

بنام آنکه جان را فکرت آموخت



یکپارچه سازی داده ها کلیاتی بر مفاهیم و روش ها

مخصوص دانشجویان درس مباحث پیشرفته در پایگاه داده ها

مقطع کارشناسی ارشد و دکتری تخصصی

فصل اول : یکپارچه سازیداده ها(کلیات)

رشد سریع اطلاعات و منابع اطلاعاتی در سالهای اخیر با توزیع شدگی بسیار زیاد سیستم ها و منابع با درجات خودکارسازی متفاوت همراه بوده و در نتیجه دسترسی به اطلاعات در چنین محیط هایی بسیار دشوار شده است. با توسعه سیستم ها، ناهمگونی معنایی و ساختاری که بصورت بالقوه در منابع وجود دارد، فرهنگ های چندگانه و ارائه فناوری های جدید همگی باعث شده اند، این پیچیدگی روز به روز بیشتر شود.

از سوی دیگر نیاز کاربران سیستم ها این است که بتوانند بدون محدودیت، اطلاعات را از منابع گوناگون، بدون پرداختن به جنبه های مختلف آنها، مورد دسترسی و استفاده قرار دهند. در این حالت پرس و جوهای کاربر باید در منابع داده ای مربوطه مطرح و سپس نتایج آنها با هم ترکیب شوند. ولی مراجعه انفرادی به هر یک از منابع و ترکیب نتایج آنها به صورت دستی کاری بسیار مشکل است. زیرا برای این کار ابتدا فرد باید پرس و جوی ترکیبی را برای هر منبع تفکیک کرده، روی هر سیستم پرس و جوی مربوطه را مطرح کند و جواب های هر یک از سیستم ها را با هم ترکیب نماید. این کار پیچیده و در برخی مواقع در بازه زمانی محدود غیر ممکن است. در نتیجه، سرویسی مورد نیاز است که دستیابی به مجموعه ای از منابع داده ای مرتبط به هم را به رغم ناهمگونی های مختلف فراهم آورد، بطوریکه کل این منابع داده ای به صورت یک منبع داده دیده شوند. در این حالت پرس و جوهای کاربر باید در فضایی صوری مطرح، پس از تبدیلات لازم به منابع حاوی داده ارسال، پردازش لازم انجام و سپس نتایج آنها با هم ترکیب شوند. سیستمی که با اجتماع منابع چند گانه امکان چنین سرویسی را فراهم می آورد، سیستم یکپارچه سازی داده نامیده می شود.

تاکنون در ارتباط با این موضوع راهکارهای مختلفی برای غلبه بر مشکلات یکپارچه سازی، که همان انواع ناهمگونی ها است، ارائه گردیده است. اما به موضوع یکپارچه سازی معنایی کمتر توجه شده و این امر باعث گردیده بدون آنکه سعی شود معنا و مفهوم داده درک گردد، مبادرت به برقراری رابطه بین آنها شود. به عبارت دیگر یکپارچه سازی معنایی از اشتباهاتی جلوگیری می کند که ناشی از نفهمیدن قصد و منظور داده است.

در این فصل ابتدا کاربردهای یکپارچه سازی داده را بررسی کرده و در ادامه مشکلات یکپارچه سازی را مطرح خواهیم نمود.

۱-۱) کاربردهای یکپارچه سازی داده

عملیات یکپارچه سازی به دو هدف انجام می شود: ابتدا ایجاد یک دید مجتمع بر روی مجموعه ای از منابع اطلاعاتی که باعث تسهیل در دسترسی به اطلاعات و استفاده مجدد از منابع اطلاعاتی می گردد. دوم استخراج مجموعه ای از داده های مورد نیاز و مشخص از میان منابع اطلاعاتی مختلف که با یکدیگر یکپارچه شده اند، باعث می گردد که کاربران هرآنچه که می خواهند به دست آورند و نیازهای اطلاعاتی خود را بر طرف نمایند. اما به طور کلی مهمترین هدف یکپارچه سازی داده این است که کاربران بر روی "آنچه که می خواهند"، نه بر روی "چگونگی بدست آمدن آن"، متمرکز شوند. برای دستیابی به این امر باید دیدی یکپارچه از منابع مستقل داده ای فراهم گردد. استفاده از یک نمایش ساخت یافته (همانند مدل رابطه ای و XML¹) به منظور نگهداری اطلاعات و به کار گیری بیش از یک نمایش، نقطه اشتراک تمام کاربرد هایی از پایگاه داده

¹ Extensible Markup Language

است که به یکپارچه سازی معنایی نیازمند هستند. این کاربردها باید به صورت خودکار ناسازگاری ها را با توجه به شماها (شما^۲ توصیفی از چگونگی ساختاردهی داده در پایگاه داده ها است) و اطلاعات برطرف نمایند تا قادر به انجام تغییرات بر روی داده ها (مانند یکپارچه سازی شماها یا محاسبه تفاوت آنها [Bernstein 2003]) و یا انتقال اطلاعات و انجام پرس و جوهای میان شماها باشند. بسیاری از این کاربرد ها در طول زمان به وجود آمده اند و توسط تیم های تحقیقاتی مختلفی به صورت فعال مورد مطالعه قرار گرفته اند.

در سال های اخیر رشد انفجاری اطلاعات برخط^۳ باعث رشد انواع کاربردهایی شده است که نیازمند یکپارچه سازی معنایی داده ها می باشند. یک دسته از این کاربردها در جهت ساخت سیستم های یکپارچه سازی داده^۴ فعالیت دارند [Lambrecht, Kambhampati, Gnanaprakasam , 1999]. این سیستمها از یک رابط یکپارچه ساز^۵ جهت ارتباط میان پرس و جوی کاربران و داده های موجود در منابع متعدد بهره می برند، بنابراین این سیستم ها کاربران را از پرس و جو در هر منبع به صورت جداگانه بی نیاز می نمایند.

شکل شماره ۱-۱ شمای کلی سیستم یکپارچه سازی داده را نشان می دهد. این سیستم، پرس و جوی کاربر را توسط شمای واسط دریافت می نماید و سپس به کمک مجموعه ای از تطابق های معنایی میان شمای واسط و شماهای محلی موجود در منابع مختلف، پرس و جو را به یک پرس و جوی محلی بر روی شماهای محلی تبدیل می کند و سپس پرس و جوها را با به کار گرفتن برنامه های پوششی متصل به منابع اجرا می کند و در پایان نتایج بدست آمده را یکپارچه کرده و به کاربر باز می گرداند. بنابراین مسئله مهم در ساخت سیستم های یکپارچه سازی داده، یافتن انطباق های معنایی^۶ است. از سوی دیگر، در عمل منابع اطلاعاتی اغلب شامل

^۲ Schema

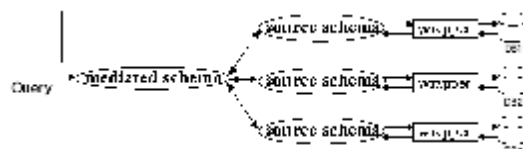
^۳ Online

^۴ Data Integration Systems

^۵ Mediated Schema

^۶ Semantic Matching

رکوردهای تکراری هستند، پس بنابراین مسئله مهم دیگر کشف این تاپل های اطلاعاتی تکراری و حذف آنها از جواب های بازگشتی است، قبل از آن که پاسخ پرس و جوها به کاربران نمایش داده شود.



شکل (۱-۱). شمای کلی سیستم یکپارچه سازی داده

یکی از کاربرد های اولیه یکپارچه سازی داده، یکپارچه سازی شما^۷ می باشد که وظیفه آن ادغام مجموعه ای از شماهای داده شده در یک شمای عمومی می باشد [Pottinger, Bernstein, 2003].

این مسئله از دهه ۱۹۸۰ مطرح گردید. این مشکل در هنگام ساخت یک سیستم پایگاه داده که شامل چندین بانک اطلاعاتی جدا از هم می باشد و یا در هنگام طراحی یک شمای واحد از میان چندین شما که توسط گروههای مختلفی تولید شده است، رخ خواهد داد. فرآیند یکپارچه سازی شامل به وجود آوردن تمام انطباق ها میان اجزاء شماها و سپس استفاده از این تطابق ها برای یکپارچه سازی شماها می باشد [Batini, Lenzerini, & Navathe-1986]

کاربردها و فواید بسیاری برای یکپارچه سازی داده وجود دارد. به عنوان نمونه در حوزه هوش تجاری^۸، یکپارچه سازی داده می تواند جهت پرس و جو و گزارش گیری بر روی اطلاعات تجاری، تحلیل های آماری، پردازش تحلیلی برخط^۹، داده کاوی^{۱۰} جهت پیش بینی، تصمیم گیری و برنامه ریزی گسترده مورد استفاده قرار گیرد. در حوزه مدیریت ارتباط با مشتری، یکپارچه سازی اطلاعات شخصی مشتریان، قوانین کاری

⁷ Schema Integration

⁸ Business intelligence

⁹ Online Analytical Process (OLAP)

¹⁰ Data Mining

و نرخ فروش می تواند در جهت بهبود سرویس دهی بهتر به مشتریان استفاده گردد. سیستم پرتال اطلاعات^{۱۱}، اطلاعات موجود در یک سازمان را به کمک وب سایت های خاص شده یکپارچه نموده و یک نقطه واحد جهت دسترسی اطلاعات را برای کارمندان سازمان، مشتریان، شرکای اقتصادی و به طور کلی برای عموم افراد فراهم می آورد.

در حوزه بازرگانی الکترونیکی^{۱۲} و تجارت الکترونیکی^{۱۳} یکپارچه سازی داده باعث تسهیل در امر تراکنشهای کاری و سرویس دهی بهتر بر روی شبکه های کامپیوتری می گردد.

همراه با گسترش استفاده از پایگاه داده، نیاز ترجمه داده میان چندین پایگاه داده نیز گسترش پیدا نمود. این مسئله زمانی بروز می کند که سازمان ها تصمیم می گیرند پایگاه داده هایشان را در هم ادغام کنند. بنابراین اطلاعات باید از پایگاه داده های قدیمی به یک پایگاه داده جدید انتقال پیدا کند. این مسئله یکی از مهمترین اجزاء انبار داده^{۱۴} و داده کاوی (دو موضوع تحقیقی و اقتصادی مهم از دهه ۱۹۹۰) می باشد. در این کاربردها اطلاعات از چندین منبع مختلف تامین شده و باید به فرم شمای نهایی تبدیل گردند تا آنکه بتوانند امکان تجزیه و تحلیل بیشتری را بر روی اطلاعات فراهم آورند. [Rahm, 2000] [Miller, Haas, Hernandez, 2000] Bernstein, 2001]

مدیریت پایاپای داده ها^{۱۵} دسته ای دیگر از کاربردها هستند که در زمینه یکپارچه سازی داده^{۱۶} مطرح می باشند [Aberer, 2003]. مدیریت پایاپای داده ها از شمای میانی^{۱۷} بهره ای نمی برد، بلکه امکان رد و بدل مستقیم

¹¹ Enterprise information portals (EIP)

¹² E-commerce

¹³ E-business

¹⁴ Data-Warehousing

¹⁵ Peer Data Management

¹⁶ Data Integration

¹⁷ Mediated Schema

و پایاپای^{۱۸} پرس و جوها و بازیابی داده ها را میان منابع داده ای ممکن می سازد. این پرس و جوها و بازیابی

داده ها نیازمند ایجاد یک تناظر معنایی میان منابع هم پایا خود خواهند داشت.

مدیریت مدل^{۱۹} یکی دیگر از کاربردهایی است که توجه خاصی به یکپارچه سازی داده دارد چرا که یکپارچه سازی و تطابق مدل ها، از اعمال اصلی جبر مدیریت مدل می باشند.

کاربردهای اشتراک گذارنده اطلاعات که در بالا توضیح داده شدند، در حوزه های متعددی از جهان واقعی فعالیت دارند که نقش مهمی را در زمینه های مختلفی از قبیل تجارت الکترونیکی^{۲۰} و محاسبات^{۲۱} ایفا می کنند. علاوه بر این سیستم یکپارچه سازی داده، کاربردهای پشتیبانی از تصمیم^{۲۲} مانند پردازش تحلیلی برخط^{۲۳} و داده کاوی را تسهیل می کند. پیشرفت های اخیر و گسترش کاربردهایی که نیازمند یکپارچه سازی معنایی هستند اهمیت یکپارچه سازی معنایی را خاطر نشان می گردند. از یک سو اینترنت میلیون ها منبع داده ای را کنار هم جمع کرده و اشتراک داده را میان آنها ممکن ساخته است و از سوی دیگر گسترش XML به عنوان یک استاندارد جهت اشتراک گذاری داده، این فرآیند را ساده تر کرده است. گسترش وب معنایی^{۲۴} استفاده از کاربردهای اشتراک گذارنده داده را افزایش خواهد داد و در نتیجه تمام این موارد دست به دست هم داده و به نقش مهم یکپارچه سازی معنایی در گسترش این کاربردها اشاره دارند.

اما برای رسیدن به این هدف باید بر مشکلات یکپارچه سازی فائق آمد. از جمله مهمترین مشکلات

یکپارچه سازی داده را در بخش بعدی بررسی خواهیم نمود.

¹⁸ Peer

¹⁹ Model Management

²⁰ E-commerce

²¹ Computing

²² Decision Support Application

²³ Online Analytical Process (OLAP)

²⁴ Semantic Web

۱-۲) چالش های یکپارچه سازی داده

یکپارچه سازی چندین سیستم اطلاعاتی عموماً از طریق ادغام آن ها به دست می آید، به گونه ای که تمام سیستم های اطلاعاتی در قالب یک سیستم واحد شکل گرفته و این حس را به کاربران القا می کنند که آن ها تنها با یک سیستم واحد در تماس می باشند. سیستم یکپارچه سازی برای کاربران یک دید منطقی از داده های همگن را فراهم می آورد در حالی که به طور فیزیکی داده ها به صورت ناهمگن و بر روی چندین منبع اطلاعاتی قرار دارند. به همین منظور، باید تمام داده ها با استفاده از یک سطح تجریدی مشخص (مدل داده ای واحد و معنای یکسان) سازماندهی گردند. این کار شامل تشخیص و تحلیل شما و رفع ناسازگاری های داده اعم از ساختار و معنا می گردد.

به طور کلی سیستم های اطلاعاتی جهت یکپارچه سازی طراحی نشده اند. بنابراین از یک سو اگر دسترسی یکپارچه به چندین منبع اطلاعاتی مختلف مورد نیاز باشد و از سوی دیگر منابع اطلاعاتی و داده های موجود در آن ها با یکدیگر منطبق نباشند، باید از طریق تلفیق قابلیت ها و ایجاد سازگاری میان منابع مختلف، آن ها را در هم ادغام نمود. توجه کنید که این تنها مشکل مطرح در زمینه یکپارچه سازی داده نمی باشد. تا زمانی که هدف، ایجاد یک دید واحد و یکپارچه بر روی منابع داده ای مختلف می باشد، باید عملیات یکپارچه سازی خاصی متناسب با:

- دید معماری موجود در سیستم اطلاعاتی
- محتوا و کارایی های سیستم های اطلاعاتی
- نوع اطلاعاتی که به وسیله سیستم های اطلاعاتی مدیریت می گردد (داده های عددی، داده

های چندرسانه ای، داده های ساخت یافته، نیمه ساخت یافته و غیرساخت یافته)

- نیازمندیهای مورد نیاز جهت استقلال منابع داده ای
 - نوع استفاده از داده های یکپارچه شده (دسترسی از نوع فقط خواندنی یا نوشتنی)
 - تمهیدات مورد نیاز جهت ارائه بهترین کارایی
 - قابل دسترس بودن منابع، بر روی سیستم های اطلاعاتی مختلف انجام گیرد [Dittrich, Jonscher, 1999].
- علاوه بر این باید با انواع مختلفی از ناهمگنی ها نیز مقابله گردد. از جمله این ناهمگنی ها می توان به موارد زیر اشاره نمود:
- سخت افزار و سیستم عامل
 - نرم افزار مدیریت داده
 - مدل های داده ای، شما و مفاهیم داده ای
 - میان افزارها^{۲۵}
 - واسط کاربران^{۲۶}
 - قوانین کاری و محدودیت های یکپارچه سازی

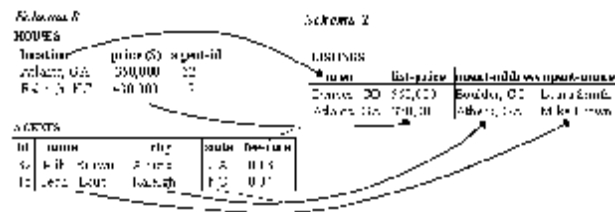
بر خلاف اهمیت و گستردگی یکپارچه سازی معنایی، متأسفانه هنوز این مسئله به عنوان یک مشکل باقی مانده است. به عنوان اولین مشکل می توان به فرآیند تطابق شما (عمل تطابق میان شمای پایگاه های داده را انجام می دهد) اشاره نمود. برای مثال به شمای دو پایگاه داده رابطه ای متعلق به لیست املاک در شکل شماره

²⁵ Middleware

²⁶ User interfaces

۲-۱ آورده شده است، توجه کنید. فرآیند انطباق، تطابق هایی را به صورت زیر پیدا می کند " Location in

" Schema S matches area in Schema T و " Name matches agent-name



شکل (۲-۱). انطباق شما از میان دو شمای متعلق به بانک املاک

در اصل تطابق شمای S و شمای T نیازمند تصمیم گیری بر روی این مسئله است که هر عنصر S از S و t از T با هم تطابق دارند یا خیر یا به عبارت دیگر آیا این مفاهیم واقعا به مفاهیم یکسانی اشاره دارند و یا آن که متعلق به دو مفهوم مختلف می باشد. این مشکل به چندین دلیل، به عنوان اساسی ترین چالش مطرح است:

- معنای عناصر فقط از تعداد کمی از منابع اطلاعاتی (معمولا سازنده اطلاعات، مستندات، شماهای مربوطه) می تواند استخراج گردد. اغلب استخراج معنای اطلاعات از سازنده های داده و مستندات بسیار مشکل است. غالبا سازنده ی داده ها تغییر مکان داده اند یا آنکه بازنشسته شده اند و یا دچار فراموشی شده اند. مستندات نیز معمولا سرسری و یا دارای اشتباه بوده و یا آنکه بروز نمی باشند. در بسیاری از مواقع مانند زمان ساخت یک سیستم یکپارچه سازی داده از روی منابع وب، سازنده های داده و مستندات به سادگی در دسترس نمی باشند.

- با توجه به اول، عناصر شما در حالت معمول بر اساس قراین موجود در شما و داده، تطابق داده می شوند. مثالهایی از این قراین: نام، نوع، مقادیر داده، ساختارهای شما و قوانین جامعیت عناصر می باشند. به هر حال اغلب این قراین غیر قابل اطمینان هستند. به عنوان مثال دو عضو با نام یکسان (مانند area

(می توانند به دو موجودیت مختلف از جهان واقعی اشاره کنند) location و square-feet area
موجود در حوزه خانه). برعکس این قضیه نیز امکان پذیر است یعنی: دو عنصر با نام مختلف (مانند
area و location) می توانند به یک موجودیت یکسان از جهان واقعی اشاره نمایند.

- شواهد و قراین موجود در شما و داده اغلب ناکامل هستند. برای مثال عنوان contact-agent فقط بیانگر ارتباط عنصر با agent می باشد اما هیچ گونه اطلاعاتی را در مورد طبیعت دقیق ارتباط در اختیار ما نمی گذارد (مانند اینکه contact-agent بیانگر شماره تلفن agent است یا عنوان agent جهت ارتباط).

- جهت تصمیم گیری درباره اینکه عنصر S از شمای S با عنصر t از شمای T انطباق دارد یا خیر، باید تمام عناصر دیگر T را نیز مورد امتحان قرار داد تا مطمئن گردیم عنصر دیگری از T بهتر از t با S منطبق نمی باشد. این ویژگی عمومی انطباق ها هزینه های زیادی را به فرآیند تطابق تحمیل می کنند.

- نکته بغرنج تر قضیه این است که انطباق در اغلب موارد وابسته به حوزه مساله است و بستگی به نوع کاربرد در حوزه دارد. یک کاربرد ممکن است که house-style را با house-description قابل انطباق تشخیص دهد در حالی که دیگری این کار را انجام ندهد. بنابراین در اغلب موارد کاربر باید در طول پروسه انطباق حضور داشته باشد. گاهی اوقات نیز ممکن است یک کاربر خاص دارای نظرات شخصی خاص خود باشد، در این گونه مواقع کل تیم باید برای تشخیص انطباق درست تصمیم بگیرند.

به خاطر چالش های بالا ، مدت هاست که ایجاد تطابق معنایی به صورت دستی، امری بسیار مشکل و آمیخته با اشتباه شناخته شده است. برای مثال اخیرا پروژه ای در شرکت ارتباطات GTE مطرح گردید که به دنبال یکپارچه سازی ۴۰ پایگاه داده با جمعا ۲۷۰۰۰ عنصر بودند [Li & Clifton-2000]، طراحان پروژه تخمین

زده بودند که بدون توسعه دهندگان اصلی پایگاه داده، تنها پیدا کردن و مستند کردن انطباق عناصر موجود به بیش از ۱۲ نفر سال کار نیاز خواهد داشت.

مسئله تطابق تاپل ها نیز با چنین چالش هایی روبه رو است. در حالت کلی هزینه های بالای تطابق شماها و داده ها به صورت دستی، باعث به وجود آمدن راه حل های زیادی در جهت خودکارسازی این فرآیند شده است و به دلیل آنکه کاربرها همواره باید در این فرآیند حضور داشته باشند این راه حل ها همگی نیمه خودکار خواهند بود. آغاز این تحقیقات برای دستیابی به راه حل های جدید، به دهه ۸۰ باز می گردد که انرژی بسیار زیادی به منظور مدیریت بسیار زیاد داده های ناهمگن و توزیع شده موجود در شرکت ها و بر روی وب، صرف شده است. از دیگر مشکلات موجود در زمینه یکپارچه سازی می توان به موارد ذیل اشاره نمود:

- ناهمگونی: متأسفانه هر منبع که در سیستم یکپارچه سازی داده قرار می گیرد، دنیای اطراف را به روش خود مدل می کند. نحوه ارائه داده هایی که مفهوم یکسانی دارند ممکن است در منابع داده ای مختلف، متفاوت باشد. به عنوان مثال، هر یک ممکن است از نامگذاری های متفاوتی برای مشخص کردن یک نمونه داده استفاده کنند و یا ممکن است دارای داده های مختلف باشند. علاوه بر ناهمگونی در سطح نمایش داده و مدل سازی، ناهمگونی منابع داده ای ممکن است در سطحی پایین تر، شامل روش های دسترسی استفاده شده در منابع و یا سیستم های عامل که منابع داده ای تحت آن عمل می کنند باشد.

- خودمختاری^{۲۷} منابع داده ای: معمولاً منابع داده ای قبل از سیستم یکپارچه ساز داده بوجود می آیند. در حقیقت بیشتر مواقع این منابع نمی دانند که جزئی از یک سیستم یکپارچه سازی داده هستند. آنها به طور مستقل تصمیم گیری می کنند و نمی توان آنها را مجبور کرد که به

روش خاصی عمل نمایند. در نتیجه منابع می توانند داده و یا عملکرد خود را بدون اطلاع دنیای خارج تغییر دهند. متأسفانه موضوع رفتار مستقل منابع هر چند اجتناب ناپذیر است ولی باعث کاهش میزان یکپارچگی می شود.

- درستی پرس و جو و کارایی آن: پرس و جوهای یک سیستم یکپارچه سازی، معمولاً طبق مدل یکنواخت شده²⁸ سیستم ساخته می شود. این پرس و جوها باید به صورت پرس و جوهای قابل فهم و قابل پردازش برای منابع داده ای مجزا ترجمه شوند. این نگاشت نباید در نتیجه پرس و جو تاثیر نادرست بگذارد. همچنین کارایی پرس و جو باید کنترل شود، زیرا پارامترهای بسیاری می توانند منجر به کاهش کارایی شوند. متأسفانه به رغم ارائه روش های مختلف برای بازنویسی پرس و جوها، تاکنون روشی جامع که در برگیرنده همه مسائل باشد ارائه نشده و روش های موجود مبتنی بر فرضیات و محدودیت های خاص خود هستند.

فصل دوم : روش های یکپارچه سازی داده

اساساً موضوع ناهمگونی و راه حل های مقابله با مشکلات آن، همراه با مسئله تعامل پذیری^{۲۹} مطرح می شود. مسئله تعاملات بین سیستم ها و مشکلات آن بصورت مختلف، در مقالات متعدد ارائه گردیده است. در [Sheth , 1998] که نسبت به سایر مقالات تقسیم بندی بهتری ارائه شده، مشکلات تعامل پذیری به سطوح زیر تقسیم گردیده است:

- سطح سیستمی^{۳۰} : عدم سازگاری سخت افزار و سیستم های عامل
- سطح نحوی^{۳۱} : اختلاف در زبان و روش ارائه و نمایش داده ها
- سطح ساختاری^{۳۲} : اختلاف در مدل داده ای
- سطح معنایی^{۳۳} : معنا و منظور اصطلاحات استفاده شده در مبادله ها. مثال بارز این دسته از مشکلات، استفاده از کلمات مترادف است که در آن یک معنا با کلمات مختلف قابل ارجاع می باشد. انواع ناهمگونی های معنایی در [Visser , 1998] طبقه بندی شده است.

[Aris , 1999] در یک تقسیم بندی کلی دیگر ناهمگونی ها را به دو دسته تقسیم می نماید: ناهمگونی اطلاعاتی و ناهمگونی سیستمی. ناهمگونی اطلاعاتی خود به سه دسته ناهمگونی معنایی، ساختاری و نحوی

²⁹ Interoperability

³⁰ System level

³¹ Syntactic level

³² Structure level

³³ Semantics level

تقسیم می شوند. همچنین ناهمگونی سیستمی به دو دسته ناهمگونی سیستم اطلاعاتی (سیستم مدیریت پایگاه داده^{۳۴}، رسانه ها^{۳۵} و ...) و بستر^{۳۶} تقسیم می شوند.

در این فصل، ابتدا مهمترین معماری های موجود جهت یکپارچه سازی را از دیدگاه [Dittrich and Jonscher, 1999] بررسی خواهیم نمود و در ادامه مهمترین روش هایی که بر اساس این معماری ها پیاده سازی گردیده اند را مطالعه می نمایم.

۲-۱) معماری های یکپارچه سازی

در این بخش، یک معماری از روش های مختلفی که برای مشکل یکپارچه سازی وجود دارد را ارائه خواهیم داد. این دسته بندی منطبق بر دیدگاه [Dittrich and Jonscher, 1999] می باشد که از نظر او، روش های یکپارچه سازی بر اساس آن که در کدامیک از سطوح تجزید قرار می گیرند، شناخته شده و عمل یکپارچه سازی را انجام خواهند داد.

سیستم های اطلاعاتی می توانند مبتنی بر معماری لایه بندی نمایش داده شده در شکل ۲-۱، شرح داده شوند. کاربران در بالاترین لایه قادر خواهند بود با استفاده از واسط کاربرانی که بر روی برنامه های کاربردی قرار خواهند گرفت، به داده ها دست یابند. برنامه های کاربردی ممکن است برای دسترسی به داده ها از میان افزارها^{۳۷} استفاده نمایند. داده ها توسط یک سیستم ذخیره سازی داده مدیریت می گردد. معمولاً از سیستم مدیریت پایگاه داده^{۳۸} جهت ادغام لایه های دسترسی داده و ذخیره سازی استفاده می گردد.

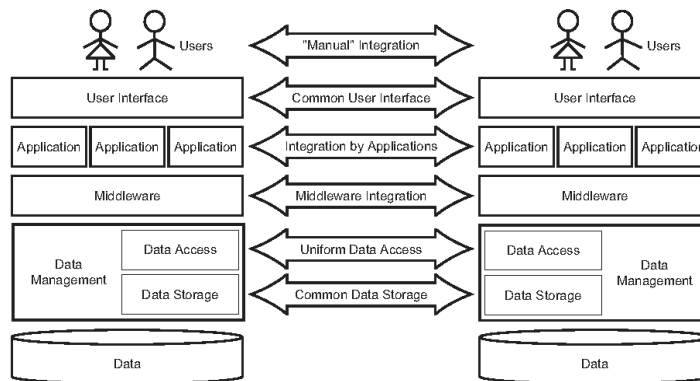
³⁴ DBMS : Data Base Management System

³⁵ Media

³⁶ Platform

³⁷ همانند میان افزار مدیریت پردازش تراکنش ها، میان افزار مبتنی بر پیام، میان افزار SQL

³⁸ Database Management System (DBMS)



شکل (۱-۲). عمومی ترین روش های یکپارچه سازی مبتنی بر لایه های مختلف معماری

به طور معمول، هر یک از مشکلات یکپارچه سازی می توانند در یکی از لایه های موجود گنجاده شوند.

به همین دلیل، در زیر روش های معمول یکپارچه سازی متناسب با معماری ارائه شده، آورده شده است:

یکپارچه سازی دستی^{۳۹}: در این حالت کاربران مستقیماً با تمامی سیستم های اطلاعاتی در تماس بوده و

به صورت دستی داده های انتخابی را یکپارچه می سازند. در این حالت کاربران مجبور خواهند بود با واسط

های کاربری گوناگون و زبان های پرس و جوی مختلفی سر و کار داشته باشند. علاوه بر این کاربران باید در

زمینه های مختلفی از قبیل: حوزه اطلاعات، نمایش منطقی و معنای داده ها، دارای دانش کافی باشند.

واسط کاربری مشترک^{۴۰}: کاربر از یک واسط کاربری مشترک جهت ایجاد یک دید مشترک بهره می برد.

داده ها همچنان بر روی سیستم های اطلاعاتی مربوطه به صورت جداگانه وجود دارند. بنابراین یکپارچه سازی

وهمگون سازی داده ها همچنان توسط کاربران انجام می شود (همانند موتورهای جستجو).

یکپارچه سازی به کمک برنامه های کاربردی^{۴۱}: در این روش از برنامه های کاربردی جهت یکپارچه

سازی داده ها استفاده می شود. این برنامه های کاربردی به منابع اطلاعاتی گوناگونی دسترسی داشته و نتایج

³⁹ Manual Integration

⁴⁰ Common User Interface

یکپارچه شده را به کاربران باز می گردانند. این روش تنها برای تعداد کمی از سیستم ها عملی است چرا که، هر اندازه که برنامه های کاربردی توسعه یافته و بزرگ گردند، به همان میزان نیز تعداد واسط های کاربری سیستم و فرمت های داده ای که جهت یکپارچه سازی و همگون سازی داده ها استفاده می گردند نیز، گسترش یافته و پیچیده تر می گردند.

یکپارچه سازی به کمک میان افزارها^{۴۲}: میان افزارها عموماً ارائه دهنده قابلیت هایی هستند که صرفاً در جهت حل مشکلات یکپارچه سازی استفاده می گردد^{۴۳}. با استفاده از میان افزارها، برنامه های کاربردی تلاشی در جهت پیاده سازی قابلیت های عمومی یکپارچه سازی صرف نمی کنند اما هنوز بسیاری از تلاش های یکپارچه سازی باید توسط برنامه های کاربردی انجام شود^{۴۴}. علاوه بر این معمولاً باید میان افزارهای گوناگون در جهت ساخت سیستم های یکپارچه نیز با یکدیگر ادغام گردند.

دسترسی متحد داده^{۴۵}: در این روش یک یکپارچه سازی منطقی بر روی سطح دسترسی داده ها انجام می شود. یک برنامه کاربردی عمومی با ایجاد یک دید سراسری بر روی داده های فیزیکی توزیع شده، این تصور را ایجاد می کند که داده ها به صورت مجازی در این سطح موجود می باشند. سیستم یکپارچه ساز داده ها بایستی بتواند دسترسی داده، یکپارچه سازی و همگون سازی داده ها را در کمترین زمان ممکن و به صورت Runtime انجام دهد.

⁴¹ Integration by Application

⁴² Integration by Middleware

⁴³ SQL Middleware مانند

⁴⁴ به عنوان نمونه SQL Middleware یک نقطه دسترسی واحد جهت ارسال پرس و جوهای SQL را برای تمامی سیستم های اتصالی فراهم می آورد اما عمل

یکپارچه سازی و همگون سازی نتایج به دست آمده را انجام نمی دهد.

⁴⁵ Uniform Data Access

ذخیره سازی داده به صورت مشترک^{۴۶}: در این روش یکپارچه سازی فیزیکی داده به وسیله انتقال داده ها به یک مکان ذخیره سازی جدید انجام می گیرد. منبع ذخیره ساز قدیم داده ها ممکن است همچنان فعال باقی مانده یا آنکه بازنشسته گردد. به طور کلی یکپارچه سازی فیزیکی داده باعث دسترسی سریعتر به داده ها می گردد. اگر منبع قدیمی ذخیره ساز داده ها بازنشسته گردد باید برنامه های کاربردی که از این منبع استفاده می کردند، از این پس به منبع جدید داده ها رجوع نمایند. اما اگر منبع قدیمی ذخیره ساز داده ها همچنان فعال باقی بماند، بایستی عملیات بروز رسانی بر روی مکان ذخیره ساز داده های مشترک به صورت متناوب انجام گیرد.

۲-۲) نمونه هایی از روش های یکپارچه سازی

پایه و اساس راه حل های یکپارچه سازی ارائه شده بر اساس شش روش معمول یکپارچه سازی است که بدان اشاره گردید. مهمترین نمونه های این روش ها عبارتند از:

- سیستم های پرس و جوی میانجی^{۴۷} این سیستم ها با فراهم آوردن یک نقطه دسترسی فقط خواندنی برای پرس و جوها، امکان دسترسی همسان به منابع داده ای گوناگون را فراهم می آورند. در این حالت از یک سیستم واسط [Wiederhold, 1992] که حاوی یک پردازشگر عمومی پرس و جوهاست جهت ارسال زیر پرس و جوها به منابع داده ای و ادغام پاسخ آن ها استفاده می گردد.

⁴⁶ Common Data Storage

⁴⁷ Mediated query systems

- پرتال‌ها^{۴۸} نوع دیگری از دسترسی همسان به منابع داده ای هستند. این سیستم ها با فراهم آوردن درگاه های اینترنتی یا اینترنتی، این امکان را برای کاربران فراهم می آورند که به اطلاعات مناسب خود، متناسب با نیاز خویش دسترسی یابند. معمولاً از وب کاوی^{۴۹} و با استفاده از تحلیل جریان کلیک های کاربران در جهت شناسایی User-Profile استفاده می گردد. اطلاعات مورد علاقه کاربران ممکن است از این راه به دست آمده و یا حاضر گردند.
- انبارهای داده^{۵۰} نوعی روش یکپارچه سازی مبتنی بر معماری ذخیره سازی داده به صورت مشترک می باشد. در این روش داده های موجود در منابع داده ای فعال به داخل یک انبار داده استخراج، تبدیل و بارگذاری^{۵۱} می گردند. پس از انجام این فرآیند، تحلیل هایی از قبیل پردازش تحلیلی برخط^{۵۲} می تواند بر روی دسته داده های یکپارچه و مجتمع شده، انجام گیرد.
- ذخیره ساز داده های عملیاتی^{۵۳} نوع دیگری از روش یکپارچه سازی مبتنی بر معماری ذخیره سازی داده مشترک می باشد. در اینجا انبار داده به همراه داده های بروز به وسیله انتشار فوری تغییرات منابع داده ای محلی به منابع داده ای، ساخته می شود. بنابراین با وجود داده های یکپارچه و بروز یافته، امکان تصمیم گیری صحیح وجود خواهد داشت. در این روش داده ها برخلاف روش انبار داده، مجتمع و پاک نبوده و امکان در اختیار داشتن سابقه داده ها وجود ندارد.

⁴⁸ Portals

⁴⁹ Web Mining

⁵⁰ Data Warehouses

⁵¹ Extracted, Transformed and Loaded (ETL)

⁵² Online Analytical Process (OLAP)

⁵³ Operational Data Stores

- سیستم پایگاه داده های متحد^{۵۴} روشی مبتنی بر معماری دسترسی متحد داده می باشد. در این روش یک یکپارچه سازی منطقی بر روی DBMS محلی و زیرین انجام می گیرد. سیستم پایگاه داده های متحد (FDBMS) یک DBMS کاملاً تکامل یافته اند. به این معنا که برای این دسته از سیستم ها مدل داده ای خاص، سیستم پشتیبان پرس و جوهای عمومی، سیستم انجام تراکنش های عمومی و سیستم کنترل دسترسی عمومی پیاده سازی می گردد. معمولاً جهت ساخت FDBMS از پنج سطح معماری ارائه شده توسط [Sheth and Larson, 1990] استفاده می گردد.
- سیستم مدیریت گردش کار^{۵۵} این سیستم امکان اجرا شدن فرآیندهای کاری را فراهم می آورد که هر کدام به تنهایی و به ترتیب توسط برنامه های کاربردی یا کاربران مختلف انجام می شوند. عموماً سیستم های مدیریت گردش کار از مدل سازی، اجرا و نگهداری فرایندهایی که میان برنامه های کاربردی و کاربران انسانی رد و بدل می گردند، پشتیبانی می نمایند. سیستم های مدیریت گردش کار مبتنی بر معماری یکپارچه سازی به کمک برنامه های کاربردی می باشند.
- یکپارچه سازی به وسیله سرویس های تحت وب^{۵۶} در این حالت یکپارچه سازی میان دو نرم افزار که بر روی ماشین هایی که از طریق یک شبکه با هم در تماس بوده و این تماس توسط پیام های مبتنی بر XML (که به کمک پروتکل های مبتنی بر اینترنت منتقل می گردند) انجام می شود، صورت خواهد گرفت. بر اساس نوع کاربردی که یکپارچه سازی ارائه خواهد داد، معماری سرویس های تحت وب می تواند مبتنی بر یکی از معماری های دسترسی متحد داده، واسط کاربری مشترک، دستی و یا یکپارچه سازی به کمک برنامه های کاربردی باشند.

⁵⁴ Federated database systems (FDBMS)

⁵⁵ Workflow Management systems (WFMS)

⁵⁶ Integration by web services

- جفت به جفت⁵⁷ این روش یکپارچه سازی یک روش غیر متمرکز جهت یکپارچگی داده هایی

است که میان دو جفت منبع داده ای مستقل و توزیع شده به اشتراک گذاشته شده و یکپارچه

گردیده اند. بر اساس نوع کاربردی که یکپارچه سازی ارائه خواهد داد، معماری روش جفت به

جفت می تواند مبتنی بر یکی از معماری های دسترسی متحد داده، واسط کاربری مشترک، دستی

و یا یکپارچه سازی به کمک برنامه های کاربردی باشد.

با توجه به اینکه کاربران از برنامه های مختلف با داده هایی که دارای ساختار و معنای متفاوت هستند

استفاده می کنند، یکپارچه سازی داده کماکان با مشکلات زیادی روبرو است [Bruijn , 2003] و اکثر راه حل

هایی که ارائه شده همگی موردی و برای حل مسئله یکپارچه سازی در یک حوزه خاص بوده است

[Tierney , 2004]. فناوری ها و روش های مختلفی برای برطرف کردن ناهمگونی های ارائه شده است.