عنوان: **بررسی داده کاوی و OLAP**
نویسنده: محمد رجبی
تاریخ: ۱۳۹۳/۰۲/۱۶ ۱۰:۳۵
آدرس: www.dotnettips.info
گروه‌ها: data mining, OLAP, SSAS

بررسی OLAP واژه OLAP در اوایل سال‌های 1990 شکل گرفت. E.F.Codd بنیانگذار مدل داده‌ی رابطه‌ای، این واژه را در فرهنگ نامه کاربران بانک‌های اطلاعاتی توصیف نمود.
مشابه یک بانک اطلاعاتی رابطه‌ای که شامل تعدادی جدول می‌باشد، یک بانک اطلاعاتی OLAP شامل تعدادی Cube است. هر Cube مجموعه ای از Dimensionها و Measure هاست. Dimension یک شیء تحلیلی است که محورهای مختصات را برای پرسش‌های تحلیلی تعریف می‌کند و از Member هایی تشکیل شده است که Member هر Dimension در قالب سلسله مراتب می‌تواند تعریف شود؛ در حالیکه Measure یک مقدار عددی است که در مختصات Cube تعریف می‌شود که این مقادیر از جداول تراکنشی بدست می‌آید (جدول Fact) که جزئیات هر رکورد تراکنشی در آنها ذخیره می‌شود. Measureها حاوی اطلاعاتی هستند که از پیش، محاسبات تجمیعی بر روی آنها براساس سلسله مراتب تعریف شده در Dimension انجام شده است.
ساختار OLAP شبیه به یک مکعب روبیک از داده‌ها است که می‌توان آنرا در جهات مختلف چرخانید تا بتوان سناریوهای «قبلا چه شده» و «چه می‌شد اگر ...» را بررسی نمود. مدل چند بعدی OLAP طریقه نمایش دادن داده‌ها را در مقایسه با بانک‌های اطلاعاتی رابطه‌ای تسهیل می‌کند. غالبا OLAP داده‌ها را از یک انباره داده استخراج می‌کند.

ابزارهای OLAP را به چند دسته تقسیم می‌کنند:

OLAP رو میزی: ابزارهای ساده و مستقل که روی کامپیوترهای شخصی نصب شده و مکعب‌های کوچکی می‌سازند و آنها را نیز بر روی سیستم به شکل فایل ذخیره می‌کنند. بیشتر این ابزارها با صفحات گسترده ای نظیر Excel کار می‌کنند. به این ترتیب کسانی که در سفر هستند قادر به استفاده از این دسته از محصولات هستند. (در حال حاضر Web OLAP در حال جایگزین کردن این محصولات است)

MOLAP : بجای ذخیره کردن اطلاعات در رکوردهای کلید دار، این دسته از ابزارها، بانک‌های اطلاعاتی خاصی را برای خود طراحی کرده‌اند؛ بطوری که داده‌ها را به شکل آرایه‌های مرتب شده بر اساس ابعاد داده ذخیره می‌کنند. در حال حاضر نیز دو استاندارد برای این نوع ابزار وجود دارد. سرعت این ابزار بالا و سایز بانک اطلاعاتی آن نسبتا کوچک است.

ROLAP : این ابزارها با ایجاد یک بستر روی بانک‌های رابطه‌ای اطلاعات را ذخیره و بازیابی می‌کنند. بطوری که اساس بهینه سازی برخی بانک‌های مانند MicreoStrategy، Red Brick و ... بر همین اساس استوار است. اندازه بانک اطلاعاتی این ابزار قابل توجه می‌باشد.

HOLAP : در اینجا منظور از hybrid ترکیبی از MOLAP و ROLAP است. ابزار دارای بانک اطلاعاتی بزرگ و راندمان بالاتر نسبت به ROLAP می‌باشد.

مقایسه گزینه‌های ذخیره سازی در OLAP:

MOLAP: این نوع ذخیره‌سازی بیشترین کاربرد در ذخیره اطلاعات را دارد. همچنین به صورت پیش فرض جهت ذخیره‌سازی اطلاعات انتخاب شده است. در این نوع تنها زمانی داده‌های منتقل شده به Cube به روز می‌شوند که Cube پردازش شود و این امر باعث تاخیر بالا در پردازش و انتقال داده‌ها می‌شود.

ROLAP: در ذخیره‌سازی ROLAP زمان انتقال بالا نیست که از مزایای این نوع ذخیره‌سازی نسبت به MOLAP است. در ROLAP اطلاعات و پیش‌محاسبه‌ها در یک حالت رابطه‌ای ذخیره می‌شوند و این به معنای زمان انتقال نزدیک به صفر میان منبع داده (بانک اطلاعاتی رابطه‌ای) و Cube می‌باشد. از معایب این روش می‌توان به کارایی پایین آن اشاره کرد زیرا زمان پاسخ برای پرس‌وجوهای اجرا شده توسط کاربران طولانی است. دلیل این کارایی پایین بکار نبردن تکنیک‌های ذخیره‌سازی چند بعدی است.

HOLAP : این نوع ذخیره‌سازی چیزی مابین دو حالت قبلی است. ذخیره اطلاعات با روش ROLAP انجام می‌شود، بنابراین زمان انتقال تقزیبا صفر است. از طرفی برای بالابردن کارایی، پیش‌محاسبه‌ها به صورت MOLAP انجام می‌گیرد در این حالت SSAS آماده است تا تغییری در اطلاعات مبداء رخ دهد و زمانی که تغییرات را ثبت کرد نوبت به پردازش مجدد پیش‌محاسبه‌ها می‌شود. با این نوع ذخیره‌سازی زمان انتقال داده‌ها به Cube را نزدیک به صفر و زمان پاسخ برای اجرای کوئری‌های کاربر را زمانی بین نوع ROLAP و MOLAP می‌رسانیم.

این سه روش ذخیره‌سازی انعطاف‌پذیری مورد نیاز را برای اجرای پروژه فراهم می‌کند. انتخاب هر یک از این روش‌ها به نوع پروژه، حجم داده‌ها و ... بستگی دارد. در پایان می‌توان نتیجه گرفت که بهتر است زمان پردازش طولانی‌تری داشته باشیم تا اینکه کاربر نهایی در هنگام ایجاد گزارشات زمان زیادی را منتظر بماند.

بررسی داده کاوی

حجم زیاد اطلاعات، مدیران مجموعه‌ها را در تحلیل و یافتن اطلاعات مفید دچار چالش کرده است. داده کاوی، ابزار مناسب برای تجزیه و تحلیل اطلاعات و کشف و استخراج روابط پنهان در مجموعه‌های داده‌ای سنگین را فراهم می‌کند. گروه مشاوره‌ای گارتنر داده کاوی را استخراج نیمه اتوماتیک الگوها، تغییرات، وابستگی‌ها، نابهنجاری‌ها و دیگر ساختارهای معنی دار آماری از پایگاه‌های بزرگ داده تعریف می‌کند. داده کاوی، تلاشی برای یافتن قوانین، الگوها و یا میل احتمالی داده به مُدلی، در بین انبوهی از داده‌ها است.

داده کاوی فرآیندی پیچیده جهت شناسایی الگوها و مدل‌های صحیح، جدید و به صورت بالقوه مفید، در حجم وسیعی از داده می‌باشد؛ به طریقی که این الگوها و مدلها برای انسانها قابل درک باشند. داده کاوی به صورت یک محصول قابل خریداری نمی‌باشد، بلکه یک رشته علمی و فرآیندی است که بایستی به صورت یک پروژه پیاده سازی شود.

به بیانی دیگر داده کاوی، فرآیند کشف الگوهای پنهان، جالب توجه، غیر منتظره و با ارزش از داخل مجموعه وسیعی از داده‌هاست و فعالیتی در ارتباط با تحلیل دقیق داده‌های سنگین بی ساختار است که علم آمار ناتوان از تحلیل آنهاست. بعضی مواقع دانش کشف شده توسط داده کاوی عجیب به نظر می‌رسد؛ مثلا ارتباط افراد دارای کارت اعتباری و جنسیت با داشتن دفترچه تامین اجتماعی یا سن، جنسیت و درآمد اشخاص با پیش بینی خوش حسابی او در بازپرداخت اقساط وام. داده کاوی در حوزه‌های تصمیم گیری، پیش بینی، و تخمین مورد استفاده قرار می‌گیرد.

پایه و اساس این تکنیک، ریشه در علوم زیر دارد:

علم آمار و احتمال
کامپیوتر (تکنولوژی اطلاعات)
هوش مصنوعی (تکنیکهای یادگیری ماشین)

ارتباط داده کاوی و OLAP

OLAP و داده کاوی فن آوری‌های تحلیلی در خانواده BI به شمار می‌آیند. OLAP در زمینه تجمیع مقادیر عظیم داده‌های تراکنشی بر پایه تعاریف ابعادی مناسب است.

سوالات موضوعی که در ادامه به آن اشاره می‌شود توسط OLAP پاسخ داده می‌شوند:
مقدار فروش کل تولیدات در سه ماهه گذشته در یک منطقه بخصوص چقدر بوده است؟

کدامیک از محصولات جزء ده محصول پر فروش تمامی فروشگاه‌ها در ماه گذشته بودند؟

کدامیک از محصولات برای مشتریان زن و مشتریان مرد فروش قابل توجهی داشته است؟

تفاوت میزان فروش روزانه در هنگام تبلیغات در مقایسه با دوره زمانی عادی چیست؟

فن آوری OLAP بر پایه محاسبات تجمیعی است. سرویس دهنده OLAP نوع خاصی از سرویس دهنده‌ی بانک اطلاعاتی محسوب می‌گردد که با داده‌های چند بعدی سروکار دارد. بسیاری از مشکلات و مخاطرات نظیر ایندکس گذاری، ذخیره سازی داده‌ها و ...

که در RDBMSها وجود دارد در سرویس دهنده‌ی OLAP نیز وجود دارد.
داده کاوی در یافتن الگوهای پنهان از یک مجموعه داده توسط تحلیل همبستگی میان مقادیر مشخصه‌ها مناسب است.

تکنیک‌های داده کاوی دو گونه هستند: نظارت شده و نظارت نشده. در داده کاوی نظارت شده کاربر می‌بایست مشخصه‌ی هدف و مجموعه داده‌ی ورودی را تعیین نماید. الگوریتم‌های داده کاوی نظارت شده شامل درخت تصمیم، نیو بیز و شبکه‌های عصبی هستند. تکنیک‌های داده کاوی نظارت نشده نیازی به تعیین مشخصه‌ی قابل پیش بینی ندارد. خوشه بندی مثال خوبی از داده کاوی نظارت نشده می‌باشد و به گروه بندی نقاط داده ای ناهمگن به زیر گروه هایی می‌پردازد که در آنها نقاط داده ای کم و بیش مشابه و همگن هستند.

در زیر نمونه ای از سوالات پاسخ داده شده توسط داده کاوی ارائه شده است: مشخصات مشتریانی که تمایل به خرید جدیدترین مدل را دارند، چیست؟

چه کالاهایی باید به این دسته از مشتریان خاص توصیه و پیشنهاد گردد؟

برآورد میزان فروش مدلی خاص در سه ماهه آینده چیست؟

چگونه باید مشتریان را تقسیم بندی کرد؟

یکی از فرآیندهای اصلی داده کاوی، تحلیل همبستگی میان مشخصه‌ها و مقادیر آنها است. محققین آمار در این موارد قرن‌ها مطالعه داشته‌اند. OLAP و داده کاوی دو فن آوری مختلف هستند اما فعالیت‌های یکدیگر را تکمیل می‌کنند. OLAP فعالیت هایی نظیر خلاصه سازی، تحلیل تغییرات در طول زمان و تحلیل‌های What If را پشتیبانی می‌نماید. همچنین می‌توان آنرا برای تحلیل نتایج داده کاوی در سطوح مختلف و مجزا استفاده کرد. داده کاوی نیز می‌تواند در ساخت Cubeهای مفیدتر سودمند باشد.

تفاوت میان OLAP و داده کاوی ارتباطی به تفاوت میان داده‌های تلخیص شده و داده‌های تشریحی ندارد. در واقع تمایز قابل توجهی میان مدل سازی توصیفی و تشریحی وجود دارد. توابع و الگوریتم هایی که معمولاً در ابزارهای OLAP یافت می‌شود، توابع مدل سازی توصیفی به شمار می‌آیند. در حالیکه توابعی که در آنچه که اصطلاحاً بسته داده کاوی نامیده می‌شود، یافت می‌شود توابع یا الگوهای مدل سازی تشریحی هستند.

الگوریتم‌های داده کاوی موجود در SSAS و زمینه کاری متناظر
این الگوریتم‌ها را به 5 دسته تقسیم می‌توان نمود:

پیش بینی توالی وقایع
برای مثال جهت تجزیه و تحلیل مجموعه ای از شرایط آب و هوایی که منجر به وقوع پدیده خاصی می‌شود. از الگوریتم زیر استفاده می‌شود:
Microsoft Sequence Clustering Algorithm
یافتن گروهی از موارد مشترک در تراکنش ها معروفترین مثال در خصوص تجزیه و تحلیل سبد بازار است. از الگوریتم‌های زیر استفاده می‌شود:
Microsoft Association Algorithm
Microsoft Decision Trees Algorithm

یافتن گروهی از موارد مشابه معمول‌ترین کاربرد زمینه بخش بندی داده‌های مشتریان به منظور یافتن گروه‌های مجزا از مشتریان است. از الگوریتم‌های زیر استفاده می‌شود:
Microsoft Clustering Algorithm
Microsoft Sequence Clustering Algorithm

پیش بینی صفات گسسته به عنوان مثال، پیش بینی اینکه یک مشتری خاص، تمایلی به خرید محصول جدید دارد یا خیر. از

الگوریتم‌های زیر استفاده می‌شود:

Microsoft Decision Trees Algorithm

Microsoft Naive Bayes Algorithm

Microsoft Clustering Algorithm

Microsoft Neural Network Algorithm

پیش بینی صفات پیوسته پیش بینی درآمد در ماه آینده مثالی از آن می‌باشد. از الگوریتم‌های زیر استفاده می‌شود:

Microsoft Decision Trees Algorithm

Microsoft Time Series Algorithm