Conceitos de mineração de dados

A mineração de dados é o processo de descoberta de informações acionáveis em grandes conjuntos de dados. A mineração de dados usa análise matemática para derivar padrões e tendências que existem nos dados. Normalmente, esses padrões não podem ser descobertos com a exploração de dados tradicional pelo fato de as relações serem muito complexas ou por haver muitos dados.

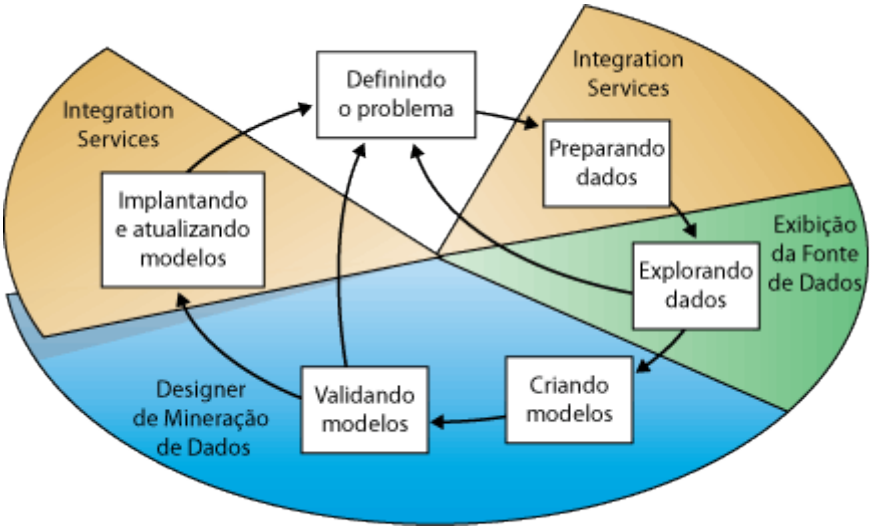
Esses padrões e tendências podem ser coletados e definidos como um modelo de mineração de dados. Os modelos de mineração podem ser aplicados a cenários específicos, como:

- **Previsão**: Estimando vendas, prevendo cargas de servidor ou tempo de inatividade de servidor
- **Risco e probabilidade**: Escolhendo os melhores clientes para malas diretas, determinando o ponto equilibrado provável para cenários de risco, atribuindo probabilidades a diagnósticos ou outros resultados
- **Recomendações**: Determinando quais produtos são mais prováveis de serem vendidos juntos, gerando recomendações
- **Localizando sequências**: Analisando seleções de cliente em um carrinho de compras, prevendo os próximos eventos prováveis
- **Agrupamento**: Separando clientes ou eventos em cluster de itens relacionados, analisando e prevendo afinidades

A criação de um modelo de mineração representa apenas uma parte de um processo maior que inclui desde perguntas sobre dados e criação de um modelo até respostas para as perguntas feitas e implantação do modelo em um ambiente de trabalho. Esse processo pode ser definido usando as seis etapas básicas a seguir:

1. Definindo o problema
2. Preparando dados
3. Explorando dados
4. Criando modelos
5. Explorando e validando modelos
6. Implantando e atualizando modelos

O diagrama a seguir descreve as relações entre cada etapa do processo e as tecnologias no Microsoft SQL Server que você pode usar para concluir cada etapa.



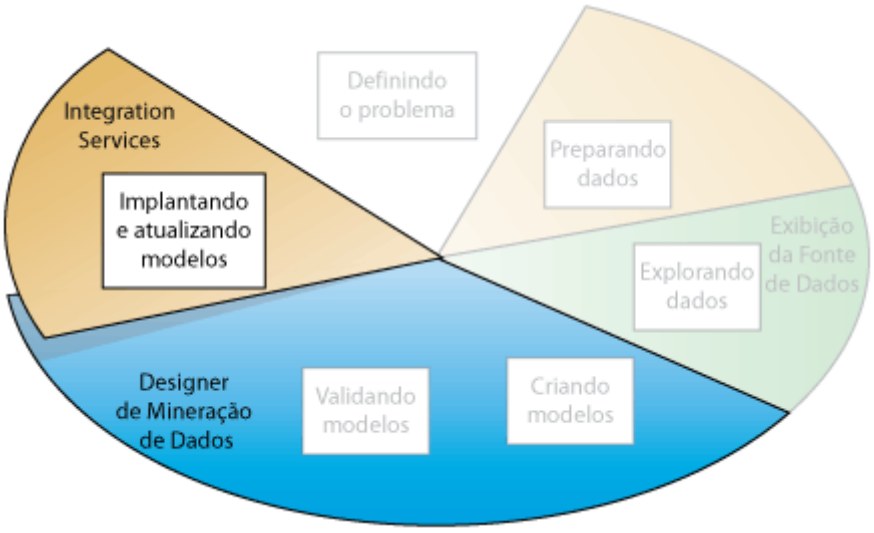
O processo ilustrado no diagrama é cíclico, ou seja, criar um modelo de mineração de dados é um processo dinâmico e repetitivo. Depois de explorar os dados, você pode descobrir que eles são insuficientes para criar modelos de mineração apropriados e que você terá, portanto, que obter mais dados. Ou você pode criar vários modelos e, depois, perceber que os modelos não respondem adequadamente o problema definido e que você deverá redefinir o problema. Talvez seja necessário atualizar os modelos depois de eles serem implantados, pois haverá mais dados disponíveis. Cada etapa do processo pode precisar ser repetida muitas vezes para criar um bom modelo.

A Mineração de Dados do Microsoft SQL Server fornece um ambiente integrado para a criação e manipulação dos modelos de mineração de dados. Este ambiente inclui o SQL Server Development Studio, que contém algoritmos de mineração de dados e ferramentas de consulta que facilitam a construção de uma solução abrangente para uma variedade de projetos e o SQL Server Management Studio, que contém ferramentas para procurar modelos e gerenciar objetos de mineração de dados. Para obter mais informações, consulte Criando modelos multidimensionais usando o SSDT (Ferramentas de Dados do SQL Server).

Para obter um exemplo de como as ferramentas do SQL Server podem ser aplicadas a um cenário de negócios, consulte Tutorial de mineração de dados básico.

Definindo o problema

A primeira etapa do processo de mineração de dados, como destacado no diagrama a seguir, é definir claramente o problema e considerar maneiras de os dados serem utilizados para fornecer respostas para ele.



Essa etapa inclui a análise dos requisitos de negócio, a definição do escopo do problema, a definição das métricas usadas para avaliar o modelo e, por fim, a definição de objetivos específicos para o projeto de mineração de dados. Essas tarefas podem ser traduzidas em perguntas, como:

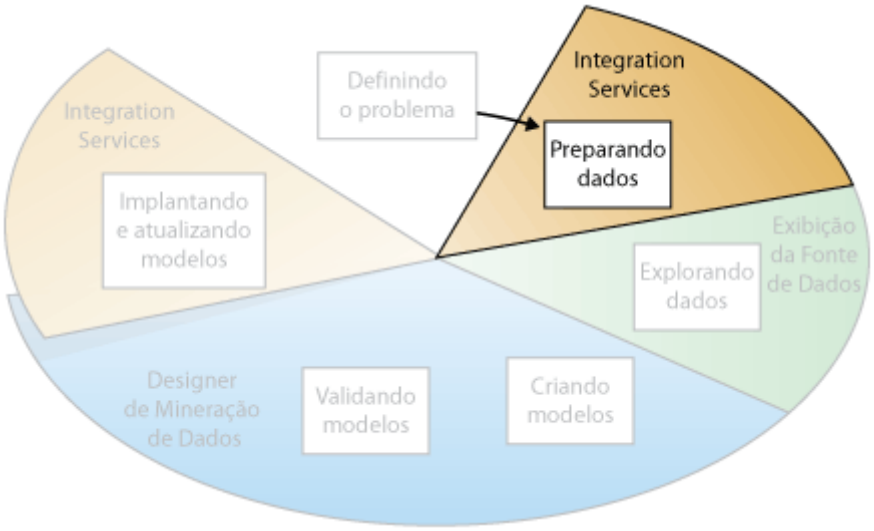
- O que você deseja? Quais tipos de relações está tentando localizar?
- O problema que você está tentando solucionar se reflete nas políticas e nos processos da empresa?
- Você deseja fazer previsões com o modelo de mineração de dados ou apenas identificar padrões e associações interessantes?
- Qual resultado ou atributo você deseja prever?
- Que tipo de dados você tem e que tipo de informações está em cada coluna? Se houver várias tabelas, como elas estão relacionadas? Você precisa executar alguma limpeza, agregação ou processamento para tornar os dados utilizáveis?
- Como os dados são distribuídos? Os dados são sazonais? Os dados representam de forma precisa os processos da empresa?

Para responder essas perguntas, talvez seja necessário realizar um estudo da disponibilidade de dados para investigar as necessidades dos usuários da empresa com relação aos dados disponíveis. Se os dados não forem suficientes para suprir as necessidades dos usuários, talvez você tenha que redefinir o projeto.

Também é necessário considerar a forma como os resultados do modelo podem ser incorporados em KPIs (indicadores chave de desempenho) usados para avaliar o progresso dos negócios.

Preparando dados

A segunda etapa do processo de mineração de dados, como destacado no diagrama a seguir, é consolidar e limpar os dados identificados na etapa Definindo o problema.



Os dados podem estar espalhados pela empresa e armazenados em diferentes formatos ou podem conter inconsistência, como entradas ausentes ou incorretas. Por exemplo, os dados podem mostrar que um cliente comprou um produto antes desse produto ser efetivamente colocado a venda no mercado ou que o cliente faz compras regularmente em uma loja localizada a 3.000 quilômetros da casa dele.

A limpeza de dados não envolve apenas a remoção de dados incorretos ou interpolação de valores ausentes, mas também a localização de correlações ocultas nos dados, a identificação de fontes de dados mais precisas e a determinação de quais colunas são mais apropriadas para a análise. Por exemplo, você deveria usar a data de envio ou a data de pedido? O melhor influenciador de vendas é a quantidade, o preço total ou o preço com desconto? Dados incompletos, incorretos e entradas que parecem independentes, mas que são muito relacionadas, podem influenciar os resultados do modelo de formas inesperadas.

Portanto, antes de iniciar a criação de modelos de mineração, você deve identificar esses problemas e determinar como solucioná-los. Para a mineração de dados, normalmente você está trabalhando com um conjunto de dados muito grande e não pode examinar a qualidade de dados de cada transação; portanto, você pode precisar usar alguma forma de criação de perfis de dados e ferramentas automatizadas de limpeza de dados e filtragem, como as fornecidas no Integration Services, Microsoft SQL Server 2012 Master Data Services ou SQL Server 2012 Data Quality Services para explorar os dados e localizar as inconsistências. Para obter mais informações, consulte estes recursos:

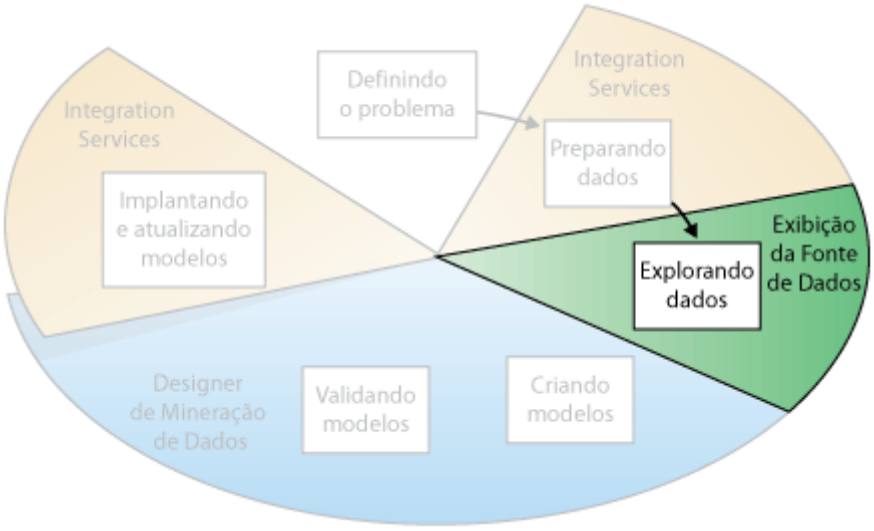
- Integration Services in Business Intelligence Development Studio
- Visão geral do Master Data Services
- Data Quality Services

É importante saber que os dados usados na mineração de dados não precisam estar armazenados em um cubo OLAP (processamento analítico online) nem mesmo em um banco de dados relacional, apesar de ambos poderem ser usados como fontes de dados. Você pode conduzir a mineração de dados usando qualquer fonte de dados definida como uma fonte de dados do Analysis Services. Isso inclui arquivos de texto, pasta de

trabalho do Excel e dados de outros provedores externos. Para obter mais informações, consulte Fontes de dados com suporte (SSAS - modelos multidimensionais).

Explorando dados

A terceira etapa do processo de mineração de dados, como destacado no diagrama a seguir, é explorar os dados preparados.



Você deve compreender os dados para tomar decisões apropriadas ao criar os modelos de mineração. As técnicas de exploração incluem cálculos dos valores máximos e mínimos, cálculos das médias e dos desvios padrões e análise da distribuição dos dados. Por exemplo, ao analisar os valores máximos, mínimos e médios, você pode determinar que os dados não são representativos para seus clientes ou processos de negócio e que você deve obter mais dados equilibrados ou revisar as suposições que determinam suas expectativas. Os desvios padrão e outros valores de distribuição podem fornecer informações útil sobre a estabilidade e precisão dos resultados. Um desvio padrão muito grande indica que incluir mais dados pode ser útil para melhorar o modelo. Os dados que desviam muito de uma distribuição padrão podem estar distorcidos ou representar uma imagem precisa do problema real, dificultando, porém, o ajuste de um modelo aos dados.

Ao explorar os dados levando em consideração o seu conhecimento do problema dos negócios, é possível decidir se o conjunto contém dados imperfeitos. Com isso, você poderá criar uma estratégia para solucionar os problemas ou compreender ainda mais os comportamentos típicos na sua empresa.

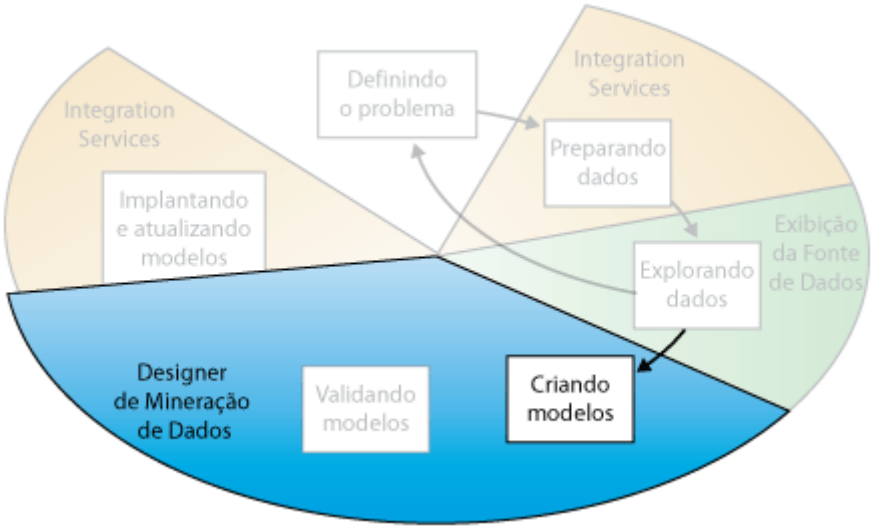
Você pode usar ferramentas como o Microsoft SQL Server 2012 Master Data Services para investigar origens disponíveis de dados e determinar a sua disponibilidade para mineração de dados. Você pode usar ferramentas como o SQL Server 2012 Data Quality Services, ou o Criador de Perfil no Integration Services, para analisar a distribuição de seus dados e corrigir problemas como dados errados ou ausentes.

Depois de definir suas origens, você as combina em uma exibição da fonte de dados usando o Designer de Exibição da Fonte de Dados no Ferramentas de dados do SQL Server. Para obter mais informações, consulte Exibições de fontes de dados em modelos multidimensionais. Este designer também contém várias ferramentas que você pode usar para explorar os dados e verificar que eles funcionarão para criar um modelo. Para obter mais informações, consulte Explorar dados em uma exibição da fonte de dados (Analysis Services).

Observe que, ao criar um modelo, o Analysis Services cria automaticamente resumos estatísticos de dados do modelo, que podem ser consultados para uso em relatórios ou para análises detalhadas. Para obter mais informações, consulte Consultas de mineração de dados.

Criando modelos

A quarta etapa do processo de mineração de dados, como destacado no diagrama a seguir, é criar o modelo ou modelos de mineração. Você usará o conhecimento obtido na etapa Explorando dados para ajudá-lo a definir e criar os modelos.



Você define as colunas de dados que você deseja usar ao criar uma estrutura de mineração. A estrutura de mineração é vinculada à origem dos dados, mas realmente não contém nenhum dado até que seja processada. Ao processar a estrutura de mineração, o Analysis Services gera agregações e outras informações estatísticas que podem ser usadas para análise. Essas informações podem ser usadas por qualquer modelo de mineração com base na estrutura. Para obter mais informações sobre como as estruturas de mineração estão relacionadas aos modelos de mineração, consulte Arquitetura lógica (Analysis Services – Mineração de Dados).

Antes de a estrutura e o modelo serem processados, também o modelo de mineração de dados é apenas um contêiner que especifica as colunas usadas para entrada, o atributo que você está prevendo e os parâmetros que indicam ao algoritmo como os dados devem ser processados. O processamento de um modelo é geralmente chamado de treinamento. Treinamento refere-se ao processo de aplicação de um algoritmo

matemático específico aos dados na estrutura com a finalidade de extrair padrões. Os padrões que você localiza no processo de treinamento dependem da seleção de dados de treinamento, o algoritmo que você escolheu e como você configurou o algoritmo. O SQL Server 2012 contém muitos algoritmos diferentes, cada um adequado a um tipo diferente de tarefa e cada um criando um tipo diferente de modelo. Para obter uma lista dos algoritmos fornecidos no SQL Server 2012, consulte Algoritmos de mineração de dados (Analysis Services – Mineração de Dados).

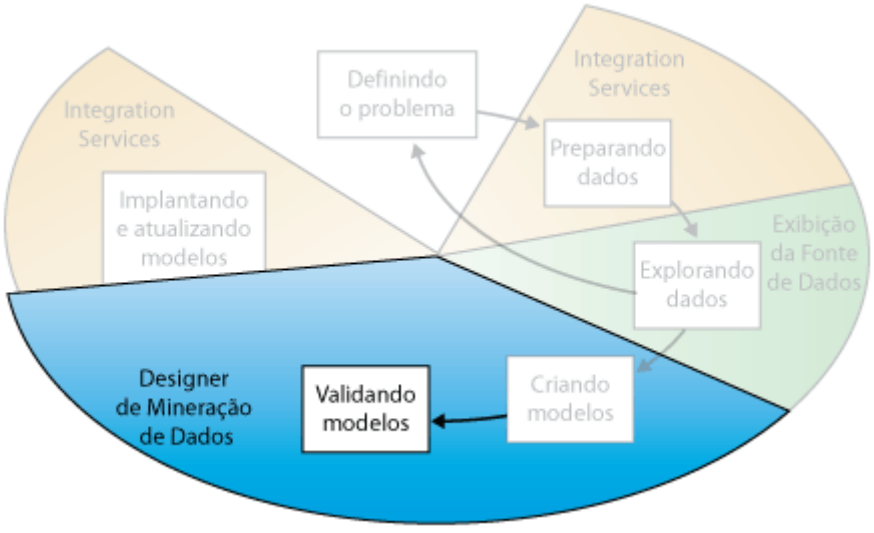
Você também pode usar parâmetros para ajustar cada algoritmo e aplicar filtros aos dados de treinamento para usar apenas um subconjunto de dados, criando diferentes resultados. Depois de passar os dados pelo modelo, o objeto do modelo de mineração conterá resumos e padrões que poderão ser consultados ou usados para previsão.

É possível definir um novo modelo usando o Assistente de Mineração de Dados do Ferramentas de dados do SQL Server ou a linguagem DMX (Data Mining Extensions). Para obter mais informações sobre como usar o Assistente de Mineração de Dados, consulte Assistente de Mineração de Dados (Analysis Services - Mineração de dados). Para mais informações sobre como usar DMX, consulte Referência DMX (Data Mining Extensions).

É importante lembrar-se de que sempre que os dados mudarem, será necessário atualizar a estrutura e o modelo de mineração. Quando você atualiza uma estrutura de mineração reprocessando-a, o Analysis Services recupera dados da origem, incluindo quaisquer dados novos caso a origem seja atualizada dinamicamente, e repopula a estrutura de mineração. Se você tiver modelos com base na estrutura, poderá optar pela atualização dos modelos com base na estrutura, o que significa que eles serão retreinados nos novos dados, ou poderá manter o modelo como está. Para obter mais informações, consulte Requisitos e considerações de processamento (mineração de dados).

Explorando e validando modelos

A quinta etapa do processo de mineração de dados, como destacado no diagrama a seguir, é explorar os modelos de mineração criados e testar a eficiência deles.



Antes de implantar um modelo em um ambiente de produção, você provavelmente o testará para avaliar o desempenho. Além disso, ao criar um modelo, você normalmente cria vários modelos com diferentes configurações e os testa para verificar qual deles gera os melhores resultados para seu problema e seus dados.

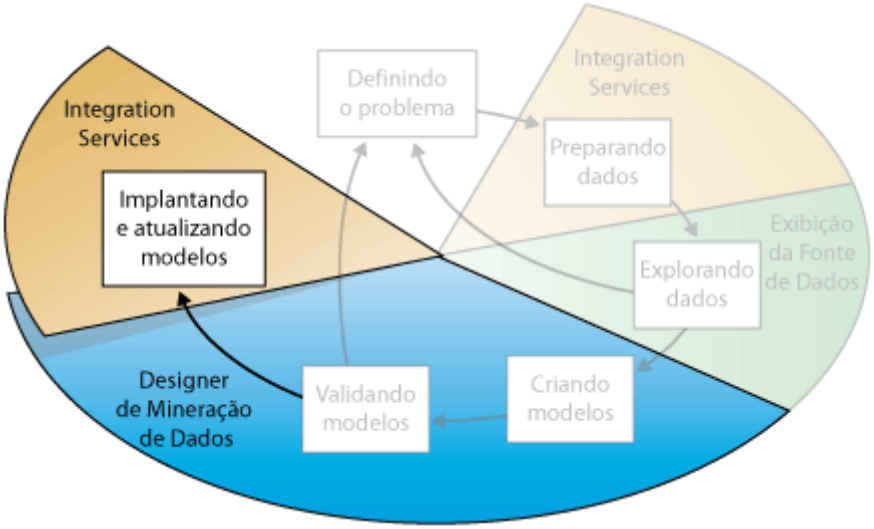
O Analysis Services fornece ferramentas que o ajudam a separar seus dados em conjuntos de dados de teste e de treinamento, de forma que você possa avaliar com precisão o desempenho de todos os modelos nos mesmos dados. Você usa o conjunto de dados de treinamento para criar um modelo e o conjunto de dados de teste para testar a precisão do modelo ao criar consultas de previsão. No SQL Server 2012 Analysis Services (SSAS), esse particionamento pode ser feito automaticamente durante a criação do modelo de mineração de dados. Para obter mais informações, consulte Teste e validação (mineração de dados).

É possível explorar as tendências e os padrões que os algoritmos descobrem usando os visualizações no Designer de Mineração de Dados do Ferramentas de dados do SQL Server. Para obter mais informações, consulte Visualizadores do Modelo de Mineração de Dados. Você também pode testar como os modelos criam previsões usando ferramentas do designer, como o gráfico de comparação de precisão e a matriz de classificação. Para verificar se o modelo é específico para seus dados ou se pode ser usado para fazer deduções na população geral, você pode usar a técnica estatística chamada validação cruzada para criar, automaticamente, subconjuntos de dados e testar modelos em cada subconjunto. Para obter mais informações, consulte Teste e validação (mineração de dados).

Se nenhum dos modelos criados na etapa Criando modelos tiver um bom desempenho, você poderá retornar a uma etapa anterior no processo e redefinir o problema ou investigar novamente os dados no conjunto de dados original.

Implantando e atualizando modelos

A última etapa do processo de mineração de dados, como destacado no diagrama a seguir, é implantar os modelos que tiveram o melhor desempenho em um ambiente de produção.



Depois que os modelos de mineração existirem em um ambiente de produção, será possível realizar várias tarefas de acordo com suas necessidades. Estas são algumas tarefas que você poderá realizar:

- Use os modelos para criar previsões, que poderão ser usadas para tomar decisões comerciais. O SQL Server fornece a linguagem DMX, que pode ser usada para criar consultas de previsão, e o Construtor de Consultas de Previsão para ajudá-lo a criar as consultas. Para obter mais informações, consulte Referência DMX (Data Mining Extensions).
- Crie consultas de conteúdo para recuperar estatísticas, regras ou fórmulas do modelo. Para obter mais informações, consulte Consultas de mineração de dados.
- Insira a funcionalidade de mineração de dados diretamente em um aplicativo. Você pode incluir Objetos de Gerenciamento de Análise (AMO) que contém um conjunto de objetos que seu aplicativo pode usar para criar, alterar, processar e excluir estruturas e modelos de mineração. Como alternativa, você pode enviar mensagens XMLA (XML for Analysis) diretamente para uma instância do Analysis Services. Para obter mais informações, consulte Development (Analysis Services - Data Mining).
- Use o Integration Services para criar um pacote no qual um modelo de mineração é usado para separar dados recebidos, de forma inteligente, em diversas tabelas. Por exemplo, se um banco de dados for atualizado continuamente com clientes potenciais, será possível usar um modelo de mineração juntamente com o Integration Services para dividir os dados recebidos entre os clientes que têm probabilidade de adquirir um produto e os que têm probabilidade em não adquirir o produto. Para obter mais informações, consulte Typical Uses of Integration Services.
- Crie um relatório que permita que os usuários consultem diretamente um modelo de mineração existente. Para obter mais informações, consulte Reporting Services nas Ferramentas de Dados do SQL Server (SSRS).
- Atualize os modelos depois da revisão e análise. As atualizações exigirão o reprocessamento dos modelos. Para obter mais informações, consulte Processando objetos de mineração de dados.

- A atualização dinâmica dos modelos, à medida que a organização gera mais dados, e alterações constantes para melhorar a eficiência da solução devem fazer parte da sua estratégia de implantação. Para obter mais informações, consulte Gerenciamento de soluções de mineração de dados e objetos.

Consulte também

Conceitos

Soluções de mineração de dados
Ferramentas de mineração de dados