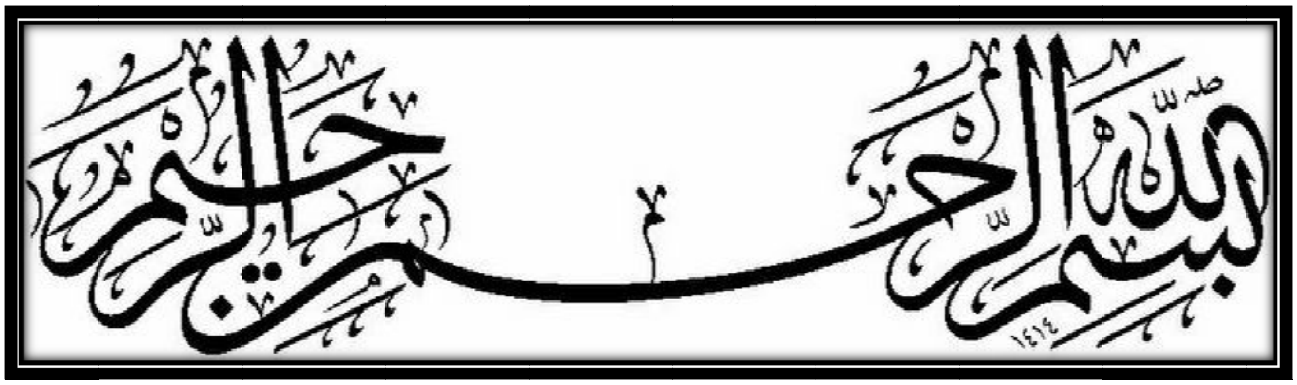

بسم الله الرحمن الرحيم

دانشگاه صنعتی اصفهان

دانشکده برق و کامپیوتر

یکپارچه سازی و کاوش مجموعه داده‌های حجیم RFID

پایان نامه کارشناسی ارشد مهندسی کامپیوتر- نرم‌افزار

لیلا حافظی

استاد راهنما

دکتر محمد حسین سرایی

۱۳۹۰

دانشگاه صنعتی اصفهان

دانشکده برق و کامپیوتر

پایان نامه کارشناسی ارشد رشته مهندسی کامپیوتر- نرم‌افزار خانم لیلا حافظی

تحت عنوان

یکپارچه سازی و کاوش مجموعه داده‌های حجیم RFID

در تاریخ ۱۳۹۰/۲/۶ توسط کمیته تخصصی زیر مورد بررسی و تصویب نهایی قرار گرفت.

۱- استاد راهنمای پایان نامه	دکتر محمد حسین سرایی
۲- استاد مشاور	دکتر محمد علی منتظری
۲- استاد داور	دکتر عبدالرضا میرزایی
۴- استاد داور	دکتر پژمان خدیوی
سرپرست تحصیلات تکمیلی	دکتر سید محمود مدرس هاشمی

تشکر و قدردانی:

نگارنده بر خود می‌داند که در ابتدا از زحمات بی‌دریغ، تلاش‌های بی‌وقفه و راهنمایی‌های ارزشمند استاد گرامی جناب آقای دکتر محمد حسین سرایی در راستای انجام این پروژه تشکر و قدردانی نماید. از جناب آقای دکتر محمد علی منتظری نیز که به عنوان استاد مشاور زحمت زیادی کشیده‌اند تشکر نموده، هم چنین از کمک خالصانه جناب آقای مهندس مسعود صفری که در ابتدای راه از راهنمایی‌هایی ارزشمند ایشان بهره گرفته شد و از تمامی افرادی که در طول انجام این پروژه به گونه‌ای کمک و مساعدت نمودند سپاسگزاری می‌نماید.

تقدیم به آنانی که زیبایی زندگیشان را ارزانی‌ام داشتند:

پدر، مادر و همسرم

فهرست مطالب

عنوان **صفحه**

فصل سوم: تکنولوژی شناسایی امواج رادیویی

فصل چهارم: پیشینه تحقیق

فصل پنجم: مدل پیشنهادی: یک مدل جدید یکپارچه سازی داده‌های RFID

فصل ششم: طراحی و پیاده سازی مدل پیشنهادی

فهرست شکل‌ها

عنوان **صفحه**

چکیده:

تکنولوژی شناسایی از طریق امواج رادیویی یا به صورت خلاصه‌تر تکنولوژی رادیو شناسه چندین سال است که به عنوان یکی از تکنولوژی‌های کارآمد و مفید مطرح شده است. این تکنولوژی در کاربردهای زیادی به کار گرفته شده و نتایج مطلوبی نیز داشته است از جمله در مدیریت زنجیره تأمین، مدیریت چمدان‌های فرودگاه‌ها، خرده فروشی‌ها، کنترل دسترسی، شناسایی حیوانات خانگی و بیماران آلزایمری. در ساده‌ترین سطح این تکنولوژی اجازه می‌دهد یک بارکد از راه دور و در خطی غیر مستقیم با استفاده از امواج رادیویی خوانده شود. بارکدها در سیستم‌های رادیو شناسه معمولاً کدهای یونیکی هستند که به هر شیء اختصاص می‌یابند. این کد روی تراشه‌ای متصل به یک آنتن ذخیره می‌شود. به مجموعه این تراشه و آنتن تگ گفته می‌شود. از آنتن برای انتقال اطلاعات تراشه به تگ‌خوان استفاده می‌شود. تگ‌خوان امواج رادیویی را از تگ‌ها دریافت و آن‌ها را به اطلاعات قابل انتقال به کامپیوتر تبدیل می‌کند. این اطلاعات یا داده‌ها در کامپیوتر ذخیره و پردازش می‌شوند. داده‌هایی که به وسیله‌ی سیستم‌های رادیو شناسه تولید می‌شوند بسیار بزرگ و حجیم هستند. متدهای قدیمی در این مورد کارساز نیست و واضح است که بایست تکنیک‌های جدید و کارآمدی برای پردازش این داده‌های حجیم به کار گرفته شود. استفاده از تکنیک‌های داده‌کاوی برای مدل کردن ارتباطات و کشف الگوهای مخفی در داده‌های بزرگ مفید به نظر می‌رسد. یکپارچه سازی داده‌های تکنولوژی رادیو شناسه آن‌ها را به صورت ابعادی، در فرمت مشخصی مرتب می‌کند و در نتیجه کار تحلیل و آنالیز راحت‌تر می‌شود. لذا قبل از به کارگیری الگوریتم‌های داده کاوی ابتدا با استفاده از قوانین یکپارچه سازی، داده‌ها را به انباره داده تبدیل خواهیم کرد. در این پروژه هدف و تمرکز را بر روی مراحل پیش پردازش قرار داده و سعی شده با بهبود این مراحل گامی مؤثر در جهت بهبود داده کاوی برداشته شود. در این پروژه با ارائه مدل جدیدی از یکپارچه سازی داده‌ها که در آن علاوه بر فازهای موجود در مدل قدیمی یعنی فاز جمع‌آوری داده، فاز پاک سازی و فاز تبدیل داده، فاز دیگری به این مدل اضافه شد که آن فاز فشرده سازی است و برای فشرده سازی داده‌های کاربرد مورد نظر در این پروژه ساختار جدیدی ارائه شد که البته این ساختار برای تمامی کاربردهایی از RFID که یک سری عملیات ثابت و تکراری بر روی داده‌ها انجام می‌شوند نیز قابل اعمال است. با به کارگیری این ساختار، داده‌هایی که از یک خط تولید موتور خودرو به دست آمده بودند و تعداد آن‌ها ۷۵,۱۲۷,۳۵۲ بود را تا ۱/۵۰ مقدار اولیه فشرده کردیم، که باعث بهبود انباره داده در هنگام اعمال الگوریتم‌های داده کاوی شد. علاوه بر فشرده سازی بسیار زیاد داده‌ها مزیت دیگر این ساختار این است که می‌توان با ذخیره روند فشرده سازی در یک جدول در دیتابیس، داده‌ها را به داده‌های اولیه قبل از فشرده سازی بازگرداند. هم چنین کار دیگری که بر روی انباره داده انجام شد این بود که از نرمال بودن جداول در انباره داده به منظور کاهش تعداد جداول صرف نظر شد، که آن چنان که نشان داده خواهد شد این کار نیز موجب بهبود زمان اجرای الگوریتم داده کاوی شد. با به کارگیری الگوریتم‌های داده کاوی بر روی انباره داده ساخته شده با استفاده از ساختار جدید میزان خطا را ۱۶ درصد به دست آوردیم که نشاندهنده این است که روش ارائه شده روش قابل قبولی است.

واژه‌های کلیدی: تکنولوژی شناسایی از طریق امواج رادیویی، تکنیک‌های داده‌کاوی، یکپارچه سازی، انباره داده.

فصل اول
مقدمه

۱-۱ معرفی

تکنولوژی شناسایی از طریق امواج رادیویی[1] یا رادیو شناسه اول بار در جنگ جهانی دوم توسط آلمان‌ها برای شناسایی هواپیماهای خودی از دشمن استفاده شد. آن‌ها هواپیماهای خودی را با استفاده از سیگنال‌های کد شده‌ای که به وسیله‌ی امواج رادیویی فرستاده می‌شد تشخیص می‌دادند. پس از آن این تکنولوژی بسیار مورد استفاده قرار گرفت و نتایج قابل قبولی نیز به جای گذاشت. RFID تکنولوژی است که به یک حس‌گر (تگ‌خوان[2]) اجازه می‌دهد تا از مسافت دور و در خطی غیر مستقیم یک شناسه یونیک، که با استفاده از سیگنال‌های رادیویی با تگ‌های[3] نسبتاً ارزان قیمت به آیتم‌ها وابسته شده‌اند را بخواند. این سیستم قابل مقایسه با سیستم‌های بارکد است ولی تفاوت آن‌ها در این است که در سیستم‌های RFID بر خلاف سیستم‌های بارکد نیاز نیست که تگ‌خوان وتگ در خط مستقیم قرار داشته باشند و هم چنین در سیستم‌های بارکد فاصله بین تگ و تگ‌خوان حداکثر چند سانتی‌متر است اما در سیستم‌های RFID این فاصله تا ۱۵ متر و حتی با تقویت آنتن‌ها به چند برابر این مقدار افزایش می‌یابد. RFID روش بهتری برای سیستم‌های شناسایی بارکد پیشنهاد می‌دهد و کاربردهایی مثل ردیابی آیتم‌ها[4]، مدیریت زنجیره‌های موجودی[5]، شناسایی حیوانات خانگی، شناسایی بیماران، ردیابی پرسنل و بسیاری موارد دیگر را تسهیل می‌کند. این تکنولوژی مدیریت زنجیره‌های موجودی، مسیریابی و توزیع محصولات را ساده می‌کند و هزینه‌ها را با بهبود کارآیی کاهش می‌دهد. شرکت‌های بزرگ مثل Walmart، Target و Albertsons سیستم‌های RFID را در انبار و مراکز توزیع خود به کار گرفته‌اند. شرکت Walmart از این سیستم برای مدیریت زنجیره تأمین خود

---

[1] RFID: Radio Frequency Identification
[2] Reader
[3] Tag
[4] Item Tracking
[5] Supply Chain

استفاده کرده است به این ترتیب که از زمانی که کالای مورد نیاز خود را از کارخانه تولیدی تحویل می‌گیرد و در کامیون یا کشتی بار می‌کند و تا زمانی که به انبار می‌آورند و بعد به قفسه‌های فروشگاه خود منتقل می‌کند این کالاها پیگیری می‌شوند. در این پروژه نیز از یکی از سیستم‌های RFID که در خط تولید موتور خودرو استفاده شده بود بهره می‌بریم. چالش اصلی این است که چگونه شرکت‌ها حجم بالای داده‌های تولیدی کاربردهای RFID را مدیریت و تفسیر کنند. یک شرکت تحقیقاتی[1] پیش بینی کرده است وقتی تگ‌های RFID در سطح آیتم‌ها استفاده شود، در حال حاضر در این شرکت به کالاها به صورت دسته‌ای تگ اختصاص می‌یابد یعنی هر دسته‌ای یک برچسب یا بارکد می‌خورد، شرکت Walmart در یک روز نزدیک به ۷ ترابایت داده تولید خواهد کرد. بنابراین برای این داده‌های حجیم بایست از تکنیک‌های یکپارچه سازی[2]، فشرده‌سازی[3] و متراکم‌سازی[4] به عنوان تکنیک‌های مهمی از پیش پردازش برای آماده سازی داده‌ها برای فرآیند داده کاوی استفاده کرد و از تکنیک‌های داده کاوی[5] برای مدل کردن ارتباطات و کشف الگوهای مخفی استفاده شود[۱].

۱-۲ هدف تحقیق

در سیستم‌های RFID همان طور که ذکر شد داده‌های سطح پایین زیادی تولید می‌شوند و تکنولوژی‌ها و روش‌های قدیمی در مقابل این حجم عظیم داده با چالش‌های اساسی رو به رو خواهند بود. لذا برای استفاده از این داده‌های انبوه در این پروژه ابتدا با استفاده از تکنیک‌های یکپارچه سازی، داده‌ها را برای به کارگیری تکنیک‌های داده کاوی آماده می‌کنیم. در این تحقیق هدف و تمرکز را بر روی مراحل پیش پردازش قرار داده و سعی شده با بهبود این مراحل گامی مؤثر در جهت بهبود داده کاوی برداشته شود. تلاش شد تا با ارائه مدل جدیدی از یکپارچه سازی که در آن بعد از فازهای جمع‌آوری، پاک سازی و تبدیل داده، فاز فشرده سازی با ساختار جدید پیشنهاد شده قرار دارد فرآیند داده کاوی بهبود یابد. با ساختار جدید فشرده سازی و در نتیجه با کاهش حجم داده‌ها زمان اجرای الگوریتم داده کاوی تا حد زیادی کاهش یافت.

با استفاده از مجموعه داده‌های واقعی RFID پروسه داده کاوی تولید الگوهای حقیقی خواهد کرد و چالش‌های داده‌ها نیز واقعی خواهند بود. بنابراین در اختیار داشتن داده‌های سیستم‌های واقعی بسیار مفید می‌باشد، پس بر آن شدیم تا فرآیند داده کاوی و یکپارچه سازی را بر روی این نوع داده‌ها به کار گیریم. از این رو از داده‌های سیستم RFID خط تولید موتور یک شرکت معتبر خودرو سازی استفاده خواهیم کرد.

۱-۳ ساختار پایان نامه

پس از معرفی مختصر در فصل اول ادامه پایان نامه به شرح ذیل ادامه می‌یابد:

در فصل دوم داده کاوی معرفی شده و مراحل پیش پردازش و پردازشی که این تکنیک‌ها به کار خواهند گرفت مورد بحث خواهد بود.

در فصل سوم به صورت تفضیلی به تکنولوژی شناسایی از طریق امواج رادیویی می‌پردازیم، تاریخچه و عناصر اساسی آن را بررسی و سپس به بیان بعضی از کاربردهای آن خواهیم پرداخت.

در فصل چهارم مروری خواهیم داشت بر کارهای انجام شده تا کنون، در زمینه کاری خود و پیشینه تحقیق را بررسی خواهیم نمود.

---

[1] Venture Development Corporation
[2] Warehousing
[3] Compression
[4] Aggregate
[5] Data Mining

در فصل پنجم مدل پیشنهادی ما که یک مدل جدیدی از یکپارچه سازی داده‌ها خواهد بود توضیح داده می‌شود.
در فصل ششم در مورد طراحی و پیاده سازی مدل پیشنهادی توضیحاتی داده خواهد شد.
در فصل هفتم ارزیابی از مدل پیشنهادی خواهیم داشت.
فصل هشتم نتیجه بحث و کارهایی که در آینده باید انجام شود مشخص شده است.

فصل دوم
داده کاوی بر روی داده‌های RFID

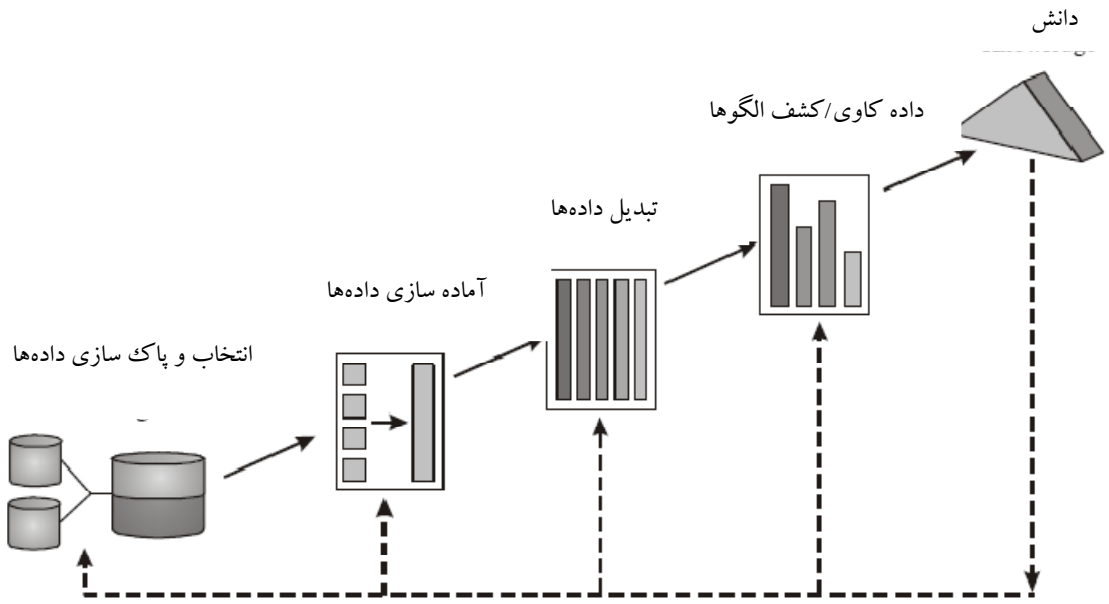
۲-۱ مقدمه

داده کاوی عبارت است از استخراج الگوها و اطلاعات جالب از داده‌ها در پایگاه داده‌های بزرگ.

داده کاوی را " کشف دانش در داده‌ها[1]" نیز می نامند. کشف دانش در داده‌ها دارای مراحل مختلفی می باشد که در اینجا به صورت خلاصه آنها را بیان می کنیم :

- استخراج اطلاعات از چندین منبع داده (پایگاه داده)[2].
- یکپارچه سازی اطلاعات و حذف داده‌های زاید[3].
- قرار دادن اطلاعات اصلاح شده در انبار داده‌ها[4].
- انجام عملیات داده کاوی توسط نرم افزار های مخصوص.
- نمایش نتایج به صورت قابل فهم مانند گزارش و گراف[5].

[1] Knowledge Discovery in Databases (KDD)
[2] Data Gathering
[3] Data Cleaning
[4] Data Warehousing
[5] Result Presentation



شکل ۲-۱ دیدگاه کلی از مراحل ترکیب فرآیند کشف دانش

۲-۲ ویژگی‌های داده‌های RFID

طبیعت داده‌های RFID اساساً از داده‌های رابطه‌ای قدیمی و تکنولوژی‌های انباره داده متفاوت است، این تفاوت چالش‌های تحقیقاتی بزرگی را مطرح می‌کند و نیاز است تا این تفاوت‌ها در سیستم‌های مدیریت داده‌های RFID در نظر گرفته شود. این ویژگی‌ها عبارتند از:

داده‌های ساده: داده‌هایی که از یک کاربرد RFID تولید می‌شود به عنوان رشته‌ای از تاپل‌های RFID به فرم (EPC, location, time) دیده می‌شود، که EPC کد شناسایی یک آیتم است. Location موقعیتی است که تگ‌خوان RFID آیتم را اسکن می‌کند و time زمان خواندن است. همان طور که مشخص است داده‌های RFID اطلاعات بیشتری را شامل نمی‌شوند. به منظور تبدیل این داده‌های خام به فرمی که برای کاربردهای مهم قابل استفاده باشد چندین سطح استنتاج بایست انجام شود.

سیل عظیم[1]: یکی از بزرگ‌ترین مشکلات در RFID نحوه برخورد با سیل عظیم داده‌ها است. به عنوان مثال، Wal-Mart در سه روز به اندازه کل کتاب‌خانه‌های آمریکا داده تولید می‌کند و این فقط مشکل شرکت Wal-Mart نیست. حتی توسعه بسیار کم RFID گیگابایت داده در هر روز تولید می‌کند و داده‌های در مورد EPC، location ، time به صورت پیوسته فرستاده می‌شود.

عدم دقت: یکی از پارامترهای اولیه که اتخاذ گسترده تکنولوژی RFID را محدود می‌سازد عدم دقت رشته داده‌هایی است که توسط تگ‌خوان تولید می‌شود. نرخ خواندن ملاحظه شده در دنیای واقعی رشد RFID اغلب در رنج ۶۰-۷۰٪ است. متأسفانه این نرخ‌های خطا که به داده‌های خام منتقل می‌شود آن‌ها را برای کاربردهای سطح بالای RFID کم فایده می‌کند. بنابراین نیاز است تا این داده‌ها قبل از تغذیه سیستم به این داده‌های غیرقابل اعتماد زدوده شوند. در نتیجه داده‌ها در RFID معمولاً بی‌دقت هستند.

[1] Large in-flood

موقتی و فاصله‌ای: کاربردهای RFID به صورت پویا مشاهدات (مثلاً تغییرات منظم موقعیت‌های آیتم‌های تگ‌دار) را تولید می‌کنند و داده‌ها تغییرات حالت‌ها را در بر دارند. بنابراین در مدیریت داده‌های RFID ضروری است که همه داده‌ها به وسیله مدل داده‌ای رسا و بامعنی مناسب برای سطح کاربرد مدل شوند. با دانستن این حقیقت که نه تنها تگ‌خوان بلکه تگ‌ها نیز داخل PDAها، تلفن‌های همراه و آبجکت‌های متحرک قرار می‌گیرند بنابراین هم آیتم تگ‌دار و تگ‌خوان حرکت ثابتی دارند[۲،۳].

۲-۳ پیش پردازش

همان طور که ذکر شد داده‌های RFID از دقت بسیار پایینی برخوردارند و شامل داده‌های خطادار(داده‌های نامتعارف[1]، مقادیر غیر دقیق و داده‌های اشتباه)، داده‌های ناسازگار(نام‌ها یا کدهای متفاوت، مقادیر غیر ممکن یا خارج از محدوده) و ناکامل(صفات از دست رفته یا مقادیر صفات از دست رفته) می‌باشند بنابراین برای بالا بردن کیفیت داده کاوی بایست داده‌های با کیفیتی نیز در اختیار داشت. در نتیجه پیش پردازش یک گام مهم در موفقیت داده کاوی به حساب می‌آید. گام‌های پیش پردازش معمولاً به قرار زیر است[۴]:

استخراج خصایص[2]: برای شناسایی صفات مرتبط برای یک کار داده کاوی با استفاده از تکنیک‌هایی مانند کشف رخدادها، انتخاب خصایص و تبدیل خصایص.

پاک سازی داده‌ها: برای بالا بردن کیفیت داده‌ها با رفع مشکلاتی مثل داده‌های خطادار، داده‌های پرت، مقادیر از دست رفته و خطاهای مقیاس‌گذاری و درجه بندی.

تغییر شکل یا تبدیل داده‌ها[3]: شامل نرمال کردن داده‌ها

کاهش داده‌ها[4]: برای بهبود زمان پردازش یا کاهش تغییر پذیری در داده‌ها با استفاده از تکنیک‌هایی مثل نمونه‌گیری آماری یا تجمع داده‌ها.

کاهش بعد: برای کاهش تعداد خصایص حاضر در الگوریتم داده کاوی؛ بعضی از تکنیک‌های کاهش بعد خطی یا غیر خطی principal component analysis (PCA)، ISOMAP، locally linear embedding (LLE) هستند.

یکپارچه سازی داده‌ها[5]: شامل سازگاری داده‌ها، تجمیع چندین منبع داده و geocoding , deduplication.

ساختن و استفاده از انباره داده یک وظیفه پیچیده است زیرا نیاز به مهارت‌هایی در زمینه‌های کسب و کار، تکنولوژی و مدیریت برنامه دارد. در راستای مهارت‌های کسب و کار ساختن یک انباره داده شامل فهمیدن این که چنین سیستم‌هایی چگونه ذخیره می‌شوند و داده‌هایشان چگونه مدیریت می‌شوند، استخراج کننده‌ها چگونه ساخته شوند تا داده‌ها از سیستم‌های عملیاتی به انباره داده منتقل شوند، نرم افزار به روز نگه داشتن انباره چگونه ساخته شود تا انباره

---

[1] Outlier
[2] Feature extraction
[3] Data Transformation
[4] Data Reduction
[5] Data integration

داده را با داده‌های سیستم عملیاتی به صورت منطقی به روز نگه دارد. استفاده از انباره داده شامل فهم اهمیت و معنی و مفهوم داده‌های آن است و هم چنین فهم وترجمه نیازمندی‌های کاری به پرس و جویی که بایست توسط انباره پاسخ داده شود. در راستای مهارت‌های تکنولوژی، تحلیل‌گرها بایست بفهمند چگونه اطلاعات کمی را ارزیابی کنند و حقایق را بر اساس نتایج از اطلاعات تاریخی در انباره داده استخراج کنند. این مهارت‌ها شامل توانایی کشف الگوها و گرایش‌ها، پیش بینی گرایش‌ها بر اساس تاریخ و نظر به غیر متعارف‌ها و ارائه توصیه‌های مدیریتی منسجم بر اساس این آنالیزهاست. نهایتاً مهارت‌های مدیریت برنامه شامل نیاز به رو به رو شدن با خیل عظیم تکنولوژی، فروشندگان و کاربران نهایی است که نتایج را با زمان وهزینه قابل قبول تحویل می‌گیرند[۵،۶].

۲-٤ انباره داده چیست

یکپارچه سازی داده سازمان و ابزاری برای اجرای کارها فراهم می‌آورد تا داده‌ها برای گرفتن تصمیمات استراتژیک به صورت سیستماتیک سازمان دهی، فهمیده و استفاده شوند. سیستم انباره داده در دنیای رقابتی امروز یک ابزار ارزشمند محسوب می‌شود. در سال‌های اخیر بسیاری از شرکت‌ها میلیون‌ها دلار برای ساخت انباره داده هزینه کرده‌اند.

اما انباره داده دقیقاً چیست؟ انباره داده به طرق مختلفی تعریف می‌شود و این باعث شده تا فرموله کردن یک تعریف کارا سخت باشد. در کلام انباره داده دیتابیسی عملیاتی و سازمان یافته است. سیستم انباره داده باعث جمع‌آوری سیستم‌های کاربردی مختلف می‌شود. آن‌ها پردازش اطلاعات را با فراهم کردن پلت فرمی از داده‌های یک شکل شده برای آنالیز پشتیبانی می‌کنند.

طبق گفته William H. Inmon انباره داده یک مجموعه‌ای از داده‌های موضوع گرا، یک شکل شده، با زمان‌های مختلف و غیرفرار است که پروسه مدیریت تصمیم گیری را پشتیبانی می‌کند. این تعریف کوتاه اما جامع طرح اصلی انباره داده را نمایان می‌سازد. چهار کلید واژه موضوع گرا، یک شکل شده، با زمان‌های مختلف و غیرفرار انباره داده را از سایر سیستم‌های ذخیره داده مثل سیستم‌های دیتابیس رابطه‌ای، سیستم‌های پردازش تراکنش و سیستم‌های فایل متمایز می‌سازد. نگاهی اجمالی به این چهار واژه داریم:

موضوع گرا: انباره داده بر اساس یک موضوع خاص مثل مشتری، محصول و فروش نسبت به تمرکز روی عملیات روز به روز و فرآیند تراکنش‌های یک سازمان سازمان دهی می‌شود. انباره داده روی مدل کردن و آنالیز داده‌ها برای تصمیم‌گیری متمرکز است. بنابراین انباره داده یک دید مختصر و ساده حول موضوع خاص با حذف داده‌هایی که در فرآیند پشتیبانی تصمیم گیری مفید نیستند فراهم می‌کند.

یک شکل شده: یک انباره داده معمولاً با گردآوری چندین منبع ناهمگن مثل دیتابیس رابطه‌ای، فایل و رکوردهای تراکنشی آنلاین ساخته می‌شود. تکنیک‌های پاک سازی داده و یک شکل سازی داده برای تأمین سازگاری در تبدیل نام‌ها، ساختار رمزگذاری، مقیاس خصیصه‌ها و غیره به کار گرفته می‌شوند.

با زمان‌های مختلف: داده‌ها به منظور فراهم آوردن اطلاعاتی از یک چشم انداز تاریخی (مثلاً ۵-۱۰ سال گذشته) ذخیره می‌شوند. هر ساختار کلیدی در انباره داده به صورت صریح یا ضمنی شامل عنصری از زمان است.

غیر فرار: یک انباره داده از لحاظ فیزیکی همیشه از داده‌های تبدیل شده‌ی داده‌های کاربردی در محیط‌های عملیاتی، جدا ذخیره شده است. به خاطر این جدا سازی یک انباره داده به پردازش تراکنش‌ها، ترمیم و مکانیسم کنترل همزمانی نیازی ندارد. فقط به دو عملیات در دسترسی به داده‌ها نیاز است: لود اولیه داده و دسترسی داده‌ها[۴].

۲-۵ یکپارچه‌سازی داده‌ها

یکپارچه‌سازی داده‌ها مجموعه‌ای از تکنولوژی‌های پشتیبان تصمیم‌گیری است که تصمیم‌ها را سریع‌تر و بهتر می‌کند. اگر داده‌های مناسب در یک انباره داده جمع‌آوری و ذخیره شوند پتانسیل داده کاوی افزایش می‌یابد. یک انباره داده یک سیستم مدیریت پایگاه داده رابطه‌ای است که برای تسهیل گزارش‌گیری و تحلیل ساخته می‌شود. یکپارچه‌سازی داده‌ها یک تکنیک جدید است که داده‌های عملیاتی را استخراج می‌کند و بر ناسازگاری میان فرمت‌های داده‌ای متفاوت غلبه می‌کند.

انباره داده پایگاه داده‌ای با خصوصیات خاص است. داده‌های آن نتیجه تغییر شکل[۱]، بهبود کیفیت و یکپارچه سازی داده‌ها می‌باشد. تغییر شکل داده پروسه تغییر فرم یا ساختار خصیصه‌های موجود است. تغییر شکل با پاک سازی یا غنی سازی[۲] داده‌ها متفاوت است زیرا خصیصه‌های موجود را تصحیح نمی‌کند و خصیصه‌ای نیز اضافه نمی‌کند بلکه فقط خصایص موجود را برای هدف داده کاوی بهبود می‌دهد.

برای طراحی انباره داده رابطه‌ای دو روش وجود دارد یکی تکنیک‌های مدل سازی ابعادی[۳] را به کار می‌گیرد و دیگری بر اساس مفهوم materials است. مدل‌های ابعادی داده را در ساختار cube نمایش می‌دهند و با مدیریت داده OLAP داده‌ها سازگارتر نمایش داده می‌شوند. هدف مدل ابعادی:

الف) تولید ساختار پایگاه داده‌ای است که برای کاربران نهایی فهم و نوشتن پرس و جو ساده باشد.

ب) حداکثر کردن کارآیی پرس و جوها

این اهداف با مینیمم کردن تعداد جداول و روابط بین آن‌ها به دست می‌آید. دیتابیس‌های نرمال خصوصیاتی دارند که برای سیستم‌های OLTP مناسب است و برای انباره داده مناسب نمی‌باشد:

الف) ساختار آن برای فهم و استفاده برای کاربران نهایی آسان نیست. در سیستم‌های OLTP این مشکل ساز نیست زیرا کاربران نهایی با دیتابیس با استفاده از یک لایه نرم‌افزاری تعامل دارند.

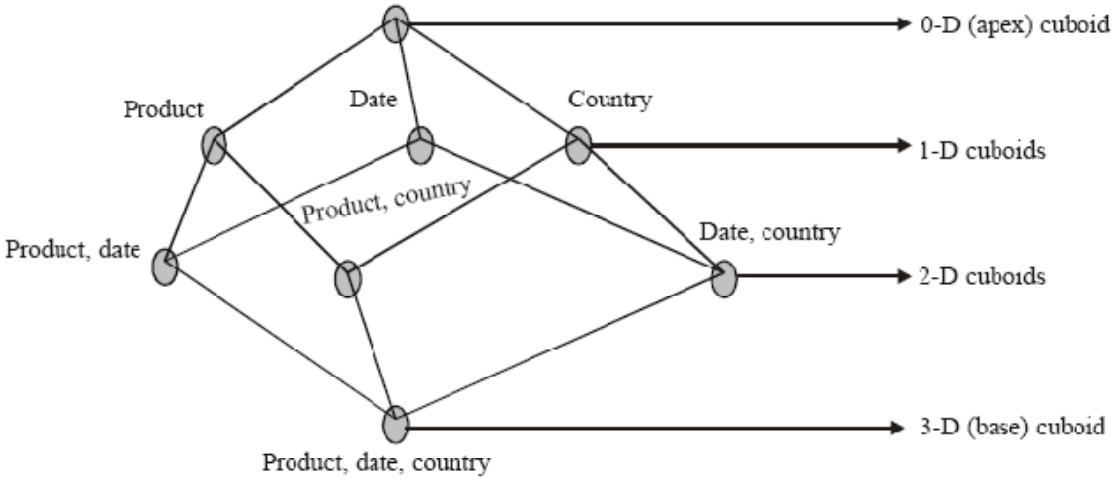
ب) افزونگی داده‌ها کاهش می‌یابد و این کارآیی به روز رسانی را حداکثر می‌کند اما بازیابی را سخت‌تر می‌کند. افزونگی در انباره داده یک مشکل نیست زیرا داده‌ها به صورت آنلاین به روز نمی‌شوند.

مفاهیم ابتدایی مدل‌های ابعادی حقایق۴، ابعاد۵ و مقیاس‌ها۱ هستند. یک حقیقت مجموعه‌ای از آیتم‌های داده‌ای مرتبط به هم است که شامل مقیاس‌ها و داده‌های مفهومی است و آیتم‌های کاری یا تراکنش‌های کاری را در بر

---

[1] Transformation
[2] Data Enrichment
[3] Dimensional Modeling
[4] Facts
[5] Dimensions

می‌گیرد. یک بعد مجموعه‌ای از داده‌هایی است که یک بعد کاری را تشریح می‌کنند. ابعاد زمینه مفهومی حقایق را تعیین می‌کنند. یک مقیاس یک خصیصه عددی از یک حقیقت است، کارآیی یا رفتار کاری وابسته به ابعاد را ارائه می‌دهد. اگر انباره داده‌ای برای پشتیبانی کاربران برای پرس و جوی داده‌ها طراحی شود نیازمندی‌های طراحی مکعب OLAP نتیجه طبیعی مدل ابعادی خواهد بود[۴،۷].



شکل۲-۲ مکعب داده

۲-٦ آبجکت‌های یکپارچه سازی داده‌ها

انواع آبجکت‌هایی که در زیر آمده معمولاً در شمای انباره داده ابعادی استفاده می‌شود:

جداول حقایق: جداول بزرگی در شمای انباره که مقیاس‌های کاری را در خود ذخیره می‌کنند. این جداول اصولاً شامل حقایق و کلیدهای خارجی به جداول ابعادی است. جداول حقایق معمولاً شامل داده‌های عددی و افزایشی است که می‌توانند آنالیز و بررسی شوند. مثال از این نوع جداول شامل فروش، هزینه و سود است.

جداول ابعادی: این نوع جداول هم چنین جداول مرجع یا lookup نیز گفته می‌شود و شامل داده‌های ایستای وابسته در انباره است. جداول ابعادی اطلاعاتی را ذخیره می‌کند که برای پرس و جوها استفاده می‌شود. این جداول معمولاً متنی و تشریحی هستند و به عنوان هدرهای سطرهای مجموعه نتایج استفاده می‌شوند. مثالی از این جداول شامل جداول مشتری، مکان، زمان، محصول و تولید کننده است[۸،۹].

۲-٦-۱ جداول حقایق

یک جدول حقیقت دو نوع ستون دارد: ستون‌هایی شامل حقایق عددی (معمولاً به آن‌ها اندازه[۲] می‌گویند) و ستون‌هایی که کلید خارجی به جداول ابعادی هستند. یک جدول حقیقت هم چنین شامل حقایقی در سطح جزئی یا حقایقی که جمع‌آوری شده است می‌باشد. جداول حقیقت که شامل حقایق جمع‌آوری شده می‌باشد اغلب جداول خلاصه[۳] نامیده می‌شوند. یک جدول حقیقت معمولاً شامل حقایق افزایشی، نیمه افزایشی یا غیر افزایشی است. حقایق

---

[1] Measures
[2] Measurement
[3] Summary Tables