عنوان: **فرآیند داده کاوی در Microsoft SQL Server (بخش یک)**
نویسنده: محمد رجبی
تاریخ: ۱۳۹۳/۰۸/۱۷ ۱۶:۱۸
آدرس: www.dotnettips.info
گروه‌ها: Analysis Services, data mining, Microsoft SQL Server

مقدمه

بطور کلی داده کاوی به دو قسمت زیر تقسیم می‌شود:

1- اهداف توصیفی (Descriptive Goal): بدنبال یافتن الگوها و روابط بین داده‌ها هستیم، بدین ترتیب مدلی برای توصیف بهتر داده‌ها بدست خواهد آمد.

2- اهداف پیش بینانه (Predictive Goal): بدنبال انجام پیش بینی با استفاده از الگوها و مدل‌های فوق هستیم.

همچنین مراحل اجرای یک پروژه داده کاوی شامل مراحل زیر است:

1- تحلیل: مهمترین فعالیت در این فاز، فهم عمیق مسئله و شناخت درست مسئله و شناسائی مفاهیم کلیدی (Key Concept) در مسئله است.

2- طراحی: مهمترین فعالیت این فاز، فرموله کردن مسئله با استفاده از مفاهیم کلیدی است.

3- پیاده سازی/ نگهداری و بهبود

مراحل کاری داده کاوی بر اساس استاندارد CRISP-DM

محصول مشترک شرکت‌های SPSS, Teradata, NCR و دایملر- کرایسلر است و یک فرآیند استاندارد Cross-Industry برای داده کاوی است که به طور گسترده ای استفاده می‌شود. مراحل کاری در این مدل به شش فاز اصلی به شرح زیر تقسیم می‌شوند:

1. درک پروژه و فهم حوزه کاربرد (Business Understanding):
به طور صریح و آشکار اهداف و نیازمندی‌ها مشخص می‌شود. ترجمه اهداف و محدودیت آن در قاعده سازی، تعریف مسئله داده کاوی و مهیا کردن استراتژی اولیه برای نائل شدن به اهداف در این مرحله تعریف می شود.

2. انتخاب داده‌ها (Data Understanding):
این مرحله شامل جمع آوری داده‌ها برای استفاده از تحلیل اکتشافی و مشخص کردن اطلاعات اولیه برای ارزیابی داده‌های با کیفیت و انتخاب داده‌های مفید و مورد نیاز می‌باشد.

3. آماده سازی داده‌ها (Data Preparation):
آماده کردن داده‌های اولیه خام به داده‌های نهایی، این دادها در کلیه مراحل بعدی استفاده می‌شود و از این نظر این مرحله تحلیل و تلاش بیشتری را می‌طلبد. انتخاب عناصر و شناسه‌های تحلیل شده را برای کاوش داده‌ها اختصاص می‌دهیم و با تمیز کردن داده‌های خام آن را برای ابزارهای مدل سازی آماده می کنیم.

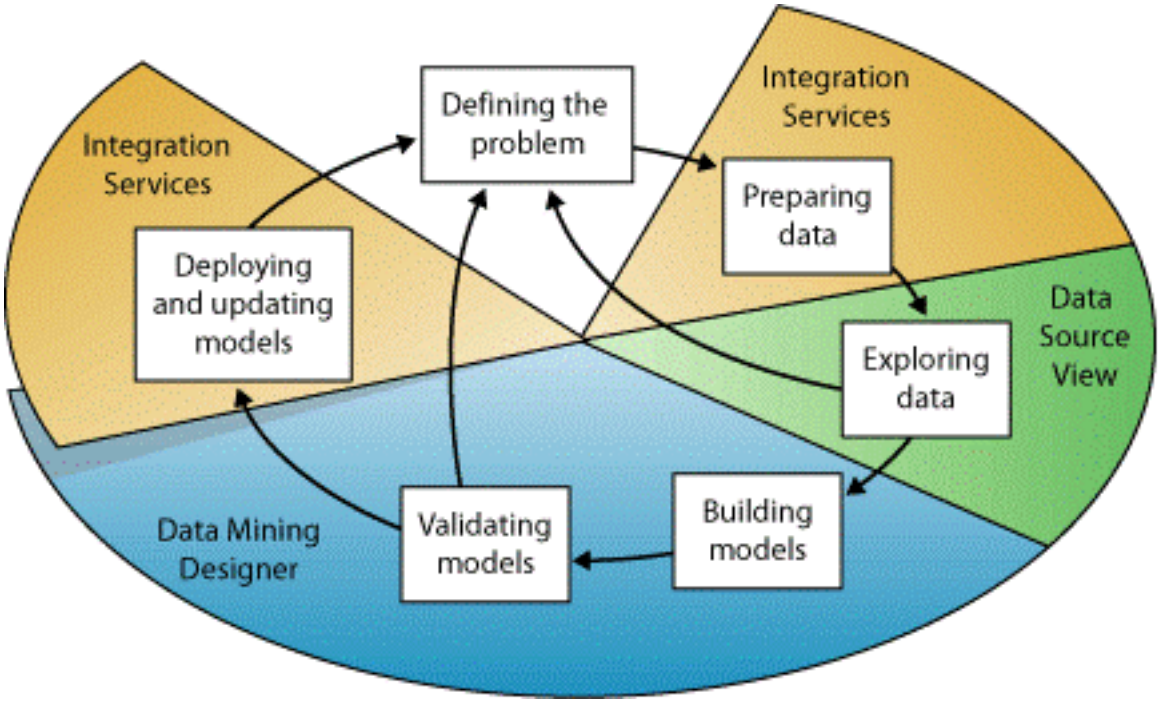
4. مدل سازی (Modeling):
با انتخاب و به کار بستن تکنیک‌های مدل سازی مناسب و روش داده کاوی معین نتایج مدل سازی را بهینه می کنیم، که در صورت نیاز می‌توانیم با برگشت به عقب تحلیل مدل سازی را بهینه‌تر نماییم.

5. ارزیابی (Evaluation):
مشخص کردن اینکه آیا مدل انتخابی، ما را به اهدافمان که در اولین مرحله تعیین کردیم، می رساند. اتخاذ تصمیم راجع به استفاده از نتایج داده کاوی برای اعتبارسنجی نیز در این مرحله انجام می شود.

6. استقرار (Deployment):
استفاده کردن از مدل ایجاد شده، برای مثال می‌تواند تولید یک گزارش ساده از خروجی‌ها را نام برد، و برای یک مثال پیچیده تکمیل کردن پردازش داده کاوی موازی در سایر حوزه‌ها می‌باشد، که این الگوها به یک دانش مفید و قابل استفاده تبدیل می‌شوند و پس از بهبود آنها، الگوهایی که کارا محسوب می شوند در یک سیستم اجرایی به کار گرفته خواهند شد.

مراحل کاری داده کاوی در بستر تکنولوژی Microsoft

داده- کاوی غالباً به عنوان فرآیند استخراج اطلاعات، الگوها و روندهای موجود در مجموعه- ی عظیمی از داده-ها یاد می- شود. این الگوها و روندها را می- توان به عنوان یک مدل کاوشی تعریف نمود. به بیانی دیگر ایجاد یک مدل کاوشی بخشی از فرآیند بزرگتری است که در برگیرنده- ی همه مراحل؛ از تعریف مسئله که مدل حل خواهد نمود تا اجرای مدل در محیط-های کاری است. می- توان این فرآیند را با استفاده از 6 مرحله اساسی زیر تعریف نمود:



باید در نظر داشت که تهیه یک مدل داده کاوی، فرآیندی چرخشی، پویا و تکرار پذیر می- باشد و ممکن است هر یک از این مراحل آن قدر تکرار شود، تا مدل مناسبی تهیه گردد.

تعریف مسئله (Defining the Problem):

تعریف روشنی از مشکل و مسئله کسب و کار است. این مرحله شامل تجزیه و تحلیل نیازمندی-های کسب و-کار، تعریف دامنه مشکل، تعریف معیارهایی که با آن مدل-ها ارزیابی خواهد شد و تعریف هدف نهایی پروژه- ی داده- کاوی است.

آماده- سازی داده-ها (Preparing Data):

یکپارچه -سازی و پالایش داده- هایی است که در مرحله- ی تعریف مسئله فرآیند معین شده است. SSIS حاوی تمامی ابزارهای ملزوم برای تکمیل این مرحله می-باشد.

بررسی داده-ها (Exploring Data):

به منظور تصمیم- گیری-های مناسب در هنگام تهیه مدل، می- بایست داده-ها را درک نمود و پس از آن می- توان تصمیم گیری در مورد وجود داده-های مخدوش در مجموعه داده و در نهایت استراتژی مناسب برای رفع این مشکلات اتخاذ نمود. Data Source view Designer موجود در BIDS حاوی ابزارهای جامعی برای بررسی و شناخت داده‌ها شامل محاسبه ارقام حداقل و حداکثر، محاسبه میانگین و انحراف معیار و بررسی توزیع داده-ها می- باشد.

تهیه مدل -ها (Building Models):

پیش از تهیه مدل باید، داده-ها را به دو دسته- ی داده-های آموزشی و اعتبارسنجی (آزمایشی) تقسیم نمود. از داده-های آموزشی برای تهیه مدل و از داده-های اعتبار-سنجی برای آزمایش صحت مدل با ایجاد سوالاتی در مورد صحت پیش- بینی-ها استفاده نمود. پس از تعریف ساختار کاوشی، می- بایست به پردازش مدل پرداخته شود و ساختارهای خالی با الگوهایی که مدل را توصیف می- نمایند، پُر شوند. این مرحله با عنوان آموزش مدل شناخته می- شود.

بررسی و ارزیابی مدل-ها (Exploring and Validating Models):

این مرحله شامل بررسی مدل-های ایجاد شده به منظور آزمودن کارایی آنهاست. می- توان مدل-ها را با ابزار-های موجود در Designer از جمله نمودار صعود و یا ماتریس دسته- بندی بررسی نمود.

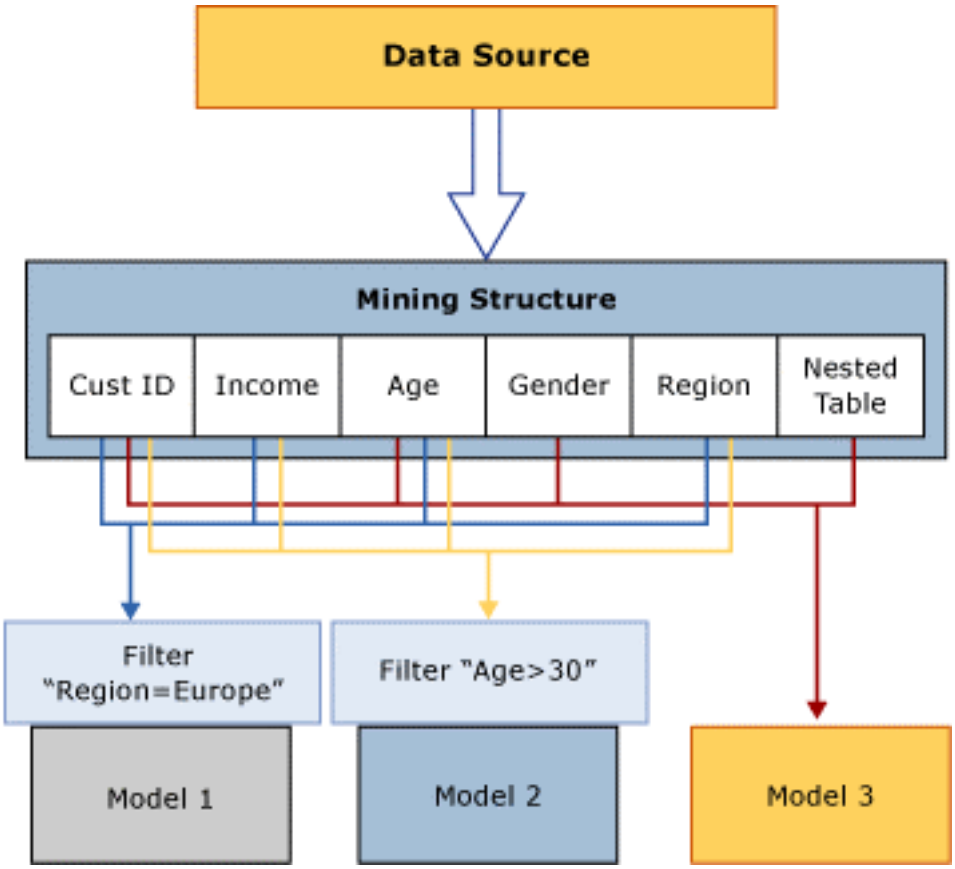
اجرا و بروزرسانی مدل-ها (Deploying and Updating Models):

این مرحله شامل اجرای مدل- هایی است که بهترین کارائی را در یک محیط عملیاتی داشته- اند. پس از استقرار مدل-های کاوشی در یک محیط عملیاتی می- توان از این مدل-ها برای پیش- بینی- هایی بهره گرفت.

مراحل سه گانه موجود در ساخت یک مدل کاوش

ایجاد ساختار کاوشی (Mining Structures): تعریف یک ساختار کاوشی شامل، تعیین تعداد ستون-های ورودی، تعداد ستون-های قابل پیش -بینی و الگوریتم وابسته به آن می-باشد. ساختار کاوشی یک ساختار داده- ای است که محدوده- ی داده- هایی را که از روی آنها مدل-های کاوش ساخته می- شود را تعریف می- نماید.

آموزش مدل (Model Training): یک مدل کاوشی، الگوریتم-های کاوش را به داده- هایی که ساختار کاوش ارائه می- نماید، اعمال می- کند. به بیان دیگر استفاده و کاربرد هر ستون و الگوریتمی که برای ساخت مدل استفاده می- شود را تعریف می- کند، پس شامل داده منبع اصلی نیست، بلکه شامل اطلاعاتی است که توسط الگوریتم کشف می- شود. به آموزش مدل، پردازش مدل نیز گفته می‌شود و زمانی که یک مدل پردازش می- شود داده- هایی که توسط ساختار کاوش تعریف شده- اند، از طریق الگوریتم-های داده- کاوی انتخابی منتقل می- شوند، الگوریتم؛ الگوها و روندها را جستجو می- کند و در ادامه این اطلاعات در مدل ذخیره می- شوند. از این رو پس از یادگیری و آموزش مدل، الگوهای بدست آمده در مدل کاوش ذخیره می- شوند.



پیش بینی مدل (Prediction): غالباً مهمترین مرحله و هدف نهایی در پروژه-های داده- کاوی است. پیش- بینی به کشف اطلاعات ناشناخته با استفاده از الگوهای یافته شده از سوابق داده-ها اشاره دارد. در پیش- بینی به یک مدل کاوشی آموزش دیده و یک مجموعه داده- ی جدید نیاز است. و در طول پیش- بینی موتور داده- کاوی، قواعد بدست آمده در مرحله یادگیری را در مورد مجموعه داده- ی جدید بکار می- برد و نتایج پیش- بینی را به هر Case ورودی تخصیص می- دهد.