به نام خدا

Datawarehousing

انبار داده (انباره داده)

استاد راهنمای کارورزی : سرکار خانم ترابی

ارائه دهنده : مینا فرنقی

1390

مقدمه

تصمیم گیری سازمانی به طور ایده آل با آنالیز عمیق اطلاعات مدیریتی آغاز می شود . سپس تصمیم گیرندگان از اطلاعات موجود برای سنجش راهکارها ، آنالیز گزینه ها، پیش بینی تأثیرات و نمایش نتایج در حوزه سازمان و محیط استفاده می کنند . اما اگر تصمیم گیرندگان ، اطلاعات مربوط ، قابل اطمینان و به هنگام در اختیار نداشته باشند ، کیفیت و اعتبار تصمیمات آنها مورد تردید خواهد بود .

مدیران در سازمان ها و آموزش عالی همیشه لازم است آنها متغیرهای فراوانی را به طور همزمان مد نظر قرار داده و تأثیر تصمیماتشان بر مؤلفه های داخلی و خارجی را بسنجند . این متغیرها و مؤلفه ها در چهار دسته یا گروه اصلی قرار می گیرند :

- پیچیدگی سازمان ؛
- تعداد فراوان افراد دخیل و ناظران ؛
- بازار رقابتی ؛
- محدودیت منابع .

■ پیدایش انبار داده

انبار داده ای در اوایل دهه 90 به عنوان یک تکنولوژی پشتیبان تصمیم گیری پدیدار شد که می توانست داده های منابع گوناگون و متمایز را با هم ادغام کند و جهت گیری خاصی را در طریقه سازماندهی و ارائه داده ها به وجود آورد . تا اواسط دهه 90 ،کاملا مشخص شده بود که ایجاد یک انبار داده سازمانی بسیار دشوار است ، و تمرکز موجود به دیتا مارت های سازمانی منتقل شد . همچنین تغییر دیگری که در حال اتفاق افتادن بود ؛ ایده ایجاد یک محیط گزارش دهی به جای یک روح ادغام کننده بود ... ؛ تمرکز بر گزارش نویسی به عنوان یک هدف منجر به افزایش فشار برای گزارش دهی عملیاتی و پشتیبانی از تصمیم شد .
یکپارچه سازی با داده های خارجی و نیز دیدگاه های موجود در طول زمان را برای آنالیز روند ها پیشنهاد می کنند .

وظیفه اصلی انبار داده ، تسهیل یکپارچه سازی چند کاربردی و افزایش یکپارچگی نتایج است. تمرکز باید از ذخیره سازی داده ها به تسهیل جریان اطلاعات تغییر کند و ماهیت یکپارچه سازی ، تغییر به همراه ایجاد ارتباط است . (Haisten , 2002)

■ رایجترین تعریف انبار داده عبارتست از:

"یک پایگاه داده ای غیر فرار ، متغیر زمانی ، یکپارچه و موضوع محور که می تواند پشتیبانی از تصمیم را انجام دهد . "

عبارت انبار داده شامل ایجاد و نگهداری یک مجموعه داده ای و فرآیند اکتساب اطلاعات سودمند از داده های ذخیره شده است . به عبارت ساده تر ، انبار داده روشی است برای فرآیند جمع آوری داده ای از انواع مختلف که برای سازمان داخلی یا خارجی هستند .

■ تفاوت انبار داده و پایگاه داده

وظیفه اصلی پایگاه داده ، پشتیبانی از تراکنش های آنلاین است که بیشتر عملیات روزمره سازمان را پوشش می دهند . در سوی دیگر انبار داده برای اهداف پشتیبانی از تصمیم است که داده های تاریخی بلند مدت را نگهداری و به کاربران خدماتی در نقش تحلیل گر داده و تصمیم گیرنده ارایه می کند . چنین سیستم هایی می توانند داده ها را در قالب های مختلف برای هماهنگ کردن نیازهای مختلف کاربران مختلف ، سازماندهی و ارایه کنند . این سیستم ها با نام سیستم های پردازش تحلیلی آنلاین(OLAP) شناخته می شوند . موارد تفاوت پایگاه داده و انبار داده به شرح زیر است :

* از لحاظ مدل های داده ؛
* از لحاظ عملیات قابل اجرا بر روی آنها ؛
* از لحاظ مقدار داده ها ؛
* از لحاظ زمان پرس و جو .

	پایگاه داده	انباره داده
مدل های داده	**از مدل داده رابطه ای استفاده می کنند ، بر اساس دو مفهوم اساسی موجودیت و رابطه**	**برای پردازش تحلیلی بر خط ، طراحی شده و امکان پردازش تعداد کمی پرس و جوی پیچیده بر روی تعداد بسیار زیادی رکورد داده فراهم می شود .**
عملیات قابل اجرا بر روی آنها	**به طور عمومی شامل عملیات روزآمدسازی**	**عمده عملیات قابل اجرا بر روی انبار داده ، عمل خواندن از انبار داده است .**
مقدار داده ها	**در حدود چند صد مگا بایت تا چند گیگا بایت**	**در حدود چند صد گیگا بایت تا چند ترا بایت**
زمان پرس و جو	**با توجه به حجم و مقدار داده ها ، سرعت پردازش بالاتری دارد**	**با استفاده از دو تکنیک تجمیع و سلسله مراتبی کردن فیلدها سرعت انجام پرس و جو ها بهبود بخشیده شده**

سیستم های پشتیبانی از تصمیم در مقابل سیستم های عملیاتی

- سیستم های عملیاتی برای تهیه اطلاعات استراتژیک تولید نشده اند بلکه برای رسیدن به این اطلاعات باید اطلاعات را از مجموع انواع مختلف این سیستم ها گرفت که این کار را فقط " سیستم های پشتیبان از تصمیم " با سیستم های اطلاعاتی خاص می توانند انجام دهند .
- سیستم های پشتیبان از تصمیم برای اجرای فعالیت های اصلی سازمان طراحی و ساخته نشده اند بلکه برای مشاهده چگونگی انجام فعالیت ها و سپس تصمیم گیری های اجرایی برای بهبود این فعالیت ها در نظر گرفته شده اند .

•تفاوت سیستم های پشتیبان از تصمیم با سیستم های عملیاتی در این است که سیستم های پشتیبان برای بیرون کشیدن اطلاعات استراتژیک از پایگاه داده ها توسعه یافته اند یا به عبارتی برای تولید اطلاعات استراتژیک می باشند اما سیستم های عملیاتی برای وارد کردن داده ها به پایگاه داده ها طراحی شده اند .

انبار داده در آموزش عالی

گان در سال 2002در مقاله ای چشمگیر در خصوص نقش انبار داده در آموزش عالی ،به بررسی چالش هایی می پردازد که دانشگاه ها و دانشکده ها در مدیریت اطلاعات ضروری برای برنامه ریزی و تصمیم گیری با آن مواجه می شوند ؛ سپس او انبار داده ای را به عنوان یک روشی برای مدیریت دانش در محیط های آکادمیک معرفی می کند . او در توصیف محیطی برای تغییر ، فشار های وارد بر سیستم های اطلاعاتی را از سه حوزه می داند :

ابتدا آنکه مجریان و سیاست گذاران در همه سطوح به دنبال استراتژی های بهینه مدیریت داده برای پشتیبانی از مدیریت منابع و برنامه ریزی استراتژیک هستند . دوم آنکه ، به واسطه وجود یک محیط رقابتی متغیر برای آموزش عالی به طور کلی ،مجریان و مدیران تشنه اطلاعاتی اند که بتوانند به مؤسسه مربوطه در جذب دانشجو کمک کنند . سوم آنکه ، سازمان های نظارتی خارجی ، مانند دولت فدرال ،دولت های محلی و سازمان های ارائه مجوز به دنبال اطلاعاتی پیرامون کارکرد سازمان ها و برنامه های آنها در دامنه وسیعی هستند .

انبار داده در کتابخانه ها

کتابخانه ها اغلب با مشکلاتی نظیر :

•نرم افزارهای کامپیوتری که غالبا فقط برای پرس و جوهای معمولی و پشتیبانی از مسائل مدیریتی و برنامه ریزی کوتاه مدت اداری جوابگو هستند

•مدیریت کارآمد بار سنگین داده ها که دائما نیز در حال افزایش است

•درون حجم عظیم داده ها، الگوها و روابط بسیار جالبی میان پارامترهای مختلف بصورت پنهان باقی میماند .

واضح است که انبار داده اساسا" برای پرس و جوهای پشتیبان تصمیم گیری ساخته شده است. بر این اساس سازماندهی وعملیات انبار داده چنان طراحی شده اند تا نیازهای اطلاعاتی روزمره یا معمولی را پاسخگو باشند. بدلیل حجم بسیار بالای چنین پایگاه اطلاعاتی یک سیستم کامپیوتری پیشرفته برای عملیات انبارسازی داده ها لازم است. همچنین یک بانک اطلاعات مجزا شامل ابرداده که مشخصه هایی نظیر نوع، فرمت، مکان و پدیدآورندگان داده های ذخیره شده در یک انبار داده ها را توصیف میکند نیز برای کمک به کاربران و مدیران داده ها ساخته میشود. مشخص شد که انبار داده بدلیل اندازه و تنوعش، اگر مبتکرانه پردازش شود میتواند به تولید اطلاعاتی منجر شود که در وهله اول آشکار نیستند. با انتخاب متناسب داده ها، بکار گرفتن فنون مختلف غربال کردن و تفسیر زمینه ای ,داده ذخیره شده میتوانست منجر به کشف الگوها یا رابطه هایی شود که بینش نویی به تصمیم گیرنده دهد .

■ انبار داده و کتابخانه دیجیتال

با انتشار اینترنت ، مقدار زیادی از اسناد برای جستجو و بازیابی بر روی وب در دسترس است و اینترنت در حال حاضر یکی از بزرگترین مخازن اطلاعات به شمار می رود . با این حال ، محتوای آن دارای آشفتگی و توزیع شده است. علاوه بر این ، سیستم عامل سخت افزار و نرم افزار های گوناگون ، و همچنین اسناد با فرمتهای مختلف و رسانه های گوناگون در دسترس ، یک پایگاه داده بزرگ ناهمگن ساخته اند ، که حاوی داده های ساختار ، نیمه ساختار یافته و غیر ساختار یافته است. همه این توزیع و ناهمگونی در جستجو و کسب محتوای وب دشواری هایی ایجاد کرده اند.

در چند دهه اخیر کتابخانه های دیجیتال (DL) ، به عنوان حوزه های پژوهشی ، با هدف سازماندهی و ارتقاء دسترسی آسان تر به اسناد موجود در وب مطرح اند .تعاریف بسیاری از نظر ادبیاتی نیز برای آنها وجود دارد و هیچ اتفاق نظری در خصوص مفهوم DL وجود ندارد ، آنچه تقریبا در میان تعاریف به آنها بسیار اشاره شده شامل : یک مجموعه بزرگ ،با فرمت های متنوع دیجیتال ، مداوم ، مدیریت شده و به خوبی سازماندهی شده با استفاده از یک شیوه فهرست نویسی و با دسترسی از طریق وب در نظر گرفته شده است .

توسعهDL به طور کلی متضمن یکپارچه سازی محتوای توزیع شده چند رسانه ای در وب است.از آنجا که ماهیت ابر رسانه وب دلالت بر هدایت از طریق متن به منظور رسیدن به اطلاعات مورد نظر است، برای سازمان دهی داده های یکپارچه باید یک طبقه بندی در یک سلسله مراتب خاص را در نظر گرفت .

در طرح های مختلف که در حال توسعه DLهای اند ، پیشنهاداتی برای حل مشکل ادغام محتوا و همچنین دسترسی به این محتویات با استفاده از طبقه بندی سلسله مراتبی وجود دارد.

برخی از پروژه های آثار مرتبط عبارتند از : فن آوری کتابخانه دیجیتال دانشگاه استنفورد ، کتابخانه دیجیتال طرح پروژه ایلینویز ، پروژه کتابخانه دیجیتال اسکندریه و پروژه کتابخانه دیجیتال دانشگاه میشیگان . با این حال لازم به ذکر است که طرح های پیشنهادی جامعی برای رسیدگی به مسائل مربوط به DL وجود نداشته است.

یکی از حوزه تحقیقی که به حل مشکلات پیچیده پایگاه داده کمک کرده است ، حوزه انبار داده ها (DWing) است.رویکرد DWing برای رسیدگی به مسائل مربوط به یکپارچه سازی داده ها و جستجو های پیچیده بسیار مفید می باشد.

روش DWing در توسعه DL

توسعه DL بر اساس رویکرد DWing بر درک معماری DWing و چگونگی استفاده و / یا فرآیندهای آن و اجزای DL اشاره دارد .

انبار داده دارای مشخصات زیر است :

موضوعی هستند : بر خلاف داده های سیستم های عملیاتی که بر حسب برنامه های کاربردیشان ذخیره می شوند ؛ داده های موجود در انبارهای داده بر حسب موضوع های کاری ذخیره می شوند . یعنی بر یک محور خاص مانند مشتری تأکید دارند .

یکپارچه هستند : یعنی همه داده های مربوط به یک موضوع با هم ترکیب و آنالیز می شوند . بدین صورت که داده ها از سیستم های عملیاتی مختلف جمع آوری ، تناقضات بین آنها رفع و به صورت مناسب ذخیره می گردند .

دارای تغییرات زمانی هستند : از آنجایی که داده های موجود در انبارهای داده به منظور تحلیل و تصمیم گیری می باشند بنابراین بر خلاف سیستم های عملیاتی که دارای ارزشهای جاری هستند داده های انبار داده به صورت تاریخی یا به عبارتی بهتر تاریخ دار می باشند یعنی تاریخچه ای از داده ها با جزئیات نگهداری می شوند . به طور مثال می توانیم

تاریخچه ای از یک نفر یا یک موضوع داشته باشیم. این بدین معناست که در ساختار همه داده های موجود در انبار داده عنصر زمان وجود دارد .

تغییر ناپذیرند : یعنی داده ها فقط قابل خواندن هستند و کاربران نمی توانند هیچ گونه تغییری در آنان ایجاد کنند .

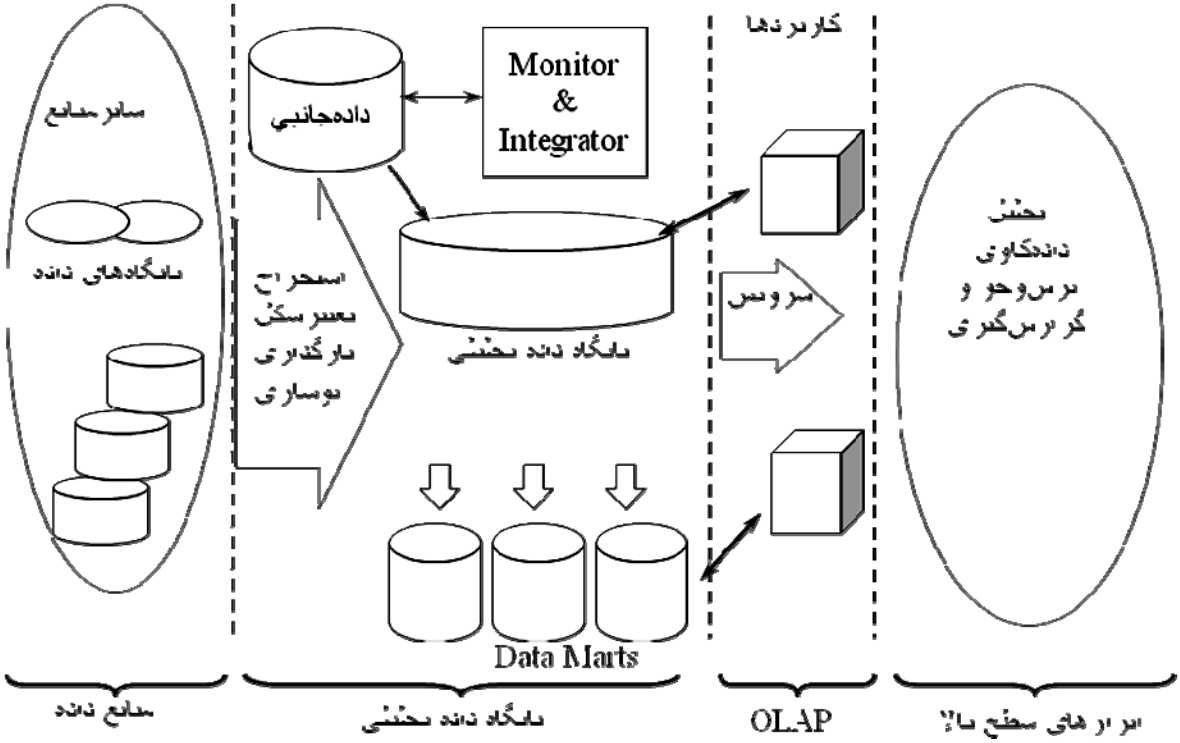
به طور کلی ، معماری سیستم های مبتنی بر DW شامل یکپارچه سازی داده های کنونی و تاریخی است. منابع داده ها می تواند داخلی باشد (سیستم های عملیاتی شرکت / موسسه) و یا خارجی (حاوی داده ها مکمل نشات گرفته از سازمان ، از جمله شاخص های اقتصادی). به طور کلی ، یکپارچه سازی داده ها با مدل های مختلف داده ها ، تعاریف و / یا سیستم عامل های مختلف طرح می شود. این ناهمگونی ، وجود برنامه های کاربردی را برای استخراج و تبدیل داده ها می طلبد که یکپارچه سازی داده ها را امکان پذیر سازد . پس از یکپارچه سازی اطلاعات، اطلاعات جدیدی را در یک پایگاه داده جدید ذخیره می شود که DW ترکیبی از نقطه های (نظر های) مختلف برای حمایت از تصمیم گیری های مدیریت است ایجاد می شود . این پایگاه داده برای تجزیه و تحلیل داده ها توسط کاربران نهایی استفاده می شود.

■ DWing از رویکرد Dling

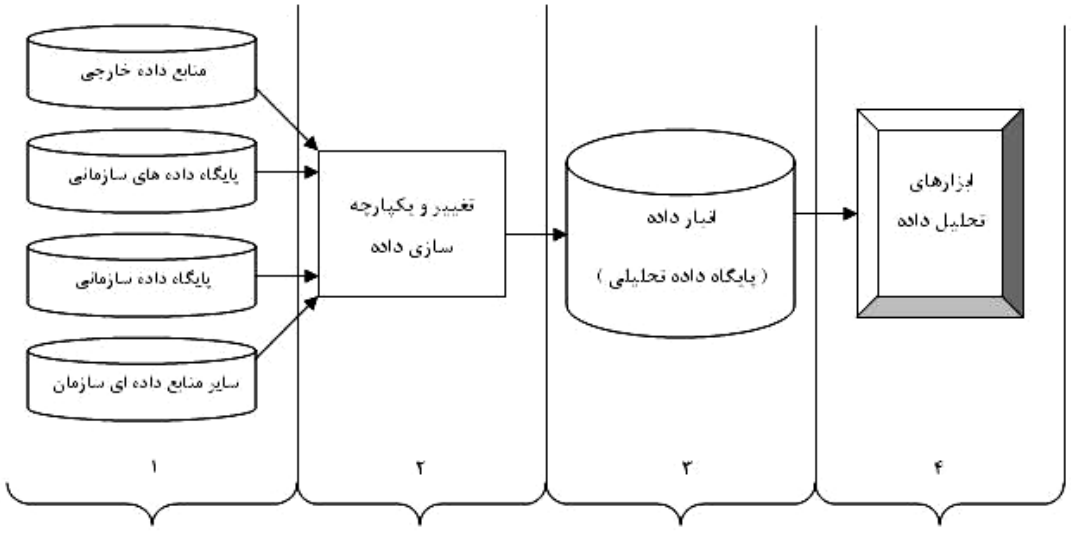
با استفاده از مفاهیم مرتبط با DWنشان داده شده بالا در روند DL ، مشاهده می کنیم که :

- یک DL باید موضوع گرا باشد ؛
- یک DL باید یک دیدگاه یکپارچه از اسناد داشته باشد ؛
- اسناد و ابرداده متناظر با آن باید فقط یک بار در DLبارگذاری شود ؛
- اسناد در DL ذخیره می شود و نسخه های دیگر می تواند افزایش یافته شوند ؛
- در نهایت ، اگر چه DL لزوما برای پشتیبانی از تصمیمات مدیریت لازم نیست ، از آن برای حمایت از روند تصمیم گیری در پژوهش می توان استفاده کرد .

■ شمای کلی معماری مخازن داده



معماری دو لایه انبار داده



معماری دو لایه

شمایی از این معماری در شکل 1 نشان داده شده است . این معماری از 4 لایه تشکیل شده است :

1. **در لایه یک منابع انبار داده شامل داده های از جمله پایگاه های داده قرار دادند که منبع تامین داده های انبار داده محسوب می شوند و داده های مورد نیاز از آنها تامین می شود .**

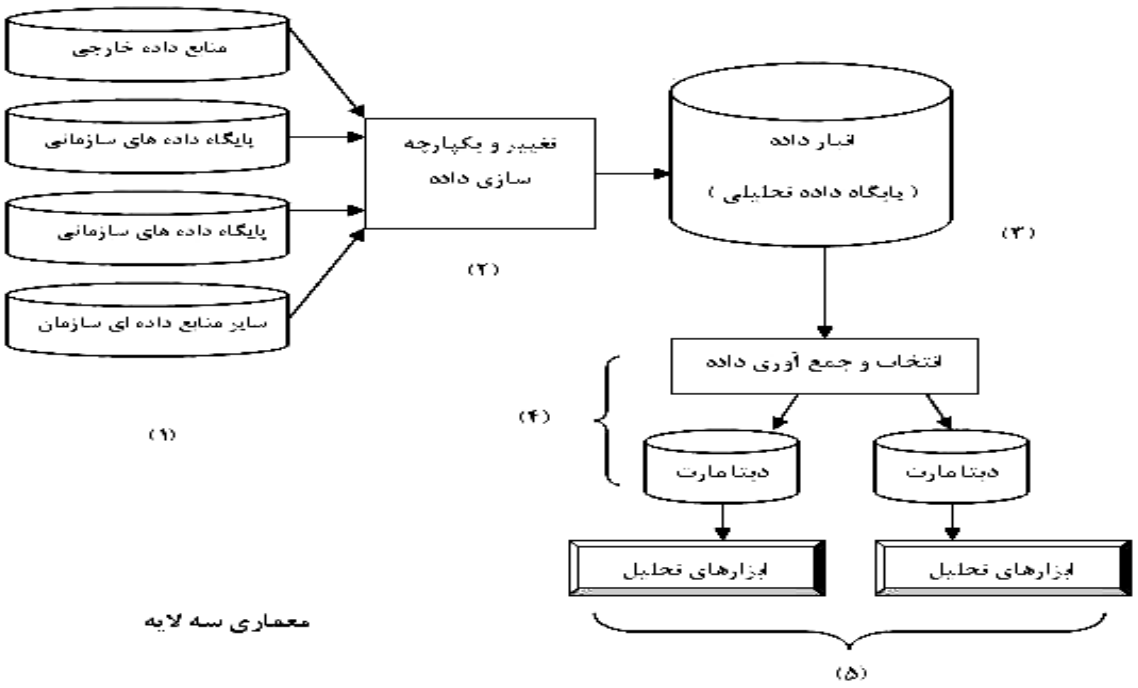
این منابع بر حسب نیاز می تواند قسمتی یا تمامی داده های داخل سازمان باشد و در صورت لزوم از داده های خارج از سازمان نیز بر حسب نیاز استفاده می شود .

2. **عملیات لازم از قبیل تمیز سازی و یکپارچه سازی بر روی داده ها قبل از بارگذاری صورت می پذیرد و در واقع فرآیند ای تی ال در این قسمت انجام می شود و در واقع این بخش در موفقیت یا شکست انبار داده تاثیر بسیار دارد و تقریبا 70 درصد فعالیت ها و منابع پروژه ، برای پیاده سازی و نگهداری انبار داده در همین بخش صرف می گردد .**
3. **انبار داده به وجود می آید که شامل داده های جزئی و خلاصه شده و ابر داده است و به آن پایگاه داده تحلیل نیز گفته می شود و در واقع مخزن اصلی داده هایی است که پس از انجام عملیات و اصلاحات لازم ذخیره می شوند .**
4. **این قسمت شامل ابزارهای تحلیل داده و داده کاوی می باشد که کاربران با استفاده از آنها الگو ها و دانش نهفته در داده ها را استخراج می نمایند .**

■ معماری سه لایه انبار داده

این معماری شامل پنج مرحله می باشد :

1. **داده ها از فایلها و بانکهای اطلاعاتی مختلف گرفته می شوند .**
2. **عملیات لازم از قبیل تمیز سازی و یکپارچه سازی بر روی داده ها قبل از انتقال داده ها صورت می پذیرد .**
3. **انبار داده به وجود می آید که شامل داده های جزئی و خلاصه شده و ابر داده است و به آن پایگاه داده تحلیلی نیز گفته می شود.**
4. **در این قسمت دیتا مارت ها قرار دارند . ابر داده یک انبار داده کوچک است که برای گروه خاصی از تقاضاها طراحی شده است و داده های آن متناسب با نیاز آن گروه می باشد و با انتخاب و خلاصه کردن داده های انبار داده به دست می آیند . برای مثال می توان برای اداری ،مالی و فروش دیتا مارت های جداگانه ای تشکیل داد .**
5. **این قسمت شامل ابزارهای تحلیل داده و داده کاوی می باشد که کاربران با استفاده از آنها الگوها و دانش نهفته در داده ها را**



معماری سه لایه

استخراج می نمایند .

■ تفاوت معماری دو لایه انبار داده و سه لایه

در معماری دو لایه ، داده ها پس از ذخیره سازی در انبار داده مستقیما توسط ابزارهای تجزیه و تحلیل داده مورد بررسی قرار می گیرند اما در معماری سه لایه ، داده ها پس از ذخیره سازی در انبار داده در دیتا مارت های کوچک تری ذخیره می شوند . دیتا مارت یک انبار داده کوچک است که برای گروه خاصی از تقاضاها طراحی شده است و داده های آن متناسب با نیاز آن گروه می باشد و با انتخاب و خلاصه کردن داده های انبار داده به دست می آید و سپس توسط ابزارهای تحلیل داده استفاده می شود . برای مثال می توان برای اداری ، مالی و فروش دیتا مارت های جداگانه ای تشکیل داد .

■ نیاز به انبار داده

هدف اصلی مدیریت اطلاعات فراهم نمودن :

اطلاعات مناسب به شکل مناسب و در زمان مناسب ؛

ذخیره سازی و بازایابی اطلاعات برای مدیران

داده هایی که در سیستم های تراکنشی تولید می شود ، قابلیت تحلیل پذیری اندکی دارند . اگرچه سیستم ها در زمینه رفع نیازهای پردازشی کاربران موفق بوده ولی در زمینه برآورد نیازهای تصمیم گیری مدیریت ضعیف بوده است .

انبار داده به عنوان زیرساختی برای سیستم های پشتیبان تصمیم گیری نیز مطرح می باشد و برای ساخت آن باید قالب سخت افزاری و نرم افزاری مطابق با ملزومات مورد نیاز در نظر گرفته شود و به طور کلی سه موضوع را باید در این خصوص در نظر داشت :

داده های مرتبط با هم باید یکپارچه و مجتمع شود .

داده باید هنگام یکپارچه شدن با سایر داده ها مفهوم خود را حفظ نماید .

تکنولوژی ذخیره سازی داده ها باید به نحوی باشد که مستقل از نوع سیستم عاملی که انبار داده بر روی آن قرار گرفته بتوان از این سیستم استفاده نمود .

■ اهداف سیستم های انبار داده

- ❖ تحلیل تفضیلی داده های سازمانی
- ❖ ایجاد ابزاری در دست مدیران برای پیش بینی بازار و بالا بردن توان رقابتی سازمان

- بالا بردن میزان سوددهی از طریق تجزیه و تحلیل داده ها و کمک به تصمیم گیری بهتر مدیریت
- شناسایی مشتریان دائمی و حفظ آنها از طریق ذخیره سازی و تحلیل داده های مربوط به روند خرید مشتریان
- بالا بردن توان برنامه ریزی دقیق در مدیران
- توانایی استفاده مناسب از منابع اطلاعاتی موجود در سازمان
- توانایی فراهم کردن اطلاعات با کیفیت بالا
- بالا بردن سرعت تصمیم گیری مدیران
- تشخیص زود هنگام تهدیدات و فرصت ها

مدیریت امنیت داده های انبار داده

مدیریت امنیت انبار داده زیر مجموعه ای از مدیریت تکنولوژی اطلاعات می باشد .امنیت انبار داده و به طبع آن امنیت داده های سازمان در واقع اصلی ترین چالش مدیران و مسئولان فناوری اطلاعات سازمان می باشد . امروزه گسترش استفاده از سیستم های تحت شبکه موجب گردیده که مباحث مربوط به امنیت انبار داده دارای اهمیت بسیار بالایی باشد و لزوم حفاظت صحیح از داده های آن ضروری تر به نظر می رسد . برای رسیدن به این اهداف هر سازمان بسته با ارزش داده های خود ، نیازمند پیاده سازی یک سیاست کلی جهت مدیریت امنیت داده ها می باشد .

امنیت داده ها به چهار مفهوم کلی قابل تقسیم است :

محرمانگی (Confidentially) : محرمانگی اطلاعات یعنی حفاظت از اطلاعات در مقابل دسترسی و استفاده غیر مجاز ، داده های محرمانه تنها توسط افراد مجاز قابل دسترسی می باشند .

یکپارچگی در تغییر داده ها (Integrity) : در بحث امنیت اطلاعات ، یکپارچگی در تغییر داده ها به این معناست که داده ها نمی توانند توسط افراد غیر مجاز ساخته ، تغییر و یا حذف گردند .

اعتبار و سندیت (Authenticity) : اعتبار و سندیت دلالت بر موثق بودن داده ها دارد .

دسترس پذیری (Availability) : دسترس پذیری به این معنی می باشد که داده های انبار داده و سیستم های حفاظت امنیت داده ها ، در مواقع نیاز به اطلاعات در دسترس باشند .

نتیجه گیری

استفاده از انبار داده مزایای بالقوه زیادی را برای موسسات دارند ؛ مانند دسترسی دائمی به داده ها . موسسات به راحتی می توانند نقاط ضعف و قوت خود و رقبایشان را بیابند . همچنین آنها قادرند رابطه این اطلاعات را با روندهای بازار و محصولات مطلوب بیابند و اهداف دقیقی را جهت بازاریابی تعیین کنند . استفاده از انبار داده به سازمان ها این امکان را می دهد که از این اطلاعات برای اتخاذ تصمیمات صحیح استفاده کنند و این هدف نهایی فرآیند انبار داده است .

انبارهای داده به نقطه عطفی برای پشتیبانی تصمیم تبدیل شده اند . این یک شاهرگ حیاتی برای ارتقای قابلیت های پشتیبانی تصمیم است . انبار داده از بهترین سیستم های امروزی پشتیبانی تصمیم است ؛ و در آینده نیز سازمان ها را قادر خواهد ساخت که علمی تر ،جامع تر و حقیقت مدار تر در تصمیم گیری هایشان عمل کرده و سرعت و کیفیت تصمیمات را ارتقا بخشند . ارتقای قابلیت تصمیم گیری یک سازمان در افزایش رقابتی آن اثر بخش است .