

امروزه با گسترش سیستم های پایگاهی و حجم بالای داده های ذخیره شده در این سیستم ها ، نیاز به ابزاری است تا بتوان داده های ذخیره شده را پردازش کرد و اطلاعات حاصل از این پردازش را در اختیار کاربران قرار داد .با استفاده از ابزارهای گوناگون گزارش گیری معمولی ، می توان اطلاعاتی را در اختیار کاربران قرار داد تا بتوانند به نتیجه گیری در مورد داده ها و روابط منطقی میان آنها بپردازند اما وقتی که حجم داده ها خیلی بالا باشد ، کاربران هر چند زبردست و با تجربه باشند نمی توانند الگوهای مفید را در میان حجم انبوه داده ها تشخیص دهند و یا اگر قادر به این کار هم باشند ، هزینه عملیات از نظر نیروی انسانی و مادی بسیار بالا است .از سوی دیگر کاربران معمولاً فرضیه ای را مطرح می کنند و سپس بر اساس گزارشات مشاهده شده به اثبات یا رد فرضیه می پردازند ، در حالی که امروزه نیاز به روشهایی است که اصطلاحاً به کشف دانش بپردازند یعنی با کمترین دخالت کاربر و به صورت خودکار الگوها و رابطه های منطقی را بیان نمایند .

داده کاوی یکی از مهمترین این روش ها است که به وسیله آن الگوهای مفید در داده ها با حداقل دخالت کاربران شناخته می شوند و اطلاعاتی را در اختیار کاربران و تحلیل گران قرار می دهند تا براساس آنها تصمیمات مهم و حیاتی در سازمانها اتخاذ شوند .

۲-۱- عامل مسبب پیدایش داده کاوی

اصلی ترین دلیلی که باعث شده داده کاوی کانون توجهات در صنعت اطلاعات قرار بگیرد، مساله در دسترس بودن حجم وسیعی از داده ها و نیاز شدید به اینکه از این داده ها، اطلاعات و دانش سودمند استخراج کنیم. اطلاعات و دانش بدست آمده در کاربردهای وسیعی مورد استفاده قرار می گیرد.

داده کاوی را می توان حاصل سیر تکاملی طبیعی تکنولوژی اطلاعات دانست، که این سیر تکاملی ناشی از یک سیر تکاملی در صنعت پایگاه داده می باشد، نظیر عملیات جمع آوری داده ها و ایجاد پایگاه داده، مدیریت داده و تحلیل و فهم داده ها.

تکامل تکنولوژی پایگاه داده و استفاده فراوان آن در کاربردهای مختلف سبب جمع آوری حجم فراوانی داده شده است. این داده های فراوان باعث ایجاد نیاز برای ابزارهای قدرتمند برای تحلیل داده ها گشته، زیرا در حال حاضر به لحاظ داده ثروتمند هستیم ولی دچار کمبود اطلاعات می باشیم.

ابزارهای داده کاوی داده ها را آنالیز می کنند و الگوهای داده ها را کشف می کنند که می توان از آن در کاربردهایی نظیر تعیین استراتژی برای کسب و کار، پایگاه دانش^۳ و تحقیقات علمی و پزشکی، استفاده کرد. شکاف موجود بین داده ها و اطلاعات سبب ایجاد نیاز برای ابزارهای داده کاوی شده است تا داده های بی ارزش را به دانشی ارزشمند تبدیل کنیم .

۱

۳- داده کاوی و مفهوم اکتشاف دانش (K.D.D)

با حجم عظیم داده های ذخیره شده در فایلها، بانکهای اطلاعاتی و سایر بانک های داده ای، توسعه ی ابزارهایی برای تحلیل و شاید تفسیر چنین داده هایی و برای استخراج علوم شگفت انگیزی که می توانند در تصمیم گیری مفید باشند، امری بسیار مهم و ضروری است. داده کاوی با عنوان کشف دانش در پایگاه های داده (KDD) شناخته می شود. کشف علمی که قبلاً نا شناخته بوده اند و اطلاعاتی که در بانکهای اطلاعاتی موجود بوده و ذاتاً بالقوه و مفید هستند.

با وجود آنکه داده کاوی و کشف دانش در پایگاه های داده مترادف همدیگر هستند، ولی در اصل، داده کاوی ذاتاً بخشی و تنها قسمتی جزئی از فرآیند کشف دانش است. فرآیند کشف دانش در برگیرنده ی چندین مرحله می باشد که از اطلاعات خام، گونه هایی از علوم جدید را بدست می دهد. مراحل کشف دانش به قرار زیر است:

۱- پاک سازی داده ها : در این فاز داده های اضافی و نامربوط از مجموعه داده ها حذف می شوند.(داده های ناکامل)[2]

۲- یکپارچه سازی داده ها؟ چندین منبع داده ترکیب می شوند،

۳- انتخاب داده ها : انبار داده ها شامل انواع مختلف و گوناگونی از داده ها است که همه آنها در داده کاوی مورد نیاز نیستند . برای فرایند داده کاوی باید داده های مورد نیاز انتخاب شوند . به عنوان مثال در یک پایگاه داده های مربوط به سیستم فروشگاهی ، اطلاعاتی در مورد خرید مشتریان ، خصوصیات آماری آنها ، تامین کنندگان ، خرید ، حسابداری و ... وجود دارند . برای تعیین نحوه چیدن قفسه ها تنها به داده های در مورد خرید مشتریان و خصوصیات آماری آنها نیاز است . حتی در مواردی نیاز به کاوش در تمام محتویات پایگاه نیست بلکه ممکن است به منظور کاهش هزینه عملیات ، نمونه هایی از عناصر انتخاب و کاوش شوند .

۴- تبدیل داده ها : هنگامی که داده های مورد نیاز انتخاب شدند و داده های مورد کاوش مشخص گردیدند، معمولاً به تبدیلات خاصی روی داده ها نیاز است. نوع تبدیل به عملیات و تکنیک داده کاوی مورد استفاده

بستگی دارد، تبدیلاتی ساده همچون تبدیل نوع داده ای به نوع دیگر تا تبدیلات پیچیده تر همچون تعریف صفات جدید با انجام عملیاتهای ریاضی و منطقی روی صفات موجود.

۵- داده کاوی : بخش اصلی فرایند ، که در آن با استفاده از روش ها و تکنیک های خاص ، استخراج الگو های مفید ، دانش استخراج می شود.

۶-زیبایی الگو^۵: مشخص کردن الگوهای صحیح و مورد نظر به وسیله معیارهای اندازه گیری.

۷-زنجایی دانش : در این بخش به منظور ارائه دانش استخراج شده به کاربر ، از یک سری ابزارهای بصری سازی استفاده می گردد.

۱-۳-۱-تعریف داده کاوی

در متون آکادمیک تعاریف گوناگونی برای داده کاوی ارائه شده اند . در برخی از این تعاریف داده کاوی در حد ابزاری که کاربران را قادر به ارتباط مستقیم با حجم عظیم داده ها می سازد معرفی گردیده است و در برخی دیگر ، تعاریف دقیقتر که در آنها به کاوش در داده ها توجه می شود. برخی از این تعاریف عبارتند از :

- داده کاوی عبارت است از فرایند استخراج اطلاعات معتبر ، از پیش ناشناخته قابل فهم و قابل اعتماد از پایگاه داده های بزرگ که شامل بهره گیری از ابزارهای آنالیز داده ها، برای کشف الگوهای موجود و روابط ناشناختهی میان داده ها در حجمی وسیع می باشد. و استفاده از آن درتصمیم گیری فعالیتهای تجاری مهم.

- اصطلاح داده کاوی به فرایند نیم خودکار تجزیه و تحلیل پایگاه داده های بزرگ به منظور یافتن الگوهای

مفید اطلاق می شود. [3]

- داده کاوی یعنی جستجو در یک پایگاه داده ها برای یافتن الگوهایی میان داده ها [4].
- داده کاوی یعنی استخراج دانش کلان ، قابل استناد و جدید از پایگاه داده ها ی بزرگ .
- داده کاوی یعنی تجزیه و تحلیل مجموعه داده های قابل مشاهده برای یافتن روابط مطمئن بین داده ها

همانگونه که در تعاریف گوناگون داده کاوی مشاهده می شود ، تقریباً در تمامی تعاریف به مفاهیمی چون استخراج دانش ، تحلیل و یافتن الگوی بین داده ها اشاره شده است .