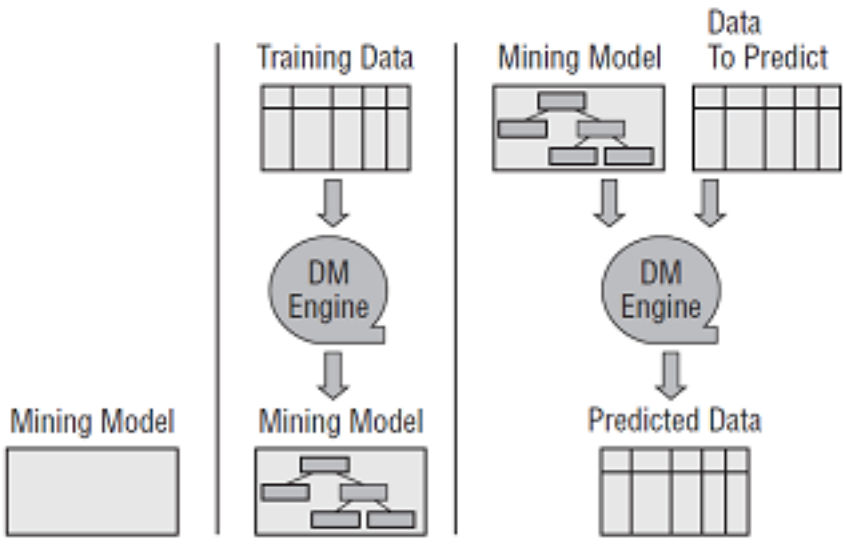
عنوان: **مروری مختصر بر زبان DMX**
نویسنده: محمد رجبی
تاریخ: ۱۳۹۳/۰۸/۲۶ ۱۸:۷
آدرس: www.dotnettips.info
گروه‌ها: Analysis Services, data mining, Microsoft SQL Server

این بخش مروری اجمالی است بر زبان DMX (Data Mining eXtensions) که به منظور انجام عملیات داده کاوی توسط شرکت ماکروسافت ایجاد شده است. (از آنجا که هدف این دوره معرفی الگوریتم‌های داده کاوی است از این رو به صورت کلی به بررسی این زبان می‌پردازیم)

برای بسیاری داده کاوی تنها مجموعه ای از تعدادی الگوریتم تعبیر می‌شود؛ به همان طریقی که در گذشته تصورشان از بانک اطلاعاتی تنها ساختاری سلسله مراتبی به منظور ذخیره داده‌ها بود. بدین ترتیب داده کاوی به ابزاری تبدیل شده که تنها در انحصار تعدادی متخصص (بویژه PhDهای علم آمار و یادگیری ماشین) قرار دارد که آشنائی با اصطلاحات یک زمینه خاص را دارند. هدف از ایجاد زبان DMX تعریف مفاهیمی استاندارد و گزارهایی متداول است که در دنیای داده کاوی استفاده می‌شود به شکلی که زبان SQL برای بانک اطلاعاتی این کار را انجام می‌دهد.

فرضیه اساسی در داده کاوی و همچنین یادگیری ماشین از این قرار است که تعدادی نمونه به الگوریتم نشان داده می‌شود و الگوریتم با استفاده از این نمونه‌ها قادر است به استخراج الگوها بپردازد. بدین ترتیب به منظور بازبینی و همچنین استنتاج از اطلاعات درباره نمونه‌های جدید می‌تواند مورد استفاده قرار گیرد.

ذکر این نکته ضروری است که الگوهای استخراج شده می‌توانند مفید، آموزنده و دقیق باشند. تصویر زیر به اختصار مراحل فرآیند داده کاوی را نمایان می‌سازد:

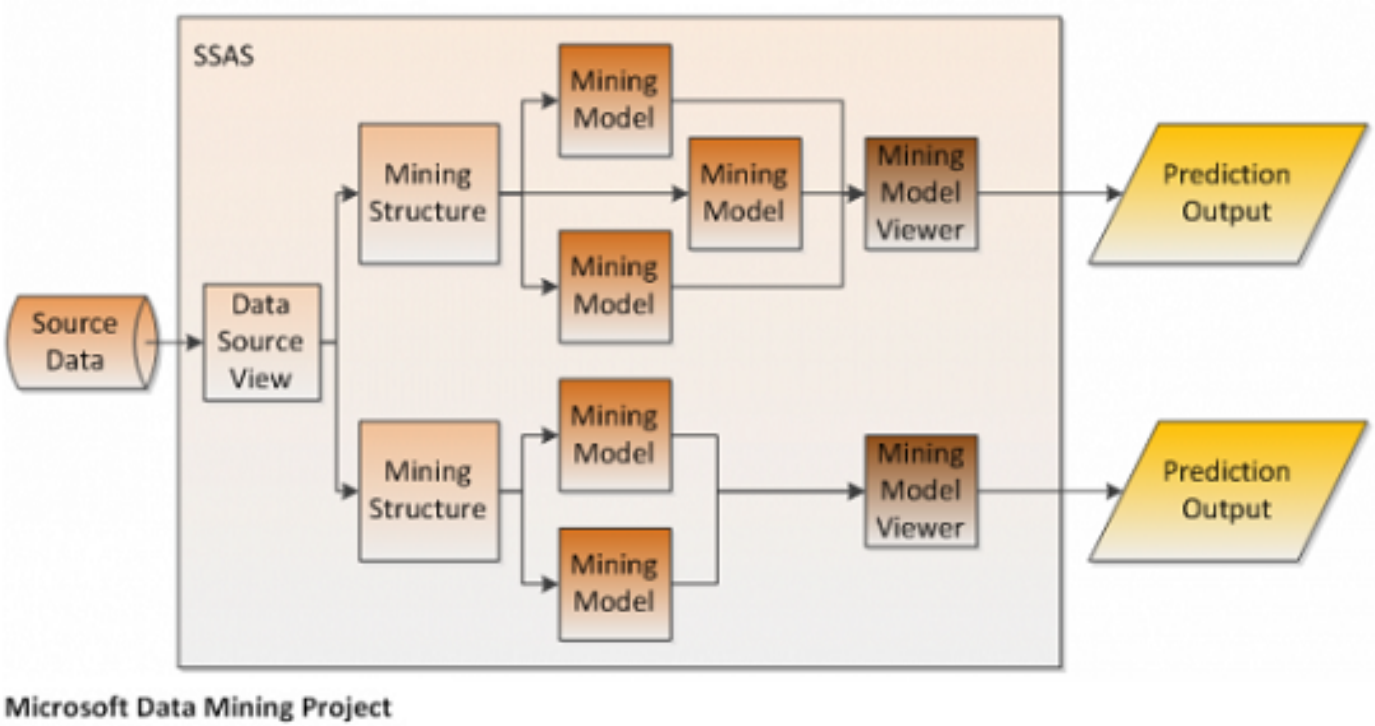


The data mining process

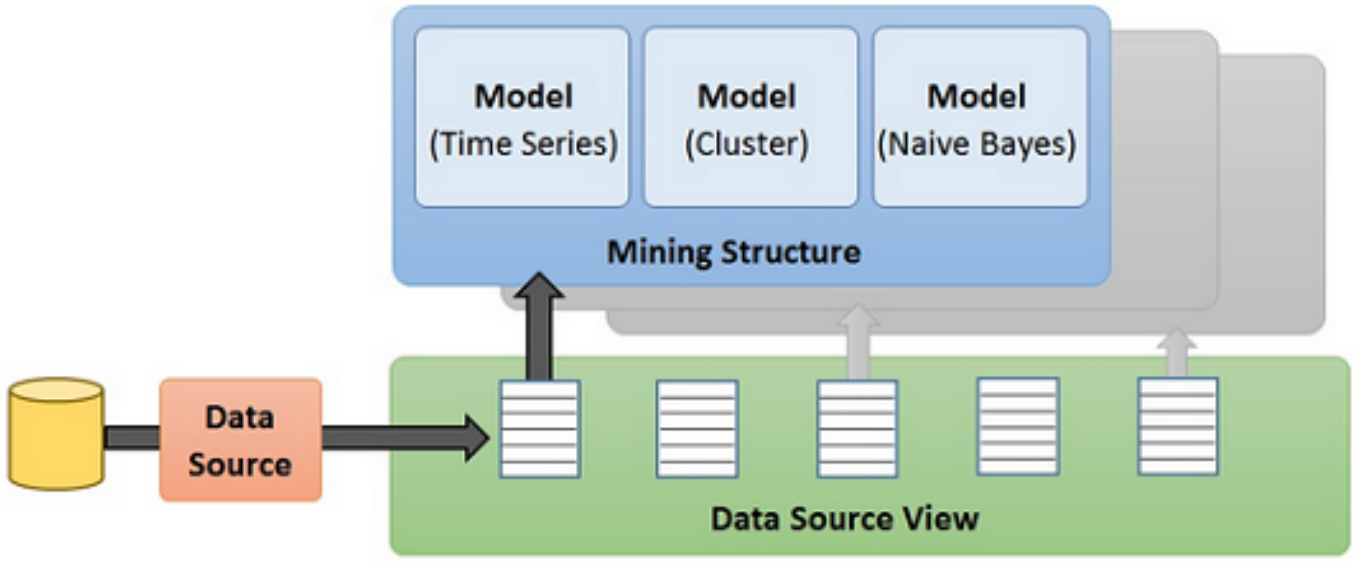
در گام نخست اقدام به تعریف مسئله و فرموله کردن آن می‌کنیم که اصطلاحاً Mining Model نامیده می‌شود. در واقع Mining Model توصیف کننده این است که داده نمونه به چه شکل به نظر می‌رسد و چگونه الگوریتم داده کاوی باید داده‌ها را تفسیر کند. در گام بعدی به فراهم کردن نمونه‌های داده برای الگوریتم می‌پردازیم، الگوریتم با بهره گیری از Mining Model به طریقی که یک لنز داده‌ها را مرتب می‌کند، به بررسی داده‌ها و استخراج الگوها می‌پردازد؛ این عملیات را اصطلاحاً Training Model می‌نامیم.

هنگامی که این عملیات به پایان رسید، بسته به اینکه چگونه آنرا انجام داده اید، می‌توانید به تحلیل الگوهایی که توسط الگوریتم از روی نمونه هایتان بدست آمده بپردازید. و در نهایت می‌توانید اقدام به فراهم کردن داده‌های جدید و فرموله کردن آنها، به همان طریقی که نمونه‌ها آموزش دیده اند، به منظور انجام پیش بینی و استنتاج از اطلاعات با استفاده از الگوهای کشف شده توسط الگوریتم پرداخت.

زبان DMX وظیفه تبدیل داده‌های موجودتان (سطرها و ستون‌های Tables) به داده‌های مورد نیاز الگوریتم‌های داده کاوی (Cases و Attributes) را دارد. به منظور انجام این تبدیل به Mining Structure و Mining Model (که در قسمت اول به شرح آن پرداخته شد) نیاز است. بطور خلاصه Mining Structure صورت مسئله را توصیف می‌کند و Mining Model وظیفه تبدیل سطرهای داده ای به درون Caseها و انجام عملیات یادگیری ماشین با استفاده از الگوریتم داده کاوی مشخص شده را بر عهده دارد.



Microsoft Data Mining Project



Syntax زبان DMX

مشابه زبان SQL دستورات زبان DMX نیز به محیطی جهت اجرا نیاز دارند که می‌توان با استفاده از SQL Server Management Studio (SSMS) به اجرای دستورات DMX اقدام نمود. ایجاد ساختار کاوش (Mining Structure) و مدل کاوشی (Mining Model) مشابه دستورات ایجاد Table در زبان SQL می‌باشد. همانطور که اشاره شد، **گام اول** (از سه مرحله اصلی در داده کاوی) ایجاد یک مدل کاوش است؛ شامل تعیین تعداد ستون‌های ورودی، ستون‌های قابل پیش بینی و مشخص کردن نام الگوریتم مورد استفاده در مدل. **گام دوم** آموزش مدل که پردازش نیز نامیده می‌شود و **گام سوم** مرحله پیش بینی است که نیاز به یک مدل کاوش آموزش

دیده و مجموعه اطلاعات جدید دارد. در طول پیش بینی، موتور داده کاوی قوانین (Rules) پیدا شده در مرحله‌ی آموزش (یادگیری) را با مجموعه اطلاعات جدید تطبیق داده و نتیجه پیش بینی را برای هر Case ورودی انجام می‌دهد. دو نوع پرس و جوی پیش بینی وجود دارد Batch و Singleton که به ترتیب چند Case ورودی دارد و خروجی در یک جدول ذخیره می‌شود و دیگری تنها یک Case ورودی دارد و خروجی در زمان اجرا ساخته می‌شود.

در زبان DMX دو روش برای ساخت مدل‌های کاوش وجود دارد:
• ایجاد یک ساختار کاوش و مدل کاوش مربوط به هم و تحت یک نام، زمانی کاربرد دارد که یک ساختار کاوش فقط شامل یک مدل کاوش باشد.
• ایجاد یک ساختار کاوش و سپس اضافه نمودن یک مدل کاوش به ساختار تعریف شده، زمانی کاربرد دارد که یک ساختار کاوش شامل چندین مدل کاوشی باشد. دلایل مختلفی وجود دارد که ممکن است نیاز به این روش باشد، برای مثال ممکن است مدل‌های متعددی را با استفاده از الگوریتم‌های مختلف ساخت و سپس بررسی نمود که کدام مدل بهتر عمل خواهد کرد و یا مدل‌های متعددی را با استفاده از یک الگوریتم ولی با مجموعه پارامترهای متفاوت برای هر مدل ساخت و سپس بهترین را انتخاب نمود.

عناصر سازنده‌ی ساختار کاوش، ستون‌های ساختار کاوشی هستند که داده هایی را که منبع اصلی داده فراهم می‌کند، توصیف می‌کند. این ستون‌ها شامل اطلاعاتی از قبیل نوع داده (Data Type)، نوع محتوا (Content Type)، ماهیت داده و اینکه داده چگونه توزیع شده است می‌باشند. نوع محتوا پیوسته و یا گسسته بودن آن را مشخص می‌کند و بدین ترتیب به الگوریتم راه درست مدل کردن ستون را نشان می‌دهیم. کلمه کلیدی Discrete برای ماهیت گسسته داده و از کلمه Continuous برای ماهیت پیوسته داده استفاده می‌شود. مقادیر نوع داده و نوع محتوا به قرار زیر می‌باشند:

Data Type	کاربرد
LONG	اعداد صحیح
DOUBLE	اعداد اعشاری
TEXT	داده‌های رشته ای
DATE	داده‌های تاریخی
BOOLEAN	داده‌های منطقی (True و False)
TABLE	برای تعریف Nested Case

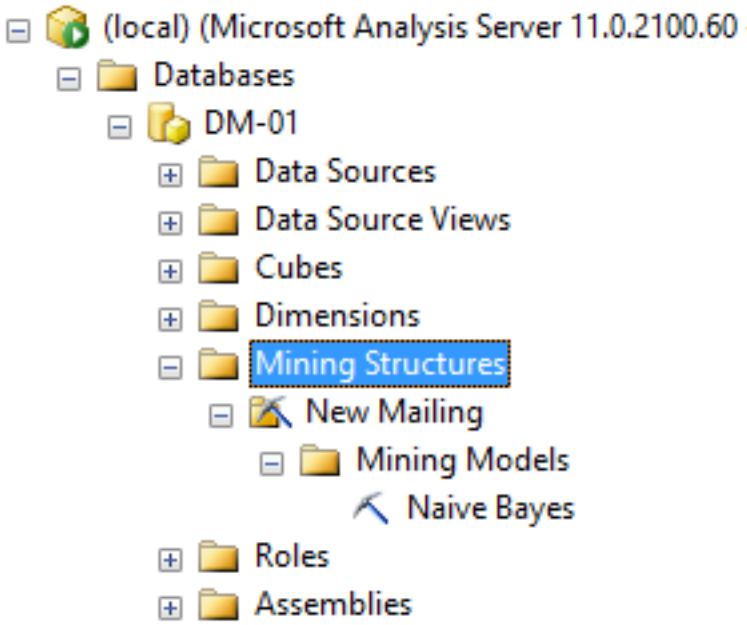
Content Type	کاربرد
KEY	مشخص کننده کلید
DISCRETE	داده‌های گسسته
CONTINUOUS	داده‌های پیوسته
DISCRETIZED	داده‌های گسسته شده
KEY TIME	کلید زمان، تنها در مدل‌های Time Series استفاده می‌شود
KEY SEQUENCE	کلید توالی، تنها در بخش Nested Table مدل‌های Sequence Clustering استفاده می‌شود

همچنین یک مدل کاوش استفاده و کاربرد هر ستون و الگوریتمی که برای ساخت مدل استفاده می‌شود را تعریف می‌کند، می‌توانید با استفاده از کلمه کلیدی Predict و یا Predict_Only خاصیت پیش بینی را به ستون‌ها اضافه نمود، برای نمونه به دستورات زیر توجه نمائید:

```
CREATE MINING STRUCTURE [New Mailing]
(
CustomerKey LONG KEY,
Gender TEXT DISCRETE,
```

```
[Number Cars Owned] LONG DISCRETE,
[Bike Buyer] LONG DISCRETE
)
GO
ALTER MINING STRUCTURE [New Mailing]
ADD MINING MODEL [Naive Bayes]
(
CustomerKey,
Gender,
[Number Cars Owned],
[Bike Buyer] PREDICT
)
USING Microsoft_Naive_Bayes
```

شکل زیر نشان دهنده ارتباط بین ساختار کاوش و مدل کاوشی پس از ایجاد در محیط SSMS می‌باشد.



به منظور آموزش یک مدل کاوش از دستور Insert به شکل زیر استفاده می‌شود:

```
INSERT INTO <mining model name>
[<mapped model columns>]
<source data query>
```

که source data query می‌تواند یک پرس و جوی Select از بانک اطلاعاتی باشد که معمولاً با استفاده از سه طریق OPENQUERY، OPENROWSET و SHAPE بدست می‌آید.

در ادامه به شکل عملی می‌توانید با طی مراحل و اجرای کوئری‌های زیر به بررسی بیشتر موضوع بپردازید.

ابتدا به سرویس SSAS متصل شوید و اقدام به ایجاد یک Database با تنظیمات پیش فرض (مثلاً با نام DM-02) نمائید و در ادامه کوئری XMLA زیر را جهت ایجاد Data Source ای به بانک AdventureWorksDW2012 موجود روی دستگاه تان، اجرا نمائید.

```
<Create xmlns="http://schemas.microsoft.com/analysisservices/2003/engine">
<ParentObject>
  <DatabaseID>DM-02</DatabaseID>
</ParentObject>
<ObjectDefinition>
  <DataSource xmlns:xsd="http://www.w3.org/2001/XMLSchema"
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xmlns:ddl2="http://schemas.microsoft.com/analysisservices/2003/engine/2"
xmlns:ddl2_2="http://schemas.microsoft.com/analysisservices/2003/engine/2/2"
xmlns:ddl100_100="http://schemas.microsoft.com/analysisservices/2008/engine/100/100"
```

```
xmlns:ddl200="http://schemas.microsoft.com/analysisservices/2010/engine/200"
xmlns:ddl200_200="http://schemas.microsoft.com/analysisservices/2010/engine/200/200"
xmlns:ddl300="http://schemas.microsoft.com/analysisservices/2011/engine/300"
xmlns:ddl300_300="http://schemas.microsoft.com/analysisservices/2011/engine/300/300"
xmlns:ddl400="http://schemas.microsoft.com/analysisservices/2012/engine/400"
xmlns:ddl400_400="http://schemas.microsoft.com/analysisservices/2012/engine/400/400"
xsi:type="RelationalDataSource">
<ID>Adventure Works DW2012</ID>
<Name>Adventure Works DW2012</Name>
<ConnectionString>Provider=SQLNCLI11.1;Data Source=(local);Integrated Security=SSPI;
Initial Catalog=AdventureWorksDW2012</ConnectionString>
<ImpersonationInfo>
<ImpersonationMode>ImpersonateCurrentUser</ImpersonationMode>
</ImpersonationInfo>
<Timeout>PT0S</Timeout>
  </DataSource>
</ObjectDefinition>
</Create>
```

و در ادامه کوئری‌های DMX زیر را اجرا نمائید و خروجی هر یک را تحلیل نمائید.

```
/* Step 1 */
CREATE MINING MODEL [NBSample]
(
CustomerKey LONG KEY,
Gender TEXT DISCRETE,
[Number Cars Owned] LONG DISCRETE,
[Bike Buyer] LONG DISCRETE PREDICT
)
USING Microsoft_Naive_Bayes
Go

/* Step 2 */
INSERT INTO NBSample (CustomerKey, Gender, [Number Cars Owned],
[Bike Buyer])
OPENQUERY([Adventure Works DW2012],'Select CustomerKey, Gender, [NumberCarsOwned], [BikeBuyer]
FROM [vTargetMail]')

/*  */
SELECT * FROM [NBSample].CONTENT

/*  */
SELECT * FROM [NBSample_Structure].CASES

/* Step 3*/

SELECT FLATTENED MODEL_NAME,
(SELECT ATTRIBUTE_NAME, ATTRIBUTE_VALUE, [SUPPORT], [PROBABILITY], VALUETYPE FROM NODE_DISTRIBUTION) AS
t
FROM [NBSample].CONTENT
WHERE NODE_TYPE = 26
```

در قسمت‌های بعد تا حدی که از هدف اصلی دوره بررسی الگوریتم‌های داده کاوی موجود در SSAS دور نیافتیم، به بررسی بیشتر دستورات DMX می‌پردازیم. جهت اطلاعات بیشتر در مورد زبان DMX می‌توانید به Books Online for SQL Server مراجعه نمائید.