معماری پایگاه داده تحلیلی تقریباً بی‌درنگ مبتنی بر هستان‌شناسی

سیدمصطفی شفائی، نگین دانشپور و سیدمجید شفائی

چکیده: پایگاه داده تحلیلی، پشتیبانی از داده‌های خارجی را که به صورت پویا بعد از ساخت و طراحی پایگاه داده تحلیلی مورد نیاز است مهیا نمی‌کند. از این رو تحلیل‌گر برای انجام تحلیل‌های مؤثر خود نیازمند پیداکردن همبستگی میان داده‌های خارجی و پایگاه داده تحلیلی است و در مواقعی نیز نیازمند مقایسه بین هر دو داده پایگاه داده تحلیلی و داده‌های خارجی با یکدیگر می‌باشد. همچنین تحلیل‌گر مجبور است که برای برخی از موقعیت‌های تکراری کارهای گذشته را تکرار کند که این کارها شامل اصطلاحات، ایجاد معیارها و مقایسه می‌باشد. برای فارغ‌شدن از این مسایل در این مقاله تلاش شده که یک معماری تقریباً بی‌درنگ مبتنی بر هستان‌شناسی پیشنهاد شود. علاوه بر این الگوریتمی نیز جهت کاهش زمان پاسخ‌دهی به پرس و جوهای تحلیلی کاربران با استفاده از دیدهای ذخیره‌شده و پردازش موازی پیشنهاد می‌شود. نمونه‌های مطالعاتی به منظور نشان‌دادن نحوه ایجاد همبستگی میان داده‌های خارجی با داده پایگاه داده تحلیلی صورت گرفت و نتایج به دست آمده کشف همبستگی میان داده‌های خارجی و داده پایگاه داده تحلیلی را نشان می‌دهد. در آزمایش‌ها استفاده از دیدهای ذخیره‌شده در دو رویکرد مستقیم و والد در پایگاه داده تحلیلی معماری موجود، باعث کاهش زمان پاسخ‌دهی به پرس و جوهای ترتیبی، مقایسه‌ای و ترکیبی موازی کاربران می‌شود.

کلیدواژه: **پایگاه داده تحلیلی تقریباً بی‌درنگ، هستان‌شناسی، داده خارجی، پرس و جوی مقایسه‌ای، پرس و جوی ترتیبی، دید ذخیره‌شده.**

۱- مقدمه

پایگاه داده تحلیلی[1] از دهه گذشته تاکنون به عنوان یکی از مهم‌ترین ابزارهای تحلیلی مورد استفاده قرار گرفته است. از این رو با بررسی فعالیت‌های تاریخی شرکت و یا تجارت می‌توان از وضعیت‌های مختلفی که شرکت در طی سال‌های گذشته تاکنون سپری کرده است آگاه شد. پایگاه داده تحلیلی معمولاً در ساختار چندبعدی[2] به نام مکعب داده[3] مدل می‌شود. هر بعد[4] معادل یک خصوصیت از مجموعه خصوصیات در شمای داده‌ای است و هر سلول[5] مقدار تعدادی معیار تجمعی[6] از قبیل تعداد یا جمع (محصول فروخته‌شده) را ذخیره می‌کند [۱]. به عنوان نمونه، تعداد فروش کالا در هر شهر و در همه سال‌ها به چه میزان می‌باشد. معماری ارائه‌شده در این مقاله با استفاده از داده‌های خارجی[7] و هستان‌شناسی[8] و با بهره‌بردن از بخش‌های بی‌درنگ[9]، سعی در قدرتمندکردن قواعد تحلیلی پایگاه داده تحلیلی می‌کند. این معماری با یکپارچه کردن موارد فوق، امکان پاسخ‌گویی مناسب به پرس و جوهای ترتیبی و مقایسه‌ای را فراهم می‌کند. مقالات ارائه‌شده در حوزه معماری بی‌درنگ صرفاً به ارائه معماری‌های بی‌درنگ به منظور کاهش زمان پاسخ‌دهی و افزودن قابلیت‌هایی از جمله کاهش تداخل عملیات پرس و جوی تحلیلی برخط و بارگذاری داده پرداخته‌اند. همچنین مقالات ارائه‌شده در حوزه هستان‌شناسی نیز به ارائه راه حل‌هایی به منظور ساخت و طراحی شمای پایگاه داده تحلیلی و یا ساخت یک شمای هستان‌شناسی محض پرداخته‌اند. داده‌های خارجی نیز به عنوان بخشی از داده‌های اصلی در کارهای گذشته در نظر گرفته شده است. همچنین این داده‌های خارجی صرفاً قابل فهم برای ماشین می‌باشند اما هیچ یک از راهکارهای قبلی به ارائه یک معماری واحد که بتواند قابلیت بی‌درنگ و قابل فهم نمودن داده و پرس و جو را برای انسان و ماشین در یک معماری یکپارچه نماید، ارائه نشده است. معماری ارائه‌شده توانایی استفاده از داده‌های بی‌درنگ و داده‌های خارجی را دارد. داده‌های بی‌درنگ، داده‌هایی هستند که هنوز به پایگاه داده تحلیلی بارگذاری نشده‌اند. از آنجایی که تجارت نیاز به داده‌های به روز دارد، این معماری این قابلیت را در خود جای داده است. داده‌های خارجی برای پیداکردن همبستگی[10] میان داده‌های پایگاه داده تحلیلی مورد استفاده قرار می‌گیرند. به عنوان مثال، یک شرکت تمایل به پیداکردن همبستگی میان میزان فروش محصولات در شرایط آب و هوایی خاص برای مشتریان دارد. این همبستگی از طریق ایجاد ارتباط با داده آب و هوا میسر می‌شود. علاوه بر این شرکت تمایل به مقایسه میزان فروش محصولات در شرایط آب و هوایی خاص برای مشتریان پیر و یا جوان نیز دارد. این معماری با فراهم‌کردن قابلیت مقایسه درخواست‌ها با شرایط مختلف از طریق یکپارچه‌سازی و استفاده از لایه‌های مختلف به این نیاز پاسخ می‌دهد. هستان‌شناسی نیز به عنوان یک دانش که از طریق تحلیل‌گران ایجاد و معرفی می‌گردد، سعی بر قابل فهم کردن مفاهیم برای انسان و ماشین و همچنین جلوگیری از تکرار و سهولت کار دارد. خاصیت داده خارجی ایجاب می‌کند که این داده‌ها به صورت جدا از پایگاه داده تحلیلی نگهداری و مدیریت شوند. برخی از دلایلی که نگهداری مجزای داده‌های خارجی را از داده‌های پایگاه داده تحلیلی مشخص می‌کند در موارد زیر شرح داده شده‌اند:

- داده‌های خارجی خاصیتی پویا برای تحلیل ایفا می‌کنند. این داده‌ها پس از ایجاد پایگاه داده تحلیلی در صورت نیاز توسط تصمیم‌گیران و یا تحلیلگران برای تحلیل مورد استفاده قرار می‌گیرند. به عنوان مثال استفاده از داده آب و هوا برای تحلیل در برخی از موارد توسط

این مقاله در تاریخ ۲۵ شهریور ماه ۱۳۹۵ دریافت و در تاریخ ۱۶ فروردین ماه ۱۳۹۶ بازنگری شد.

سیدمصطفی شفائی، دانشکده مهندسی کامپیوتر، دانشگاه تربیت دبیر شهید رجایی، تهران، (email: s.m.shafaei@outlook.com).
نگین دانشپور، دانشکده مهندسی کامپیوتر، دانشگاه تربیت دبیر شهید رجایی، تهران، (email: ndaneshpour@srttu.edu).
سیدمجید شفائی، دانشکده فنی و مهندسی، دانشگاه آزاد واحد تهران مرکزی، تهران، (email: sm.shafaei71@gmail.com).

1. Data Warehouse
2. Multidimensional
3. Data Cube
4. Dimension
5. Cell
6. Aggregate Measure
7. External Data
8. Ontology
9. Real-Time Partitions
10. Correlation

گروه تحلیل‌گران توصیه می‌شود و همیشه جزء لاینفک تحلیل به شمار نمی‌آید.

- اضافه و حذف‌کردن داده‌های خارجی نیازی به تغییر شمای[1] پایگاه داده تحلیلی ندارد.
- اگر تصمیم تحلیل‌گران برای حذف یک یا چند داده خارجی از لیست تحلیل باشد، این تصمیم هیچ آسیبی به داده‌های پایگاه داده تحلیلی نمی‌رساند و به راحتی قابل حذف هستند. از این رو حجم پایگاه داده تحلیلی افزایش پیدا نمی‌کند.
- ذخیره مقدار در جدول حقیقت[2] به ازای هر تراکنش همراه با داده خارجی تا حدودی بی‌معنی می‌شود. به عنوان مثال به ازای فروش یا سفارش هر محصول، نمی‌توان شرایط آب و هوایی را همراه با داده فروش ذخیره کرد.
- با ذخیره‌کردن داده‌های خارجی در هستان‌شناسی، حجم و تکرار کاهش پیدا می‌کند. به عنوان مثال اگر در امروز و در ساعات مشخصی نوع آب و هوا برفی است، کافی است که فقط از کلاس برفی برای ذخیره تمامی خصوصیات آب و هوای برفی استفاده شود. این روش ذخیره‌سازی نیز به خوبی با پرس و جوهای SPARQL هماهنگ است. در صورتی که اگر در پایگاه داده تحلیلی عبارت Snowy ذخیره گردد، نیاز به اضافه‌نمودن این دانش به یک منبع هستان‌شناسی است و سپس برای واکشی مفهوم آن نیاز به اجرای پرس جوی SPARQL است که این کار مستلزم اجرای پرس و جوی SQL برای واکشی مجموع ابعاد دیگر مورد نیاز به همراه واکشی خصوصیت Snowy به وسیله SPARQL است.
- حجم و سرعت دریافت برخی از این داده‌های خارجی بسیار زیاد است. به عنوان مثال داده‌های حسگر، حجیم و پرسرعت است و این زمان استخراج، تغییر و بارگذاری (ETL)[3] و همچنین زمان پاسخ‌گویی به پرس و جوهای پردازش تحلیلی برخط (OLAP)[4] تأثیرگذار است.
- در صورت مختل‌شدن عملیات ETL در منابع داده خارجی و یا در دسترس‌نبودن آنها، پایگاه داده تحلیلی کار خود را به صورت مستقل انجام می‌دهد. همچنین داده‌های خارجی دیگر نیز به طور مستقل عمل خواهند کرد.

ممکن است داده‌های خارجی از طریق منابع دولتی و یا خصوصی از طریق واسطهای وب معنایی[5] و سرویس‌های وب[6] در دسترس باشند و نیازی به انجام عملیات استخراج این نوع داده‌ها نباشد.

در معماری پیشنهادی، پایگاه داده تحلیلی به صورت رابطه‌ای بوده و دارای شمای ستاره‌ای[7] است. این مدل داده‌ای و شما از معروف‌ترین و رایج‌ترین مدل‌های داده‌ای و شما برای ساخت یک پایگاه داده تحلیلی به شمار می‌روند [۱]. برای اضافه‌نمودن دانش نیز معماری از یک هستان‌شناسی برای پایگاه داده تحلیلی استفاده می‌کند. به عنوان مثال، شهرهای پرجمعیت که جمعیت آنها بیشتر از شش میلیون نفر است از طریق هستان‌شناسی داخلی برای بعد منطقه و گروه شهر در پایگاه داده تحلیلی مورد استفاده قرار می‌گیرد. معماری با فراهم‌کردن سه لایه عمودی و مؤلفه‌های داخلی آن، تمامی عملیات و مدیریت درخواست‌های تحلیلی را انجام می‌دهد. این سه لایه عبارتند از لایه موتور معماری، یکپارچه‌ساز و داده. پس از آماده‌شدن نتایج، نتایج به دو بخش داخلی و خارجی[8] تقسیم می‌شوند. نتایج داخلی و خارجی از مقایسه داخلی و خارجی خارجی همبستگی بین داده‌ها به وجود می‌آید. سپس اقدام مناسب که شامل انتخاب و ارسال گزارش به واحدهای مختلف و دیگر عملیات مورد نیاز واحدهای سازمان است، صورت می‌گیرد.

استفاده از دیدهای ذخیره‌شده در پایگاه داده تحلیلی باعث کاهش چشم‌گیر زمان پاسخ‌دهی به پرس و جوهای تحلیلی کاربران می‌شود. از این رو با کاهش زمان پاسخ‌دهی به پرس و جوهای تحلیلی برخط، زمان اجرای پرس و جوها در بخش داده‌های خارجی نیز به طور کلی کاهش پیدا می‌کند زیرا در بیشتر مواقع، پرس و جوهای داده خارجی نیاز به نتایج پرس و جوهای تحلیلی برخط به منظور واکشی داده‌های خارجی بر طبق آن نتایج دارند. بنابراین با کاهش زمان پاسخ‌دهی به پرس و جوهای تحلیلی، نتایج این پرس و جوها برای پرس و جوهای خارجی سریع‌تر مهیا می‌شود. زمانی که پرس و جوی تحلیلی ترتیبی، مقایسه‌ای و یا ترکیبی به سیستم وارد شوند، کاهش زمان اجرای پرس و جوی تحلیلی برخط به منظور کاهش زمان پاسخ‌دهی کلی حایز اهمیت است. زیرا در بیشتر مواقع، داده‌های خارجی نیاز به نتایج به دست آمده از پرس و جوی تحلیلی برخط به منظور اجرای پرس و جو در داده‌های خارجی دارند. بنابراین با کاهش زمان پاسخ‌دهی به پرس و جوهای تحلیلی برخط، زمان کل پاسخ‌دهی به پرس و جوها در معماری کاهش پیدا می‌کند. علاوه بر این در بیشتر موارد، پرس و جوهای تحلیلی بر خط شامل یک یا چندین معیار هستند. محاسبه مقادیر معیارها، زمان پاسخ‌دهی را افزایش می‌دهد. اگر چنانچه پرس جوی تحلیلی برخط در پایگاه داده تحلیلی و پرس و جوی SPARQL در داده‌های خارجی به صورت هم‌زمان رخ دهد، زمان پرس و جوی تحلیلی بر خط به دلیل محاسبه معیارها بیشتر خواهد بود. به این دلیل نتیجه کلی وابسته به پرس و جوی تحلیلی برخط می‌باشد. همچنین پایگاه داده تحلیلی شامل پرس و جوهای تحلیلی برخط عادی نیز می‌باشد که حجم بارکاری قسمت پایگاه داده تحلیلی معماری را افزایش می‌دهد. احتمال استفاده از داده‌های خارجی بدون استفاده از داده پایگاه داده تحلیلی بسیار پایین است زیرا بیشتر تحلیل‌ها به منظور کشف همبستگی میان داده خارجی با داده پایگاه داده تحلیلی صورت می‌گیرد. با افزایش حجم بارکاری قسمت پایگاه داده تحلیلی معماری، نسبت به داده‌های خارجی، نیاز به استفاده از راه حل‌هایی جهت کاهش زمان پاسخ‌دهی در این قسمت از معماری خواهیم داشت. استفاده از دیدهای ذخیره‌شده در رویکردهای مستقیم و والد در این مقاله، باعث کاهش زمان پاسخ‌دهی به پرس و جوهای تحلیلی برخط می‌شود. تأثیر کاهش زمان پاسخ‌دهی به پرس و جوهای تحلیلی برخط، باعث کاهش زمان پاسخ‌دهی کامل پرس و جوهای تحلیلی کاربران که خواهان پیداکردن همبستگی بین داده پایگاه داده تحلیلی با داده‌های خارجی هستند می‌شود.

این مقاله با بیان داده‌ها به صورت معنادار از طریق هستان‌شناسی و اجرای پرس و جو به نحوی معنایی سعی بر ارائه یک رویکرد جدید در حوزه معنادار کردن پایگاه داده تحلیلی و ابزارهای تحلیلی دارد. این مقاله همچنین الگوریتمی با دو رویکرد به منظور کاهش زمان پاسخ‌دهی به این گونه از پرس و جوها را ارائه کرده است. در این مقاله سه نوع پرس و

1. Schema
2. Fact Table
3. Extraction, Transformation, and Loading
4. Online Analytical Processing
5. Semantic Web
6. Web Services
7. Star Schema
8. Internal and External Values

جوی تحلیلی معنایی که برای انسان و ماشین قابل فهم است به منظور آزمایش معماری پیشنهاد شده است. همچنین مدیریت و کاهش زمان پاسخ‌دهی به این نوع از پرس و جوها در معماری پیشنهادی نیز بیان شده است. معماری پیشنهادی با تفکیک نحوه ذخیره‌سازی و نگهداری داده‌های خارجی و داخلی و شرح مزایای این روش، یک معماری جدید در حوزه معناسازی پایگاه داده تحلیلی را ارائه کرده است. همچنین این روش به بالابردن قابلیت تحمل خطا از طریق ذخیره‌سازی داده‌های خارجی و داخلی به صورت مجزا و بسط‌پذیری معماری پایگاه داده تحلیلی کمک کرده است. این مقاله، یک نمونه مطالعاتی را از طریق بیان یک پرس و جوی تحلیلی معنایی به منظور ارزیابی نوآوری ارائه می‌کند. هدف از ارائه این نمونه مطالعاتی، نمایش چگونگی بیان یک پرس و جوی تحلیلی معنایی می‌باشد.

مقاله در بخش‌های پیش رو سازمان‌دهی شده است: بخش دوم، کارهای گذشته و بخش سوم تعاریف پایه الگوریتم را شرح داده است. بخش چهارم به معرفی معماری پیشنهادی می‌پردازد. بخش پنجم الگوریتم کاهش زمان پاسخ‌دهی به پرس و جوهای ترتیبی را تشریح کرده است. بخش ششم نمونه‌های مطالعاتی معماری در بخش‌های طراحی و مقایسه را شامل می‌شود و در بخش هفتم، آزمایش‌های الگوریتم ارائه و مورد بررسی قرار می‌گیرند.

۲- پیشینه تحقیق

در سال ۲۰۱۲ [۲] یک معماری مرجع برای پایگاه داده تحلیلی ارائه کردند که این معماری از چندین لایه تشکیل شده است. اولین لایه، لایه تملک داده است که وظیفه دریافت داده‌ها را از منابع داده‌ای بر عهده دارد. لایه‌های دیگر برای تغییر داده، تعریف قوانین تجاری و در نهایت انتقال داده تمیز داده شده به پایگاه داده‌های تحلیلی خاص است. معماری پیشنهادی با اضافه‌کردن بخش‌های بی‌درنگ قابلیت استفاده از داده‌های اخیر به همراه داده‌های تاریخی و داده‌های خارجی را مهیا کرده است. معماری پیشنهادی این قابلیت را دارد که از هر بخش، داده‌ها را مستقیماً واکشی و در نهایت یکپارچه کند. معماری [۲] فقط اجازه دستیابی به داده‌های تمیز داده شده و داده‌هایی که منطق تجاری روی آنها اعمال شده است را می‌دهد در حالی که معماری پیشنهادی در این مقاله، شامل داده تمیز داده شده و قابل استفاده در هر بخش به منظور تحلیل در بازه‌های زمانی مختلف است.

در سال ۲۰۱۲ [۳] یک معماری مبتنی بر سرویس در بستر وب را ارائه کرده است. این معماری از آن جهت معرفی شده که ممکن است یک سازمان یک زیربنای بزرگ تجهیزات سخت‌افزاری به منظور ساخت و استفاده از پایگاه داده تحلیلی فراهم کرده باشد. از این رو، این مقدار سرمایه و توان پردازشی مورد نیاز سازمان نیست، بنابراین با ساخت یک بستر اینترنتی با استفاده از سرویس‌های وب، می‌توان یک بستر اشتراکی برای ساخت پایگاه داده‌های تحلیلی برای دیگر استفاده‌کنندگان مهیا کرد. یکی از معایب این معماری انتقال حجم انبوه داده است. حجم انبوه داده قابل انتقال به وسیله اینترنت نیست و بایستی از طریق جابه‌جایی فیزیکی به سایت فراهم‌کننده انتقال داده شود. در صورتی که معماری ارائه‌شده می‌تواند از سرویس‌های وب و وب معنایی به منظور اجرای پرس و جوهای تحلیلی استفاده کند. همچنین قابلیت بی‌درنگ معماری اجازه بارگذاری داده را در هر زمان به منابع داده‌ای فراهم می‌کند.

در سال ۲۰۰۹ [۴] یک مدل باینری رابطه‌ای را برای ساخت و طراحی پایگاه داده تحلیلی معرفی کرد. نویسندگان استفاده از این مدل را به جای مدل رابطه‌ای برای طراحی شماهای مختلف پایگاه داده تحلیلی از قبیل ستاره‌ای و دانه برفی پیشنهاد کرده‌اند. مشکلات این روش عبارتند از ۱) رایج‌نبودن مدل باینری رابطه‌ای، ۲) در دسترس‌نبودن نرم‌افزارهای کار با این مدل، ۳) نبود زبان پرس و جوی مختص این مدل، ۴) مشکلات مربوط به ترجمه SQL به مدل رابطه باینری و ۵) عدم وجود سامانه‌های مدیریت پایگاه داده رابطه باینری. معماری پیشنهادی این مقاله، ساختار و شمای پایگاه داده تحلیلی را تغییر نمی‌دهد و همچنین جهت افزایش سرعت پاسخ‌دهی به پرس و جوهای تحلیلی، الگوریتم‌های دید مستقیم و والد را ارائه کرده است.

در سال ۲۰۰۸ [۵] یک معماری تقریباً بی‌درنگ پایگاه داده تحلیلی را ارائه کرد. در این معماری از چندین کش برای نگهداری داده‌های موقت در یک زمان محدود و از سرویس‌های وب برای واکشی داده از منابع داده‌ای استفاده شده است. همچنین از XML نیز به عنوان فرمت انتقال داده در بین کش‌ها استفاده گردیده است. این مقاله رویکردی را به منظور اجرای عملیات OLAP و نحوه اجرای پرس و جو پیشنهاد نکرده است.

در سال ۲۰۰۸ [۶] یک رویکرد جدید در نگهداری داده‌ها در معماری تقریباً بی‌درنگ ارائه کرد. در این معماری، یک بخش بی‌درنگ وجود دارد که داده از منابع داده‌ای ابتدا به این مکان انتقال داده می‌شود. از آنجایی که در طول روز ممکن است برای یک تراکنش چندین عملیات از قبیل حذف و به روز رسانی به وجود آید، آخرین عمل انجام‌شده برای تراکنش به عنوان آخرین داده مانا در پایگاه داده تحلیلی ثبت می‌شود. این عمل فقط بر روی بخش داده‌ای بی‌درنگ که حجم داده به مراتب خیلی کمتر از بخش پایگاه داده تحلیلی است رخ می‌دهد و سپس آخرین حالت تراکنش به بخش پایگاه داده تحلیلی منتقل می‌شود. بنابراین از نگهداری داده‌های تکراری و افزایش حجم داده‌ای جلوگیری به عمل می‌آید. عیب این روش در این است که اگر به روز رسانی‌های یک رکورد جدید دیرتر از زمان بخش بی‌درنگ وارد شوند، دیگر قابل‌شناسایی نیستند زیرا رکورد قبلی وارد پایگاه داده تحلیلی شده است. این رویکرد قابل انجام برای معماری‌های چندبخشی نیست و از این رو معماری پیشنهادی از دو رویکرد دیدهای ذخیره‌شده مستقیم و والد برای استفاده در معماری‌های تک و چندبخشی بهره برده است.

در سال ۲۰۱۴ نویسنده [۷] یک معماری برای به وجود آوردن قابلیت تقریباً بی‌درنگ و روشی برای انجام عملیات OLAP ارائه کرده که این معماری از دو بخش پویا و ایستا تشکیل شده است. بخش ایستا برای نگهداری داده‌های مانا که معرف یک پایگاه داده تحلیلی است استفاده می‌شود. بخش پویا یک بخش بی‌درنگ برای نگهداری موقت داده‌ها است. داده‌ها ابتدا وارد این قسمت می‌شوند و پس از مدتی معین وارد بخش ایستا می‌شوند. اگر پرس و جویی نیاز به تحلیل داده‌های اخیر همراه با داده‌های تاریخی داشته باشد از عملگر UNION ALL پایگاه داده برای ادغام نتایج هر دو بخش ایستا و بی‌درنگ استفاده می‌شود. این معماری قابلیت اضافه‌کردن داده‌های خارجی به منظور تحلیل را ندارد. اگر چنانچه داده خارجی به عنوان داده تحلیل مد نظر قرار گیرد، کلیه شما و ساختار در دو بخش ایستا و پویا دچار تغییر خواهند شد. در معماری پیشنهادشده، داده خارجی در هر زمان قابل اضافه‌شدن به معماری می‌باشد و همچنین داده خارجی قابل فهم برای انسان و ماشین نیز است.

در سال ۲۰۱۳ نویسندگان [۸] یک معماری بی‌درنگ که از یک بخش بی‌درنگ استفاده می‌کند، ارائه کردند. در این رویکرد از سرویس‌های وب به منظور تملک داده استفاده می‌شود. داده‌های اخیر وارد بخش بی‌درنگ می‌شوند و پس از سپری‌شدن مدت زمانی مشخص به عنوان مثال یک

روز، وارد بخش پایگاه داده تحلیلی می‌شود. زمانی که پرس و جوی تحلیلی برخط وارد شود با بررسی محدوده تاریخی مورد نیاز برای واکشی داده، داده از بخش‌های بی‌درنگ و پایگاه داده تحلیلی بازگردانی می‌شود. آنها همچنین یک رویکرد سه‌مرحله‌ای برای تمیزکردن داده‌ها نیز ارائه کرده‌اند. در گام اول برای هر جدول OLTP یک فایل گزارش ساخته می‌شود. گام دوم، آخرین وضعیت داده‌ها را از فایل گزارش مشخص می‌کند. برای مثال، یک داده در ابتدای محدوده روز اضافه و سپس حذف می‌شود و در این لحظه آخرین وضعیت داده حذف است و این گام این داده را حذف می‌کند. در گام سوم، تجمع بی‌درنگ در هنگام تقاضا محاسبه می‌شود. این معماری قابلیت استفاده در معماری‌های بی‌درنگ چندبخشی را به دلیل استفاده از گام تمیزکاری داده ندارد.

مقاله [۹] در سال ۲۰۱۱ به ارائه یک رویکرد جدید در حوزه معماری تقریباً بی‌درنگ به منظور کاهش تداخل بین عملیات استخراج، تغییرات و بارگذاری و عملیات OLAP پرداخته است. بر حسب این رویکرد، چندین بخش بی‌درنگ و بخش توقفگاه[1] به همراه یک بخش پایگاه داده تحلیلی به معماری اضافه شده است. هر بخش بی‌درنگ داده‌ها را برای مدت معینی نگهداری می‌کند و پس از اتمام این زمان، داده‌ها به بخش بعدی منتقل می‌شوند. در این مقاله رویکرد خاصی برای انجام عملیات OLAP ارائه نشده است. این مقاله روش صریحی را برای پاسخ‌دهی به پرس و جوهای تحلیلی ارائه نکرده است. در صورتی که معماری پیشنهادی، نحوه پاسخ‌دهی به پرس و جوهای تحلیلی را به همراه استفاده از الگوریتم دیدهای ذخیره‌شده مستقیم و والد به منظور کاهش زمان پاسخ‌دهی پیشنهاد کرده است.

مقاله [۱۰] نیز رویکردی برای مقابله با تداخل بین عملیات استخراج، تغییرات و بارگذاری، و عملیات OLAP ارائه کرده است. آنها دو روش ETL را برای رویکرد خود ارائه کرده‌اند. یک ETL برای حالت بی‌درنگ معرفی شده و ETL دیگر برای حالت تاریخی نیز مشخص گردیده است. مقالات [۱۱] و [۱۲] نیز کارهایی نظیر این مقاله انجام داده‌اند. معماری پیشنهادی با تفکیک داده‌های خارجی و داخلی به بخش‌هایی مجزا به بهبود تداخل عملیات ETL و پرس و جوهای تحلیلی اقدام کرده است.

نویسندگان در [۱۳] یک روش برای نگهداری پویای دیدهای ذخیره‌شده ارائه کرده‌اند. این روش زمان اجرای پرس و جو، فضای ذخیره‌سازی، تناوب دسترسی و بارکاری را به عنوان معیارهایی در روش پیشنهادیشان مورد ملاحظه قرار داده‌اند. این معماری رویکردی را برای نگهداری دیدهای ذخیره‌شده در معماری پایگاه داده تحلیلی بی‌درنگ- بخش‌بندی شده- ارائه نکرده است. معماری پیشنهادی الگوریتمی را برای نگهداری دیدهای ذخیره‌شده در معماری پایگاه داده تحلیلی تقریباً بی‌درنگ- بخش‌بندی شده- در مقایسه با این معماری پیشنهاد داده است.

مقاله [۱۴] رویکردی را برای تقسیم‌بندی داده ورودی به پایگاه داده تحلیلی اتخاذ کرده‌اند. بر این اساس اگر داده ورودی ثابت باشد به عنوان مثال جدول محصولات که تغییرات به مراتب کمتری نسبت به جدول سفارش‌ها دارد کش می‌شود که موجب افزایش سرعت می‌گردد. مقالاتی در این راستا که قابلیت بی‌درنگ را مهیا کرده باشند در [۱۵] تا [۱۷] ارائه شده‌اند.

مقالات [۱۸] و [۱۹] عواملی از قبیل بارکاری پرس و جوی ورودی، عملیات ETL مانند نوع استراتژی بارگیری داده، حجم داده و اندیس‌گذاری را مورد بررسی و تحلیل قرار داده‌اند.

Tjoav و Nguyen در سال ۲۰۰۳، راه حلی را برای داده‌های جریانی به منظور نگهداری در پایگاه داده تحلیلی پیشنهاد کرده‌اند. راه حل‌های مبتنی بر جریان، نتایج را به صورت تقریبی محاسبه می‌کنند. فراهم‌کردن نتایج دقیق، هزینه‌ای به مراتب بیشتر از نتایج تقریبی دارد [۲۰]. مقالات [۲۱] و [۲۲] به ارائه معماری‌های جریانی پرداخته‌اند. تمامی راه حل‌های مبتنی بر جریان قابل تطبیق با معماری‌های تقریباً بی‌درنگ نیستند. معماری پیشنهادی برای معماری‌های تقریباً بی‌درنگ مناسب است زیرا معماری‌های مبتنی بر جریان، نتایج و داده‌های ورودی را به صورت تقریبی محاسبه می‌کنند که باعث از بین رفتن مقداری از داده‌های ورودی می‌شود در صورتی که برای داده‌های تراکنشی چنین رفتاری نامناسب است.

مرجع [۲۳] در سال ۲۰۰۷ میان‌افزار سرویس تجاری مبتنی بر معماری سرویس‌گرا را معرفی کرد. این معماری، لایه مدیریت که ارتباط کاربران با لایه‌های انتهایی و جلویی SOA اتصال داده را برقرار کرده است. لایه مدیریت شامل یک پایگاه داده تحلیلی تقریباً بی‌درنگ، مسیریابی، امنیت، پردازش پیام و توانایی مدل‌سازی است. این معماری توانایی استفاده‌شدن در بستر وب معنایی را ندارد. قابلیت‌های وب معنایی تنها محدود به استفاده از XML و سرویس‌های وب نیست بلکه استفاده از هستان‌شناسی جزء اصلی ساختار وب معنایی است. در معماری پیشنهادی استفاده از قابلیت هستان‌شناسی امکان استفاده از معماری پیشنهادی را در بستر وب معنایی فراهم می‌کند.

در سال ۲۰۱۱ نویسندگان [۲۴] یک ابزار برای مقایسه تحلیل داده به نام Semantic Cockpit ارائه کرده‌اند. این ابزار مبتنی بر هستان‌شناسی است. آنها با استفاده و توسعه فناوری‌های معنایی از قبیل هستان‌شناسی و قوانین تجاری، مقایسه نتایج تحلیل داده را پیشنهاد نموده‌اند. این ابزار دانش‌های مختلف از قبیل دانش داخلی و خارجی سازمان را استدلال می‌کند. همچنین معناهایی از معیار، امتیازدهی و دانش قبلی به دست آمده را برای راهنمایی و کمک به کاربران مورد استفاده قرار می‌دهد. نویسندگان [۲۵] در مورد چگونگی استفاده از هستان‌شناسی به منظور استنباط و خلاصه‌سازی در عملیات پردازش تحلیلی برخط بحث کرده‌اند. این معماری فقط قابلیت تحلیل داده‌های داخلی به کمک هستان‌شناسی را مهیا کرده است. معماری پیشنهادی علاوه بر استفاده از هستان‌شناسی به منظور تحلیل داده‌های داخلی از هستان‌شناسی به منظور ذخیره‌سازی داده، انجام پرس و جو و قابل فهم نمودن داده‌های خارجی بهره برده است.

در سال ۲۰۱۴ نویسندگان [۲۶] یک رویکرد هوشمند مبتنی بر هستان‌شناسی را برای مقایسه نتایج تحلیل داده که در یک پروژه به نامه Semantic Cockpit توسعه داده می‌شود ارائه کرده‌اند. کار اصلی آنها در این مقاله، رسم یک بصیرت از گراف تحلیلی و قوانین راهنمای مرتبط به منظور تحلیل می‌باشد. آنها ابتدا به بررسی یک ابزار برای مقایسه نتایج تحلیل داده به نام Semantic Cockpit پرداخته‌اند. سپس با معرفی مورد مطالعاتی درمان دارویی برای شرکت‌های بیمه، کار خود را شرح داده‌اند. آنها با ساخت و طراحی گراف تحلیل هوش تجاری مدل ارتباطی را ایجاد می‌کنند و سپس این گراف به منظور تحلیل مورد بررسی قرار می‌گیرد. این معماری نیز فقط به تحلیل داده‌های داخلی و ایستا پرداخته است. علاوه بر این معماری رویکردی را برای کاهش زمان پاسخ‌دهی پرس و جوها پیشنهاد نکرده است. معماری پیشنهادی مقاله، علاوه بر استفاده از داده‌های خارجی و داخلی به منظور کشف همبستگی به ارائه الگوریتم‌هایی به منظور کاهش زمان پاسخ‌دهی پرس و جوها پرداخته است.

مقالات [۲۷] تا [۳۰] طراحی پایگاه داده تحلیلی مبتنی بر هستان‌شناسی را به منظور خودکارسازی عملیات ساخت پایگاه داده

1. Stage

تحلیلی و ETL انجام داده‌اند. این عملیات می‌تواند شامل ساخت خودکار شمای پایگاه داده تحلیلی و همچنین خودکارسازی نحوه عملیات ETL باشد و به این منظور از هستان‌شناسی برای انجام این عملیات استفاده می‌کنند.

نویسندگان [۳۱] یک مدل رسمی برای داده‌های ترتیبی ارائه کرده‌اند که امکان ساخت مجموعه‌ای از رویدادها و سازمان‌دهی آنها را مانند پرس و جوهای پردازش تحلیلی برخط فراهم می‌کند. برای دست‌کاری و ذخیره داده‌های به فرم ترتیبی، رویکردهای استفاده از مدل رابطه‌ای یا الحاقات آن استفاده می‌شود. به منظور پشتیبانی کامل از این نوع داده، نویسندگان یک مدل رسمی را معرفی کرده‌اند. این مقاله فقط یک مدل ترتیبی برای اجرای پرس و جو در یک شمای مشخص را ارائه کرده است در صورتی که معماری پیشنهادی، علاوه بر ارائه یک مدل به منظور اجرای پرس و جوهای ترتیبی، الگوریتمی به منظور کاهش زمان پاسخ‌دهی به این‌گونه از پرس و جوها را نیز ارائه کرده است. همچنین مقاله پیشنهادی مدل اجرای ترتیبی پرس و جوهای ترکیبی که شامل استفاده از داده‌های داخلی و خارجی است را نیز پیشنهاد کرده است. علاوه بر این، مقاله ارائه‌شده به ارائه دو نوع پرس و جو به نام‌های پرس و جوهای تحلیلی مقایسه‌ای و ترکیبی نیز پرداخته و برای هر یک از این پرس و جوها نیز الگوریتم کاهش زمان پاسخ‌دهی نیز پیشنهاد شده است.

نویسندگان [۳۲] یک رویکرد جدید برای تولید پرس و جوهای مشخص از زیرمکعب‌ها[1] ارائه کرده‌اند. زیرمکعب داده‌ای است که از الگوی پرس و جوی ترتیبی داده‌شده تشکیل گردیده است. این زیرمکعب‌ها در ابزارهای پردازش تحلیلی برخط نیز قابل استفاده هستند و بنابراین کاربران قادر خواهند بود که تمامی عملیات از قبیل برش عمقی[2]، برش جمعی[3] و برش ورقه‌ای[4] را استفاده کنند. این مقاله نیز با مبنا قرار دادن برخی از فناوری‌های معنایی از قبیل هستان‌شناسی درصدد ساخت یک معماری جدید پایگاه داده تحلیلی تقریباً بی‌درنگ مبتنی بر هستان‌شناسی است. معماری پیشنهادی از معیار و امتیازدهی و دانش قبلی به دست آمده نیز استفاده می‌کند.

علاوه بر این روشی برای کشف همبستگی میان داده‌های خارجی با داده پایگاه داده تحلیلی در بخش کارهای پیشین، ارائه نشده است. همچنین یک معماری جامع که بتواند کارهایی از قبیل تسهیل کار کاربر در محیط‌های عملیاتی، قابلیت استفاده از داده بی‌درنگ، کاهش زمان پاسخ‌دهی به پرس و جوهای تحلیلی و قابلیت افزودن داده‌های خارجی به صورت پویا به منظور کشف همبستگی میان این داده‌ها را در یک معماری مشخص کند، ارائه نشده و از این رو در مقاله پیش رو از هستان‌شناسی به منظور افزایش سطح کیفی استفاده از معماری در محیط‌های عملیاتی، ذخیره‌سازی و بیان داده‌های خارجی بهره برده شده است. همچنین از دیدهای ذخیره‌شده به منظور کاهش زمان پاسخ‌دهی به پرس و جوهای تحلیلی برخط کاربران که تأثیر آن کاهش زمان پاسخ‌دهی در معماری است استفاده شده است.

۳- تعاریف پایه

پرس و جوهای تحلیلی ترتیبی، مقایسه‌ای و ترکیبی شامل چندین پرس و جو به صورت تو در تو هستند که با یکدیگر ارتباط دارند. این پرس و جوها می‌توانند شامل چندین پرس و جوی تو در تو باشند. پرس و جوهای ترتیبی شامل چندین ترتیب هستند. ترتیب‌ها در برخی از خصیصه‌ها و معیارها با یکدیگر مشترک هستند. اگر چنانچه چندین پرس و جو تحلیلی از انواع ترتیبی، مقایسه‌ای و ترکیبی به طور هم‌زمان وارد معماری شوند با استفاده از دیدهای ذخیره‌شده و رویکرد پردازش موازی ارائه‌شده، می‌توان زمان پاسخ‌دهی را کاهش داد. به این منظور در این مقاله الگوریتمی ارائه می‌شود. برای بیان الگوریتم از چندین تعریف و قواعد نگارشی در این مقاله استفاده می‌شود که این تعاریف و قواعد در ادامه معرفی می‌گردند.

ترتیب: یک ترتیب پرس و جویی شامل چندین معیار، ابعاد و شرایط است. مکان قرارگیری ترتیب‌ها بر حسب نیاز تحلیلی در یک $SequenceQuery$ مرتب می‌شوند. ترتیب $S_i \in S$ است که S یک مجموعه از ترتیب می‌باشد و تابع $Sequence(D,M,W)$ ترتیب را ایجاد می‌کند. پارامترهای این تابع به شرح زیر است:

D: مجموعه ابعاد را مشخص می‌کند.

M: مجموعه معیارهای موجود در ترتیب را مشخص می‌کند که برخی از ترتیب‌ها فاقد مجموعه معیار هستند.

W: مجموعه شرایط را برای ترتیب مشخص می‌کند. این پارامتر یک فیلتر برای واکشی داده‌ها است.

پرس و جوی ترتیبی: هر پرس و جوی ترتیبی $(SequenceQuery)$ شامل ترتیبی از پرس و جوهای $S_1, S_2, \ldots, S_n$ است که با شرط پرس جوی $SequenceQuery$ به یکدیگر متصل می‌شوند. یک پرس و جوی ترتیبی شامل چندین ترتیب مرتب است که هر ترتیب از یک سری خصیصه (ابعاد) و معیار تشکیل شده است. انواع پرس و جوهای ترتیبی عبارتند از پرس و جوهای تحلیلی ترتیبی، مقایسه‌ای و ترکیبی. چندین پرس و جوی ترتیبی به طور هم‌زمان برای اجرا وارد معماری می‌شوند.

رویداد: رویداد e_i یک مجموعه مقدار[5] خصیصه $v_{i1}, v_{i2}, \ldots, v_{in} \in A_j$ در n سطر است. به طور رسمی $e_i \in E$ است که E مجموعه رویدادها است. رویدادهای متعلق به یک مجموعه ترتیب در یک یا چندین مقدار خصیصه مشترک هستند که این خصیصه‌های مشترک برای ایجاد ارتباط میان ترتیب‌ها مورد استفاده قرار می‌گیرند. رویدادهای متعلق به یک پرس و جوی ترتیبی شامل مقادیر به دست آمده از مجموعه ترتیب‌های داخل آن است.

به عنوان مثال، رویدادهای زیربخشی از مجموع فروش کالای LG CASE را به ازای همه مشتریان و تمامی سال‌ها نشان می‌دهد

$$e_1 = \{11, Customer\#0005768\text{۳}, LG-CASE, 19931002\}$$
$$e_2 = \{4, Customer\#0003453\text{۲۵}, LG-CASE, 19960210\}$$
$$\ldots$$

خصیصه: هر بعد $a_{i1}, a_{i2}, \ldots, a_{in} \in A_j$ با یک یا مجموعه‌ای از خصوصیات در شما متناظر است. به عنوان مثال، مجموعه خصیصه‌ها در A نشان داده شده است

$$A = \{c_name, c_address, c_city, s_name, s_address, p_name, p_category, d_year, d_month, d_day\}$$

سلسله‌مراتب خصیصه: خصیصه‌ها می‌توانند شامل سلسله‌مراتب باشند. مجموعه $L = \{L_1, L_2, \ldots, L_n\}$ سطح‌های خصیصه‌های موجود در رویداد

1. Query Specific Subcubes
2. Drill-Down
3. Roll Up
4. Slice
5. Value

است. سلسله‌مراتب یک خصیصه شامل ترتیب $l_{Ai} \rhd_{A_i}$ است. به عنوان مثال، سلسله‌مراتب بعد تاریخ به صورت زیر است

$$\{day, month, quarter, year\} \rhd_{Date} \{day \rightarrow month \rightarrow quarter \rightarrow year\}$$

جدول حقیقت: جدول حقیقت F شامل معیارها و کلیدهای خارجی ابعاد است. نمونه‌ای از جدول حقیقت در زیر نشان داده شده است

$$F = \{lo_orderkey, lo_linenumber, lo_custkey, lo_partkey, lo_\sup pkey, lo_orderkey, lo_quantity, lo_discount, lo_revenue, lo_tax\}$$

معیار: پرس و جوهای تحلیلی ترتیبی، مقایسه‌ای و ترکیبی شامل یک یا چندین معیار $m_i \in M$ هستند که M مجموعه‌ای از معیارهاست. در پرس و جوهای تحلیلی ترتیبی، حداقل یک معیار ساخته می‌شود. در ترتیب اول یک معیار ایجاد می‌شود و در ترتیب‌های بعدی ممکن است معیار ایجاد نشود. در پرس و جوهای تحلیلی مقایسه‌ای و ترکیبی، نیز حداقل یک معیار وجود دارد. معیار موجود در پرس و جوهای تحلیلی مقایسه‌ای و ترکیبی به صورت یک تابع خارجی معرفی می‌شود.

برای ساخت یا معرفی یک معیار از تابع $Measure$ استفاده می‌شود. پارامترهای ورودی تابع به دو صورت برای پرس و جوهای تحلیلی ترتیبی، مقایسه‌ای و ترکیبی معرفی می‌شود: $Measure(Name, f_i)$ و $Measure(Name, F(M))$. پارامترهای ورودی به شرح زیر است:

$Name$: نام معیار را برای ساخت مشخص می‌کند. نام معیار مشخص‌کننده نیاز تحلیلی استفاده‌کننده است.

f_i: خصیصه iام در جدول حقیقت f که باید تابع تجمعی روی آن محاسبه شود را مشخص می‌کند.

$F(M)$: تابعی را برای خروجی معیار مشخص می‌کند. این تابع محاسبات مورد نیاز را بر روی معیارهای $[m_1, \ldots, m_n]$ انجام می‌دهد. این پارامتر برای پرس و جوی تحلیلی مقایسه‌ای و ترکیبی صورت می‌گیرد.

بعد: نماد d_{ji} به خصیصه iام از مجموعه سلسله‌مراتب بعد d_j و D نیز به مجموعه ابعاد اشاره می‌کند، بنابراین $d_j \in D$ است. تابع $Dimension(Name, f_{dji}, L_{dji})$ یک بعد را نشان می‌دهد.

$Name$: نام بعد که سطح مورد نظر در آن قرار دارد را مشخص می‌کند.

a_{di}: خصیصه مورد نظر را برای بعد مشخص می‌کند.

L_{di}: سطح سلسله‌مراتب بعد را نشان می‌دهد.

به عنوان مثال، $Dimension(Date, a_4, quarter \rightarrow year)$ سطح فصل در بعد تاریخ را نشان می‌دهد.

دید ذخیره‌شده: دید ذخیره‌شده $v_i \in V$ نتایج پرس و جوی ذخیره‌شده است که V مجموعه دیدهای ذخیره‌شده است.

موجودیت: در این مقاله برای واکشی کلاس یا موجودیت از تابع $Entity(repository, class_{name})$ و $Entity(repository, \{condition\})$ استفاده گردیده است که در این توابع، $repository$ مشخص‌کننده نام منبع هستان‌شناسی است و $class_{name}$ نام موجودیت که به صورت یک خصیصه داده‌ای[1] است را مشخص می‌کند. به عنوان مثال، $k_1 \leftarrow Entity(SocialOnto, HighAdvertisment)$ تعداد پست‌های به اشتراک گذاشته شده برای تبلیغات را از موجودیت $HighAdvertisment$ واکشی می‌کند. در صورت دوم تابع، $\{condition\}$ می‌تواند شامل چندین عبارت شرطی به همراه خصیصه شیء[2]، خصیصه داده‌ای و توابع باشد. به عنوان مثال

$$c_2 \leftarrow Entity(SocialOnto, \{\uparrow hasPost\{\uparrow subject = 'toothpast' or \uparrow tag = 'toothpast' and count() \geq k_1\}\})$$

c_2 خصیصه شیء $hasPost$ را با شرایط لازم برای خصوصیات موجودیت $Post$ از هستان‌شناسی در شبکه اجتماعی نشان می‌دهد.

زیرکلاس‌ها نیز به صورت نماد $\downarrow$ نشان داده می‌شوند. به عنوان مثال $c_3 = Entity(WeatherOnto, city \downarrow region \downarrow nation)$ زیرکلاس‌های کلاس $nation$ را نشان می‌دهد. کلاس $region$ زیرکلاس کلاس $nation$ و کلاس $city$ زیرکلاس کلاس $region$ است.

خصیصه موجودیت: در این مقاله از نماد $\uparrow$ به منظور نمایش خصیصه موجودیت استفاده گردیده است. به عنوان مثال، خصوصیت $\{\uparrow hasWeatherState = 'SnowyWeatherState'\}$، خصیصه شیء $hasWeatherState$ را نشان می‌دهد.

۴- معرفی معماری پیشنهادی

در این بخش، معماری پیشنهادی شرح داده می‌شود. به این منظور ابتدا فرایند ETL معماری، معماری کلی پیشنهادی، فرایند اجرای پرس و جو در معماری و در آخر نیز ۳ نوع پرس و جو تحلیلی به همراه الگوریتم دید ذخیره‌شده ارائه می‌شود.

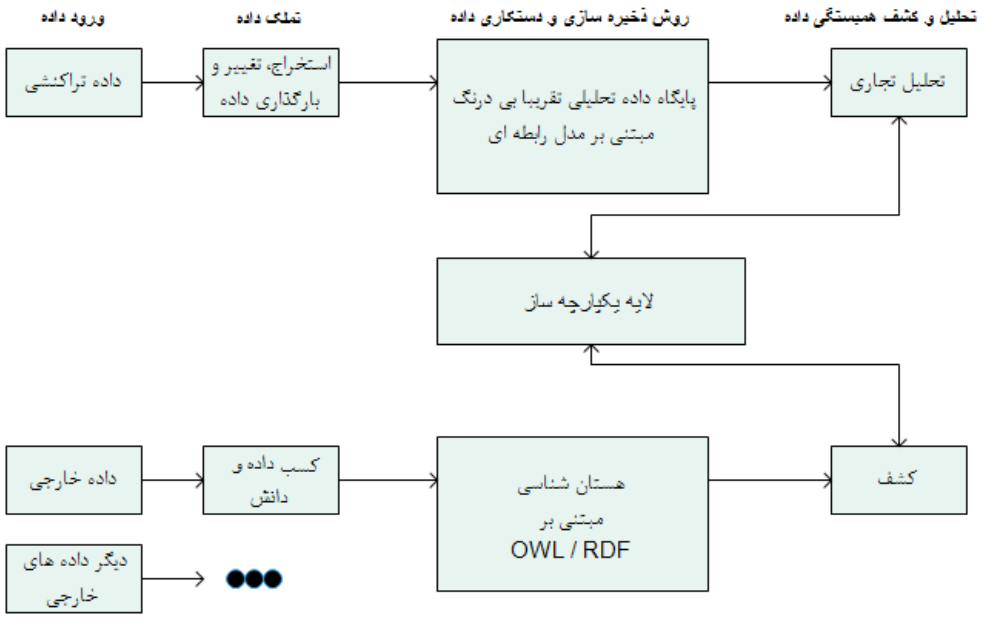
۴-۱ فرایند ETL معماری

فرایند ETL معماری در شکل ۱ نشان داده شده است. این فرایند برای داده‌های تراکنشی پایگاه داده تحلیلی و داده‌های خارجی به صورت مجزا عمل می‌کند. برای داده‌های تراکنشی که پایگاه داده تحلیلی به منظور آن ساخته شده است از فرایند استخراج، تغییر و بارگذاری استفاده می‌شود. پایگاه داده تحلیلی از یک یا چندین بخش بی‌درنگ برای نگهداری داده‌ها در درجه‌های متفاوت تازگی و در دسترس‌بودن این داده‌ها استفاده می‌کند. برای استخراج داده‌های هر منبع خارجی به عنوان مثال آب و هوا و شبکه اجتماعی، از فرایند کسب داده و دانش از یک پایگاه داده مجزا استفاده می‌شود. به عنوان مثال شکل ۱ به صورت سه‌لایه نشان داده شده است. لایه اول داده‌های داخلی است که در پایگاه داده تحلیلی ذخیره می‌شوند. لایه دوم یک لایه مدیریتی از قبیل ساخت دید ذخیره‌شده، گزارش‌دهی و دیگر موارد مدیریتی است. لایه سوم، داده خارجی است که به فرم شمای هستان‌شناسی و مستقل از لایه اول و دیگر لایه‌های داده‌ای ذخیره می‌گردد.

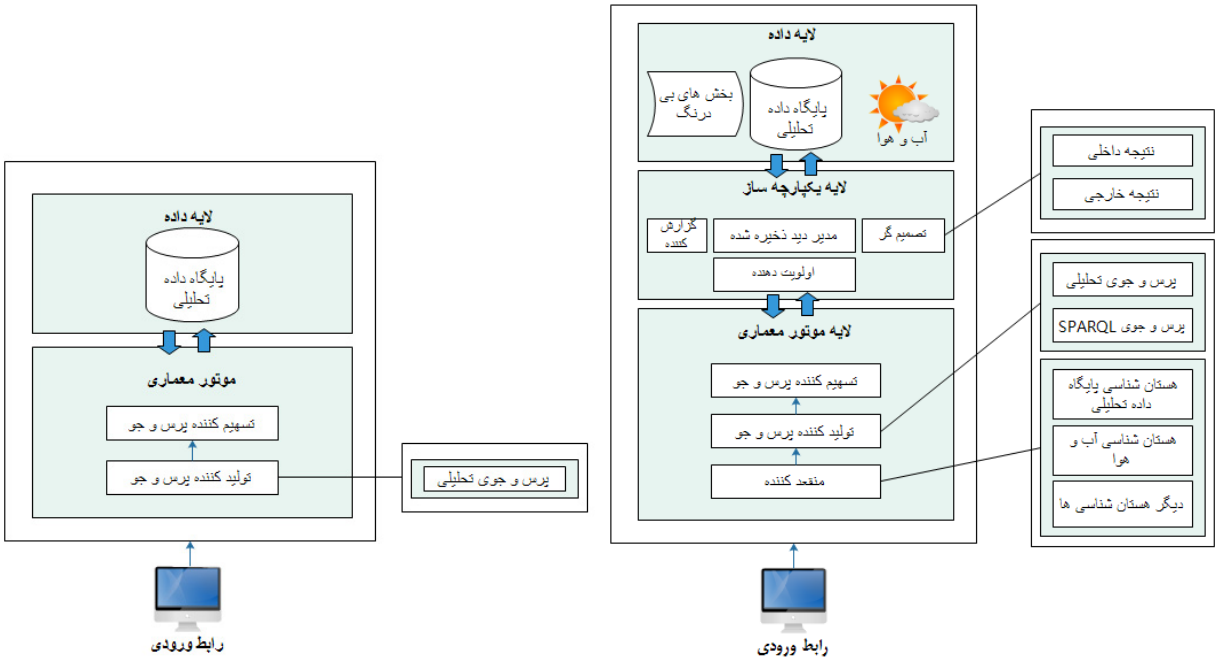
در این معماری منابع داده خارجی پس از ساخت پایگاه داده تحلیلی در صورت نیاز به معماری اضافه و یا از آن حذف می‌گردند. به همین دلیل برای جلوگیری از تداخل عملیات ETL و کسب داده از داده‌های خارجی، هر بخش به صورت مجزا عمل می‌کند. در صورت در دسترس نبودن هر یک از منابع، پایگاه داده تحلیلی می‌تواند مستقلاً به کار خود ادامه دهد. داده‌های خارجی با ساختار هستان‌شناسی که پیشتر برای آن شمای مناسب ایجاد شده است، ذخیره می‌گردند. برای استفاده از داده پایگاه داده تحلیلی از یک پایگاه داده رابطه‌ای استفاده می‌شود و شمای آن ستاره‌ای در نظر گرفته می‌شود. زبان‌های پرس و جو برای پایگاه داده تحلیلی و

1. Data Type Property

2. Object Property



شکل ۱: فرایند ETL معماری.



شکل ۲: (الف) معماری پایگاه داده تحلیلی تقریباً بی‌درنگ مبتنی بر هستان‌شناسی و (ب) معماری یک پایگاه داده تحلیلی سنتی.

داده‌های خارجی به ترتیب SQL و SPARQL است.

سازمان یا تجارت بسته به شرایط در دسترس بودن داده و نحوه دسترسی به آن ممکن است روش‌های گوناگونی را در نظر بگیرند. برای مثال اگر داده آب و هوا از طریق وب معنایی و سرویس‌های وب توسط سازمان آب و هوایی کشور در دسترس باشد، شرکت از کسب این داده و انتقال آن به سیستم ذخیره‌سازی خود صرف نظر کرده و تنها از واسطهایی که توسط سازمان آب و هوای کشور در اختیار است برای انجام پرس و جوهای خود استفاده می‌کند. اگر شرکت به عنوان مثال داده‌های حسگر رفت و آمد به فروشگاه‌ها را می‌خواهد مورد تحلیل قرار دهد تا همبستگی میان داده‌های پایگاه داده تحلیلی و یا دیگر داده‌های خارجی با یکدیگر را کشف کند، داده‌ها را از درون سرورهای خود کسب می‌کند.

۲-۴ معماری کلی

شکل سمت چپ ۲، معماری یک پایگاه داده تحلیلی را که به عنوان یک سیستم تحلیلگر مورد استفاده قرار می‌گیرد نشان می‌دهد در حالی که شکل سمت راست ۲، معماری پایگاه داده تحلیلی تقریباً بی‌درنگ مبتنی بر هستان‌شناسی را نشان می‌دهد. به طورمعمول، تحلیلگر، مدیر و یا کاربر پرس و جوی تحلیلی خود را از طریق یک واسط به پایگاه داده تحلیلی ارسال می‌کنند. این پرس و جوهای تحلیلی به منظور پاسخ‌دهی به سؤالات تحلیلی کاربران از وضعیت‌های مختلف شامل میزان فروش در درجه‌های مختلف دانه‌ای بودن ارائه می‌شوند. سپس مدیر و یا کاربر از طریق مقادیر به دست آمده، تصمیم مناسب را اتخاذ می‌کنند. معماری پایگاه داده تحلیلی تقریباً بی‌درنگ مبتنی بر هستان‌شناسی قابلیت فراهم‌نمودن داده‌های اخیر و تاریخی را در پاسخ‌دهی به پرس و جوهای

کاربر مهیا می‌کند. علاوه بر این، معماری از داده‌های خارجی (به عنوان مثال داده‌های آب و هوا و شبکه‌های اجتماعی) به منظور تحلیل نیز استفاده می‌کند. معماری پایگاه داده تحلیلی تقریباً بی‌درنگ مبتنی بر هستان‌شناسی به خودی خود هوشمند نیست و باید تحلیلگر یا مدیر پایگاه داده، دانش مورد نیاز برای معماری را مهیا کند. دانش در دو حوزه مورد نیاز است: ۱) حوزه داخلی و ۲) حوزه خارجی. در حوزه داخلی دانش مورد نیاز برای خود پایگاه داده تحلیلی از طریق هستان‌شناسی داخلی معرفی می‌گردد که در شکل سمت راست ۲ با عنوان هستان‌شناسی پایگاه داده تحلیلی قید شده است. دانش حوزه خارجی به صورت داخلی در داده‌های خارجی که به ساختار هستان‌شناسی هستند و در شکل با عنوان هستان‌شناسی داده آب و هوا و دیگر هستان‌شناسی‌های خارجی نشان داده شده است، قرار دارد. افزودن دانش به صورت دستی به منظور مهیاکردن اطلاعات قابل فهم برای انسان و ماشین صورت می‌پذیرد و همچنین این دانش برای سهولت کاربر و جلوگیری از تکرار مورد استفاده قرار می‌گیرد. پرس و جوهایی که کاربران اعم از مدیران و تحلیلگران بسته به شرایط مد نظرشان آماده می‌کنند، شامل اصطلاحات (مانند "شهرهای بزرگ") و پرس و جوهای شبیه به یکدیگر است. با نگهداری مفهوم هر اصطلاح و شمای پرس و جوهای پرتکرار از طریق هستان‌شناسی، این روش می‌تواند به عنوان یک دانش قابل فهم برای سیستم و انسان مورد استفاده قرار گیرد.

شکل سمت راست با افزودن قابلیت بی‌درنگ، داده‌های خارجی در لایه داده و دانش مورد نیاز برای آنها از طریق هستان‌شناسی در لایه موتور معماری یک معماری تقریباً هوشمند و بی‌درنگ ایجاد می‌کند. این معماری با بهره‌گرفتن از سه لایه عمودی که در شکل به نام‌های لایه موتور معماری، یکپارچه ساز و داده مشخص شده است، قابلیت‌های افزوده‌شده را مدیریت می‌کند. لایه اول، موتور معماری را تشکیل می‌دهد که این موتور از سه مؤلفه تشکیل شده است. مؤلفه منعقدکننده مسئول نگاشت اصطلاحات و دانش مورد نیاز برای لایه بعدی است. این لایه با استفاده از دانش‌هایی که از طریق ساختار هستان‌شناسی در دسترس است، اقدام به جستجو و واکشی اطلاعات مورد نیاز می‌کند. به عنوان مثال زمانی که اصطلاح large-city در درخواست کاربر بیان شده است، موتور معماری با استفاده از تحلیلگر شهرهای پرجمعیت را از طریق هستان‌شناسی پایگاه داده تحلیلی جستجو و واکشی می‌کند و اگر اصطلاح Snowy در درخواست قید شده باشد، موتور با استفاده از هستان‌شناسی آب و هوا مقدار یا مقادیر مناسب را استخراج می‌کند. پس از این که اطلاعات از مؤلفه منعقدکننده استخراج شدند، مؤلفه تولیدکننده پرس و جو اقدام به ساخت پرس و جو برای هر یک از داده‌های در دسترس در لایه داده می‌کند. پرس و جوهای مورد نیاز برای پایگاه داده تحلیلی از نوع SQL بوده و برای داده‌های خارجی از نوع SPARQL است. این مؤلفه برای هر یک از داده‌های مورد تحلیل کاربر، یک پرس و جو با ساختار مناسب تولید می‌کند. پس از مهیاشدن پرس و جو، پرس و جو از طریق مؤلفه تسهیم‌کننده پرس و جو به پایگاه داده تحلیلی و داده‌های خارجی جهت اجرا ارسال می‌شود. لایه یکپارچه‌ساز همانند یک فیلتر دوطرفه عمل می‌کند. این لایه مدیریت جریان داده از ورودی به خروجی (نتیجه به دست آمده از داده‌ها) را بر عهده دارد و شامل مؤلفه‌های اولویت‌دهنده، مدیر دید ذخیره‌شده، تصمیم‌گر و گزارش‌کننده است. مؤلفه اولویت‌دهنده اولویت اجرای پرس و جوها را در لایه داده مشخص می‌کند. به عنوان مثال اگر پرس و جوی مربوط به داده آب و هوا نیاز به نام شهر داشته باشد و کاربر در درخواست، نام شهر(های) مد نظر خود را وارد نکرده باشد به این دلیل که نتیجه شامل نام شهر(ها) است، بنابراین نخست پرس و جوی OLAP پاسخ‌دهی می‌شود و نام شهر از نتیجه پرس و جوی OLAP به پرس جوی SPARQL اضافه می‌گردد و پرس و جوی SPARQL جهت اجرا ارسال می‌گردد. مؤلفه مدیر دید ذخیره‌شده نیز در صورت نیاز، نتیجه به دست آمده از اجرای پرس و جو را ذخیره می‌کند. به عنوان مثال، مؤلفه اولویت‌دهنده نتیجه به دست آمده از پرس و جوی OLAP را تا زمانی که نتیجه SPARQL آماده شود، ذخیره می‌کند. سپس بر روی دید ذخیره‌شده نسبت به نتایج به دست آمده از SPARQL، فیلتر اعمال می‌شود. مؤلفه تصمیم‌گر نتیجه نهایی از تحلیل پرس و جوها را اعلام می‌کند. این مؤلفه دو نوع مقدار نسبت به درخواست کاربر ارائه می‌کند. مقدار نتیجه داخلی، نتیجه داخلی از مقایسه یک پرس و جوی واحد را مشخص می‌کند و نتیجه خارجی نتیجه بین دو پرس و جو و بیشتر را مشخص می‌کند. لایه داده، داده‌های مورد نیاز را برای تحلیل فراهم می‌کند. این لایه شامل داده پایگاه داده تحلیلی، داده‌های بخش‌های بی‌درنگ که هنوز وارد پایگاه داده تحلیلی نشده‌اند و داده‌های خارجی از قبیل آب و هوا، شبکه‌های اجتماعی، حسگر و دیگر داده‌هایی که تحلیل‌گران نیاز به پیداکردن همبستگی میان آنها با داده‌های پایگاه داده تحلیلی دارند است.

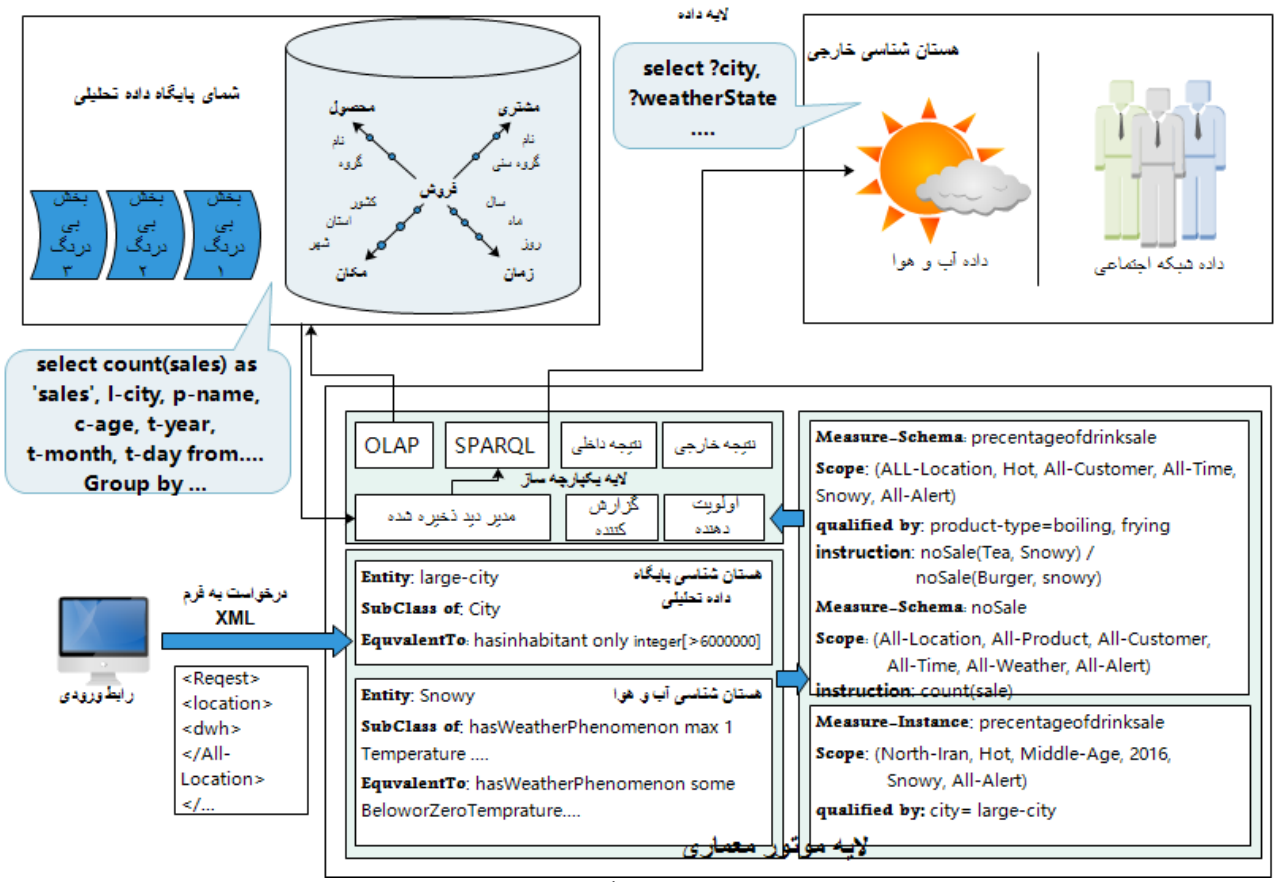
۴-۳ فرایند اجرای پرس و جو

شکل ۳ معماری پایگاه داده تحلیلی تقریباً بی‌درنگ مبتنی بر هستان‌شناسی را برای فرایند پاسخ‌دهی به پرس و جو که در بخش مورد مطالعاتی توضیح داده شده است نشان می‌دهد.

ابتدا درخواست کاربر وارد لایه موتور معماری می‌شود. در قسمت هستان‌شناسی پایگاه داده تحلیلی پرس و جوی کاربر از طریق هستان‌شناسی داخلی به مقادیر مناسب نگاشت می‌شود. به عنوان مثال large-city به شهرهای با جمعیت بیشتر از شش میلیون اشاره می‌کند. نام این شهرها استخراج و در جای مناسب برای ساخت پرس و جو قرار می‌گیرد. اگر پرس و جوی تحلیلی کاربر نیاز به داده‌های خارجی داشته باشد، همانند هستان‌شناسی داخلی، مقادیر مناسب از هستان‌شناسی خارجی استخراج می‌شود. به عنوان مثال، کاربر همبستگی میان فروش و شرایط آب و هوایی خاص برفی[1] و یا خصوصیاتی که شرایط آب و هوایی برفی را توصیف می‌کند مد نظر قرار داده است، سپس مقدار(های) مناسب از هستان‌شناسی آب و هوا استخراج می‌گردد. در قسمت Measure-Schema شمای معیار[2] کاربر ساخته می‌شود. این مؤلفه شامل چهار قسمت به نام‌های Measure-Schema، Scope، qualified by و Instruction است. Measure-Schema نام معیار را مشخص می‌کند. Scope شامل ابعاد پایگاه داده تحلیلی و موجودیت‌های داده‌های خارجی است که با یک یا چندین مشخصه، مشخص می‌شود. Instruction نیز توابع و عملگر مورد نیاز برای انجام عمل مناسب را مشخص می‌کند.

Measure-Instance نمونه مناسب برای تحلیل را مشخص می‌کند. هر دو قسمت Measure-Schema و Measure-Instance برای ساخت پرس و جو مورد استفاده قرار می‌گیرند. در لایه یکپارچه‌ساز تدابیر لازم برای اجرا، مدیریت و گزارش‌گیری درخواست کاربر انجام می‌شود. اولین فاز در این لایه، اولویت‌بندی اجرای پرس و جوها در صورت نیاز است. پس از آن پرس و جوهای OLAP و SPARQL جهت اجرا ارسال

1. Snowy
2. Measure



شکل ۳: فرایند پاسخ‌دهی به پرس و جو در معماری تحلیلی مقایسه‌ای پایگاه داده تحلیلی تقریباً بی‌درنگ مبتنی بر هستان‌شناسی.

می‌شوند. اگر نیاز به ساخت دید ذخیره‌شده جهت انجام پرس و جوی بعدی باشد، نتیجه ذخیره می‌گردد. پس از اتمام تمامی فازها، مقادیر داخلی و خارجی تحلیل و مقایسه آماده می‌شود و برای گزارش‌دهی و اقدام مناسب ارسال می‌شوند.

۴-۴ پرس و جوهای تحلیلی ترتیبی

پرس و جوی تحلیلی ترتیبی نوعی از انواع پرس و جوها است که در آن رویداد وقایع به چندین ترتیب معین بستگی دارد. مجموعه‌ای از ترتیب‌های معین در یک پرس و جو که معرف ترتیب این وقایع است، نشان‌دهنده یک پرس و جوی تحلیلی ترتیبی است. به عنوان مثال پرس و جوی تحلیلی ترتیبی ۱، دو ترتیب را نشان می‌دهد. این پرس و جو میزان فروش محصول LG CASE به همه مشتریانی که قبلاً محصول MED DRUM را خریداری کرده‌اند در همه سال‌ها مشخص می‌کند

$$D = \{d_1, d_2, d_3\}$$
$$d_1 \leftarrow \textbf{\textit{Dimension}}(Customer, f_3, C_NAME)$$
$$d_2 \leftarrow \textbf{\textit{Dimension}}(Part, f_4, P_CONTAINER)$$
$$d_3 \leftarrow \textbf{\textit{Dimension}}(Date, f_6, day \rightarrow month \rightarrow year)$$
$$M = \{m_1\}$$
$$m_1 \leftarrow \textbf{\textit{Measure}}(SaleValue, sum(f_9))$$
$$W = \{w_1\}$$
$$w_1 \leftarrow d_2 = 'LG_CASE'$$
$$s_1 \leftarrow \textbf{\textit{Sequence}}(D, M, W)$$
$$W = \{w_2\}$$
$$w_2 \leftarrow d_2 = 'MED\ DRUM' \text{ and } \textbf{\textit{Distinict}}(d_1)$$
$$s_2 \leftarrow \textbf{\textit{Sequence}}(D, \varnothing, W)$$
$$S = \{s_1, s_2\}$$
$$J \leftarrow (s1_{d_1} = s2_{d_1}) \text{ and } (s1_{d_3} > s2_{d_3})$$
$$\textbf{\textit{SequenceQuery}}1(S, D, M, J)$$

پرس و جوی ۱

ترتیب اول شامل محصول LG CASE به همراه ابعاد نام مشتری، نام محصول، تاریخ خرید و معیار میزان فروش است. ترتیب دوم شامل محصول MED DRUM و ابعاد نام مشتری، نام محصول و تاریخ است. ترتیب دوم فاقد معیار است ولی شامل دستور یکتابودن برای نام مشتری است. این یکتایی باعث می‌شود که فقط نام هر مشتری یک بار در خروجی مهیا شود و عمل جستجو کوتاه‌تر و دقیق‌تر انجام شود. پرس و جوی ۲ میزان فروش به همراه اسامی مشتریانی که محصول LG CASE را خریداری کرده‌اند نشان می‌دهد

$$D = \{d_1, d_2, d_3\}$$
$$d_1 \leftarrow \textbf{\textit{Dimension}}(Customer, f_3, C_NAME)$$
$$d_2 \leftarrow \textbf{\textit{Dimension}}(Part, f_4, P_CONTAINER)$$
$$d_3 \leftarrow \textbf{\textit{Dimension}}(Date, f_6, day \rightarrow month \rightarrow year)$$
$$M = \{m_1\}$$
$$m_1 \leftarrow \textbf{\textit{Measure}}(SaleValue, sum(LO_REVENUE))$$
$$W = \{w_1\}$$
$$w_1 \leftarrow d_2 = 'LG\,CASE'$$
$$s_1 \leftarrow \textbf{\textit{Sequence}}(D, M, W)$$
$$S = \{s_1\}$$
$$\textbf{\textit{SequenceQuery}}2(S, D, M, \varnothing)$$

پرس و جوی ۲

اگر این دو پرس و جو تحلیلی به طور هم‌زمان توسط دو کاربر به سیستم وارد شوند، سیستم می‌تواند با ساخت و نگهداری دید ذخیره‌شده و تغییر در روند موازی اجرای پرس و جوها، زمان اجرای کلی را کاهش دهد. پرس و جوی ۳ روند ذخیره‌سازی و استفاده از دیدهای ذخیره‌شده را نشان می‌دهد

$$\textbf{SequenceQuery}۱(S, D, M, J)$$
$$v_۱ \leftarrow s_۱$$
$$\textbf{SequenceQuery}۲ \leftarrow v_۱$$

پرس و جوی ۳

در این روش، نتیجه ترتیب اول از پرس و جوی تحلیلی ترتیبی اول که در پرس و جوی تحلیلی ترتیبی دوم نیز ظاهر شده است را در قالب یک دید، ذخیره می‌کنیم. از اجرای پرس و جوی تحلیلی دوم، مادامی که ترتیب اول از پرس و جوی تحلیلی اول آماده نشده باشد ممانعت به عمل می‌آید. سپس پس از آماده‌شدن نتیجه و ذخیره آن، پرس و جوی دوم از دید ذخیره‌شده جهت تولید خروجی استفاده می‌کند. اگر مجموعه پرس و جوهای کاربران شامل ترتیب‌هایی باشد که این ترتیب‌ها در پرس و جوها تکرار شده باشد، می‌توان با ذخیره و نگهداری دیدهای ذخیره‌شده، زمان اجرای پرس و جوهای تحلیلی موازی کاربران را کاهش داد. بخشی از نتیجه اجرای پرس و جوی تحلیلی ترتیبی در خروجی ۱ نشان داده شده است

$$e_۱ = \begin{Bmatrix} ۲, Customer\#۰۰۰۰۰۱۹۹, Customer\#۰۰۰۰۰۱۹۹, \\ LG\,CASE, MED\,DRUM, ۱۹۹۷۰۱۳۰, ۱۹۹۳۰۷۰۱ \end{Bmatrix}$$
$$e_۲ = \begin{Bmatrix} ۲, Customer\#۰۰۰۰۰۲۲۰, Customer\#۰۰۰۰۰۲۲۰, \\ LG\,CASE, MED\,DRUM, ۱۹۹۷۰۲۰۲, ۱۹۹۳۰۹۱۶ \end{Bmatrix}$$
$$\dots$$
$$e_{۱۰۰۱۴۹۲} = \begin{Bmatrix} ۶, Customer\#۰۰۸۹۹۷۸۲, Customer\#۰۰۸۹۹۷۸۲, \\ LG\,CASE, MED\,DRUM, ۱۹۹۳۰۹۲۸, ۱۹۹۳۰۳۲۸ \end{Bmatrix}$$

خروجی ۱

۴-۵ پرس و جوهای تحلیلی مقایسه‌ای

پرس و جوی تحلیلی مقایسه‌ای شامل چندین ترتیب و هر ترتیب دارای یک یا چندین معیار مشترک است. از این معیارهای مشترک برای محاسبه مقدار نهایی پرس و جوی تحلیلی مقایسه‌ای استفاده می‌شود. پرس و جوی ۴ یک پرس و جوی تحلیلی مقایسه‌ای است.

میزان فروش محصول LG CASE در همه شهرها نسبت به فروش محصول MED DRUM در همه شهرها چه قدر اختلاف دارد

$$D = \{d_۱, d_۲\}$$
$$S = \{s_۱, s_۲\}$$
$$d_۱ \leftarrow \textbf{Dimension}(Part, f_۴, P_CONTAINER)$$
$$d_۲ \leftarrow \textbf{Dimension}(Date, f_۶, day \rightarrow month \rightarrow year)$$
$$M = \{m_۱\}$$
$$m_۱ \leftarrow \textbf{Measure}(SaleValue, sum(LO_REVENUE))$$
$$W = \{w_۱\}$$
$$w_۱ \leftarrow d_۱ = 'LG\,CASE'$$
$$s_۱ \leftarrow \textbf{Sequence}(D, M, W)$$
$$w_۲ \leftarrow d_۱ = 'MED\,DRUM'$$
$$W = \{w_۲\}$$
$$s_۲ \leftarrow \textbf{Sequence}(D, M, W)$$
$$J \leftarrow (s۱_{d۲} = s۲_{d۲})$$
$$F \leftarrow (s۱_{m۱} - s۲_{m۱})$$
$$\textbf{SequenceQuery}۳(S, D, J, F)$$

پرس و جوی ۴

ترتیب اول از پرس و جوی تحلیلی اول میزان فروش محصول LG CASE و ترتیب دوم میزان فروش محصول MED DRUM را مشخص می‌کند. شرط اتصال دو ترتیب نیز فقط بر روی بعد تاریخ صورت پذیرفته است. برای محاسبه خروجی نهایی، تابع F از محاسبه اختلاف مقدار فروش ترتیب اول از مقدار فروش ترتیب دوم محاسبه می‌شود. نتیجه پرس و جو در خروجی ۲ نشان داده شده است

$$e_۱ = \{۳, LG\,CASE, ۱۲۲, MED\,DRUM, ۱۱۹, ۱۹۹۲۰۱۰۱\}$$
$$e_۲ = \{-۱۹, LG\,CASE, ۱۱۷, MED\,DRUM, ۱۳۶, ۱۹۹۲۰۱۰۲\}$$
$$\dots$$
$$e_{۳۴۰۶} = \{۱۰, LG\,CASE, ۱۱۵, MED\,DRUM, ۱۰۵, ۱۹۹۸۰۸۰۲\}$$

خروجی ۲

اگر پرس و جوی تحلیلی ۵ که میزان فروش محصول LG CASE را در همه شهرها درخواست کرده است هم‌زمان وارد شود، برای کاهش زمان پاسخ‌دهی به پرس و جوها از دید ذخیره‌شده استفاده می‌شود. پرس و جوی تحلیلی ۶ نحوه استفاده از دید ذخیره‌شده را نشان می‌دهد

$$D = \{d_۱, d_۲\}$$
$$S = \{s_۱\}$$
$$d_۱ \leftarrow \textbf{Dimension}(Part, f_۴, P_CONTAINER)$$
$$d_۲ \leftarrow \textbf{Dimension}(Date, f_۶, day \rightarrow month \rightarrow year)$$
$$M = \{m_۱\}$$
$$m_۱ \leftarrow \textbf{Measure}(SaleValue, sum(LO_REVENUE))$$
$$W = \{w_۱\}$$
$$w_۱ \leftarrow d_۲ = 'LG\,CASE'$$
$$s_۱ \leftarrow \textbf{Sequence}(D, M, W)$$
$$\textbf{SequenceQuery}۴(S, D, M, \varnothing)$$

پرس و جوی ۵

$$\textbf{SequenceQuery}۱(S, D, M, J)$$
$$v_۱ \leftarrow s_۱$$
$$\textbf{SequenceQuery}۴ \leftarrow v_۱$$

پرس و جوی ۶

پرس و جوی *SequenceQuery*۴ مادامی که نتیجه ترتیب اول پرس و جوی *SequenceQuery*۱ ذخیره نشده باشد اجرا نمی‌شود.

۴-۶ پرس و جوهای تحلیلی ترکیبی

پرس و جوهای تحلیلی ترکیبی از ترکیب پرس و جوهای مقایسه‌ای و ترتیبی به وجود می‌آیند. قالب کلی پرس و جو به صورت پرس و جوی تحلیلی مقایسه‌ای است که شامل پرس و جوهای تحلیلی ترتیبی برای مقایسه آنها است. به عنوان مثال، میزان اختلاف بین فروش محصول LG CASE به مشتریانی که قبلاً محصول MED DRUM را خریداری نموده‌اند با میزان فروش محصول WRAP BOX به مشتریانی که قبلاً محصول JUMBO PKG را خریداری نموده‌اند در همه شهرها چه قدر است. این پرس و جوی تحلیلی ترکیبی در مثال پرس و جوی ۷ نشان داده شده است

$$\textbf{\textit{SequenceQuery}}۱(S,D,M,J)$$
$$w_۱ \leftarrow d_۲ = 'WRAP\ BOX'$$
$$s_۳ \leftarrow \textbf{\textit{Sequence}}(D,M,W)$$
$$W = \{w_۲\}$$
$$w_۲ \leftarrow d_۲ = 'JUMBO\ PKG' \text{ and } \textbf{\textit{Distinict}}(d_۱)$$
$$s_۴ \leftarrow \textbf{\textit{Sequence}}(D,\varnothing,W)$$
$$S = \{s_۳, s_۴\}$$
$$J \leftarrow (s۳_{d۱} = s۴_{d۱}) \text{ and } (s۳_{d۲} > s۴_{d۲})$$
$$\textbf{\textit{SequenceQuery}}۵(S,D,M,J)$$
$$J \leftarrow (s۱_{d۲} = s۳_{d۲})$$
$$F \leftarrow (s۱_{m۱} - s۳_{m۱})$$
$$\textbf{\textit{SequenceQuery}}۶(S,D,J,F)$$

پرس و جوی ۷

*SequenceQuery*۱ از پرس و جوی تحلیلی ترتیبی ۱ به دست می‌آید. *SequenceQuery*۵ نیز از میزان فروش محصول WRAP BOX توسط همه مشتریان در همه شهرها که قبلاً محصول JUMBO PKG را خریده‌اند به دست می‌آید. *SequenceQuery*۶ نیز مقایسه دو پرس و جوی تحلیلی ترتیبی *SequenceQuery*۱ و *SequenceQuery*۵ است.

نتیجه در خروجی ۳ نشان داده شده است

$$e_۱ = \begin{Bmatrix} -۵, Customer\#۰۰۰۵۰۵۵۴۰, Customer\#۰۰۰۵۰۵۵۴۰, \\ LG\ CASE, ۲, WRAP\ BOX, ۷, ۱۹۹۷۰۳۱۷ \end{Bmatrix}$$
$$e_۲ = \begin{Bmatrix} ۳, Customer\#۰۰۰۳۴۶۶۳۶, Customer\#۰۰۰۳۴۶۶۳۶, \\ LG\ CASE, ۰, WRAP\ BOX, ۳, ۱۹۹۵۰۵۱۳ \end{Bmatrix}$$
$$\dots$$
$$e_{۳۴۰۶} = \begin{Bmatrix} ۷, Customer\#۰۰۰۸۲۸۵۳۳, Customer\#۰۰۰۸۲۸۵۳۳, \\ LG\ CASE, ۷, WRAP\ BOX, ۰, ۱۹۹۴۰۹۲۵ \end{Bmatrix}$$

خروجی ۳

اگر پرس و جوی ۷ هم‌زمان با پرس و جوهای دیگر وارد سیستم شود، می‌توان از ساختن دیدهای ذخیره‌شده از ترتیب‌ها، زمان پاسخ‌دهی به پرس و جوها را کاهش داد. به عنوان مثال اگر پرس و جوی تحلیلی ۴ هم‌زمان با این پرس و جو وارد شده باشد می‌توان با ساخت دید ذخیره‌شده از ترتیب اول، زمان پاسخ‌دهی را کاهش داد.

شکل ۴ شبه‌کد نحوه اجرای پرس و جوهای تحلیلی ترتیبی، مقایسه‌ای و ترکیبی را نشان می‌دهد.

ورودی الگوریتم، یک پرس و جوی تحلیلی ترتیبی، مقایسه‌ای و یا ترکیبی است. خروجی الگوریتم نیز لیست پرس و جوهای در حال اجرا و لیست پرس و جوهای معلق است. در خط ۱، ترتیب‌های داخل پرس و جو در یک ارائه استخراج می‌شوند. خطهای ۲ تا ۴ به ازای ترتیب‌های استخراج‌شده، بررسی می‌شود که آیا یک ترتیب یکسان در حال اجرا است. اگر ترتیبی در حال اجرا پیدا شود که با ترتیب فعلی یکسان باشد، ترتیب فعلی معلق شده و پس از پایان اجرای ترتیب در حال اجرا، اجرا می‌شود. در خطهای ۵ تا ۷ اگر ترتیبی در حال اجرا یافت نشود، ترتیب‌های پرس و جوی فعلی اجرا می‌شوند و ترتیب‌ها به لیست ترتیب‌های در حال اجرا منتقل می‌گردند. مرتبه اجرایی الگوریتم در دو حالت قابل بررسی است. حالت اول فقط با در نظر گرفتن یک پرس و جوی ورودی است که مرتبه اجرایی الگوریتم $O(n)$ خواهد بود و حالت دوم با در نظر گرفتن n پرس و جوی ورودی است که مرتبه اجرایی الگوریتم برابر با $O(n^۲)$ خواهد بود.

```
Job Scheduler Algorithm
Input:
   Q ← Query
Outputs:
                          ▷ suspended queries
SSR ← {}
                          ▷ Running queries
SR ← {}
                          ▷ Extarct sequnces of query
1. sequences ← Extract _ Sequecnces(Q)
2. for all s ← sequences do
3.    if s ∈ SR then
4.       SSR ← SSR ∪ {Q}
5.    else
6.       SR ← SR ∪ {Q}
7.    end if
8. end for
```

شکل ۴: شبه‌کد الگوریتم اجرای پرس و جوهای تحلیلی ترتیبی، مقایسه‌ای و ترکیبی.

۵- نمونه مطالعاتی معماری

در این بخش یک سناریو به منظور نمایش کاربرد معماری پیشنهادی و صحت راهکار پیشنهادی مطرح می‌شود. در این سناریو کشف همبستگی میان داده خارجی آب و هوا و داده داخلی فروش نشان داده می‌شود. در هستان‌شناسی آب و هوا کلمه "برفی" یک سری ویژگی را برای این نوع از پدیده آب و هوایی مشخص می‌کند. این ویژگی‌ها که توسط هستان‌شناسی بیان شده‌اند، قابل فهم برای انسان و ماشین می‌باشند. بنابراین پرس و جوهای ارائه‌شده در معماری پیشنهادی، قابلیت معنا را در خود جای داده‌اند. چنین راهکاری نقطه شروع برای معنابخشیدن به داده‌ها و حتی پرس و جوهای تحلیلی در معماری‌های پایگاه داده تحلیلی و دیگر ابزارهای تحلیلی می‌باشد. هدف از ارائه نمونه مطالعاتی به شکل یک پرس و جو، نمایش چگونگی بیان یک پرس و جوی معنایی و تحلیل نتایج به دست آمده از آن است. پرس و جوی ۸، پرس و جوی تحلیلی مقایسه‌ای تحلیلگر را که به‌صورت "میزان فروش محصول چای در شرایط آب و هوایی برفی در همه هشدارهای هواشناسی در شمال ایران توسط همه مشتریان در سال ۲۰۱۵ نسبت به فروش محصول همبرگر در شرایط یکسان فروش چای چه قدر است؟" نشان می‌دهد.

در این پرس و جو، ۲ ترتیب $s_۱$ و $s_۲$ وجود دارد. ترتیب اول واکشی میزان فروش محصول چای در سال ۲۰۱۵ در شمال ایران و توسط همه مشتریان از بخش پایگاه داده تحلیلی را مشخص می‌کند. به منظور کسب دانش از هستان‌شناسی داخلی از عملگر *Entity* استفاده شده است. $k_۱$ و $k_۲$ به ترتیب اسامی شهرهای شمال ایران و پرجمعیت را مشخص می‌کنند. پس از این که پرس و جو در بخش پایگاه داده تحلیلی اجرا شد، نتایج در مجموعه $E_۱$ نگهداری می‌شود. سپس برای به دست آوردن نام شهرها و تاریخ‌هایی که شرایط آب و هوایی برفی و هشدارهای هواشناسی در آنها رخ داده است از داده خارجی آب و هوا استفاده می‌شود. به این منظور $c_۱$ و $c_۲$ به ترتیب موجودیت‌های تاریخ و منطقه را مشخص می‌کنند. نتایج داده خارجی نیز در $c_۳$ مشخص می‌شود. همان طور که در شرایط $c_۳$ مشخص است، مقادیر ابعاد تاریخ و منطقه پایگاه داده تحلیلی باید با مقادیر موجودیت‌های تاریخ و ابعاد داده خارجی مطابقت داشته باشند.

سپس مقادیر مشترک مجموعه نتایج $E_۱$ از ترتیب اول با مجموعه نتایج $c_۳$ از نتایج داده خارجی، به عنوان نتیجه نهایی ترتیب اول مشخص می‌شود. برای ترتیب دوم نیز عملیات مشابه ترتیب اول صورت می‌گیرد و

$$D = \{d_1, d_2, d_3, d_4\}$$
$$S = \{s_1, s_2\}$$
$$d_1 \leftarrow \textbf{Dimension}(Part, f_4, P_CONTAINER)$$
$$d_2 \leftarrow \textbf{Dimension}(Date, f_6, day \rightarrow month \rightarrow year)$$
$$d_3 \leftarrow \textbf{Dimension}(Date, f_2, city \rightarrow region \rightarrow nation)$$
$$d_4 \leftarrow \textbf{Dimension}(Date, f_2, C_NAME)$$
$$M = \{m_1\}$$
$$m_1 \leftarrow \textbf{Measure}(SaleValue, sum(LO_REVENUE))$$
$$W = \{w_1\}$$
$$k_1 \leftarrow \textbf{Entity}(DWHOnto, NorthIran)$$
$$k_2 \leftarrow \textbf{Entity}(DWHOnto, LargeCity)$$
$$w_1 \leftarrow d_1 = 'Tea' \textbf{ and } d_2 = '۲۰۱۵' \textbf{ and } d_3 \in k_1 \cap k_2$$
$$s_1 \leftarrow \textbf{Sequence}(D, M, W)$$
$$c_1 = \textbf{Entity}(WeatherOnto, day \downarrow month \downarrow year)$$
$$c_2 = \textbf{Entity}(WeatherOnto, city \downarrow region \downarrow nation)$$
$$c_3 \leftarrow \textbf{Entity}(WeatherOnto, \begin{Bmatrix} \uparrow hasWaetherState = 'SnowyWeatherState' \textbf{ and} \uparrow hasAlerts = \\ 'Warning \textbf{ or } Advisory \textbf{ or } Watch' \textbf{ and } c_1 \in E_1 \textbf{ and } c_2 \in E_1 \end{Bmatrix})$$
$$E_1 \leftarrow E_1 \cap c_3$$
$$w_2 \leftarrow d_1 = 'Burger' \textbf{ and } d_2 = '۲۰۱۵'$$
$$W = \{w_2\}$$
$$s_2 \leftarrow \textbf{Sequence}(D, M, W)$$
$$c_4 \leftarrow \textbf{Entity}(WeatherOnto, \begin{Bmatrix} \uparrow hasWaetherState = 'SnowyWeatherState' \textbf{ and} \uparrow hasAlerts = \\ 'Warning \textbf{ or } Advisory \textbf{ or } Watch' \textbf{ and } c_1 \in E_2 \textbf{ and } c_2 \in E_2 \end{Bmatrix})$$
$$E_2 \leftarrow E_2 \cap c_4$$
$$J \leftarrow (s1_{d2} = s2_{d2})$$
$$F \leftarrow (s1_{m1} - s2_{m1})$$
$$\textbf{SequenceQuery}(S, D, J, F)$$

پرس و جوی ۸

در نهایت مجموعه نتایج نهایی E_1 و E_2 با شرط J به یکدیگر متصل می‌شوند و نمایه F محاسبه می‌گردد.

همان طور که از مقایسه این دو گروه مشخص است، مقایسه برای گروه اول (مورد علاقه): "میزان فروش محصول چای در شرایط آب و هوایی برفی در **همه هشدارهای هواشناسی** در شمال آمریکا توسط **مشتریان در رده سنی جوان** در سال ۲۰۱۵ **تاکنون** نسبت به فروش محصول همبرگر در شرایط یکسان فروش چای"

نسبت به

گروه دوم (مقایسه‌ای): "میزان فروش محصول چای در شرایط آب و هوایی برفی در هشدار **اخطار** هواشناسی در شمال آمریکا توسط مشتریان در رده سنی **میان‌سال** در سال ۲۰۱۶ **تاکنون** نسبت به فروش محصول همبرگر در شرایط یکسان فروش چای" صورت پذیرفته است. مقایسه داخلی بین فروش محصول چای در رده سنی جوان و همبرگر در رده سنی جوان و شرایط هشدار عادی در گروه مورد علاقه است. مقایسه خارجی بین فروش چای و همبرگر بین سن‌های جوان و میان‌سال و در شرایط هشدارهای عادی و اخطار است. از تحلیل و مقایسه گروه مورد علاقه و مقایسه‌ای موارد زیر به دست می‌آید (ممکن است برخی از موارد تکرار شده باشد):

۱) میزان فروش چای در رده سنی جوان بیشتر از فروش همبرگر در رده سنی جوان در شرایط برفی است (مقدار داخلی).

۲) میزان فروش چای در رده سنی میان‌سال از همبرگر در رده سنی میان‌سال در شرایط برفی بیشتر است (مقدار داخلی).

۳) میزان فروش چای در رده سنی جوان از همبرگر در رده سنی میان‌سال در شرایط برفی کمتر است (مقدار خارجی).

۴) میزان فروش همبرگر در رده سنی جوان از چای در رده سنی میان‌سال در شرایط برفی بیشتر است (مقدار خارجی).

۵) میزان فروش چای و همبرگر در رده سنی جوان بیشتر از فروش چای و همبرگر در رده سنی میان‌سال در شرایط برفی است (مقدار خارجی).

۶) میزان فروش همبرگر در رده سنی جوان بیشتر از رده سنی میان‌سال در شرایط برفی است (مقدار خارجی).

۷) میزان فروش چای در رده سنی میان‌سال بیشتر از رده سنی جوان در شرایط برفی است (مقدار خارجی).

میزان فروش چای و همبرگر در شرایط بدون هشدار بیشتر از شرایط اخطار است (مقدار خارجی).

زمانی که گزارش به واحد تبلیغات تحویل داده شود، آنها از این گزارش برای اهداف تبلیغاتی خود استفاده می‌کنند:

۱) در فصل زمستان که در شمال ایران میزان بارش برف زیاد است و رده سنی جوان محصول همبرگر را بیشتر مصرف می‌کند، تبلیغات محصولات همبرگر باید برای رده سنی جوان در رسانه مناسب این

جدول ۱: مقایسه معماری پیشنهادی و معماری یک پایگاه داده تحلیلی سنتی.

مسئله	معماری پیشنهادی	معماری پایگاه داده تحلیلی سنتی
نیاز به تغییرات در شما و نگهداری خود پایگاه داده تحلیلی		*
انجام عملیات به صورت خودکار	*	
صرفه‌جویی در وقت برای انجام پرس و جوهای تحلیلی	*	
قابلیت تحمل خطا	*	
آیا واکشی تمامی داده‌های منابع داده‌ای خارجی به درون بخش‌های داده‌ای خارجی معماری مورد نیاز است؟		*
پشتیبانی از قابلیت تقریباً بی‌درنگ	*	*
قابل فهم بودن داده‌ها برای انسان و ماشین	*	

رده سنی، نسبت به محصول همبرگر بیشتر صورت بگیرد.

۲) تبلیغاتی که برای افراد میان‌سال صورت خواهد پذیرفت، محصول چای را مد نظر قرار می‌دهد. افراد میان‌سال در فصل‌های بارش برف و شرایطی برفی تمایل به نوشیدن چای نسبت به خوردن همبرگر دارند.

۳) تبلیغات همبرگر برای رده سنی جوان بیشتر صورت می‌پذیرد. جوانان تمایل بیشتری به خوردن همبرگر در شرایط برفی نسبت به نوشیدن چای در این شرایط دارند.

۴) تبلیغات همبرگر برای رده سنی جوان نسبت به رده سنی میان‌سال بیشتر صورت می‌پذیرد.

۵) از آنجایی که فروش چای و همبرگر در رده سنی جوان بیشتر است، تبلیغات بیشتری باید در این رده سنی در رسانه تبلیغاتی شبکه‌های اجتماعی و سایت‌های اینترنتی صورت بگیرد.

۶) تبلیغات اینترنتی باید برای محصول همبرگر که رده سنی جوان بیشتر از آن استفاده می‌کنند، در رسانه مناسب این رده سنی صورت بگیرد.

۷) تبلیغ چای باید در رسانه تلویزیون که گروه میان‌سال از آن بیشتر استفاده می‌کنند صورت بگیرد.

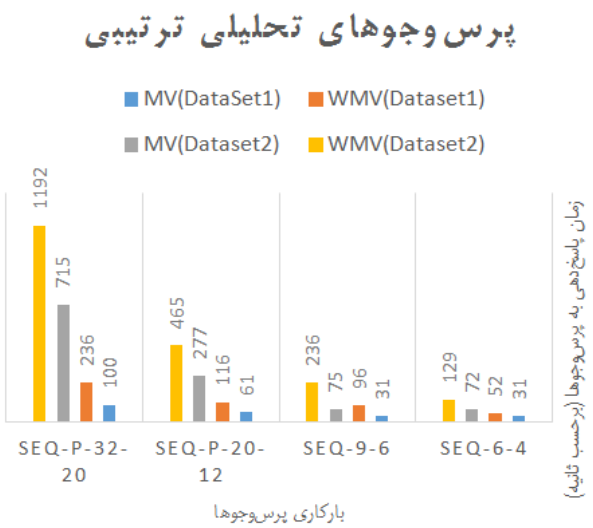
در شرایط آب و هوایی که اخطار وجود دارد، تبلیغات کمتری باید صورت بپذیرد زیرا تمایل خرید دو گروه سنی در شرایط آب و هوایی اخطار نسبت به شرایط آب و هوایی بدون هشدار کمتر است.

به طور معمول تحلیل‌گران یک یا چندین مقایسه را برای تولید گزارش‌های متفاوت برای بخش‌های مختلف تجارت یا سازمان اجرا می‌کنند. این گزارش‌ها به منظور تصمیم‌گیری و تغییر استراتژی‌های تجارت یا سازمان صورت می‌گیرد. پس از این سازمان با اتخاذ تصمیمات مناسب و یا تغییر روند استراتژیک سازمانی خود، میزان فروش محصولات را افزایش می‌دهد. تحلیل‌گران می‌توانند در صورت نیاز و یا تصمیمات مدیران و سرمایه‌گذاران، مقایسه و همبستگی میان داده‌های دیگر از قبیل فعالیت‌های شبکه‌های اجتماعی، و نظرسنجی‌های فروش را برای تحلیل اضافه کنند. همچنین تحلیلگران با بررسی اختلافات مقادیر داخلی و خارجی، گزارش‌های مناسب را برای بخش‌های مختلف سازمان آماده و ارسال می‌کنند. از این طریق مدیران و سرمایه‌گذاران توانایی اتخاذ تصمیمات و استراتژی‌های مناسب را به دست می‌آورند. جدول ۱ مقایسه‌ای از معماری پیشنهادی و معماری یک پایگاه داده تحلیلی سنتی را نشان می‌دهد.

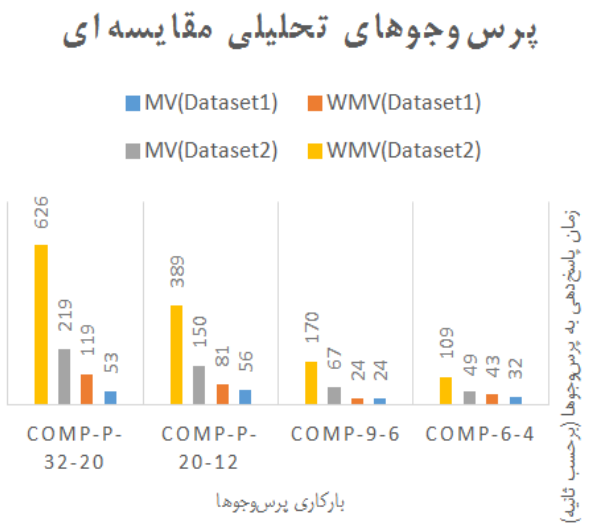
۶- آزمایش‌ها

سناریوی آزمایش شامل اجرای سه نوع از پرس و جوهای تحلیلی معنایی به صورت هم‌زمان است. این پرس و جوها از طریق کاربران به صورت هم‌زمان وارد معماری پیشنهادی می‌شوند. هدف از سناریو کاهش زمان پاسخ‌دهی به پرس و جوهای تحلیلی معنایی با استفاده از الگوریتم ارائه‌شده در بخش ۴ می‌باشد. همچنین سناریو دو رویکرد دید والد و مستقیم را برای نمایش کاهش زمان پاسخ‌دهی به وسیله الگوریتم نشان خواهد داد. آزمایش‌های انجام‌شده تأثیر کاهش زمان پاسخ‌دهی به پرس و جوهای تحلیلی موازی ورودی کاربران را از طریق دیدهای ذخیره‌شده نشان می‌دهد. ارزیابی‌های انجام‌شده توسط ارزیاب SSB صورت پذیرفته است. ارزیاب SSB شامل ۴ بعد (Part, Supplier, Customer, Date) و یک جدول حقیقت (Lineorder) است. از ابزار TPC-H dbgen به منظور پرکردن داده در ابعاد و جدول حقیقت استفاده شده است. در کلیه آزمایشات انجام‌شده، دو نوع آزمون صورت گرفته که آزمون اول به ازای فقط بخش آخر (بخش پایگاه داده تحلیلی) انجام شده و آزمون دوم به ازای همه بخش‌ها (بخش پایگاه داده تحلیلی و بخش‌های بی‌درنگ) صورت گرفته است. برای انجام آزمون از دو مجموعه داده‌ای به حجم‌های متفاوت استفاده شده است. مجموعه اول ۳ گیگابایت حجم ذخیره‌سازی و مجموعه دوم ۷ گیگابایت حجم ذخیره‌سازی را اشغال کرده است. آزمون‌های انجام‌شده به ازای دو رویکرد دیگر نیز مورد بررسی قرار گرفته‌اند. در رویکرد اول، دیدهای ذخیره‌شده به ازای هر یک از ترتیب‌ها ساخته می‌شود و به نام دید مستقیم نام‌گذاری شده است. در رویکرد دوم، برخی از دیدهای ذخیره‌شده از یک دید بزرگ‌تر ساخته می‌شوند که به نام دید والد معرفی شده است.

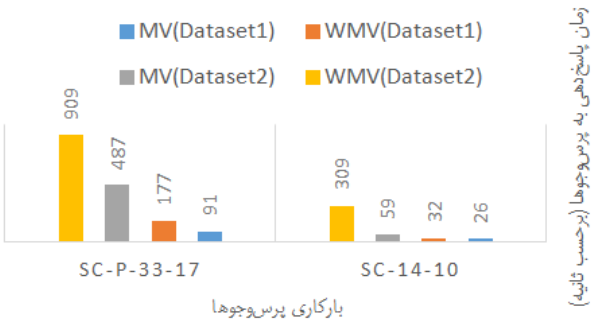
نمودار شکل ۵ آزمون پرس و جوهای تحلیلی ترتیبی موازی کاربران را نشان می‌دهد و در این نمودار زمان اجرای پرس و جوها به ازای پرس و جوهای تحلیلی ترتیبی موازی کاربران در دو مجموعه داده‌ای آمده است. در این نمودار، دو مجموعه پرس و جوی ۴-۶- SEQ و ۶-۹- SEQ، آزمون را به ازای تنها بخش آخر پایگاه داده تحلیلی نشان می‌دهد. در این رویکرد واقعاً ۶ پرس و جو اجرا می‌شود، ۴ پرس و جو، پرس و جوی کاربر است و ۲ پرس و جوی جدید نیز دید ذخیره‌شده را می‌سازند، پس در مجموع ۶ پرس و جو اجرا می‌شود. مجموعه پرس و جوی ۴-۶- SEQ نیز شامل ۴ پرس و جو است. در این مجموعه پرس و جو، ۲ پرس و جوی تحلیلی معادل در ترتیب‌ها وجود دارد. بنابراین در رویکرد این مقاله، ۲ دید ذخیره‌شده از نتیجه این ترتیب‌ها ساخته می‌شود و از این دو دید برای پاسخ‌دهی به پرس و جوهای تحلیلی معادل استفاده می‌شود. مجموعه پرس و جوی ۶-۹- SEQ نیز شامل ۶ پرس و جو است. در این مجموعه پرس و جو، ۳ پرس و جوی تحلیلی معادل در ترتیب‌ها وجود دارد. بنابراین در رویکرد این مقاله، ۳ دید ذخیره‌شده از نتیجه این ترتیب‌ها ساخته می‌شود و از این ۳ دید برای پاسخ‌دهی به پرس و جوهای تحلیلی معادل استفاده می‌شود. آزمون بعدی شامل اجرای دو مجموعه پرس و جوی ۱۲-۲۰- SEQ-P و ۲۰-۳۲- SEQ-P به ازای بخش پایگاه داده



شکل ۵: نتایج آزمون پرس و جوهای تحلیلی ترتیبی.



شکل ۶: نتایج آزمون پرس و جوهای تحلیلی مقایسه‌ای.

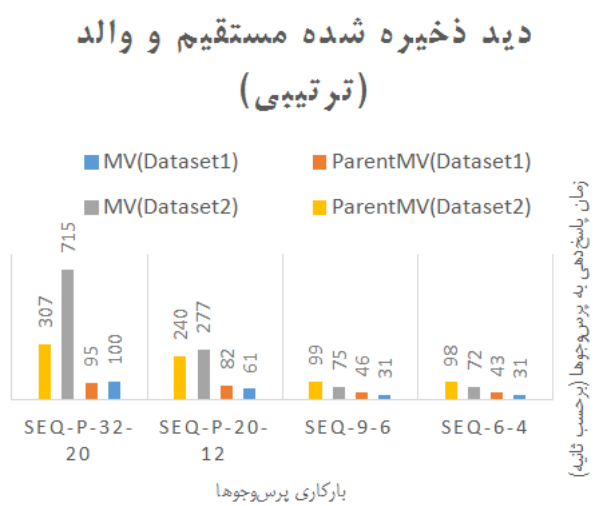


شکل ۷: آزمون پرس و جوهای تحلیلی ترکیبی.

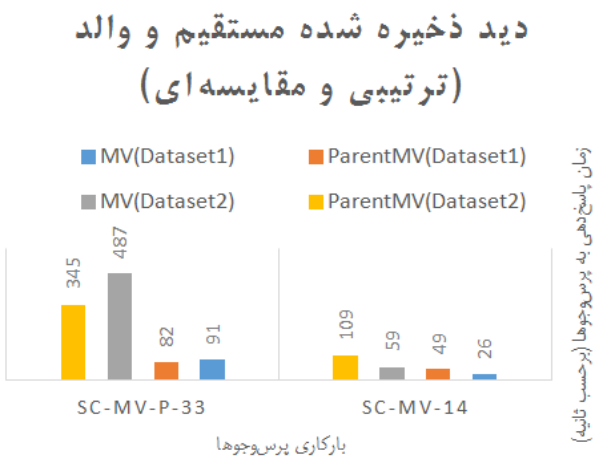
تحلیلی و بخش‌های بی‌درنگ می‌باشد. هر دو مجموعه پرس و جو بر روی دو مجموعه داده‌ای مورد آزمایش قرار گرفته‌اند. مجموعه پرس و جوی SEQ-P -۱۲-۲۰ شامل ۱۲ پرس و جو است. در این مجموعه پرس و جو، ۸ پرس و جوی تحلیلی معادل در ترتیب‌ها وجود دارد. بنابراین در رویکرد این مقاله، ۸ دید ذخیره‌شده از نتیجه این ترتیب‌ها ساخته می‌شود و از این هشت دید برای پاسخ‌دهی به پرس و جوهای تحلیلی معادل استفاده می‌شود. مجموعه پرس و جوی SEQ-P -۲۰-۳۲ شامل ۲۰ پرس و جو است. در این مجموعه پرس و جو، ۱۲ پرس و جوی تحلیلی معادل در ترتیب‌ها وجود دارد. بنابراین در رویکرد این مقاله، ۱۲ دید ذخیره‌شده از نتیجه این ترتیب‌ها ساخته می‌شود و از این ۱۲ دید برای پاسخ‌دهی به پرس و جوهای تحلیلی معادل استفاده می‌شود. در رویکرد بدون استفاده از دید ذخیره‌شده، اگر از نتایج ذخیره‌شده ترتیب‌هایی که معادل پرس و جوهای تحلیلی هستند استفاده نشود، مجموع کل پرس و جوها به طور هم‌زمان در یک زمان اجرا می‌شوند. WMV(Dataset۱) زمان اجرای پرس و جوهای تحلیلی ترتیبی موازی کاربران را بدون ساخت دیدهای ذخیره‌شده در مجموعه داده‌ای اول مشخص می‌کند. MV(Dataset۱) زمان اجرای پرس و جوهای تحلیلی ترتیبی موازی کاربران را با ساخت دیدهای ذخیره‌شده در مجموعه داده‌ای اول مشخص می‌کند. WMV(Dataset۲) و MV(Dataset۲) نیز زمان اجرای پرس و جوهای تحلیلی ترتیبی را بدون ساخت دید ذخیره‌شده و با ساخت دید ذخیره‌شده در مجموعه داده‌ای دوم مشخص می‌کند.

همان طور که از نمودار شکل ۵ مشخص است، استفاده از دیدهای ذخیره‌شده در دو مجموعه داده‌ای و در حالت‌های تک‌بخشی و چندبخشی باعث بهبود زمان پاسخ‌دهی به پرس و جوهای تحلیلی ترتیبی کاربران شده است. زمان اجرای پاسخ‌دهی به پرس و جوهای تحلیلی ترتیبی در حالتی که از دیدهای ذخیره‌شده در مجموعه داده‌ای حجیم‌تر و در حالت چندبخشی استفاده گردیده باعث کاهش چشم‌گیری در زمان اجرا شده است. در بقیه حالات نیز بهبود زمان پاسخ‌دهی به پرس و جوها حاصل شده است. نمودار شکل ۶ زمان اجرای پاسخ‌دهی به پرس و جوهای تحلیلی مقایسه‌ای کاربران را نشان می‌دهد. دو مجموعه پرس و جوی COMP -۶-۴ و COMP -۹-۶ آزمون را به ازای تنها بخش آخر پایگاه داده تحلیلی نشان می‌دهد. ۲ مجموعه پرس و جوی COMP-P -۲۰-۱۲ و COMP-P -۳۲-۲۰ آزمون را به ازای بخش پایگاه داده تحلیلی و بخش‌های بی‌درنگ نشان می‌دهد.

همان طور که در نمودار شکل ۶ مشاهده می‌شود، زمان اجرای پرس و جوهای تحلیلی مقایسه‌ای با ساخت دیدهای ذخیره‌شده، کاهش یافته است. در رویکرد استفاده از دید ذخیره‌شده در مجموعه داده‌ای دوم و حالت چندبخشی، زمان اجرا افت شدیدی داشته که باعث بهبود چشم‌گیر زمان پاسخ‌دهی به پرس و جوها شده است. نمودار شکل ۷ زمان اجرای پاسخ‌دهی به پرس و جوهای تحلیلی ترکیبی شامل پرس و جوهای تحلیلی ترتیبی و مقایسه‌ای را نشان می‌دهد. در این آزمون از دو مجموعه داده‌ای استفاده شده است. مجموعه داده‌ای اول (SC -۱۰-۱۴) برای آزمایش در حالت تک‌بخشی و مجموعه داده‌ای دوم (SC-P -۱۷-۳۳) برای آزمایش در حالت چندبخشی مورد استفاده قرار گرفته است. مجموعه پرس و جوی SC -۱۰-۱۴ شامل ۱۰ مجموعه پرس و جو است. در این مجموعه، ۴ ترتیب معادل در ۱۰ پرس و جوی تحلیلی وجود دارد و بنابراین ۴ دید ذخیره‌شده برای پاسخ‌دهی به پرس و جوها ساخته می‌شود. مجموعه پرس و جوی SC-P -۱۷-۳۳ نیز شامل ۱۶ ترتیب معادل در مجموعه پرس و جوها است و بنابراین ۱۶ دید ذخیره‌شده از نتیجه این ترتیب‌ها ساخته می‌شود و از این ۱۶ دید برای پاسخ‌دهی به ۱۷ پرس و جوی تحلیلی معادل استفاده می‌شود. در رویکرد بدون استفاده از دید ذخیره‌شده، اگر از نتایج ذخیره‌شده ترتیب‌هایی که معادل پرس و جوهای تحلیلی هستند استفاده نشود، مجموع کل پرس و جوها به طور هم‌زمان در یک زمان اجرا می‌شوند. همان طور که از نمودار شکل ۷ مشاهده می‌شود، زمان اجرا در حالت ساخت دیدهای ذخیره‌شده کاهش یافته است.



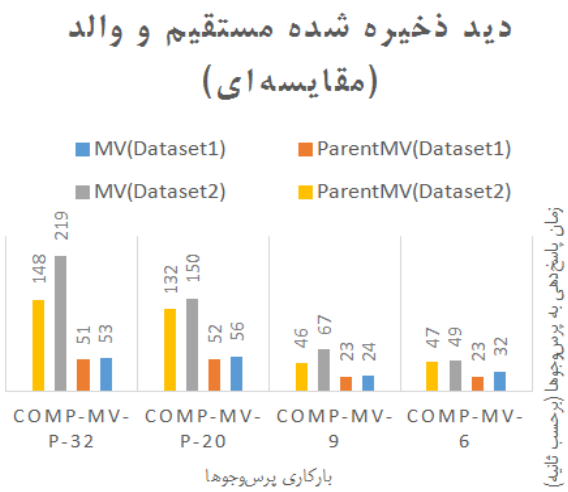
شکل ۸: نتایج آزمون پرس و جوهای تحلیلی ترتیبی در دو رویکرد دید مستقیم والد.



شکل ۱۰: نتایج آزمون پرس و جوهای تحلیلی ترکیبی در دو رویکرد دید مستقیم والد.

مقایسه سه نمودار شکل‌های ۵ تا ۷ نشان می‌دهد که زمان اجرا در حالت استفاده از دیدهای ذخیره‌شده و در حجم بالای داده‌ای در حالت چندبخشی، بهترین کارایی را خواهد داشت. به هر میزان که حجم داده بیشتر باشد، استفاده از دیدهای ذخیره‌شده تأثیر بیشتری خواهند داشت. به هر میزان که ترتیب‌های موجود در پرس و جوهای تحلیلی ورودی کاربران تکرار بیشتری داشته باشد، زمان اجرا به دلیل استفاده بیشتر از دیدهای ذخیره‌شده، کاهش بیشتری پیدا می‌کند.

نمودار شکل ۸ زمان اجرای پرس و جوهای تحلیلی ترتیبی کاربران را برای رویکردهای دید مستقیم و والد نشان می‌دهد. در نمودار شکل ۸، ParentMV زمان اجرای پرس و جو را در رویکرد دید والد نشان می‌دهد. همچنین MV زمان اجرای پرس و جو را در رویکرد ساخت مستقیم دیدهای ذخیره‌شده نشان می‌دهد. چهار مجموعه پرس و جو برای این آزمون مورد بررسی قرار گرفته است. مجموعه پرس و جوهای SEQ –۴–۶ و SEQ –۶–۹ برای حالت تک‌بخشی و مجموعه پرس و جوهای SEQ-P –۱۲–۲۰ و SEQ-P –۲۰–۳۲ در حالت چندبخشی مورد بررسی قرار گرفته‌اند. در مجموعه پرس و جوی SEQ –۴–۶ که شامل ۷ پرس و جو در رویکرد دید والد است، نخست یک دید ذخیره‌شده از ترتیب‌های معادل پرس و جوهای تحلیلی ترتیبی در مجموعه پرس و جوها ساخته می‌شود و سپس دو دید ذخیره‌شده دیگر از دید ذخیره‌شده ای که پیشتر ساخته شده بود ایجاد می‌گردد. نهایتاً پرس و جوها همانند رویکرد پرس و جوی تحلیلی ترتیبی اجرا می‌شوند. در مجموعه پرس و



شکل ۹: نتایج آزمون پرس و جوهای تحلیلی مقایسه‌ای در دو رویکرد دید مستقیم والد.

جوی SEQ –۶–۹ که شامل ۱۰ پرس و جو در رویکرد دید والد است، نخست یک دید ذخیره‌شده از ترتیب‌های معادل پرس و جوهای تحلیلی ترتیبی در مجموعه پرس و جوها ساخته می‌شود و سپس سه دید ذخیره‌شده دیگر از این دید ذخیره‌شده که پیشتر ساخته شده بود، ساخته می‌شوند. نهایتاً پرس و جوها همانند رویکرد پرس و جوی تحلیلی ترتیبی اجرا می‌شوند. در مجموعه پرس و جوهای SEQ-P –۱۲–۲۰ که شامل ۲۴ پرس و جو در رویکرد دید والد است، ۴ دید والد ساخته می‌شود که ۱۲ دید بعدی که در روش مستقیم، مستقیماً از پایگاه داده ساخته می‌شد دیگر از این ۴ دید بزرگ‌تر ساخته می‌شوند. نخست ۴ دید والد ساخته می‌شود و سپس ۸ دید ذخیره‌شده بعدی مورد نیاز از ۴ دید والد ایجاد می‌گردند. در روش دید والد برای این آزمون به جای این که ۸ دید، مستقیم از بخش‌های داده‌ای ساخته شوند، از ۴ دید والد که پیشتر از بخش‌های داده‌ای ساخته شده‌اند ایجاد می‌شوند. در مجموعه پرس و جوهای SEQ-P –۲۰–۳۲ که شامل ۳۶ پرس و جو در رویکرد دید والد است، نخست ۴ دید والد ساخته می‌شود و سپس ۱۲ دید ذخیره‌شده بعدی مورد نیاز از ۴ دید والد ایجاد می‌گردند. در روش دید والد برای این آزمون به جای این که ۱۲ دید مستقیم از بخش‌های داده‌ای ساخته شوند، از ۴ دید والد که پیشتر از بخش‌های داده‌ای ساخته شده‌اند ایجاد می‌شوند. نهایتاً رویکرد ارائه‌شده برای پرس و جوهای تحلیلی ترتیبی برای اجرای بقیه پرس و جوها انجام می‌شود.

زمان اجرای پرس و جو در رویکرد دید والد، در حجم داده‌ای بالا و در حالت چندبخشی کارایی خوبی دارد. مجموع پرس و جوهای SEQ-P –۲۰–۳۲ و SEQ-P –۱۲–۲۰ در رویکرد دید والد زمان اجرای بهتری را نسبت به دید مستقیم در مجموعه داده‌ای دوم دارند.

نمودار شکل ۹ زمان اجرای پرس و جوهای تحلیلی مقایسه‌ای موازی کاربران را در رویکردهای استفاده از دیدهای ذخیره‌شده مستقیم و والد نشان می‌دهد. مجموع پرس و جوهای COMP –۴–۶ و COMP –۶–۹ برای حالت تک‌بخشی و مجموعه پرس و جوهای COMP-P –۱۲–۲۰ و COMP-P –۲۰–۳۲ در حالت چندبخشی مورد بررسی قرار گرفته‌اند.

نتایج حاصل از آزمون نمودار شکل ۹ نشان می‌دهد که استفاده از دید والد در حجم بالای داده‌ای و حالت چندبخشی کارایی بهتری نسبت به دید مستقیم برای پرس و جوهای تحلیلی مقایسه‌ای دارد. البته در حجم پایین‌تری از داده مانند مجموعه داده‌ای اول نیز کاهش زمان پاسخ‌دهی وجود دارد.

نمودار شکل ۱۰ زمان اجرای پرس و جوهای تحلیلی ترتیبی و

مقایسه‌ای کاربران را در رویکردهای دید والد و مستقیم نشان می‌دهد. در این آزمون از دو مجموعه پرس و جو استفاده شده که مجموعه پرس و جوی اول SC-MV-۱۴ و مجموعه پرس و جوی دوم SC-MV-P-۳۳ است. مجموعه پرس و جوی SC-MV-۱۴ شامل ۱۰ پرس و جوی تحلیلی است که ۴ دید ذخیره‌شده مورد نیاز از یک دید ذخیره‌شده والد که پیشتر ساخته شده است ایجاد می‌شوند. در این آزمون به جای این که ۴ دید ذخیره‌شده از بخش‌های داده‌ای مستقیماً ساخته شوند از یک دید ذخیره‌شده والد که پیشتر از بخش‌های داده‌ای ساخته شده است ایجاد می‌شوند. مجموعه پرس و جوی SC-MV-P-۳۳ شامل ۱۷ پرس و جوی تحلیلی است که ۱۶ دید ذخیره‌شده مورد نیاز از ۴ دید والدی که پیشتر ساخته شده است ایجاد می‌شوند. در این آزمون به جای این که ۱۶ دید ذخیره‌شده از بخش‌های داده‌ای مستقیماً ساخته شوند از ۴ دید ذخیره‌شده والد که پیشتر از بخش‌های داده‌ای ساخته شده‌اند، ایجاد می‌شوند.

نتایج نمودار نشان می‌دهد که استفاده از دید والد در رویکرد چندبخشی و در مجموعه داده دوم که حجم بالاتری دارد، بهترین کارایی را داشته است. مقایسه نمودار شکل‌های ۸ تا ۱۰ نشان می‌دهد که استفاده از دید والد در مجموعه داده‌هایی با حجم بیشتر و همچنین در رویکرد چندبخشی بیشترین کارایی را خواهد داشت. در بقیه حالات استفاده از رویکرد دید مستقیم پیشنهاد می‌شود.

۷- نتیجه‌گیری

این مقاله نقش یک معماری پایگاه داده تحلیلی تقریباً بی‌درنگ همراه با داده‌های خارجی و هستان‌شناسی را برای تحلیل و مقایسه پرس و جوهای مدیران، تحلیلگران و کاربران با تحلیل همبستگی میان داده‌های پایگاه داده تحلیلی با داده‌های خارجی مشخص کرده است. این معماری با بهره‌بردن از مزیت‌های استفاده از هستان‌شناسی در ذخیره و توصیف داده‌های خارجی و نیز افزودن دانش در داده داخلی و خارجی سعی بر تسهیل کار کاربران، جلوگیری از تکرار و قابل فهم کردن اصطلاحات که منجر به کاهش زمان پاسخ‌دهی به پرس و جوهای تحلیلی می‌شود، یک معماری مبتنی بر هستان شناسی را پیشنهاد کرده است. نمونه‌های موردی که در این مقاله مورد بررسی قرار گرفتند، نقش کاربردی آنها را برای اتخاذ تصمیمات مناسب و تغییر استراتژی تبلیغ محصولات به روشنی مشخص کرده است. این معماری با پشتیبانی از سطح وسیعی از انواع پرس و جوها مانند پرس و جوهای ترتیبی، مقایسه‌ای و ترکیبی، تجارت و سازمان را در تحلیل، مقایسه و کشف همبستگی میان انواع داده‌ها و پرس و جوهای تحلیلی در رسیدن به اهداف تجاری و سازمانی خود نزدیک می‌کند. علاوه بر این در این مقاله با معرفی الگوریتم کاهش زمان پاسخ‌دهی به پرس و جوهای ترتیبی، شاهد بهبود زمان پاسخ‌دهی به پرس و جوهای کاربران بودیم.

کارهای آتی در بهبود لایه یکپارچه ساز و عملیاتی که منجر به کاهش زمان پاسخ‌دهی در این معماری می‌شود، انجام خواهد شد. همچنین برای کاهش زمان پاسخ‌دهی در داده‌های خارجی از دیدهای ذخیره‌شده استفاده خواهد شد. همچنین مقایسه‌ای برای ذخیره‌سازی داده‌های خارجی در پایگاه داده‌هایی که از قابلیت معنا پشتیبانی می‌کنند، صورت می‌پذیرد. کار آتی دیگر، در حوزه شناسایی ترتیب‌های غیر یکسان خواهد بود. در پرس و جوهای تحلیلی ترتیبی و مقایسه‌ای، ترتیب‌هایی که یکسان بودند، مورد بررسی قرار می‌گرفتند. به عنوان کار آتی بعدی، محاسبه برخی دیگر از معیارهایی است که ممکن است در آینده مورد استفاده قرار گیرند. به عنوان مثال، زمانی که یک دید برای میزان فروش محاسبه می‌شود (پرس و جوهای هم‌زمان همان لحظه به این معیار نیاز دارند)، ممکن است پرس و جوهای بعدی به همین معیار و همچنین معیار تعداد فروش نیز نیاز داشته باشند. اگر این دو معیار هم‌زمان با یکدیگر محاسبه شوند، نیازی به ساخت دید جدید که نیازمند محاسبه جدید و فضای ذخیره‌سازی است، نخواهد بود.

مراجع

[1] J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*, Morgan Kaufmann, 3rd Ed., Jun. 2011.

[2] T. Winsemann, V. Koppen, and G. Saake, "A layered architecture for enterprise data warehouse systems," in *Proc. Int. United Information Systems Conf., UNISCON'12*, pp. 192-199, 2012.

[3] Y. Sharma, R. Nasri, and K. Askand, "Building a data warehousing infrastructure based on service oriented architecture," in *Proc. IEEE Int. Conf. on Cloud Computing Technologies, Applications and Management, ICCCTAM'12,* pp. 82-87, Dubai, United Arab Emirates, 8-10 Dec. 2012.

[4] V. Gonzalez-Castro, L. M. MacKinnon, and M. Del Pilar Angeles, "An alternative data warehouse reference architectural configuration," in *Proc. British National Conf. on Databases.* vol. 5588, pp. 33-41, 2009.

[5] Y. Zhu, L. An, and S. Liu, "Data updating and query in real-time data warehouse system," in *Proc. Int. Conf. on. Computer Science and Software Engineering,* vol. 5, pp. 1295-1297, Hubei, China, 12-14 Dec. 2008.

[6] R. J. Santos and J. Bernardino, "Real-time data warehouse loading methodology," in *Proc. of the 2008 Int. Symp. on Database Engineering & Applications,* pp. 49-58, Coimbra, Portugal, 10-12 Sept. 2008.

[7] A. Cuzzocrea, N. Ferreira, and P. Furtado, "Enhancing traditional data warehousing architectures with real-time capabilities," in *Proc. Int. Symp. on Methodologies for Intelligent Systems*, *ISMIS'14*, pp. 456-465, 2014.

[8] M. Obali, B. Dursun, Z Erdem, and A. K. Görur., "A real time data warehouse approach for data processing," in *Proc. IEEE Signal Processing and Communications Applications Conf., SIU'13,* 4 pp., Haspolat, Turkey, 24-26 Apr. 2013.

[9] J. Zuters, "Near real-time data warehousing with multi-stage trickle and flip," in *Proc. Int. Conf. on Business Informatics Research*, vol. 90, pp. 73-82, 2011.

[10] Y. Mao, W. Min, J Wang, B. Jia, and Q. Jie., "Dynamic mirror based real-time query contention solution for support big real-time data analysis," in *Proc. IEEE 2nd. Int. Conf. on Information Technology and Electronic Commerce, ICITEC'14*, pp. 229-233, Dalian, China, 20-21 Dec. 2014.

[11] W. Qu, V. Basavaraj, S. Shankar, amd S. Dessloch, "Real-time snapshot maintenance with incremental ETL pipelines in data warehouses," in *Proc. Int. Conf. on Big Data Analytics and Knowledge Discovery*. vol. 9263, pp. 217-228, 2015.

[12] Z. Lin, Y. Lai, C. Lin, Y. Xie, and Q. Zou, "Maintaining internal consistency of report for real-time OLAP with layer-based view," in *Proc. Asia-Pacific Web Conf.* vol. 6612, pp. 143-154, 2011.

[13] I. Hamdi, E. Bouazizi, and J. Feki, "Dynamic management of materialized views in real-time data warehouses," in *Proc. IEEE 6th Int. Conf. on Soft Computing and Pattern Recognition, SoCPaR'14*, pp. 168-173, Tunis, Tunisia, 11-14 Aug. 2014.

[14] T. Jain, "Refreshing datawarehouse in near real-time," *International J. of Computer Applications*, vol. 46, no. 18, pp. 24-29, May 2012.

[15] M. A. Naeem, G. Dobbie, and G. Webber, "An event-based near real-time data integration architecture," in *Proc. IEEE 12th Enterprise Distributed Object Computing Conf. Workshops*, pp. 401-404, Munich, Germany, 16-16 Sept. 2008.

[16] S. YiChuan and X. Yao, "Research of real-time data warehouse storage strategy based on multi-level caches," *Physics Procedia*, vol. 25, pp. 2315-2321, 2012.

[17] R. J. Santos, J. Bernardino, and M. Vieira, "24/7 real-time data warehousing: a tool for continuous actionable knowledge," in *Proc. IEEE 35th Annual Computer Software and Applications Conf.,* pp. 279-288, Munich, Germany, 18-22 Jul. 2011.

[18] N. Ferreira, P. Martins, and P. Furtado, "Near real-time with traditional data warehouse architectures: factors and how-to," in *Proc. of the 17th Int. Database Engineering & Applications Symp*., pp. 68-75, Barcelona, Spain, 9-13 Oct. 2013.

[19] P. O'Neil, E. O'Neil, X. Chen, and S. Revilak, "The star schema benchmark and augmented fact table indexing," in *Proc. Technology Conf. on Performance Evaluation and Benchmarking,* vol. 5895, pp. 237-252, 009.
[20] T. M. Nguyen and A. Min Tjoa, "Zero-latency data warehousing for heterogeneous data sources and continuous data streams," in *Proc. 5th Int. Conf. on Information Integration and Web-Based Applications Services*, pp. 55-64, Jakarta, Indonesia, 2003.
[21] L. Golab and T. Johnson, "Data stream warehousing," in *Proc. IEEE 30th Int. Conf. on Data Engineering*, pp. 949-952, Chicago, IL, USA, 31 Mar.-4 Apr. 2014.
[22] M. Gorawski and A. Gorawska, "Research on the stream ETL process," in *Proc. Int. Conf. Beyond Databases, Architectures and Structures,* vol. 424, pp. 61-71, 2014.
[23] R. Abrahiem, "A new generation of middleware solutions for a near-real-time data warehousing architecture," in *Proc. IEEE Int. Conf. on Electro/Information Technology*, pp. 192-197, Chicago, IL, USA, 17-20 May 2007.
[24] B. Neumayr, M. Schrefl, and K. Linner, "Semantic cockpit: an ontology-driven, interactive business intelligence tool for comparative data analysis," in *Proc. Int. Conf. on Conceptual Modeling*, vol. 6999, pp. 55-64, 2011.
[25] T. Niemi and M. Niinimaki, "Ontologies and summarizability in OLAP," in *Proc. of the ACM Symp. on Applied Computing*, pp. 1349-1353, Sierre, Switzerland, 22-26 May 2010.
[26] T. Neubock, B. Neumayr, M. Schrefl, and C. Schutz., "Ontology-driven business intelligence for comparative data analysis," *Business Intelligence,* vol. 172, pp. 77-120, 2014.
[27] S. Khouri and B. Ladjel, "A methodology and tool for conceptual designing a data warehouse from ontology-based sources," in *Proc. of the ACM 13th Int. Workshop on Data Warehousing and OLAP,* pp. 19-24, Toronto, ON, Canada, 30-30 Oct. 2010.
[28] M. Niinimaki and T. Niemi, "An ETL process for OLAP using RDF/OWL ontologies," *J. on Data Semantics XIII,* vol. 5530, pp. 97-119, 2009.
[29] O. Romero and A. Abello, "Automating multidimensional design from ontologies," in *Proc. of the ACM 10th Int. Workshop on Data Warehousing and OLAP,* 8 pp, 9-9 Nov. 2007.
[30] D. Skoutas, A. Simitsis, and T. Sellis, "Ontology-driven conceptual design of ETL processes using graph transformations," *J. on Data Semantics XIII.* vol. 5530, pp. 120-146, 2009.
[31] C. Koncilia, J. Eder, and T. Morzy, "Analyzing sequential data in standard OLAP architectures," in *Proc. 18th East European Conf., Advances in Databases and Information Systems, ADBIS'14,* Ohrid, Macedonia, vol. 8716, pp. 56-69, Sept. 2014.
[32] B. Bebel, M. Morzy, T. Morzy, Z. Krolikowski, and R. Wrembel, "OLAP-like analysis of time point-based sequential data," in *Proc. Advances in Conceptual Modeling*, vol. 7518, pp. 153-161, Florence, Italy, Oct. 2012.

سیدمصطفی شفائی تحصیلات خود را در مقطع کاردانی کامپیوتر- نرم‌افزار در سال ۱۳۹۰ با کسب رتبه دوم در دانشگاه فنی و حرفه‌ای شهید جباریان، و کارشناسی مهندسی کامپیوتر- نرم‌افزار در سال ۱۳۹۲ با کسب رتبه اول در دانشگاه علم و فرهنگ به پایان رسانده است، و در سال ۱۳۹۵ کارشناسی ارشد خود در رشته مهندسی کامپیوتر- نرم‌افزار را از دانشگاه تربیت دبیر شهید رجایی اخذ کرده است. زمینه‌های تحقیقاتی مورد علاقه ایشان عبارتند از: پایگاه داده تحلیلی، هستان‌شناسی و هوش تجاری.

نگین دانشپور استادیار دانشکده مهندسی کامپیوتر دانشگاه تربیت دبیر شهید رجایی می‌باشد. نام‌برده تحصیلات خود را در مقطع کارشناسی مهندسی کامپیوتر- سخت‌افزار در سال ۱۳۷۸ با کسب رنبه اول در دانشگاه شهیدبهشتی، و کارشناسی ارشد مهندسی کامپیوتر- نرم‌افزار در سال ۱۳۸۱ در دانشگاه صنعتی امیرکبیر به پایان رسانده است، و در سال ۱۳۸۹ دکتری خود در رشته مهندسی کامپیوتر- نرم‌افزار را از دانشگاه صنعتی امیرکبیر اخذ کرده است. زمینه‌های تحقیقاتی مورد علاقه ایشان عبارتند از: پایگاه داده تحلیلی، سیستم‌های تصمیم‌یار، پیش پردازش داده‌ها، و داده- کاوی.

سیدمجید شفائی تحصیلات خود را در مقطع کاردانی کامپیوتر-نرم‌افزار در سال ۱۳۹۰ در دانشگاه فنی و حرفه‌ای خوارزمی ملایر، و کارشناسی مهندسی کامپیوتر- نرم‌افزار در سال ۱۳۹۲ در دانشگاه علم و فرهنگ به پایان رسانده است و در سال ۱۳۹۵ کارشناسی ارشد خود در رشته مهندسی کامپیوتر- نرم‌افزار را از دانشگاه آزاد اسلامی واحد تهران مرکزی اخذ کرده است. زمینه‌های تحقیقاتی مورد علاقه ایشان عبارتند از: پایگاه داده تحلیلی، هستان‌شناسی و هوش تجاری.