کل کشور و استان منتشر می‌شود و تا کنون در سطح شهرستان نتایج و اطلاعاتی استخراج نشده است. هدف این تحقیق استفاده از چارچوب هدوپ برای توزیع و پردازش داده‌های خانواری در سطح شهرستان‌های استان است، سپس اطلاعات استخراج‌شده برای تجزیه و تحلیل مورد استفاده قرارمی‌گیرد.

بر اساس مدل پیشنهادی، خوشه‌بندی داده‌های ۳۱ استان کشور در ۴ خوشه انجام و برای راه‌اندازی ۴ سرور ماشین مجازی با ۴ گره در نظر گرفته شد. داده خام از sql به csv تبدیل و در فایل‌های HDFS بارگذاری و عملیات نگاشت/کاهش انجام شد. بر اساس اهداف این تحقیق، خروجی‌های مورد نظر و شاخص‌های برخورداری یک خانوار، مانند استفاده از اینترنت در سطح شهرستان‌های استان ۰۱ استخراج شد و مورد مقایسه و تجزیه و تحلیل قرار گرفت. بدیهی است که همین اطلاعات و شاخص‌ها می‌تواند در سطح

---

* نویسنده عهده‌دار مکاتبات

دریافت: ۱۴۰۱/۹/۲۳، پذیرش: ۱۴۰۲/۱/۲۸.

وسیع‌تر و در سطح شهرستان‌های دیگر استان‌ها و حتی در سطح روستایی نیز استخراج شده و مورد تجزیه و تحلیل قرار گیرد. با توجه به نتایج این تحقیق پیشنهاد می‌شود، با استفاده از سیستم فایل توزیع‌شده هدوپ، مه‌داده خانواری را سریع‌تر از گذشته آماده کرده و با ارایه بهنگام خروجی‌ها و اطلاعات، تحلیل‌های سریع‌تر و بهتری را نسبت به گذشته انجام داد. همچنین پیشنهاد می‌شود با بکارگیری سیستم توزیع‌شده هدوپ بتوان بین اطلاعات استخراج‌شده سالانه خانواری در سطح شهرستان با اطلاعات سرشماری جمعیتی کشور ارتباط برقرار کرده و خلأی آماری و شاخص‌های برخورداری خانوار را تکمیل کرد.

واژه‌گان کلیدی: چارچوب هدوپ، سیستم فایل توزیع‌شده، نگاشت کاهش، مه‌داده، داده‌های خانواری.

۱- مقدمه

پیشرفت‌های اخیر در فن‌آوری‌های سخت‌افزاری و نرم‌افزاری، مانند رسانه‌های اجتماعی، اینترنت اشیا، حسگرهای پوشیدنی، فن‌آوری‌های موبایل و نظایر آن منجر به افزایش حجم داده‌ها و اطلاعات در این دنیای مدرن شده است و هر روز نیز به آن افزوده می‌شود. این داده‌ها می‌تواند از منابع متنوع از جمله ایمیل‌ها و تراکنش‌های آنلاین، اطلاعات چندرسانه‌ای مانند صدا و تصاویر ویدئویی، پایگاه‌های داده‌های بزرگ حاوی سوابق سلامت و داده‌های فیزیولوژیکی مانند ضربان قلب، رسانایی پوست و سایر اطلاعاتی که در طول تعامل کاربر با رسانه‌های اجتماعی گرفته شده است، جمع‌آوری شود. از موارد فوق به راحتی می‌توان نتیجه گرفت که در دنیای مدرن، داده‌ها با نرخ فزاینده‌ای تولید می‌شوند [۱۴] و [۲۹].

امروزه در شهرهای هوشمند، حجم داده‌هایی که می‌توان آن‌ها را جمع‌آوری و استفاده کرد، افزایش یافته است.

این واقعیت به‌عنوان ظهور مه‌داده[1]، مورد توجه دانشگاه و صنعت قرار گرفته است. تجزیه و تحلیل این داده‌های حجیم نیازمند تکنیک‌ها و روش‌هایی است که به پردازش و کشف الگوهای پنهان منجر شود و سپس روابط جالب بین متغیرها تفسیر شود. چارچوب هدوپ، محاسبات با کارایی بالا، محاسبات ابری، تکنیک‌های داده‌کاوی و الگوریتم‌های یادگیری ماشین کمک زیادی به ذخیره‌سازی آسان مه‌داده و تجزیه و تحلیل داده‌های حجیم کرده است [۱۴].

اهمیت داده‌ها در اقتصاد و جامعه اطلاعات محور را می‌توان در این جمله بیان کرد که «مه‌داده همانند نفت دارای ارزش است» که تأثیر زیادی بر زندگی مدرن داشته است [۱۳] و [۱۴]. مه‌داده را می‌توان به‌عنوان مجموعه عظیمی از داده‌ها با ساختاری متنوع و پیچیدگی فزاینده تعریف کرد. مشکل ذاتی در مواجهه با این حجم زیاد داده‌ها، چالش‌هایی نظیر پردازش، ذخیره‌سازی و تجزیه و تحلیل آنها است. علاوه بر این، اطلاعات باید به روشی علمی قابل تفسیر باشد و در زمان مناسب ارائه شود.

مه‌داده «در حال تبدیل شدن به یک دارایی شرکتی مهم، یک ورودی اقتصادی حیاتی، و پایه و اساس مدل‌های کسب و کار جدید است» [۱۳]. تکامل سریع فن‌آوری‌های اطلاعات، تجزیه و تحلیل داده‌های بزرگ و پروتکل‌های ارتباطی کارآمد، شتاب جدیدی را برای کسب و کارهای الکترونیکی و اتصال جهانی فراهم کرده است [۲۴]. تجزیه و تحلیل مه‌داده همچنین می‌تواند تلاش‌های دولت و مقامات محلی شهر هوشمند را برای ارائه خدمات بهتر به شهروندان خود تسهیل کند. مه‌داده‌ها می‌توانند به دولت‌ها در بهبود مراقبت‌های بهداشتی، حمل و نقل عمومی، آموزش و سایر زمینه‌های زندگی اجتماعی کمک کنند و در نتیجه به شکل‌گیری جامعه مدرن کمک کنند. به‌عنوان مثال، داده‌های سوابق ترافیکی را می‌توان برای بهبود خدمات حمل و نقل عمومی ارائه‌شده توسط دولت به مردم مورد استفاده قرار داد. به‌ویژه برنامه‌های مه‌داده مرتبط با اینترنت اشیا مانند خانه‌های هوشمند، ترکیبی از دستگاه‌های بهداشتی پوشیدنی و شهرهای هوشمند به‌طور گسترده به‌عنوان عوامل کلیدی شناخته می‌شوند که می‌توانند در توسعه اقتصادی و اجتماعی کشورهای در حال توسعه کمک کنند [۲۱].

۲- چارچوب نظری و بیان مسئله

در عصر دیجیتال، داده‌های متنوعی از منابع گوناگون تولید می‌شوند که رشد سریع فناوری‌ها، منجر به افزایش نرخ تولید و ایجاد حجم انبوهی از داده‌ها شده است. وجود این حجم از داده‌ها، امکان پیشرفت‌هایی را در زمینه‌های گوناگون علمی و صنعتی برای محققان و تصمیم‌گیران فراهم کرده است. یکی از این داده‌های مهم مه‌داده می‌باشد.

مه‌داده از منابع مهم در دنیای امروز است، که با استفاده از تجزیه و تحلیل‌های گوناگونی که روی آن انجام می‌گیرد اطلاعات و دانش ارزشمندی از آن بدست می‌آید [۱۴]. طی دو

دهه اخیر حجم این داده‌ها در حال گسترش بوده و رفته‌رفته برحجم آن نیز افزوده می‌شود. قسمتی از این داده‌ها ساختاریافته[۲] و قسمت اعظم آن ساختارنیافته[۳] هستند که این عامل باعث دشوار شدنِ فرایند تحلیل شده و چالش‌هایی را با خود به همراه آورده است [۱]. ویژگی‌هایی مانند سرعت[۴]، حجم[۵]، تنوع[۶]، صحت[۷] و ارزش[۸] در این نوع داده‌ها، پردازش داده‌ها را بسیار پیچیده‌تر کرده و این تغییرات همزمان، چالش‌هایی را برای تحلیل بیشتر داده‌ها ایجاد کرده است [۷] و [۱].

به‌منظور بهره‌برداری از مزایای داده‌های بزرگ در جامعه‌ای که دانش محور است، نیاز به توسعه راه‌حل‌هایی است که پیچیدگی و بار پردازش این حجم عظیم از داده‌ها را کاهش دهد [۲۰]. چالش‌های بزرگ استفاده از مه‌داده در دنیای واقعی از پیچیدگی روابط بین عوامل و متغیرهای مختلف مرتبط با داده‌ها ناشی می‌شود. یکی از عوامل درجه بالای ابعادی است که یک مجموعه داده ممکن است داشته باشد که دشواری پردازش و تجزیه و تحلیل داده‌ها را به دنبال دارد. فعل و انفعالات، روابط مشترک و اثرات علی این متغیرها بر یکدیگر، اغلب بسیار پیچیده‌تر از آن است که بدون نیاز به فناوری مدرن و چارچوب مناسب، تجزیه و تحلیل و درک شوند. لذا یکی از این ابزارهای مناسب که می‌تواند بر این پیچیدگی‌ها فایق آید و به تجزیه و تحلیل کمک کند استفاده از چارچوب هدوپ است.

چارچوب هدوپ برای توزیع و پردازش مه‌داده یکی از پرکاربردترین ابزاری ست که با زبان برنامه‌نویسی جاوا نوشته شده است. هدوپ یک ابزار مناسب است که این امکان را می‌دهد تا پردازش بر روی مجموعه داده‌های بزرگ با خوشه‌بندی انجام پذیرد و مدیریت داده‌های نیمه‌ساختاریافته[۹] و ساختارنیافته را تسهیل کند [۷] و [۱۴].

در حوزه آمارهای رسمی، داده‌های جمعیتی، سرشماری و خانواری در اقتصاد کلان و محاسبات ملی نقش اساسی و مهمی دارند و کشورها هر ساله برای جمع‌آوری داده‌ها و استخراج اطلاعات از آن هزینه زیادی می‌پردازند. یکی از این متغیرهای کلان، تولید ناخالص داخلی[۱۰] است که برای محاسبه آن ناگزیر به گردآوری اطلاعات خانواری است که بیش از ۵۰ درصد تولید ناخالص داخلی را تشکیل می‌دهد. در داده‌های خانواری نظیر هزینه و درآمد، هر پرسشنامه‌ی استاندارد به‌طور متوسط شامل صفات ذیل می‌باشد:

الف) خصوصیات اجتماعی اعضای خانوار نظیر سن، جنس، سطح سواد، وضع فعالیت و وضع زناشویی (۱۰ سوال)؛

ب) مشخصات محل سکونت و تسهیلات و لوازم عمده زندگی (۹۰ سوال)؛

ج) هزینه‌های خوراکی و غیرخوراکی خانوار (۱۵۰۰ سؤال)؛
د) درآمد اعضای خانوار (۵۰ سؤال).

از آنجایی‌که در جمع‌آوری داده‌های مربوط به سرشماری جمعیتی و خانواری، اطلاعات متنوعی سالانه در سطح هر کشور جمع‌آوری می‌شود، چالش‌هایی را از نظر مدیریت کردن داده‌های حجیم برای تحلیل در سطح خرد ایجاد می‌کند. یکی از این عوامل مهم و چالش‌زا نیز به جمع‌آوری داده‌ها، ذخیره‌سازی و پردازش آن برمی‌گردد [۳]، [۴] و [۲۳].

در ایران نیز هر ساله، داده‌های آماری خانواری با حدود ۴۰۰۰۰ نمونه شهری و روستایی گردآوری شده و نتایج آن در سطح کل کشور استخراج و سپس مورد تجزیه و تحلیل قرار می‌گیرد. تولید این نوع داده‌ها از یک سابقه طولانی نزدیک به ۶۰ سال برخوردار است، لذا با توجه به حجم انبوه داده‌ها در کشور، استخراج اطلاعات در سطح شهرستان‌ها می‌تواند کمک زیادی به محققان و تصمیم‌گیران استانی کرده و با تحلیل بهتری از وضعیت شاخص‌های اجتماعی اقتصادی، برای آینده برنامه‌ریزی کنند.

۲-۱- مه‌داده

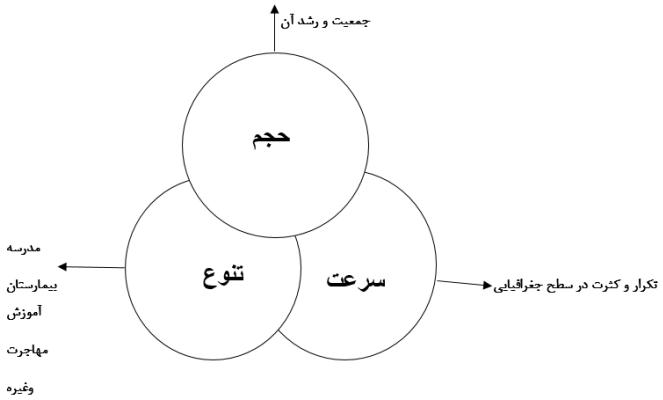
امروزه در عصری زندگی می‌کنیم که شاهد تولید حجم زیادی از داده و بخصوص مه‌داده هستیم که حاصل فن‌آوری اطلاعات و ارتباطات است. داده‌ها به‌عنوان یک دارایی با ارزش اقتصادی کاربردهای متفاوتی دارند. این داده‌ها برای کاربردهای مختلفی نظیر پیش‌بینی، نظارت بر بحران‌های طبیعی، تغییرات اقلیمی، تحلیل‌های جمعیتی، صنعتی و اقتصادی مورد استفاده قرار می‌گیرند [۲۷].

مه‌داده به مجموعه‌ای از داده‌های حجیم و پیچیده اطلاق می‌شود که پردازش آن‌ها با استفاده از سیستم‌های پایگاه داده یا نرم‌افزارهای پردازش سنتی کاری دشوار است. به طور رسمی مه‌داده را با مؤلفه‌های حجم، سرعت، تنوع، ارزش و صحت می‌شناسند. حجم در واقع مقدار داده‌هایی است که هر روز تولید می‌شوند و به آن نیز افزوده می‌شود. در حالی‌که سرعت، به نرخ رشد داده‌ها و سرعتی که برای تحلیل، گِرد هم می‌آیند اشاره دارد.

سازمان‌ها و شرکت‌های بزرگ مدت‌ها است با عبارت مه‌داده یا همان حجم عظیمی از داده‌ها که ذخیره و مورد بهره‌برداری قرار می‌گیرد با چالش مواجه شده‌اند. این داده‌ها به قدری بزرگ و حجیم هستند که با ابزارهای مدیریتی و پایگاه‌های داده سنتی و معمولی

قابل مدیریت نیستند، از این‌رو مشکلات اصلی در کار با این نوع داده‌ها مربوط به برداشت و جمع‌آوری، ذخیره‌سازی، جستجو، اشتراک‌گذاری، تحلیل و نمایش آن‌ها است.
حجم داده‌های ذخیره‌شده در مجموعه مه‌داده عموماً به دلیل تولید و جمع‌آوری داده‌ها از مجموعه تجهیزات و ابزارهای مختلف مانند موبایل، تبلت، انواع حسگرهای محیطی، شبکه‌های اجتماعی، دوربین‌ها، میکروفون‌ها، دستگاه‌های تشخیص و غیره با سرعت خیره‌کننده‌ای در حال افزایش است.
برای به‌کارگیری و تحلیل مه‌داده به مجموعه‌ای از تکنیک‌ها و روش‌ها نیاز است تا بتوانند ارزش‌ها و اطلاعات متنوعی که در این حجم از داده‌ها پنهان شده‌اند را آشکار سازند. از این‌رو با رشد روز افزون داده‌ها و نیاز به بهره‌برداری و تحلیل از این داده‌ها، به‌کارگیری زیرساخت‌های مه‌داده از اهمیت ویژه‌ای برخوردار شده است. شرکتی نظیر گوگل در سال‌های اخیر با شناخت و درک این موضوع، توانسته است گام‌های مؤثری را در راستای توسعه این حوزه برداشته و افتخار دارد یکی از مؤسسات پیشرو در این زمینه است [۸].
با پیشرفت‌های روزافزون در فن‌آوری اطلاعات، هر روز جذابیت و مقبولیت استفاده از مه‌داده بیشتر از قبل احساس می‌شود. بیشترین نیاز در پردازش مه‌داده، توسط دانشمندان در زمینه‌هایی مانند اخترشناسی، هواشناسی، ژنتیک، تغییرات آب و هوایی، شبیه‌سازی‌ها، فیزیک، زیست‌شناسی، زیست‌محیطی و تجاری انجام می‌گیرد. در نتیجه با استفاده از حجم بیشتری از داده‌ها، می‌توان تحلیل‌های بهتر و پیشرفته‌تری را برای اهداف مختلف نظیر تجاری، پزشکی، دفاعی و امنیتی دریافت کرد.
مه‌داده چند ویژگی اصلی دارد. اولاً، همانطور که از نام آن پیداست، خود داده‌ها متنوع هستند. ثانیاً، نمی‌توان داده‌ها را در پایگاه‌های اطلاعاتی رابطه‌ای منظم طبقه‌بندی کرد و در نهایت جریان‌های داده به سرعت ایجاد، ذخیره و تجزیه و تحلیل می‌شوند [۱۲]. همانطور که گرهارد ذکر می‌کند، داده‌های بزرگ یک جهش انقلابی به جلو نسبت به تحلیل سنتی است که دارای ویژگی‌های اصلی نظیر حجم، تنوع و سرعت است [۱۱]. حجم به مقدار داده‌ای اشاره دارد که ایجاد و ذخیره می‌شود. تنوع به انواع مختلف داده‌های جمع‌آوری شده مرتبط است و سرعت را می‌توان به‌عنوان سرعت تولید، جریان و تجمیع داده‌ها تعریف کرد [۱۷]. در مطالعه کایسلر و همکاران، ارزش و پیچیدگی داده نیز به‌عنوان ویژگی‌های مه‌داده ذکر شده است. ارزش داده معیاری برای سودمندی داده‌ها در

فرآیندهای تصمیم‌گیری است، در حالی که پیچیدگی به میزان وابستگی متقابل صفات و به‌هم پیوستگی آنها در ساختارهای مه‌داده اشاره دارد [۱۷].
شکل ۱ به مؤلفه‌هایی از داده‌های خانواری و سرشماری اشاره دارد که به‌عنوان مه‌داده مورد تجزیه و تحلیل قرار گرفته است. به دلیل ابعاد زیاد صفات آماری در پرسشنامه‌های خانواری، سرشماری‌ها و همچنین حجم جامعه و حجم نمونه‌ها، این نوع داده‌ها به‌عنوان مه‌داده به شمار می‌رود [۷] و [۸].



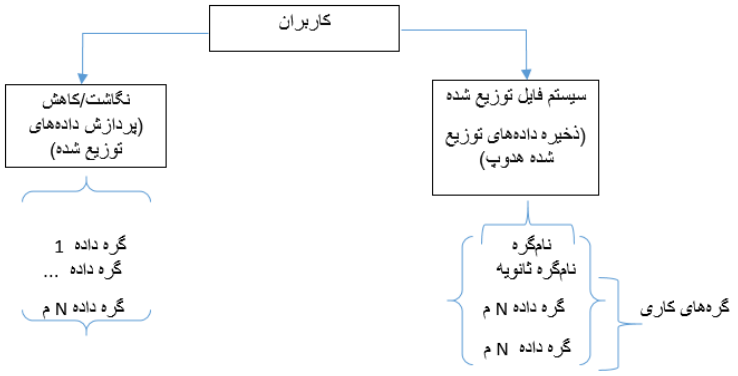
شکل ۱- مه‌داده جمعیتی [۷]

۲-۲- چارچوب هدوپ و مه‌داده

هدوپ دارای دو قسمت اصلی به نام الف) سیستم فایل توزیع‌شده هدوپ و ب) نگاشت/کاهش[11] است که اولی وظیفه ذخیره‌سازی داده‌ها را بر عهده دارد، و دومی رایانش موازی، محاسبات و بازیابی را انجام می‌دهد.
نگاشت/کاهش قلب سیستم هدوپ در نظر گرفته شده است که پردازش موازی را روی مجموعه مه‌داده در اندازه ترابایت و پتابایت انجام می‌دهد. هدوپ مبتنی بر پردازش خوشه‌ای است و داده‌های حجیم غیر ساختاریافته و نیمه‌ساختاریافته را مدیریت می‌کند که در مقایسه با سیستم‌های پایگاه داده رابطه‌ای سنتی که تنها بر روی داده‌های ساختاریافته کار می‌کند دارای برتری است [۹] و [۱].

سیستم فایل توزیع‌شده هدوپ

سیستم فایل توزیع‌شده هدوپ را یاهو بر اساس فایل سیستم گوگل پیاده‌سازی کرده است. همان‌طور که از اسم آن مشخص است، سیستم فایل توزیع‌شده هدوپ، دارای تحمل‌پذیری اشکال[۱۲] و قابل اتکا می‌باشد. سیستم فایل توزیع‌شده هدوپ را می‌توان به شکل گره اصلی[۱۳] و گره کاری[۱۴] معماری‌شده در نظر گرفت که شامل نام‌گره[۱۵]، گره‌داده[۱۶] و نام‌گره ثانویه[۱۷] است. نام‌گره یک گره اصلی است که همه گره‌داده‌ها را کنترل و عملیات سیستم فایل را مدیریت می‌کند [۱].



شکل ۲- معماری هدوپ [۱]

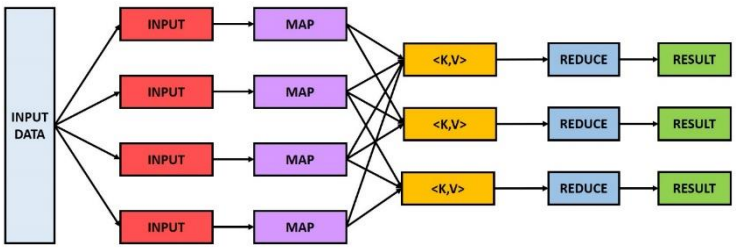
همانگونه که در شکل ۲ نشان داده شده است گره‌داده، گره‌های کاری هستند که کار واقعی مانند عملیات بلوک را انجام می‌دهند. همچنین تنها یک نام‌گره ثانویه هدوپ وجود دارد که همانند یک گره پشتیبان عمل می‌کند. قسمت اصلی، داده‌ها را به بلوک‌هایی تقسیم می‌کند که بعداً در گره‌داده‌ها در یک خوشه ذخیره می‌شوند. با پیش فرض ۳ تکثیر، سیستم فایل توزیع‌شده هدوپ نسخه‌هایی از بلوک را در گره محلی، گره‌داده‌های دوم و سوم قرار می‌دهد. این تکثیر بلوک‌ها برای جلوگیری از دست رفتن داده‌ها در صورت مواجه با شکست گره‌داده[۱۸] استفاده می‌شود که می‌تواند به دلخواه و بر اساس نیاز ما تنظیم شود.

اندازه پیش‌فرض بلوک در سیستم فایل توزیع‌شده هدوپ ۶۴ مگابایت تعریف شده است که در صورت نیاز هم می‌تواند افزایش یابد [۱].

نگاشت/کاهش

امروزه چارچوب نگاشت/کاهش در بیشتر پردازش داده‌های حجیم مورد استفاده قرار می‌گیرد. نوآوری مهم نگاشت/کاهش امکان انجام پرس و جو از یک مجموعه داده، تقسیم و اجرای آن به صورت موازی بر روی گره‌های متعدد است. نگاشت/کاهش تکنیکی است که توسط گوگل برای پردازش موازی داده‌ها ابداع شده است. نگاشت/کاهش روی مجموعه داده‌های حجیم ذخیره‌شده در سیستم فایل توزیع‌شده هدوپ کار می‌کند. مقداردهی اولیه کار، واحدهای کاری تعریف‌شده، به‌روزرسانی وضعیت، اجرای دستورات، پیشرفت و تحویل کار فعالیت‌هایی هستند که نگاشت/کاهش انجام می‌دهد. تمام این فعالیت‌ها توسط JobTracker مدیریت و سپس توسط TaskTracker انجام می‌شود.

نگاشت/کاهش دارای دو وظیفه تفکیک‌شده است. قسمت نگاشت[۱۹] که به نقشه و طراحی برمی‌گردد و قسمت کاهش[۲۰] که عملیات استخراج و نتایج را به عهده دارد. در انجام عملیات نگاشت داده‌ها به چند قسمت تقسیم می‌شوند که خروجی را به‌صورت جفت‌ها یا تاپل[۲۱] تحویل می‌دهد. سپس این تاپل‌ها به مرحله کاهش منتقل می‌شوند که بعد از ترکیب، یک خروجی واحد مورد نظر و از پیش تعیین‌شده از آن استخراج می‌گردد [۱] و [۹].



شکل ۳- چارچوب نگاشت/کاهش [۹]

همانگونه که در شکل ۳ نشان داده شده است در معماری چارچوب نگاشت/کاهش، ابتدا ورودی و قسمت‌بندی داده‌ها طراحی شده است. در فاز اول نگاشت، کلید K1 و مقدار V1 ارائه می‌شود سپس در فاز دوم کاهش، عمل پردازش و محاسبات و به‌دست آوردن نتایج انجام می‌شود [۹].

۳- پیشینه تحقیق

همانطور که در بخش‌های قبلی این تحقیق آمده است، چارچوب هدوپ در پاسخ به سؤالات و حل چالش‌های پیش روی مه‌داده توسط بنیاد آپاچی ساخته شده است. در ادامه این بخش سعی شده است مطالعات انجام‌شده، تحقیقات و تجارب بدست آمده در حوزه مه‌داده و چارچوب هدوپ ارائه شود.

آگاوان و همکاران [۱] برای تجزیه و تحلیل مجموعه داده‌های سرشماری ایالات متحده امریکا از چارچوب هدوپ استفاده کردند. برای شروع ابتدا داده‌های خام سرشماری که در فرمت CSV هستند در سیستم فایل توزیع‌شده هدوپ بارگذاری شده و بعد از آن عمل ویژه نگاشت/کاهش بر روی مجموعه داده‌ها برای یافتن نتایج مورد نظر انجام شد و سپس نتایج تولیدشده در سیستم فایل توزیع‌شده هدوپ ذخیره شد. پس از ذخیره شدن داده‌ها و عمل نگاشت/کاهش، فایل‌های خروجی به فرمت مناسب تبدیل شده و نتایج این تحقیق نشان داده است که، با افزایش تعداد گره‌ها در خوشه هدوپ و توزیع داده‌ها در گره‌های مختلف، زمان اجرا کاهش یافته و موجب افزایش کارایی سیستم می‌شود. مطالعه دیگر بر روی ویروس کرونا انجام شده است. این روزها دنیا شاهد یکی از بزرگترین بیماری همه‌گیری جهانی[۲۲] با نام کرونا می‌باشد و داده‌های حجیم و متنوعی نیز جمع‌آوری شده است. این داده‌ها ساختاریافته، نیمه‌ساختاریافته و ساختارنیافته می‌باشد که به دلیل پیچیدگی آن محققان را با چالش‌هایی مواجه کرده است. استخراج و تحلیل نتایج یکی از این چالش‌ها است. برای حل چالش این مه‌داده که مربوط به ویروس کووید ۱۹ است، آزروال و فابر [۳] از چارچوب هدوپ استفاده کردند. محققان به منظور پردازش سریع داده‌هایی که ترکیبی از ساختاریافته و ساختارنیافته است چارچوب نگاشت/کاهش را به کار گرفتند. رویکردی که در این تحقیق به کار گرفته شد، ارتباط و استفاده از دانش موجود

در سطح صنعت و دانشگاه بود. نتیجه این تحقیق نشان داده است که با استفاده از چارچوب هدوپ و انجام نگاشت/کاهش می‌توان در تجزیه و تحلیل مه‌داده به برنامه‌نویسان و کاربران به دلیل ذخیره‌کردن زمان کمک کرد. مقادیر ورودی در حجم بزرگ به مقادیر جزیی کوچک‌تر محاسبه شده و نتیجه نهایی را تسریع می‌کند. این تحقیق نشان داده است در مواجه با مه‌داده‌ای مانند کووید ۱۹ تهدیدهایی هم وجود دارد که استخراج و تحلیل اطلاعات را با مشکل روبرو کرده و نتایج را شکننده می‌کند. این تهدیدها شامل موارد زیر است:

الف) مشکل کیفیت و سطح استاندارد این نوع مه‌داده به دلیل ساختارنیافته بودن آن؛

ب) ناهمگن بودن داده‌ها، به دلیل جمع‌آوری آنها از منابع مختلف [۵] و [۶].

بخاری و همکاران برای تحلیل مه‌داده جمعیتی از چارچوب هدوپ استفاده کردند [۷]. این مجموعه از داده‌ها با افزایش جمعیت همواره در حال رشد است و در نتیجه بقیه صفات آماری نظیر مدرسه، بیمارستان و دانشگاه نیز در حال اضافه شدن به این پایگاه داده می‌باشد. برای تحلیل‌های جمعیت‌شناسی مانند نرخ مرگ و میر، باروری، میزان باسوادی و حتی کاهش فقر، نیاز به منابع داده‌ای دیگر هم هست. لذا یکی از چالش‌های پیش روی محققان استفاده از پایگاه‌های داده مختلف است که داده‌های مرتبط با جمعیت به طور سنتی در آن ذخیره شده است. برای انتقال داده از پایگاه داده‌های مختلف به چارچوب هدوپ از ابزار اسکوپ[۲۳] استفاده شده است. از آنجایی که پایگاه داده سنتی به دلیل سرعت پایین در پردازش حجم بالای داده‌های جمعیتی یک چالش برای این تحقیق محسوب می‌شد و از انعطاف‌پذیری کمتری نیز برخوردار بود، لذا برای تسریع در نتایج و تحلیل اطلاعات، از معماری چارچوب هدوپ برای انتقال داده استفاده شد. نتیجه این تحقیق نشان داده است، علاوه بر انعطاف‌پذیری چارچوب هدوپ نسبت به پایگاه داده‌های سنتی، کارایی آن بالا و کمترین زمان را برای پردازش صرف کرده است.

نیوبری و ژانگ اثربخشی کاربرد سیستم فایل توزیع‌شده هدوپ بر مه‌داده را بررسی کردند [۲۲]. آنها در این تحقیق به این نتیجه رسیده‌اند که اگر بجای معماری سنتی از چارچوب هدوپ و سیستم فایل توزیع‌شده هدوپ استفاده کنند منجر به آزادسازی پهنای باند شده و دسترسی به آن را افزایش می‌دهد. یانگ و همکاران [۳۰] در مطالعات دیگری که مربوط

به تحلیل داده‌های زمین‌شناسی است از چارچوب نگاشت/کاهش هدوپ استفاده کرده‌اند. محققان نشان داده‌اند که بهره‌گیری از چارچوب نگاشت/کاهش هدوپ و بهره‌گیری از زیرساخت محاسبات با کارایی بالا، منجر به کارایی بالا و کاهش هزینه شده است. مطالعات سوکشینی و همکاران در حوزه نجوم و اخترشناسی نیز نشان داده است استفاده از زیرساخت محاسبات با کارایی بالا و به کارگیری هوش مصنوعی و چارچوب هدوپ منجر به افزایش کارایی و تسریع در پردازش اطلاعات و کاهش هزینه‌ها شده است [۲۷]. همچنین مطالعه دیگری که توسط راتاناپس و کاوکری [۲۵] برای شمارش لغات[۲۴] در مجموعه داده‌های حجیم استفاده شده، بکارگیری چارچوب هدوپ بوده به طوری‌که بهترین زمان اجرا را نیز به دنبال داشته است. مطالعه‌ای مقایسه‌ای که توسط شارما و کار برای پردازش مه‌داده انجام شد [۲۶] نشان داد برای پردازش داده‌های حجیم باید از ابزارها و چارچوب‌های مناسب استفاده کرد. آنها در این تحقیق چارچوب و ابزارهای مختلفی برای پردازش و تجزیه و تحلیل مه‌داده مانند اسپارک، پیگ، هایو، هدوپ، سیستم فایل توزیع شده و نگاشت/کاهش را معرفی می‌کنند تا بتوانند چالش‌های پیش‌روی پایگاه داده سنتی را حل کنند. این چالش‌ها شامل سرعت پایین در پردازش حجم بالای داده‌ها و عدم همگنی داده‌ها است که از منابع مختلف جمع‌آوری می‌شود. هریک از چارچوب‌های معرفی شده داده‌های اولیه را به‌عنوان ورودی دریافت کرده و با پردازش به خروجی‌های میانی تبدیل کرده و در اختیار مرحله بعد قرار می‌دهد تا استخراج نهایی انجام پذیرد. در این تحقیق نشان داده شد، چنانچه از نظر حافظه محدودیتی وجود نداشته باشد اسپارک به مراتب نسبت به پیگ و نگاشت/کاهش از سرعت پردازش سریع‌تری برخوردار است. جین فنگ ما و همکاران از چارچوب هدوپ برای محاسبات ابری و یافتن مدل بهینه در مطالعات زیست‌محیطی آب و خاک استفاده کرده‌اند. در این روش به تعداد قابل توجهی مدل نیاز بود که بار محاسباتی را کاهش دهند. تکنیک‌های بهینه‌سازی مبتنی بر مدل‌های متوالی متداول مورد بررسی و مطالعه قرار گرفت و در نهایت از بهینه‌سازی بیزی برای کالیبره کردن یک مدل استفاده شد. الگوریتم بهینه‌سازی بیزی بر اساس چارچوب هدوپ بار کاری را به‌طور خودکار متعادل کرده و زمان محاسبات ابری را بطور بهینه کاهش می‌دهد [۱۸].

مطالعه‌ای دیگر که توسط جایسوال و همکاران انجام شده [۱۶] به مزایا و معایب هدوپ، اسپارک و دیگر ابزارها پرداخته است. در این تحقیق نشان داده شد برای پردازش مه‌داده هر یک از ابزارها دارای مزایایی نسبت به دیگری هستند. برای مثال اسپارک از سرعت پردازش بالا و تأخیر زمانی کمتری برخوردار است ولی هدوپ نسبت به بقیه ابزارها هم از مقیاس‌پذیری بالا و هم از امنیت[۲۵] بالایی برخودار است که به وسیله آن و به طور موازی حجم زیادی از داده‌ها را می‌توان پردازش کرد [۲] و [۱۵].

مطالعه‌ای دیگر نیز توسط محمود و همکاران انجام شده است [۱۹] که در آن به مزایا و معایب هدوپ در محاسبات ابری و پردازش مه‌داده پرداخته‌اند. در این تحقیق نشان داده شد که برای پردازش مه‌داده، هدوپ نسبت به بقیه ابزارها هم از هزینه پایین‌تر وهم از مقیاس‌پذیری و امنیت بالایی برخودار است که حجم زیادی از داده‌ها را می‌توان پردازش کرد [۱۰] و [۱۹]. در جدول زیر به‌طور خلأصه مزایا و معایب چارچوب‌ها و ابزارهای پردازش مه‌داده استخراج شده است که مشاهده می‌کنید.

جدول ۱- مقایسه و مزایای چارچوب‌های مناسب در پردازش مه‌داده

شرح	هدوپ	اسپارک	فلینگ و بقیه
تحمل‌پذیری اشکال (خطا)	بسیار بالا	بالا	بالا
مقیاس‌پذیر	بسیار بالا	بالا	متوسط
زمان اجرا	بالا	بسیار کم	کم
امنیت	بسیار بالا	بالا	بالا
کارایی	بالا	بسیار بالا	بسیار بالا
هزینه	خیلی پایین	متوسط	متوسط

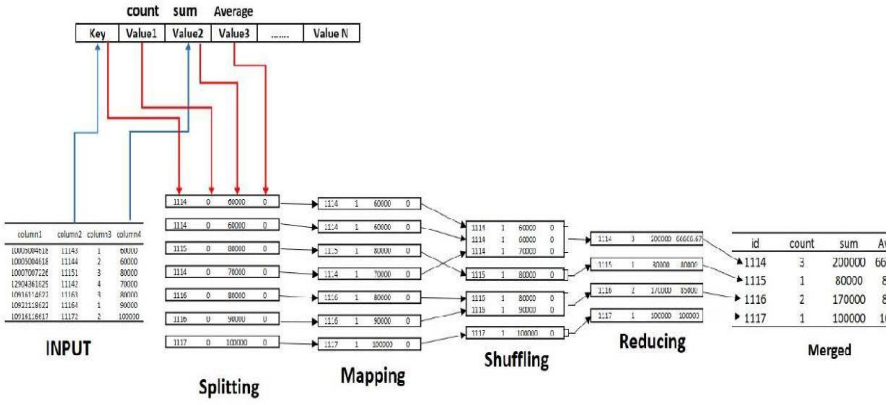
منبع: [۲، ۱۰، ۱۵، ۱۶، ۱۸، ۱۹، ۲۶]

۴- مدل پیشنهادی تحقیق و داده‌ها

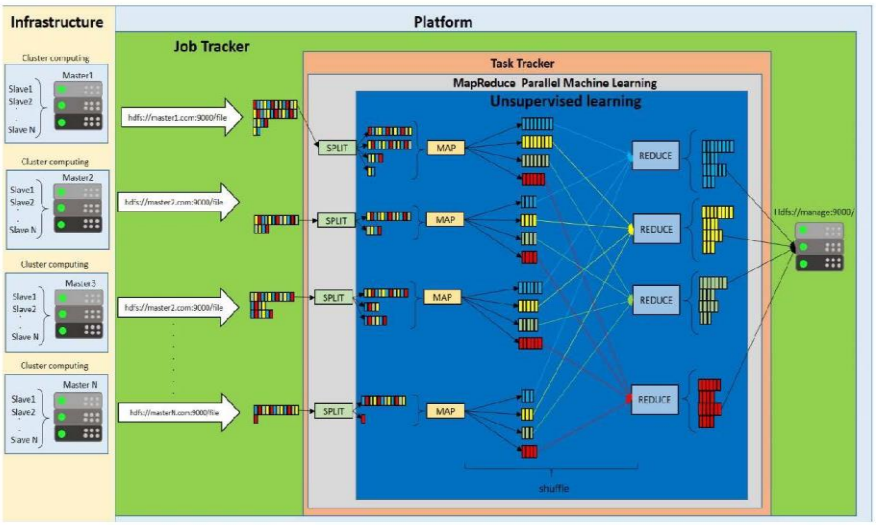
سیستم فایل توزیع‌شده هدوپ یکی از سریع‌ترین سکو (پلتفرم)های شناخته‌شده برای پردازش و ذخیره‌سازی داده‌های بزرگ غیر متمرکز در سراسر خوشه‌های سیستم است [۲۸]. با توجه به نقاط قوتی که این پلتفرم در استخراج داده‌های حجیم دارد، لذا این مقاله

بر بکارگیری سیستم فایل توزیع‌شده هدوپ و نگاشت/کاهش تمرکز دارد. در این قسمت جزییات مدل پیشنهادی و روش استخراج از داده‌های خانواری ارائه می‌شود.

در این مدل پیشنهادی، قسمتی از ورودی‌ها و جفت‌های کلید-مقدار خانواری برای نمونه در شکل‌های ۴ و ۵ آورده شده است. کلید شامل ستون کد خانوار با کد استان و مقدار نیز مانند تعداد خانوار، هزینه، درآمد و اشتغال است که در شکل نشان داده شده است. داده‌ها بعد از بلوک‌بندی و قسمت‌بندی شدن با Splitting از سطرها جدا شده و برای تولید خروجی میانی آماده می‌شود که در این مرحله با انجام عملیات نگاشت داده‌ها برای تولید مقادیر خروجی نهایی آماده می‌شود. بعد از عمل نگاشت Shuffling انجام شده سپس خروجی shuffling به‌عنوان ورودی برای مرحله کاهش ارسال می‌شود و می‌توان عملیات average، sum، max، min و Count را به ترتیب برروی استان‌ها و شهرستان‌ها انجام داد و خروجی مورد نظر را استخراج کرد.



شکل ۴- ورودی‌ها وکلید-مقدار داده‌های خانوار



شکل ۵- مدل پیشنهادی نگاشت/کاهش داده‌های خانوار

شکل ۵ مدل پیشنهادی برای پیاده‌سازی و اجرای داده‌های خانواری را نشان می‌دهد. در مرحله اول راه‌اندازی، سرورهای ماشین مجازی برای ۴ تا گره‌ها در نظر گرفته شده است که داده خام از sql به csv تبدیل شده را در فایل‌های HDFS اپلود می‌کند. در زیر کدهای نوشته‌شده مرتبط را مشاهده می‌کنید.

```
FileInputFormat.addInputPath(job,new  Path("hdfs://192.168.1.145:9000
/U99P3S01.csv"));
FileInputFormat.addInputPath(job,new
Path("hdfs://192.168.1.146:9000/U99P3S01.csv"));
FileInputFormat.addInputPath(job,new
Path("hdfs://192.168.1.147:9000/U99P3S01.csv"));
FileInputFormat.addInputPath(job,new
Path("hdfs://192.168.1.149:9000/U99P3S01.csv"));
```

U99P3S01 فایل خانواری است که به ترتیب حروف از چپ به راست معرف داده شهری، سال ۱۳۹۹، قسمت ۳ و بخش ۱ هزینه را نشان می‌دهد. برای آشنایی بیشتر به راهنمای داده خام و پرسشنامه طرح هزینه و درآمد خانوار مرکز آمار ایران رجوع شود.

۴-۱- داده‌های تحقیق

یکی از این منابع مهم آماری، داده‌های آمارهای هزینه و درآمد خانوار است که کاربرد مهمی در اقتصاد کلان دارد. این منبع مهم آماری دارای تعداد زیادی صفات است که اطلاعات بسیار مفیدی را برای انواع تحلیل‌ها در سطح کل کشور، استانی و شهرستانی ارائه می‌دهد.

جدول ۲- حجم فایل‌های داده‌خام خانواری (هزینه و درآمد خانوارشهری) کشور، خوشه‌بندی شده

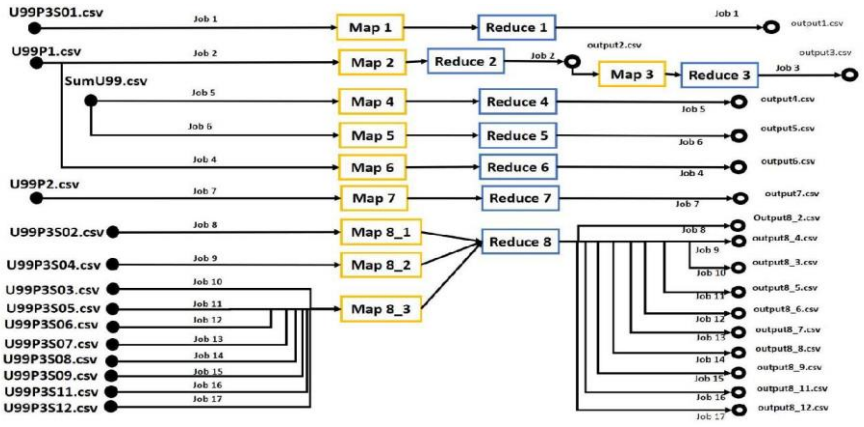
شماره job	خوشه‌بندی فایل کشور به چهار خوشه (بایت)			یک فایل کشوری (بایت)	نام فایل	
۱	۵،۰۲۶،۳۰۶	۶،۷۷۴،۸۰۷	۵،۲۹۵،۱۱۷	۵،۶۰۵،۸۳۴	۲۲،۷۰۲،۰۶۴	U99P3S01.csv	۱
۲،۴	۴۳۵،۴۳۴	۶۰۰،۳۲۹	۵۳۳،۵۵۹	۴۷۳،۹۳۳	۲،۰۴۳،۲۵۵	U99P1.csv	۲
۵،۶	۱،۹۹۲،۸۸۸	۲،۶۹۳،۲۴۶	۲،۳۱۷،۱۸۹	۲،۲۲۹،۰۷۰	۹،۲۹۴،۰۰۰	SumU99.csv	۳
۷	۳۴۸،۶۱۶	۴۶۸،۷۹۸	۴۰۴،۸۵۴	۳۹۱،۶۵۶	۱،۶۱۳،۹۲۴	U99P2.csv	۴
۸	۲۱،۹۳۸	۲۹،۷۲۰	۳۳،۶۴۰	۳۳،۱۳۱	۱۵۳،۸۱۹	U99P3S02.csv	۵
۹	۵۴۰،۵۵۴	۷۰۸،۴۲۷	۶۰۴،۹۶۸	۶۱۷،۸۳۲	۲،۸۶۷،۹۷۹	U99P3S04.csv	۶
۱۰	۱۸۳،۰۴۸	۲۲۶،۱۸۰	۲۲۰،۷۲۲	۲۱۰،۸۳۹	۹۷۸،۰۵۴	U99P3S03.csv	۷
۱۱	۶۳۳،۹۵۶	۷۲۹،۴۷۴	۶۷۵،۴۴۹	۷۲۶،۹۶۰	۳،۲۸۹،۵۳۹	U99P3S05.csv	۸
۱۲	۲۱۶،۵۷۷	۲۵۱،۹۶۷	۲۴۴،۳۸۶	۲۴۳،۲۴۵	۱،۱۱۸،۸۳۲	U99P3S06.csv	۹
۱۳	۲۵۱،۹۸۷	۳۰۰،۸۴۲	۲۵۰،۹۳۹	۲۶۵،۶۵۱	۱،۲۴۸،۴۵۶	U99P3S07.csv	۱۰
۱۴	۲۵۶،۰۲۱	۳۴۵،۰۶۵	۲۸۸،۱۵۱	۲۸۴،۹۲۴	۱،۳۸۳،۳۲۷	U99P3S08.csv	۱۱
۱۵	۶۱،۸۹۸	۸۴،۰۶۲	۶۲،۶۳۹	۶۲،۱۰۷	۳۲۰،۹۵۸	U99P3S09.csv	۱۲
۱۶	۷۳،۹۶۸	۹۹،۱۳۸	۴۷،۷۴۸	۷۳،۵۷۸	۳۳۵،۳۱۳	U99P3S11.csv	۱۳
۱۷	۵۲۶،۱۳۴	۵۹۳،۱۴۱	۵۳۷،۳۶۴	۵۵۳،۸۱۰	۲،۲۱۰،۴۷۰	U99P3S12.csv	۱۴

در این تحقیق از داده‌های هزینه و درآمد کشور برای سال ۱۳۹۹ که در سایت مرکز آمار ایران [۳۱] موجود بود استفاده شد. این مجموعه داده‌های خانواری شامل حدود ۴۰۰۰۰ نمونه شهری و روستایی است که اندازه و حجم فایل در جدول ۲ نشان داده شده است. در

این جدول، شماره هر job معرف جداول خروجی از داده‌های خام است که شامل ۱۷ job است.

۵- استخراج نتایج

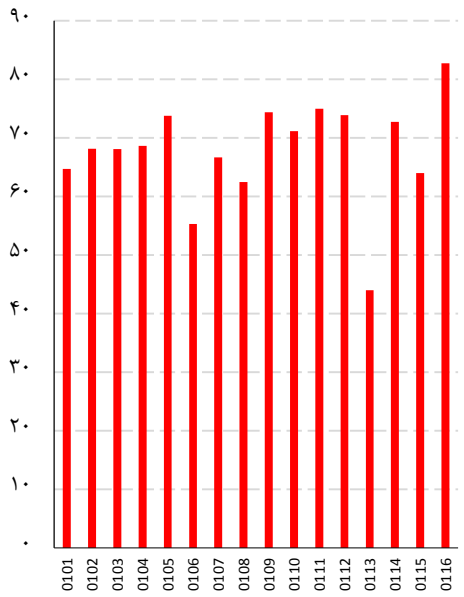
همانگونه که پیشتر اعلام شد، هدف از این تحقیق، استفاده از سیستم‌های توزیع‌شده و چارچوب هدوپ در تسریع پردازش آمارهای رسمی و مه‌داده‌های خانواری است، به‌طوریکه با ارائه بهنگام خروجی‌ها بتوان تحلیل بهتری را نسبت به گذشته ارائه داد. این خروجی‌ها اطلاعات بسیار مفیدی را برای انواع تحلیل‌ها در سطح کل کشور، استانی و شهرستانی ارائه می‌دهد و می‌تواند کمک شایانی به ارزیابی و تحلیل شاخص‌های اجتماعی و اقتصادی کند. همانگونه که اشاره شد فایل‌ها با نام‌های U99P3S01 و مشابه آن حاوی داده‌های خام می‌باشند که تبدیل‌یافته آن با پسوند CSV ذخیره شده است. شکل ۶ پیاده‌سازی و اجرای مدل نگاشت/کاهش با داده‌های خانواری را نشان می‌دهد. فایل‌های CSV از مبدأ و به‌عنوان ورودی نگاشت بکارگرفته‌شده و با مشخص شدن هر job، آن را برای محاسبه و عملیات کاهش آماده می‌کند [۱۳] و [۲۹]. پس از عملیات job اول، عملیات روی job دوم و تا ۱۷ job ادامه پیدا می‌کند. در اینجا منظور از هر job، جدول و خروجی نهایی می‌باشد که در فرمت Excel استخراج می‌شود.



شکل ۶- پیاده‌سازی و اجرای مدل نگاشت/کاهش با داده‌های خانواری

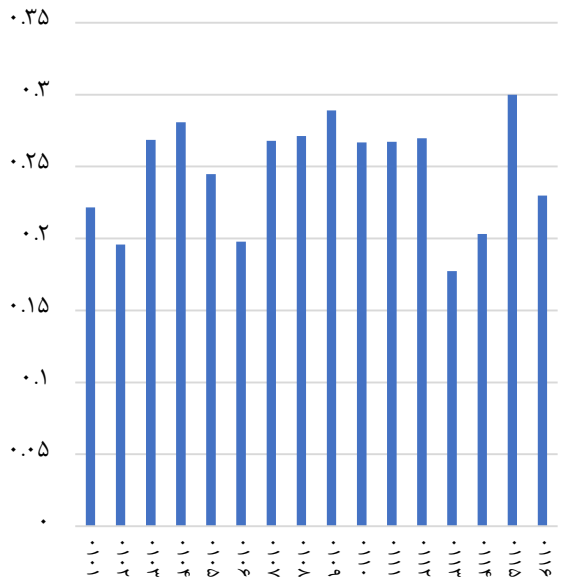
صفات موجود در داده‌های خانواری شامل هزینه‌های خوراکی، تحصیل و آموزش، بهداشت و سلامت، هزینه ارتباطات، دسترسی به اینترنت، فعالیت‌های شغلی، درآمدی و نظایر آن است که استخراج اطلاعات آن در سطح شهرستان‌های کشور می‌تواند کمک شایانی به ارزیابی و تحلیل شاخص‌های اجتماعی و اقتصادی کند که تاکنون نجام نشده است. از آنجایی‌که خروجی‌ها این تحقیق در فرمت اکسل است و دارای اطلاعات تفصیلی‌تر می‌باشد، بنابراین تعدادی از جداول در اینجا آمده است و قسمت کوچکی از خروجی اطلاعات تحلیل می‌شود.

استان و شهرستان	درصد خانوار برخوردار از اینترنت
۰۱۰۱	۶۴/۷۰۵۸۸۲۳۵
۰۱۰۲	۶۸/۱۸۱۸۱۸۱۸
۰۱۰۳	۶۸/۱۱۵۹۴۲۰۳
۰۱۰۴	۶۸/۶۲۷۴۵۰۹۸
۰۱۰۵	۷۳/۷۸۰۴۸۷۸
۰۱۰۶	۵۵/۳۱۹۱۴۸۹۴
۰۱۰۷	۶۶/۶۶۶۶۶۶۶۷
۰۱۰۸	۶۲/۵
۰۱۰۹	۷۴/۳۵۸۹۷۴۳۶
۰۱۱۰	۷۱/۱۸۶۴۴۰۶۸
۰۱۱۱	۷۵
۰۱۱۲	۷۳/۹۱۳۰۴۳۴۸
۰۱۱۳	۴۴
۰۱۱۴	۷۲/۷۲۷۲۷۲۷۳
۰۱۱۵	۶۴
۰۱۱۶	۸۲/۷۵۸۶۲۰۶۹



شکل ۷- درصد خانوارهای برخوردار از اینترنت به تفکیک شهرستان‌های استان ۰۱ (%)

استان و شهرستان	متوسط شاغل خانوار
۰۱۰۱	۰/۲۲۱۵۶۸۶۲۷
۰۱۰۲	۰/۱۹۵۸۳۳۳۳۳
۰۱۰۳	۰/۲۶۸۵۹۹۰۳۴
۰۱۰۴	۰/۲۸۰۷۱۸۹۵۴
۰۱۰۵	۰/۲۴۴۶۷۱۸۹۳
۰۱۰۶	۰/۱۹۷۸۷۲۳۴
۰۱۰۷	۰/۲۶۷۹۰۱۲۳۵
۰۱۰۸	۰/۲۷۱۲۵
۰۱۰۹	۰/۲۸۸۸۸۸۸۸۹
۰۱۱۰	۰/۲۶۶۶۶۶۶۶۷
۰۱۱۱	۰/۲۶۷۲۲۲۲۲۲
۰۱۱۲	۰/۲۶۹۵۶۵۲۱۷
۰۱۱۳	۰/۱۷۷۳۳۳۳۳۳
۰۱۱۴	۰/۲۰۳۰۳۰۳۰۳
۰۱۱۵	۰/۳
۰۱۱۶	۰/۲۲۹۸۸۵۰۵۷



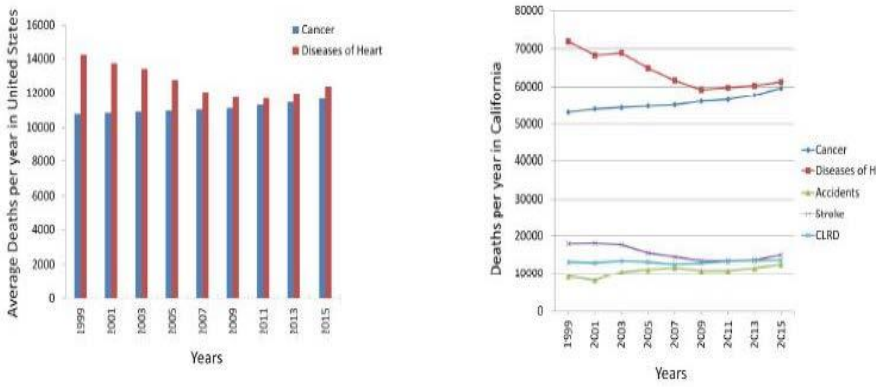
شکل ۸- متوسط افراد شاغل در خانوار به تفکیک شهرستان‌های استان ۰۱

شکل‌های ۷ و ۸ نمونه‌ای از شاخص‌های استخراج‌شده است که به ترتیب درصد خانوارهای برخوردار از اینترنت و متوسط افراد شاغل در خانوار به تفکیک شهرستان‌های استان ۰۱ را نشان می‌دهد. مقایسه و تحلیل نتایج استخراج‌شده شکل‌های ۷ و ۸ حاکی از این است که شهرستان ۱۳ از شاخص‌های اجتماعی اقتصادی پایین‌تری نسبت به بقیه شهرستان‌های دیگر این استان برخوردار است. لازم به یادآوری است که این شاخص‌ها هر ۵ سال یا ۱۰ سال یک‌بار از طریق سرشماری‌ها به‌دست می‌آید و فاقد اطلاعات شهرستانی سالانه است.

همچنین در بخش هزینه‌ای که کاربرد بسیار زیادی در حساب‌های منطقه‌ای دارد، نتایج آن استخراج نمی‌شود. لذا با تولید بخش هزینه‌ای در سطح شهرستان‌ها، اطلاعات ارزشمندی

در اختیار تصمیم‌گیران و برنامه‌ریزان منطقه‌ای قرار می‌گیرد. بدیهی است که همین اطلاعات و اطلاعات هزینه‌ای دیگر، در سطح وسیع‌تر و در سطح شهرستان‌های دیگر استان‌ها و حتی در سطح روستایی نیز می‌تواند استخراج شود و مورد تجزیه و تحلیل قرار گیرد.

استخراج چنین اطلاعاتی نقش اساسی و مهمی در اقتصاد کلان و محاسبات ملی دارند. کشورها هر ساله برای جمع‌آوری داده‌ها و استخراج اینگونه اطلاعات هزینه زیادی می‌پردازند. در ایران نیز هر ساله، داده‌های آماری خانواری با حدود ۴۰۰۰۰ نمونه شهری و روستایی گردآوری شده و نتایج آن در سطح کل کشور استخراج و سپس مورد تجزیه و تحلیل قرار می‌گیرد. با آنکه تولید این نوع داده‌ها از یک سابقه طولانی نزدیک به ۶۰ سال برخوردار است، استخراج سری زمانی اطلاعات در سطح شهرستان‌های کشور می‌تواند کمک زیادی به محققان و تصمیم‌گیران استانی کرده تا با تحلیل بهتری از وضعیت شاخص‌های اجتماعی اقتصادی، برای آینده برنامه‌ریزی کنند.



شکل ۹- متوسط مرگ و میر سالانه در امریکا و علل مرگ و میر در کالیفرنیا [۷]

۵-۱- تحلیل نتایج با داده‌های سرشماری مشابه

به منظور مقایسه با نتایج دیگر محققان که از چارچوب هدوپ برای تحلیل مه‌داده سرشماری و جمعیتی استفاده کرده‌اند می‌توان به پژوهش‌های آگاوان و همکاران [۱]، نیکولیچ و همکاران [۲۳] و بخاری و همکاران [۷] اشاره کرد. برای مقایسه بهتر به پژوهش بخاری و همکاران با جزییات بیشتری اشاره خواهد شد. این محققان برای تحلیل جمعیت‌شناسی از داده‌های سرشماری و پایگاه‌های داده‌ی مختلف جمعیتی که صفات آماری نظیر مدرسه و بیمارستان می‌باشد، استفاده کرده‌اند. سپس تحلیل‌های جمعیت‌شناسی مانند باروری و مرگ و میر و علل آن را در سطح ایالت‌های مختلف آمریکا مورد مطالعه قرار داده‌اند. محققان بدنبال این پاسخ بوده‌اند که کدام ایالت بیشترین مرگ و میر را دارند و علت آن چیست؟ برای پاسخ به این سوال از داده‌های سرشماری جمعیتی اخیر و دیگر پایگاه داده جمعیتی از سال ۱۹۹۹ تا ۲۰۱۵ استفاده کرده‌اند. محققان دریافتند که بیشترین مرگ و میر اتفاق افتاده مربوط به ایالت کالیفرنیا بوده و علل مرگ و میر به‌ترتیب مربوط به سرطان، بیماری قلبی، تصادف و سایر می‌باشد. در نمودار زیر نتایج آن را ملاحظه می‌کنید.

۶- نتیجه‌گیری و پیشنهادات

همانگونه که اشاره شد، هدف از این تحقیق ارائه یک مدل مفهومی از مه‌داده، استفاده از سیستم فایل توزیع‌شده و چارچوب هدوپ در تسریع پردازش آمارهای رسمی و مه‌داده خانواری نظیر هزینه و درآمد خانوار کشور است، بطوریکه با ارائه بهنگام خروجی‌ها بتوان تحلیل بهتری را نسبت به گذشته ارائه داد. استفاده از چارچوب هدوپ این امکان را داده است تا با پردازش موازی و ارتباط با پایگاه‌های داده‌های متفاوت، اطلاعات بسیار مفیدی را برای انواع تحلیل‌ها در سطح کل کشور، استانی و شهرستانی ارائه دهد و می‌تواند کمک شایانی به ارزیابی و تحلیل شاخص‌های اجتماعی و اقتصادی کند. سیستم فایل توزیع‌شده هدوپ این قابلیت را دارد، حجم زیادی از داده‌ها را در سطح استان‌ها و شهرستان‌ها در ماشین‌های متعدد توزیع کرده و به‌طور موازی پردازش و اجرا کند و سیستم

را برای مدیریت بیشتر داده‌ها آماده کند. پیشنهاداتی که از این تحقیق به‌دست می‌آید کاربردی است و استفاده از آن مزایای زیر را به دنبال دارد.

- استفاده از چارچوب هدوپ و سیستم فایل توزیع‌شده هدوپ، این امکان را می‌دهد تا حجم زیادی از داده‌ها را در سطح استان‌ها و شهرستان‌ها در ماشین‌های متعدد توزیع، بطور موازی پردازش و اجرا، و سیستم را برای مدیریت بیشتر داده‌ها آماده کند.
- پاسخ به درخواست کاربران و تصمیم‌گیران برای استفاده از داده‌های سال‌های گذشته و استخراج اطلاعات به تفکیک کدهای شهرستان با استفاده از سیستم فایل توزیع‌شده هدوپ که هم اکنون فقط در سطح کشور یا استان استخراج می‌شود.
- با آنکه تولید این نوع داده‌ها از یک سابقه طولانی نزدیک به ۶۰ سال برخوردار است، هنوز در سطح شهرستان‌های کشور اطلاعاتی از آن استخراج نشده است. با توجه به حجم انبوه داده‌ها در کشور، استخراج اطلاعات در سطح شهرستان‌ها می‌تواند کمک زیادی به محققان و تصمیم‌گیران استانی کرده و با تحلیل بهتری از وضعیت شاخص‌های اجتماعی اقتصادی، برای آینده برنامه‌ریزی کنند.
- برقراری ارتباط اطلاعات استخراج‌شده خانواری در سطح شهرستان با اطلاعات سرشماری جمعیتی کشور و پر کردن خلأ تولید شاخص‌های برخورداری نظیر، اینترنت وشاخص‌هایی که می‌توان سالانه تولید کرد.
- با استفاده از سیستم فایل توزیع‌شده هدوپ می‌توان داده‌های خانواری را سریع‌تر از گذشته که اکنون به‌صورت متمرکز و آفلاین جمع‌آوری می‌شود را ارائه کرد و با ارائه بهنگام خروجی‌ها و اطلاعات بتوان تحلیل‌های سریع‌تر و بهتری را نسبت به گذشته انجام داد.

با توجه به حجم انبوه داده‌های خانواری در کشور، به‌عنوان یک مه‌داده، استخراج اطلاعات در سطح شهرستان‌ها می‌تواند کمک زیادی به محققان و تصمیم‌گیران استانی کرده و با تحلیلی بهتر از وضعیت شاخص‌های اجتماعی اقتصادی، برای آینده برنامه‌ریزی کنند. البته لازم است اشاره شود که از نظر اجرایی ممکن است با موانع و چالش‌هایی روبرو شود که در ذیل آمده است.

- ارتباط و اتصال با پایگاه داده سایر سازمان‌ها با مشکل مواجه شود، لذا به هماهنگی در سطح کلان نیازمند است.
- اختلاف در زیرساخت‌ها، سخت‌افزاری و نرم‌افزاری در سطح استان‌ها می‌تواند اجرا را با مشکل روبه رو کند.بنابراین برای پیاده‌سازی به هماهنگی و آماده‌سازی نیروی انسانی در سطح سازمان‌ها و استان‌های کشور نیاز است.

یکی از این منابع مهم آماری، داده‌های آمارهای هزینه و درآمد خانوار است که کاربرد مهمی در اقتصاد کلان دارد. این منبع مهم آماری دارای تعداد زیادی صفات است که اطلاعات بسیار مفیدی را برای انواع تحلیل‌ها در سطح کل کشور، استانی و شهرستانی ارائه می‌دهد. در این تحقیق از داده‌های هزینه و درآمد کشور برای سال ۱۳۹۹ که در سایت مرکز آمار ایران [۳۱] موجود بود استفاده شد. این مجموعه داده‌های خانواری شامل حدود ۴۰۰۰۰ نمونه شهری و روستایی است که اندازه و حجم فایل به شرح ذیل در جدول ۲ نشان داده شده است. در این جدول، شماره هر job معرف جداول خروجی از داده‌های خام است که شامل ۱۷ job است.

توضیحات

1- Big Data
2- Structured
3- Unstructured
4- Velocity
5- Volume
6- Variety
7- Veracity
8- Value
9- Semi-structured
10- Gross Domestic Product
11- MapReduce
12- Fault Tolerant
13- MasterNode

14- SlaveNode
15- NameNode
16- DataNode
17- Secondary NameNode
18- DataNode Failure
19- Map
20- Reduce
21- Tuple
22- Pandemic
23- Sqoop
24- Wordcount
25- Security

مرجع‌ها

[1] Agawane, D., Pawar, R., Purohit, P., and Agre, G. (2016). Finding Insights & Hadoop Cluster Performance Analysis over Census Dataset Using Big-Data Analytics, *International Journal of Research in Advance Engineering*, **2**, 28-33.

[2] Alange, N., and Mathur, A. (2021). Access efficiency of small sized files in Big data using various techniques on Hadoop distributed file system platform. *International Journal of Computer Science & Network Security*, **21**, 359-364.

[3] Azeroual, O. and Fabre, R. (2021). Processing Big Data with Apache Hadoop in the Current Challenging Era of COVID-19. *Big Data Cognitive Computing*, *5*, 1-18.

[4] Bagui, S., and Dhar, P.C. (2019). Positive and negative association rule mining in Hadoop's MapReduce environment. *Journal of Big Data*, **6**, 1-16.

[5] Bawankule, K., Laxman, K., Dewang, R.K., and Singh, A.K. (2021). Historical data based approach for straggler avoidance in a heterogeneous Hadoop cluster. *Journal of Ambient Intelligence and Humanized Computing*, **12**, 9573-9589.

[6] Bawankule, K.L., Dewang, R.K., and Singh, A.K. (2022). A classification framework for straggler mitigation and management in a

heterogeneous Hadoop cluster: A state-of-art survey. *Journal of King Saud University-Computer and Information Sciences*, 34, 7621-7644.

[7] Bukhari, S.S., Park, J., and Shin, D.R. (2018). Hadoop based demography big data management system. *In 2018 19th IEEE/ACIS International Conference on Software Engineering*, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD) (pp. 93-98). IEEE.

[8] Chawda, M., Rane, R., and Giri, S. (2018). Demographic Progress Analysis of Census Data Using Data Mining. Proceedings of the 2nd International Conference on Inventive Communication and Computational Technologies (ICICCT 2018) IEEE Xplore Compliant, Part Number: CFP18BAC-ART, pp. 1894-1897.

[9] Deshai, N., Sekhar, B.V.D.S., Venkataramana, S., Srinivas, K., and Varma, G.P.S. (2019). Big data Hadoop MapReduce job scheduling: A short survey. In Information Systems Design and Intelligent Applications: Proceedings of Fifth International Conference INDIA 2018 Volume 1 (pp. 349-365). Springer Singapore.

[10] Ergüzen, A., and Ünver, M. (2018). Developing a file system structure to solve healthy big data storage and archiving problems using a distributed file system. *Applied Sciences*, **8**, 1-20.

[11] Gerhardt, B., Griffin, K., and Klemann, R. (2012). Unlocking Value in the Fragmented World of Big Data Analytics, Cisco Internet Business Solutions Group, June 2012

[12] Hashem, I.A.T., Yaqoob, I., Anuar, N.B., Mokhtar, S., Gani, A., and Khan, S.U. (2015). The rise of "big data" on cloud computing: Review and open research issues. *Information systems*, **47,** 98-115.

[13] Hirsch, D.D. (2013). The glass house effect: Big Data, the new oil, and the power of analogy. *Me. L. Rev.*, **66,** 373.

[14] Hsu, J.B., Lin, C.F., Chang, Y.C., and Pan, R.H. (2020). Using independent resource allocation strategies to solve conflicts of Hadoop distributed architecture in virtualization. *Cluster Computing*, **24,** 1583-1603.

[15] Jach, T., Magiera E., and Froelich, W. (2015) Application of HADOOP to Store and Process Big Data Gathered from an Urban Water Distribution System. 13th Computer Control for Water Industry Conference, CCWI. Procedia Engineering, 119, 1375-1380.

[16] Jaiswal, A., Dwivedi, V.K., and Yadav, O.P. (2020). Big Data and its Analyzing Tools: A Perspective. In 2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS) (pp. 560-565). IEEE.

[17] Kaisler, S., Armour, F., Espinosa, J.A., and Money, W. (2013). Big data: Issues and challenges moving forward. In 2013 46th Hawaii international conference on system sciences (pp. 995-1004). IEEE.

[18] Ma, J., Rao, K., Li, R., Yang, Y., Li, W., and Zheng, H. (2022). Improved Hadoop-based cloud for complex model simulation optimization: calibration of SWAT as an example. *Environmental Modelling & Software*, **149**, 105330.

[19] Mahmoud, H., Hegazy, A., and Khafagy, M.H. (2018). An approach for big data security based on Hadoop distributed file system. 2018 International Conference on Innovative Trends in Computer Engineering (ITCE). IEEE.

[20] Maniak, T., Jayne, C., Iqbal, R., and Doctor, F. (2015). Automated intelligent system for sound signalling device quality assurance. *Inf. Sci.*, **294**, 600–611.

[21] Mital, M., Chang, V., Choudhary, P., Papa, A., and Pani, A.K. (2018). Adoption of internet of things in India: a test of competing models using a structured equation modeling approach. *Technol. Forecast. Soc. Change.*, **136**, 339-346.

[22] Newberry, E., and Zhang, B. (2019). On the Power of In-Network Caching in the Hadoop Distributed File System. In 6th ACM Conference on Information-Centric Networking (ICN ’19), September 24–26, Macao, China. ACM, New York, NY, USA, 11 pages. PP. 89-99.

[23] Nikolić, A., Sladić, G., Milosavljević, B., Gostojić, S., and Konjović, Z. (2014). Hadoop and Pig for Internet Census Data Analysis. *ICIST*.

[24] Qureshi, F.F., Iqbal, R., Qasim, M., Doctor, F., and Chang, V. (2017). Integration of OMNI channels and machine learning with smart technologies. *Journal of Ambient Intelligence and Humanized Computing*, 1-17.

[25] Rattanaopas, K., and Kaewkeeree, S. (2017). Improving Hadoop MapReduce performance with data compression: A study using wordcount job. In 2017 14th International Conference on Electrical

Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON) (pp. 564-567). IEEE.

[26] Sharma, M., and Kaur, J. (2019). A Comparative Study of Big Data Processing: Hadoop vs. Spark. 6th International Conference on Computing for Sustainable global Development (Indiacom). IEEE.

[27] Sukeshini, P., Sharma, M. Ved, J.C., and Supriya, N. (2021). Big Data Analytics and Machine Learning Technologies for HPC Application. Springer Nature Singapore Pte Ltd.

[28] Veeraiah, D., and Rao, J.N. (2020). An efficient data duplication system based on hadoop distributed file system. In 2020 International Conference on Inventive Computation Technologies (ICICT) (pp. 197-200). IEEE. Xplore Part Number: CFP20F70-ART; ISBN:978-1-7281-4685-0.

[29] Villars, R.L., Olofson, C.W., and Eastwood, M. (2011). Big data: What it is and why you should care. *White paper, IDC*, **14**, 1-14.

[30] Yang, X., Liu, S., Feng, K., Zhou, S., and Sun, X.H. (2016). Visualization and adaptive subsetting of earth science data in HDFS: A novel data analysis strategy with Hadoop and Spark. In 2016 IEEE International Conferences on Big Data and Cloud Computing (BDCloud), Social Computing and Networking (SocialCom), Sustainable Computing and Communications (SustainCom) (BDCloud-SocialCom-SustainCom) (pp. 89-96). IEEE.

[31] Data Source: www.amar.org.ir, Accessed: 1st March 2021.

رضا علی‌پور
کارشناسی ارشد مهندسی کامپیوتر
تهران، دانشگاه علم و صنعت ایران، دانشکده‌ی مهندسی کامپیوتر، گروه نرم‌افزار.
رایانشانی: rezaalipour955@gmail.com

رضا انتظاری ملکی
دکترای مهندسی کامپیوتر-نرم‌افزار
تهران، دانشگاه علم و صنعت ایران، دانشکده‌ی مهندسی کامپیوتر، گروه نرم‌افزار.
رایانشانی: entezari@iust.ac.ir