Kapitola 26

Analysis Services

Tato kapitola obsahuje následující témata:

- Vytváření základních datových skladů
- Seznámení s modelem dimenzí
- Sestavení krychle
- Konstrukce rozměrů a veličin
- Konstrukce hierarchií
- Procházení a zavádění krychlí
- Výběr vhodných algoritmů dolování dat
- Vytváření modelů a struktur dolování
- Procházení modelu dolování

O službě SQL Server Analysis Services (SSAS) byly napsány celé knihy, které však pouze naznačují celý rozsah problematiky. V této kapitole získáte velmi zběžný přehled dvou klíčových funkcí modulu Analysis Services: OLAP (online analytic processing) a dolování dat. Dozvíte se, jak databázi SQL2008SBS změnit na strukturu, která umožňuje využít modul Analysis Services. Naučíte se také, jak vytvářet základní krychle a modely dolování dat.

Poznámka: V této kapitole naleznete základní informace, ale neočekávejte vyčerpávající rozbor principů návrhu datových skladů. Struktura databáze SQL2008SBSDW slouží výhradně k předvedení funkcí modulu Analysis Services. Máte-li zájem o podrobnou informaci o návrhu datových skladů, můžete využít desítek knih, které o tomto tématu vznikly. Uvedenou tematiku nebylo možné zhustit do jediné kapitoly této knihy.

Další informace: Podrobný popis služby SSAS naleznete v knize *SQL Server 2008 Analysis Services Step by Step*.

Přehled datových skladů

Službu Analysis Services sice můžete použít přímo s transakční databází, ale normalizované struktury nejsou pro analytické aktivity vhodné. Chcete-li využít plné možnosti služby Analysis Services, potřebujete převést data extrahovaná z transakčního zdroje do analytického formátu v rámci datového skladu. Hlavní překážka pro použití služby Analysis Services často nespočívá v konstrukci krychlí, ale v návrhu datového skladu, který leží v pozadí krychlí a struktur dolování dat.

Při konstrukci datového skladu je potřeba rozumět perspektivě dat z pohledu koncového uživatele.

Účelem databáze SQL2008SBS, se kterou jste pracovali v celé této knize, bylo umožnit zákazníkům zadávat objednávky. Pokud byste studovali datový diagram databáze SQL2008SBS, zjistili byste, že všechny tabulky v databázi se točí kolem tabulky Customers.Customer. Pokud chcete zadat objednávku, musíte projít přes tabulku Customer. Na tabulku Customer jsou směrovány i dotazy na informace o objednávkách. Centrální entitou celé databáze je zákazník.

Zákazníci jsou sice důležití, ale když chcete analyzovat informace v databázi SQL2008SBS, neprovádíte rozbor zákazníků. Místo toho analyzujete objednávky. Při transformaci transakčních dat v databázi SQL2008SBS je centrální entitou databáze objednávka. Informace o zákaznících poté slouží pouze jako další část popisných údajů.

V následujícím cvičení obnovíte předem vytvořenou databázi s názvem SQL2008SBSDW, kterou budete používat ve zbývající části této kapitoly.

Nastavení databáze SQL2008SBSDW

1. Obnovte databázi SQL2008SBSDW ze záložního souboru Kapitola 26\SQL2008SBSDW.bak v doprovodných ukázkách knihy.
2. Ověřte, zda je databáze online a dostupná.
3. Prohlédněte si obsah databáze SQL2008SBSDW generované z databáze OLTP s názvem SQL2008SBS.

OLAP (Online Analytic Processing)

Oblast Business Intelligence (BI) poskytuje procesy, postupy návrhu a technologie, které mají usnadnit získání přehledu o podnikových operacích. Hlavním účelem BI je řešení problému, jak „vidět přes stromy les“. Když se začnete utápět ve stále větším objemu podrobných dat, je velmi těžké přijímat řešení, která mají dopad na činnost organizace. Stejně jako nedosáhnete požadovaných výsledků tím, že budete dohlížet na každý krok svých podřízených, ani masivní objemy podrobných dat vám nedovolí splnit podnikové cíle.

Pokud chcete zjistit, jak velký je celý les, musíte odstoupit od jeho okraje. Stejně tak platí, že chcete-li se rozhodovat na základě svých dat, musíte omezit úroveň podrobností, aniž byste ztratili celkový obraz.

Problém velkých objemů detailních informací lze vyřešit pomocí technik agregace. Místo toho, abyste si prohlíželi miliony jednotlivých řádků s údaji o objednávkách, můžete podrobnosti objednávek agregovat v závislosti na různých faktorech. Na základě dat uložených v databázi SQL2008SBS můžete vytvořit následující pohledy, které poslouží pro podnikové rozhodování:

- Kolik objednávek denně dostáváme?
- Roste počet objednávek nebo klesá?
- Zvyšují se prodeje nebo snižují?
- Kolik produktů prodáváme za měsíc?
- Které produkty každý měsíc prodáváme?
- Pokud si zákazník koupí Produkt X, které další produkty si pravděpodobně koupí?

Pravděpodobně lze přijít se stovkami různých otázek, které můžete svým datům položit. Jestliže byste se však snažili najít odpověď na tyto otázky listováním v tisících nebo milionech

řádků položek objednávek, rychle byste byli podrobnostmi zahlceni. Jak si můžete všimnout, lze všechny možné podnikové dotazy vyřešit načtením a agregací malé podmnožiny dat.

Naneštěstí nelze napsat všechny dotazy SELECT...GROUP BY, které by uživatelé mohli požadovat v časovém intervalu, kdy jsou výsledky potřeba. Funkce OLAP umožňují překlenout mezeru mezi ručním psaním tisíců či milionů agregačních dotazů, aby bylo možné koncovým uživatelům zpřístupnit informace pro podnikové rozhodování. Bez ohledu na všechny koncepce probírané v této kapitole lze říci, že jedinou hlavní funkcí modulu OLAP je generovat všechny možné permutace agregačních funkcí dat, aby analýza nemusela vyčkávat na výpočet každé agregované hodnoty.

Model dimenzí

Aby bylo možné z dat odvodit platné závěry a efektivně zpracovávat velké objemy dat, spoléhá se modul OLAP na základní model dimenzí. Model dimenzí zahrnuje několik základních struktur:

- Dimenze
- Atributy
- Veličiny
- Hierarchie

Dimenze se velmi podobá entitě v relačním modelu. *Dimenze* definují základní podnikové struktury, které chcete analyzovat, např. zákazníky, produkty a čas. *Atributy* v rámci dimenze určují sloupce v dimenzi, které se používají k analýze, např. CustomerName, ProductName, City, PostalCode a OrderDate.

Hierarchie dovolují definovat strukturu navigace v rámci dimenze, např.:

- Zákazníci ve městech v oblastech v zemích
- Kalendářní měsíce v kalendářních čtvrtletích v kalendářních letech
- Fiskální měsíce ve fiskálních čtvrtletích ve fiskálních letech

Veličiny (measure) jsou sady agregačních hodnot, které chcete vypočítat, jako např. součet částek objednávek nebo jejich počet.

Zdrojem všech struktur OLAP je tabulka ve zdroji dat. OLAP se na zdrojové tabulky odkazuje dvěma termíny – tabulkou faktů a tabulkou dimenzí. *Tabulka dimenzí* (dimension table) uchovává data, která slouží k definici dimenzí, atributů dimenzí a hierarchií. *Tabulka faktů* (fact table) se používá k definici veličin.

Většina uživatelů sice nemá žádné potíže s definicí tabulek dimenzí, ale problémy obvykle způsobuje struktura tabulek faktů. Fakt je jediná instance něčeho, co chcete analyzovat, např. výskyt vady na výrobní lince, položka řádku objednávky, doručená zásilka, nemoc ve zdravotní dokumentaci nebo žádost o půjčku. Tabulka faktů je ústředním prvkem v databázové struktuře, na které je založen model dimenzí. Obsahuje sloupec, který propojuje každý řádek faktu ke každé dimenzi, která poskytuje další popisné informace.

V databázi SQL2008SBS je centrální entitou tabulka zákazníků, a chcete-li načíst všechny potřebné informace, může být nutné spojit více tabulek. Když databázi SQL2008SBS transformujete na databázi SQL2008SBSDW, přidáte ke každé položce řádku objednávky klíče k objednávce, zákazníkovi, umístění, produktu, kategorii, dílčí kategorii atd. Poté agregujete položky řádků pro libovolné podnikové prvky, aniž byste museli používat spojení s jakoukoli

jinou tabulkou. Případné spojení s jinou tabulkou by umožňovalo načtení textu souvisejícího s klíčem příslušné entity.

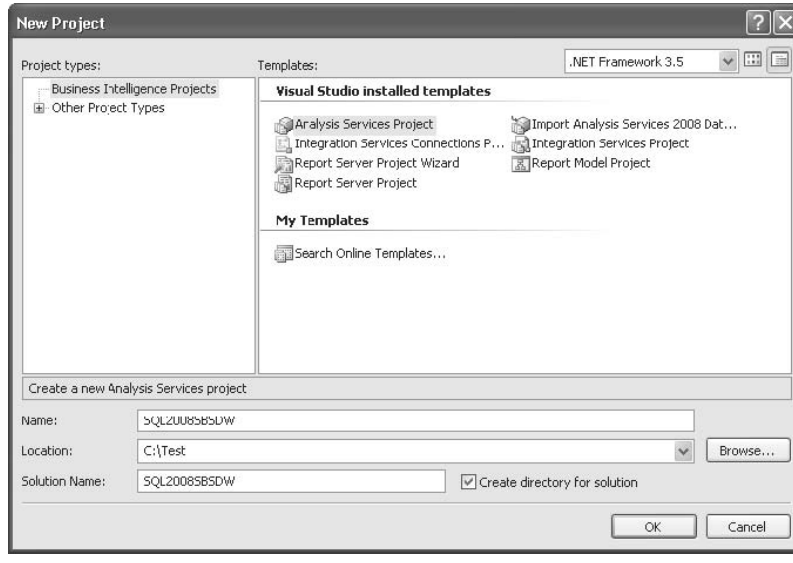
Krychle

Podobně jako relační databáze poskytují strukturní kontejner i bezpečnostní hranici, obsahuje *krychle* (cube) všechny definované analytické objekty a určuje rovněž nejvyšší úroveň zabezpečení, kterou lze přiřadit. Databáze ve službě Analysis Services zahrnuje celou instanci SSAS. V rámci databáze SSAS můžete pro své uživatele vytvořit více krychlí. Na rozdíl od geometrické krychle může mít krychle SSAS prakticky neomezený počet dimenzí a každá dimenze může mít jinou délku.

V následujícím cvičení vytvoříte první krychli podle databáze SQL2008SBSDW.

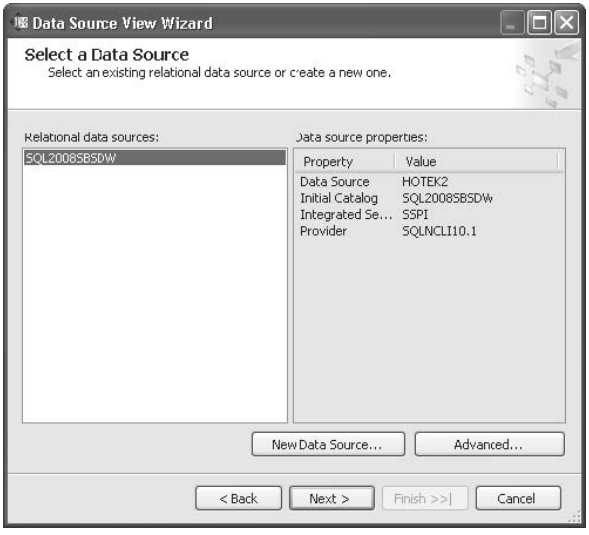
Sestavení první krychle

1. Spusťte nástroj BIDS.
2. Vyberte příkaz File → New → Project, zvolte možnost Analysis Services Project, zadejte název a umístění a klepněte na tlačítko OK.

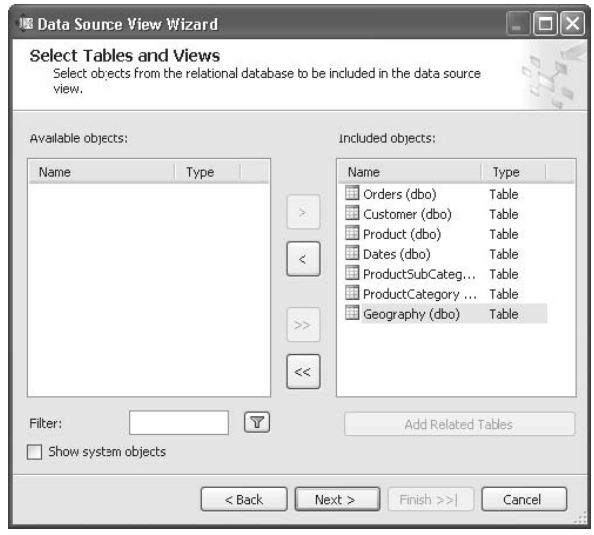


3. V okně Solution Explorer klepněte pravým tlačítkem myši na složku Data Sources a vyberte příkaz New Data Source.
4. Klepnutím na tlačítko Next přijměte výchozí možnost Create A Data Source... pro databázi SQL2008SBSDW.
5. Na stránce Impersonation Information vyberte přepínač Use The Service Account. Dokončete zbývající kroky průvodce.

6. V okně Solution Explorer klepněte pravým tlačítkem myši na složku Data Source Views, vyberte příkaz New Data Source View (DSV) a klepněte na tlačítko Next.
7. Vyberte právě vytvořený zdroj dat a klepněte na tlačítko Next.

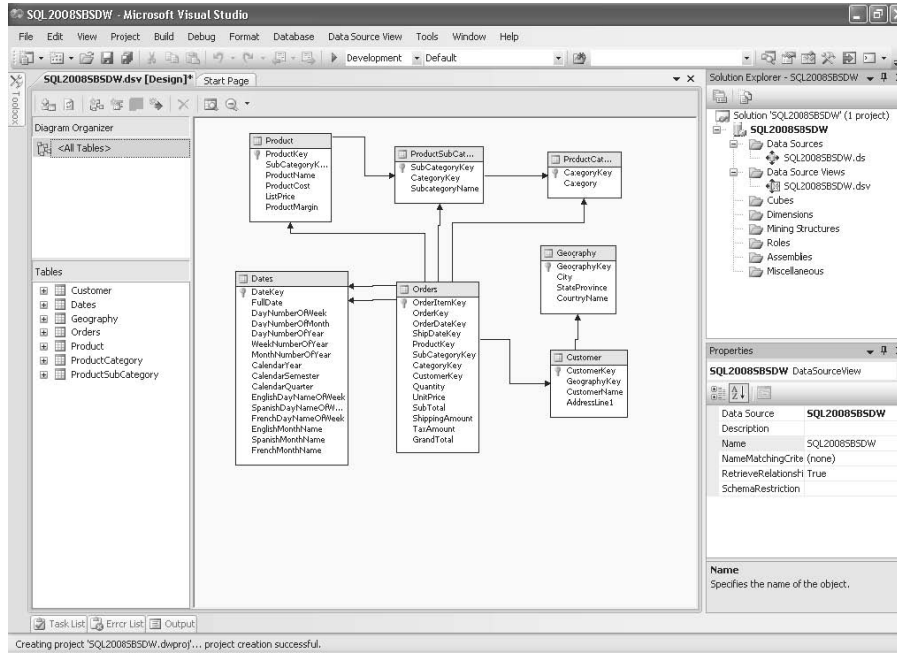


8. Přidejte všechny tabulky do seznamu Included Objects, klepněte na tlačítko Next a poté klepněte na tlačítko Finish.

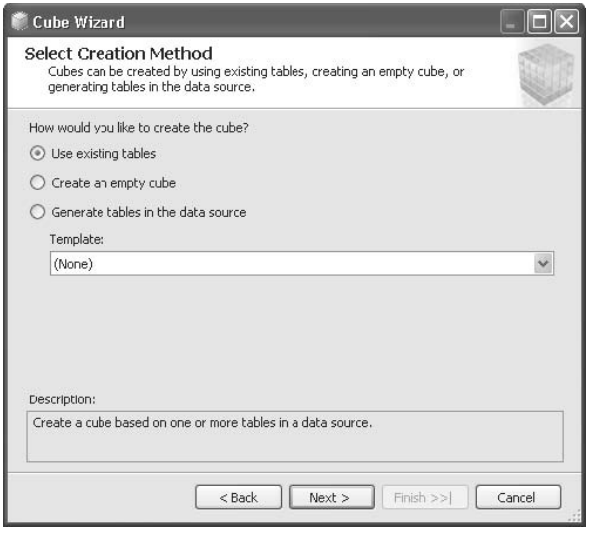


9. Zkontrolujte obsah pohledu DSV.

10. Klepněte pravým tlačítkem myši na složku Cubes, vyberte příkaz New Cube a klepněte na tlačítko Next.



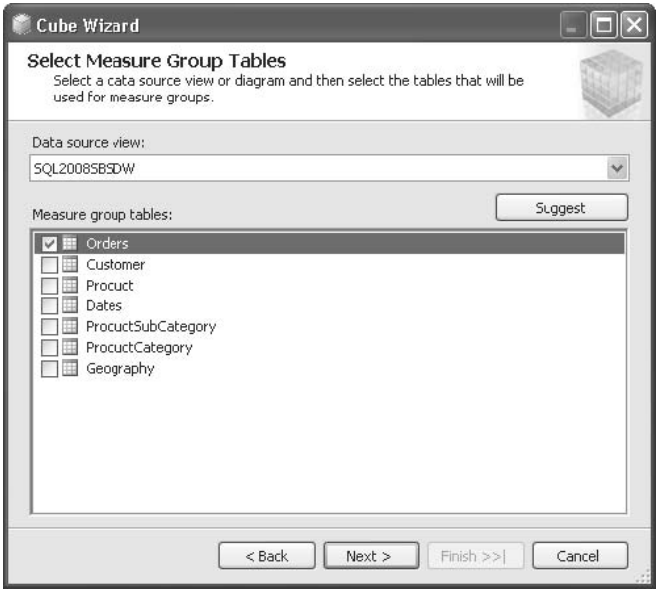
11. Zvolte přepínač Use Existing Tables a klepněte na tlačítko Next.

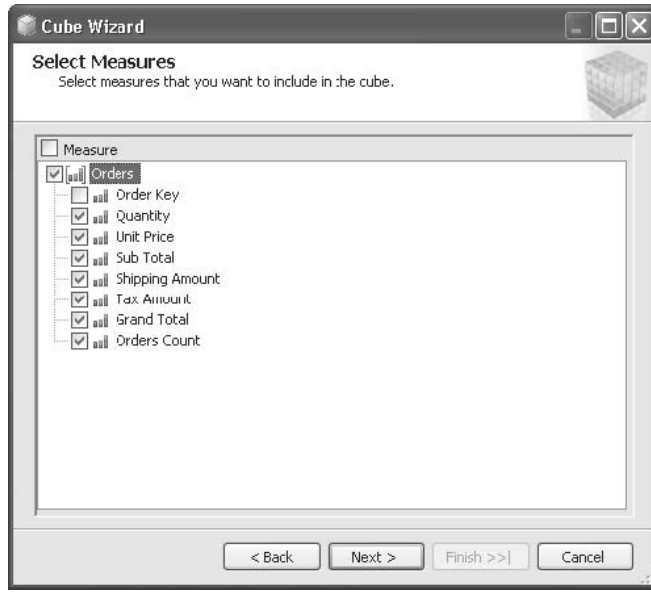


12. Jako Measure Group Tables nastavte tabulku Orders a klepněte na tlačítko Next.

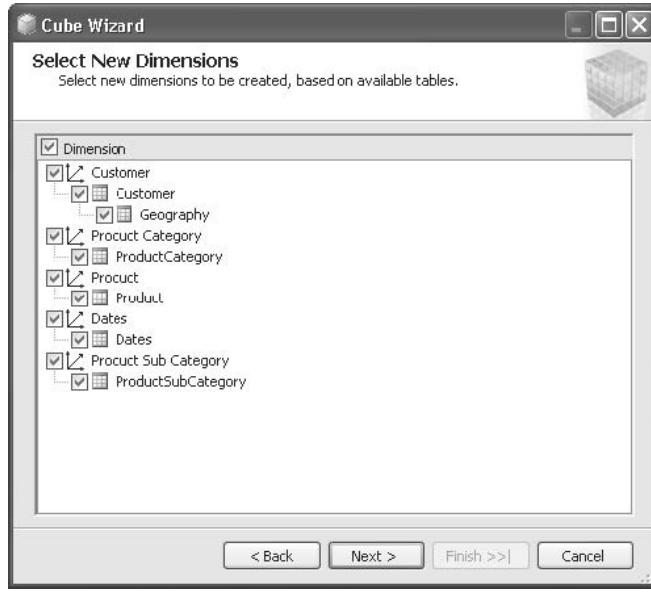
13. Zrušte výběr sloupce OrderKey a klepněte na tlačítko Next.

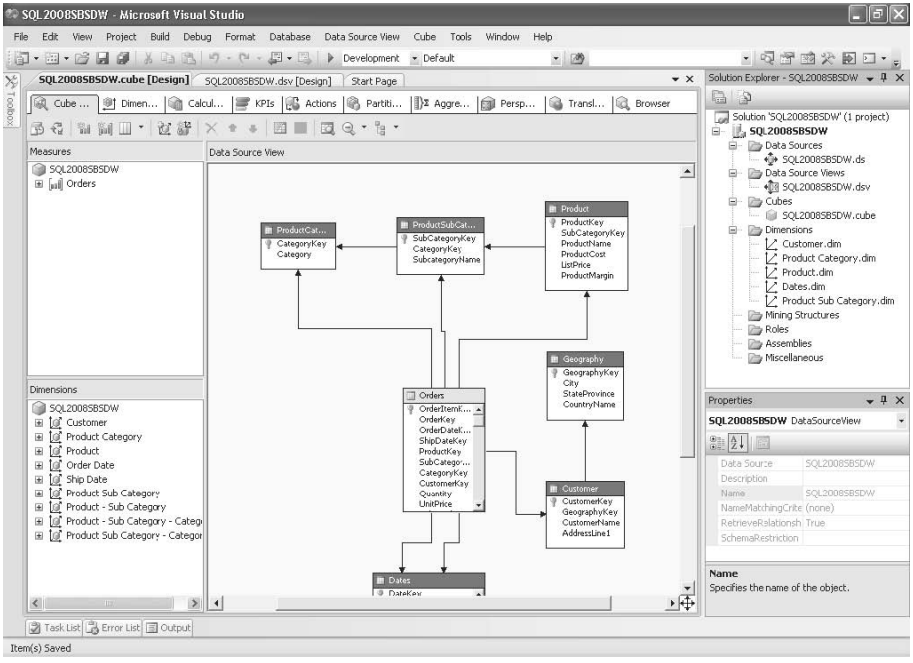
14. Potvrďte výchozí nastavení na stránce Select New Dimensions a klepněte na tlačítko Next.



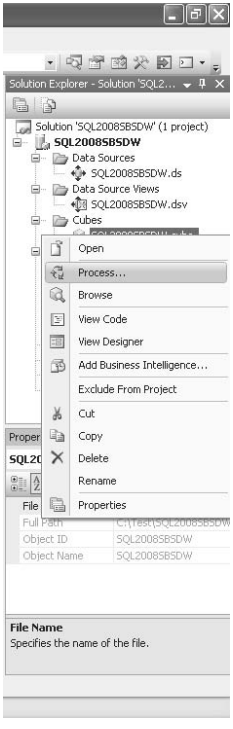


15. Klepněte na tlačítko Finish.
16. Přidejte do instance systému SQL Server účet služby SSAS, pokud zatím nemá přístup pro přihlášení.
17. Přidejte účet služby SSAS jako uživatele databáze SQL2008SBSDW a udělte účtu oprávnění SELECT k této databázi.

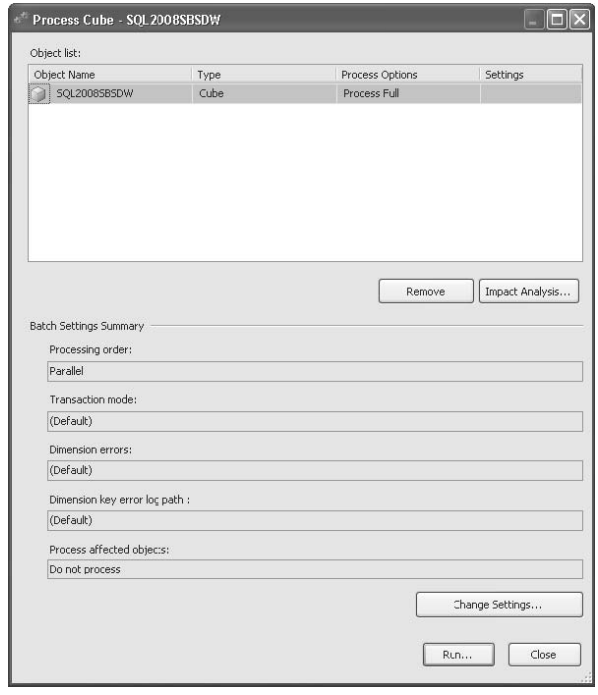




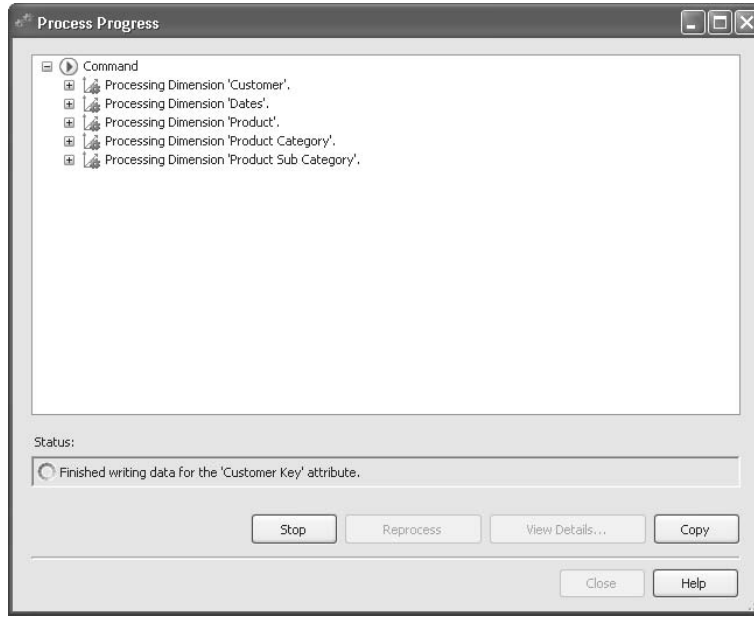
18. Klepněte na právě vytvořenou krychli pravým tlačítkem myši v okně Solution Explorer a vyberte příkaz Process.



19. Klepnutím na tlačítko Yes sestavíte a zavedete projekt.

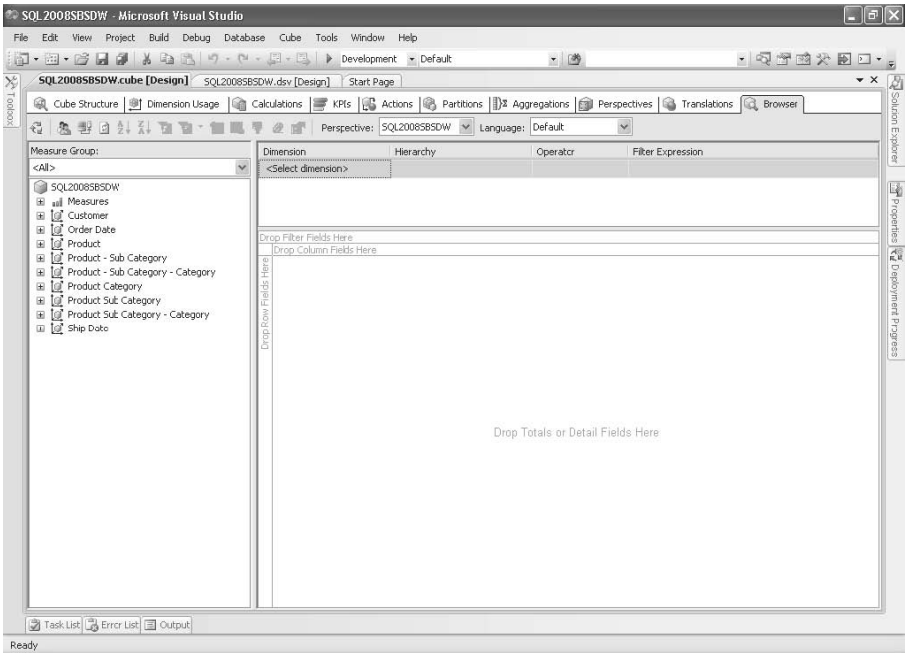


20. Chcete-li krychli zpracovat, klepněte na tlačítko Run.

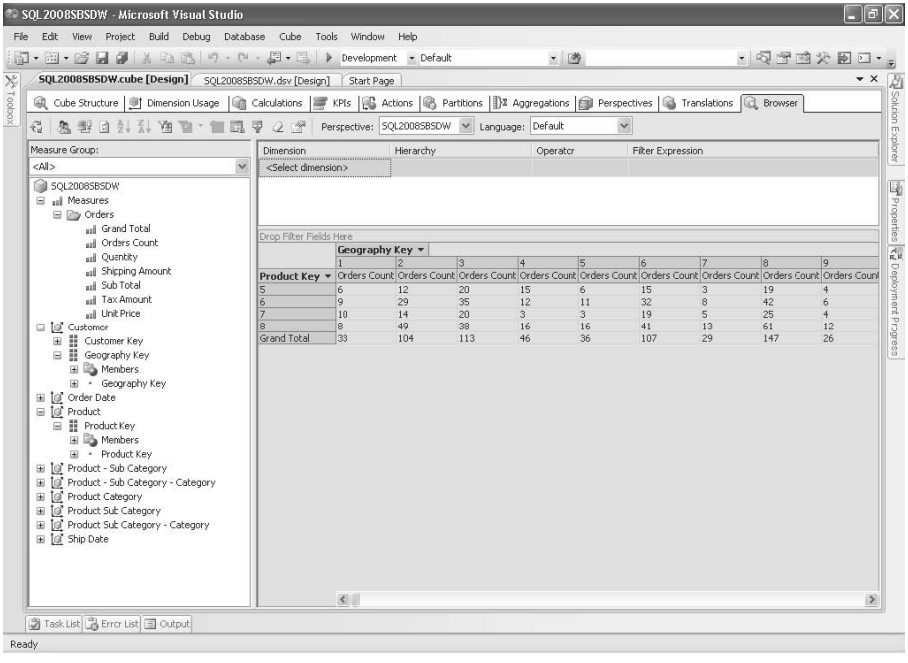


21. Klepněte na tlačítko Close.

22. Vyberte kartu Browser.



23. Věnujte chvíli času procházení obsahu krychle. V následujících cvičeních se budete zabývat všemi formátovacími prvky, které umožní zlepšit použitelnost krychle.



Dimenze, veličiny a výpočty

Dimenze poskytují základní analytické prvky v rámci krychle. Dimenze obsahují jeden nebo více atributů, které definují datové sloupce používané při analýze. Každý atribut má sadu vlastností, které lze konfigurovat. K nejčastějším patří:

- Sloupec klíče
- Název
- Datový typ a volitelně velikost dat
- Řazení
- Formátovací řetězec pro zobrazení
- Sloupec názvu
- Pořadí řazení

Sloupec klíče identifikuje atribut v rámci struktury krychle a název slouží k odkazování na atribut v elementech, které v krychli navrhnete. Atribut bude propojen se zdrojovou tabulkou a sloupcem s datovým typem a volitelně velikostí dat. Můžete nastavit pořadí řazení, formátovací řetězec a možnosti řazení, které určují, jak se dimenze zobrazí v rámci libovolného nástroje, který slouží k procházení krychle. Vlastnost názvu sloupce umožňuje definovat sloupec ze zdroje dat, který je mapován na atribut kvůli poskytování informací, které se zobrazí při procházení krychle, jako je např. převod číselné hodnoty CustomerID na skutečné jméno zákazníka.

Nejjemnější úrovní dimenze je člen. Dimenze definují analyzovanou podnikovou entitu. Atributy definují sloupce, které lze v dimenzi analyzovat. Členy jsou vlastní datové hodnoty v rámci daného atributu. Při výběru atributu dimenze pro analýzu bude koncový uživatel uvažovat skutečné členy atributu, např. Spojené státy, Produkt X nebo Zákazník Y.

Veličiny definují dostupné výpočty v krychli a jsou propojeny se sloupcem v pohledu zdroje dat krychle. Lze definovat agregační funkci, jako je *SUM*, *MIN*, *MAX* či *COUNT*. Veličiny také dovolují specifikovat formátovací řetězec, který se používá ke kontrole zobrazení dat při procházení krychle.

Některé agregační funkce, které chcete s krychlí používat, nelze definovat na základě sloupce ve zdroji dat. Do krychle je možné přidat vlastní výpočty, které zajistí libovolné požadované agregace, např. násobení množství a ceny za jednotku kvůli výpočtu celkové ceny za položku.

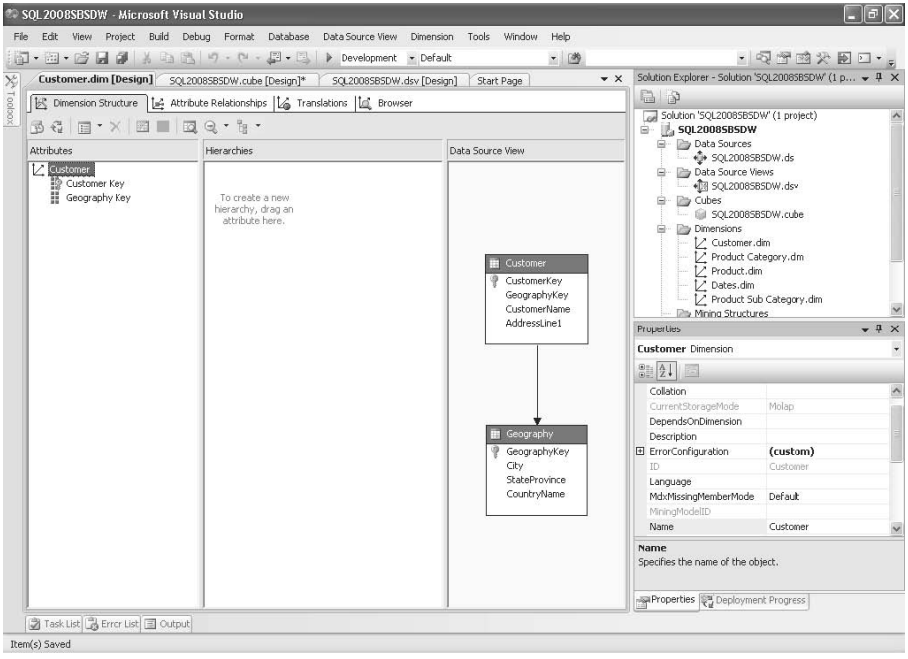
V následujícím cvičení upravíte dimenze a veličiny vytvořené průvodcem krychle, abyste uživatelům nabídli pohodlné možnosti procházení. Přidáte do krychle také výpočet celkové hodnoty řádku objednávky, aby bylo možné agregovat prodeje pro jednotlivé produkty.

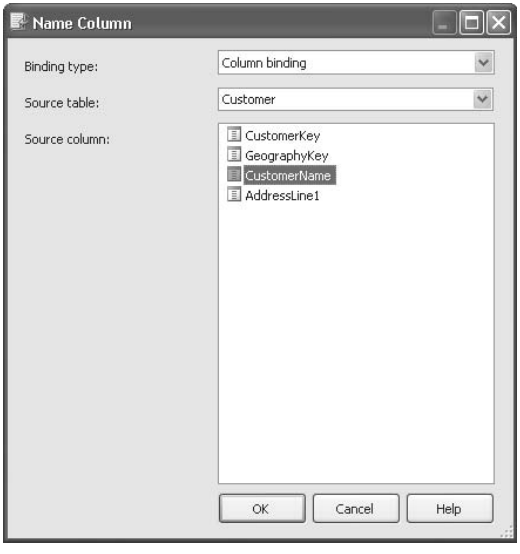
Poznámka: Návrhář se zabývá definicí prvků krychle. Každý prohlížeč se připojí ke službě Analysis Services a použije uloženou strukturu a data. Chcete-li zobrazit provedené změny návrhu, musíte zavést změny do služby SSAS a zpracovat je. V následujícím textu vynecháme snímky obrazovek a kroky vývoje a zpracování, které je nutné provést před spuštěním každého prohlížeče v nástroji BIDS.

Návrh dimenzí, veličin a výpočtů

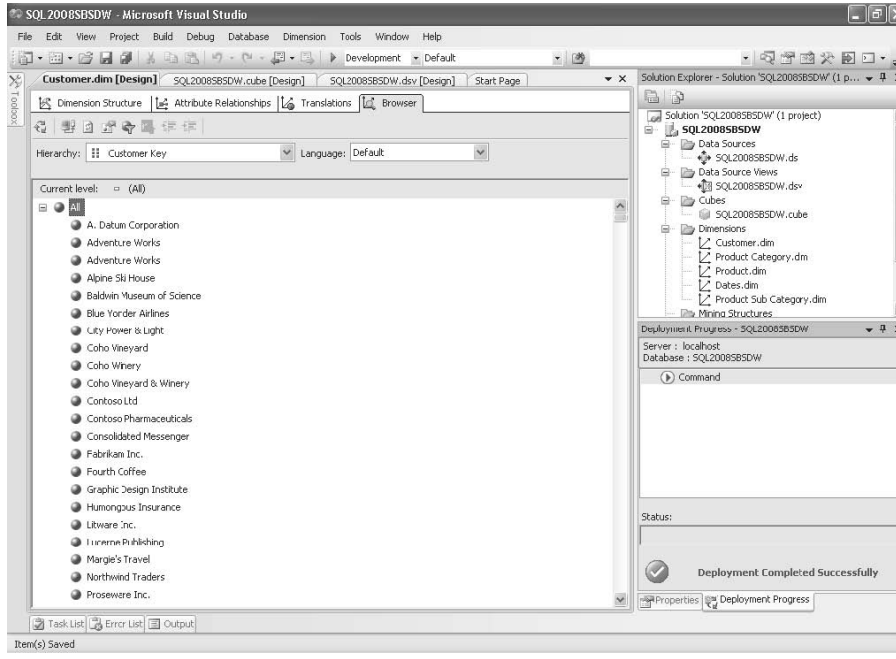
1. V okně Solution Explorer poklepejte na dimenzi Customer.

2. Vyberte možnost Customer Key a posuňte se na vlastnost NameColumn v okně Properties.
3. Klepněte na tlačítko se třemi tečkami v pravém sloupci, nastavte sloupec Source na hodnotu CustomerName a klepněte na tlačítko OK.
4. Rozbalte vlastnost NameColumn a zkontrolujte aktualizovaná nastavení.
5. Nastavte vlastnost OrderBy na hodnotu *Name*.

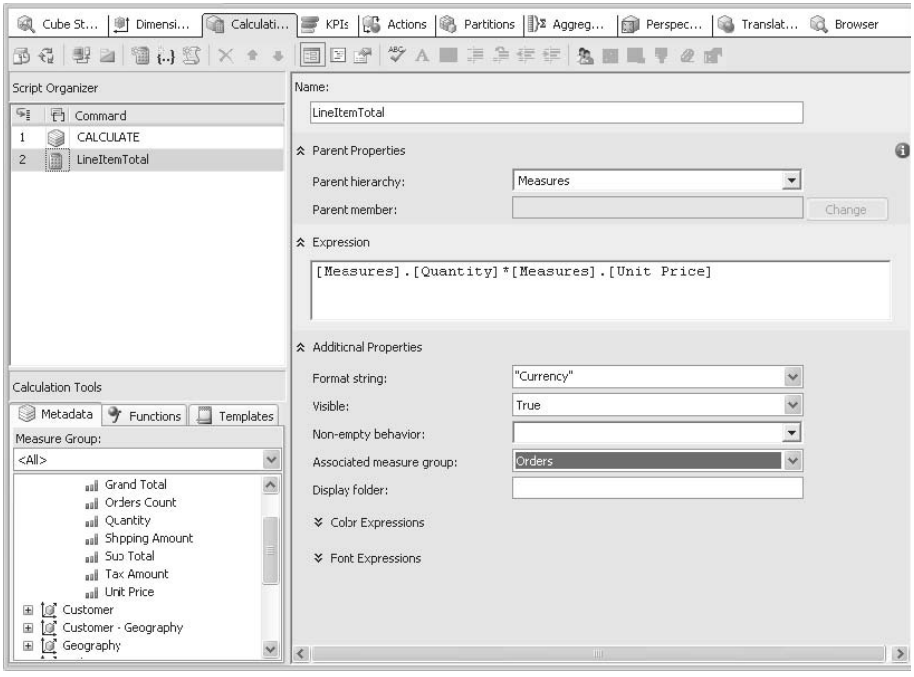




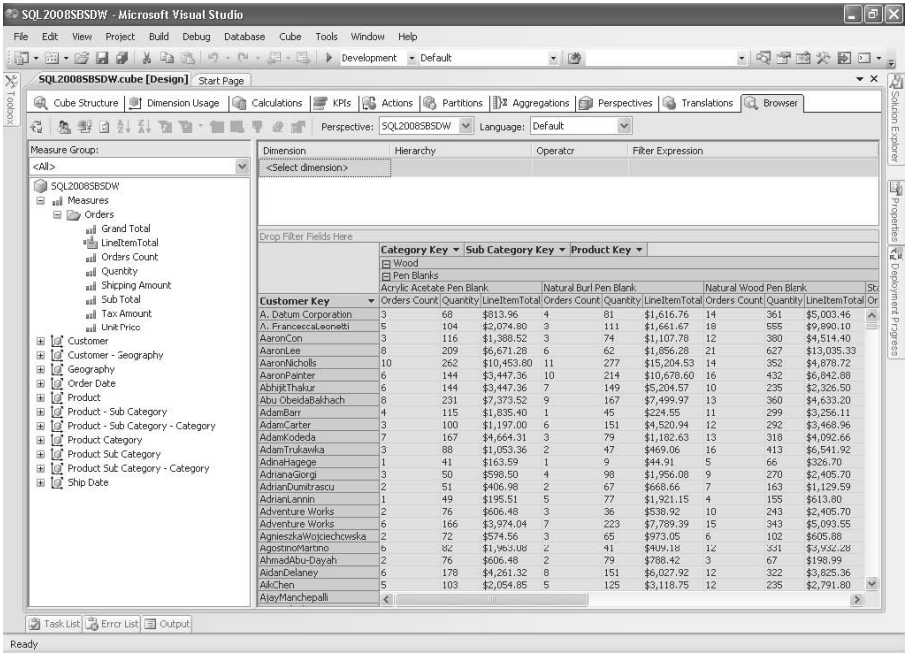
6. Klepněte na stránku Browser a zkontrolujte nové změny dimenze Customer.
7. Zavřete návrhář dimenzí Customer.
8. Nastavte vlastnost NameColumn pro Category Key v dimenzi Product Category na hodnotu *Category*. Nastavte vlastnost OrderBy na hodnotu *Name*.



9. Nastavte vlastnost NameColumn pro Product Key v dimenzi Product na hodnotu *ProductName*. Nastavte vlastnost OrderBy na hodnotu *Name*.
10. Nastavte vlastnost NameColumn pro Sub Category Key v dimenzi Product na hodnotu *SubcategoryName* z tabulky ProductSubCategory. Nastavte vlastnost OrderBy na hodnotu *Name*.
11. Nastavte vlastnost NameColumn pro Date Key v dimenzi Dates na hodnotu *FullDate*. Nastavte vlastnost Format pod vlastností NameColumn na hodnotu m/d/yyyy. Nastavte vlastnost OrderBy na hodnotu *Name*.
12. Nastavte vlastnost NameColumn pro Sub Category Key v dimenzi Product Sub Category na hodnotu *SubcategoryName*. Nastavte vlastnost OrderBy na hodnotu *Name*.
13. Nastavte vlastnost NameColumn pro Category Key v dimenzi Product Sub Category na hodnotu *Category* z tabulky ProductCategory. Nastavte vlastnost OrderBy na hodnotu *Name*.
14. Vyberte kartu Cube Structure. Nastavte vlastnost FormatString pro každou položku Measures v rámci krychle.
15. Klepněte na kartu Calculations a klepněte na tlačítko New Calculated Member.
16. Vytvořte výpočet, který vynásobí hodnoty Quantity a Unit Price a poskytne částku řádku objednávky a přidejte jej do skupiny veličin Orders.



17. Projděte krychli.



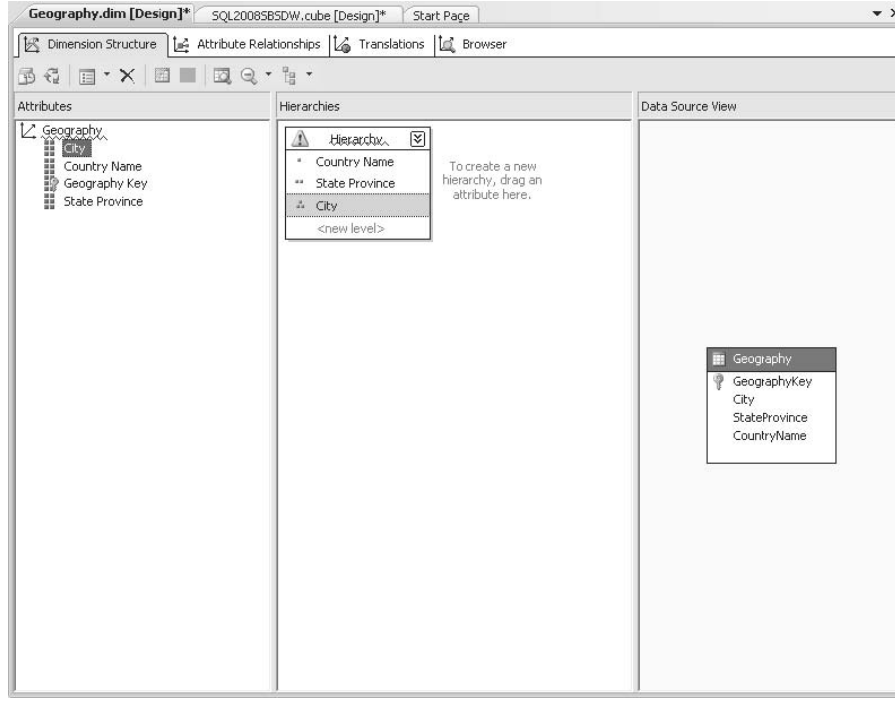
Hierarchie

Hierarchie sice nejsou pro krychli povinné, ale umožňují uživatelům procházet data intuitivnějším způsobem, protože napodobují běžné uspořádání dat v rámci podniku. Místo toho, aby uživatelé například museli samostatně načítat do prohlížeče krychlí zemi, oblast, město a jméno zákazníke, můžete jim nabídnout geografickou hierarchii zákazníků, kterou stačí přetáhnout a procházet.

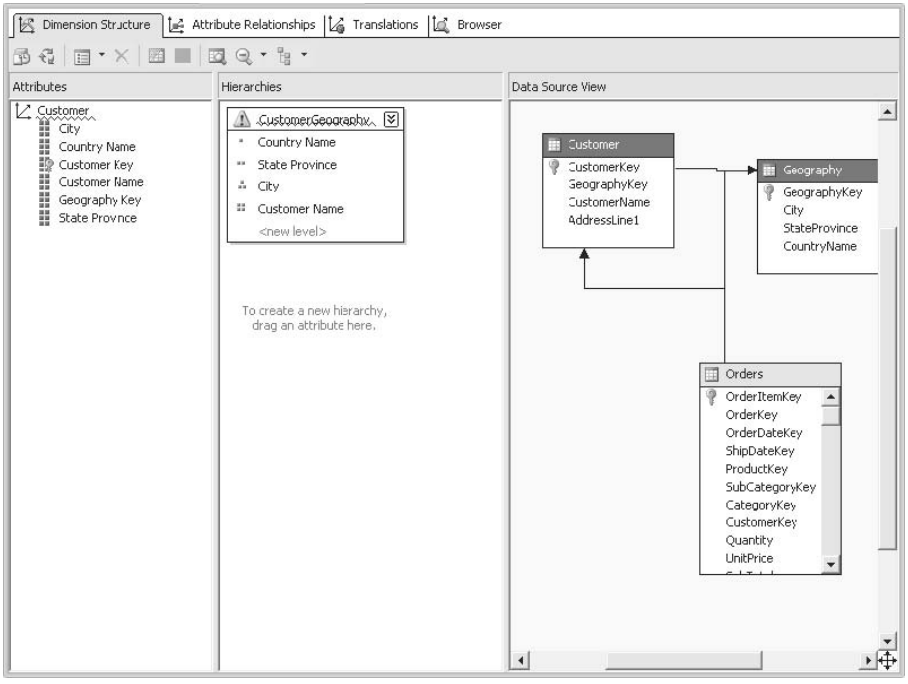
V následujícím cvičení definujete hierarchie navigace pro geografii, data a produkty.

Návrh hierarchií

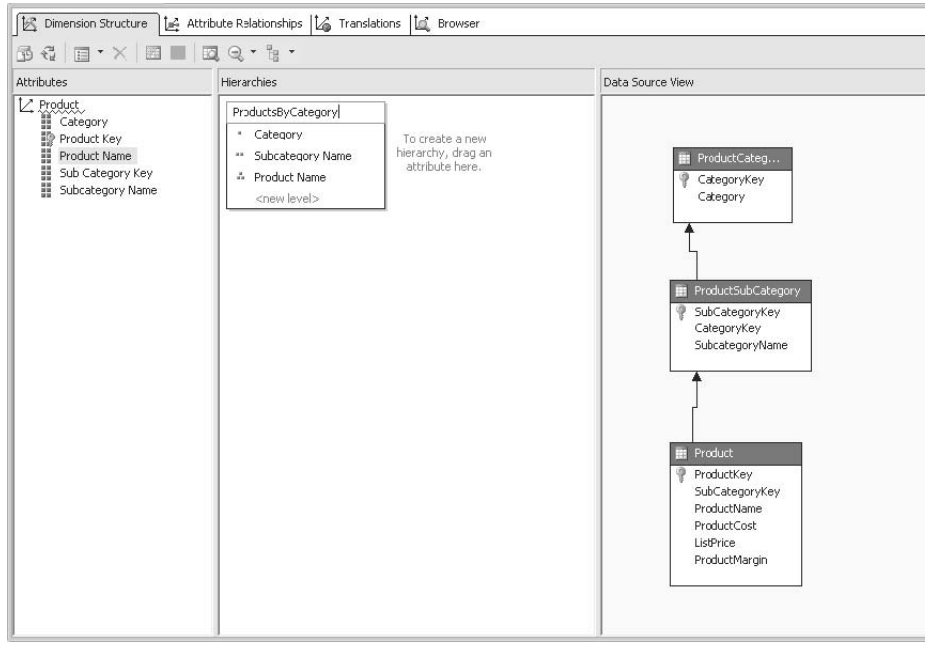
1. Poklepejte v okně Solution Explorer na dimenzi Geography.
2. Přidejte sloupce City, StateProvince a CountryName jako atributy dimenze Geography.
3. Přetáhněte atributy Country Name, State Province a City přes panel Hierarchies a vytvořte hierarchii geografie.

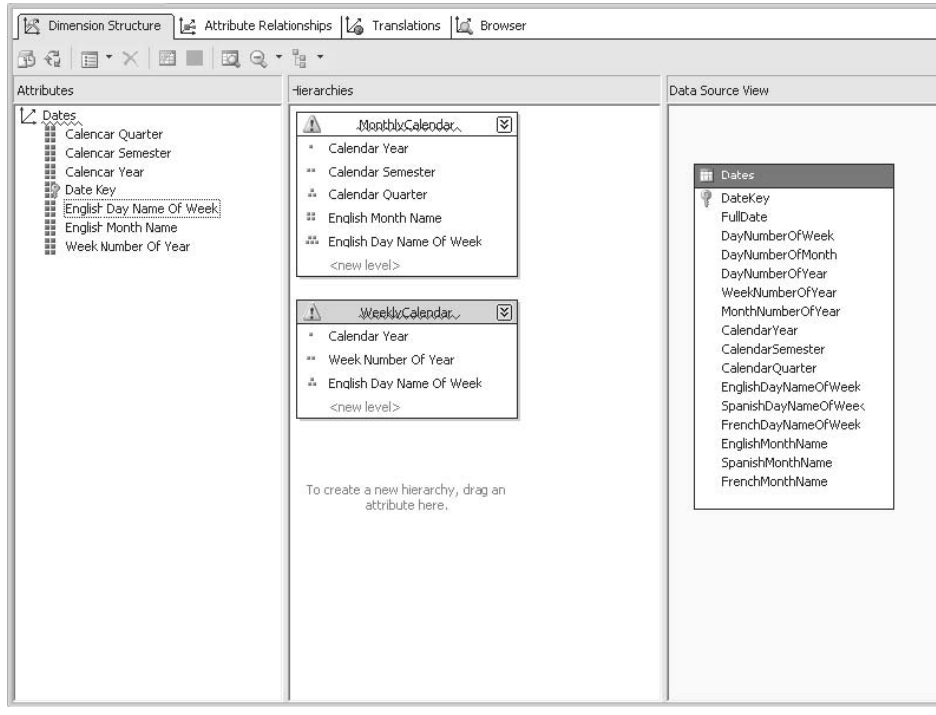


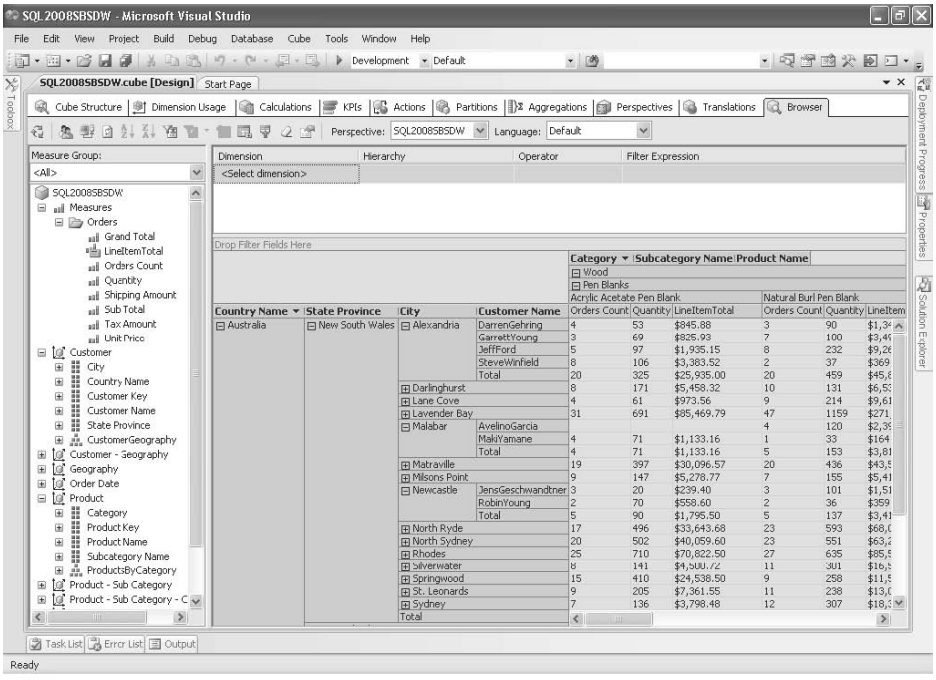
4. Poklepejte na dimenzi Customer. Klepněte pravým tlačítkem myši na tabulku Customer v podokně Data Source View a vyberte příkaz Show Related Tables.
5. Přidejte sloupce City, Country Name a State Province z tabulky Geography do atributů dimenze. Přidejte sloupec Customer Name z tabulky Customer do atributů dimenze.
6. Přetáhněte sloupce Country Name, State Province, City a Customer Name do podokna Hierarchies, abyste vytvořili hierarchii CustomerGeography.



7. Poklepejte na dimenzi Product. Přidejte související tabulky pro tabulky Product a poté ProductSubCategory.
8. Přidejte sloupce ProductName, SubCategoryName a Category do atributů dimenze.
9. Vytvořte hierarchii produktů pro sloupce Category, Sub Category Key a Product Name.
10. Poklepejte na dimenzi Dates a přidejte sloupce CalendarYear, CalendarSemester, CalendarQuarter, EnglishDayNameOfWeek, EnglishMonthName a WeekNumberOfYear do atributů dimenze.
11. Vytvořte hierarchii sloupců Calendar Year, Calendar Semester, Calendar Quarter, English Month Name a English Day Name Of Week.
12. Vytvořte hierarchii sloupců Calendar Year, Week Number Of Year a English Day Name Of Week.
13. Prohlédněte si ve své krychli nově vytvořené hierarchie.


Dimension Structure
Attribute Relationships
Translations
Browser
Attributes
Product
Category
Product Key
Product Name
Sub Category Key
Subcategory Name
Hierarchies
ProductsByCategory
Category
Subcategory Name
Product Name
<new level>
To create a new hierarchy, drag an attribute here.
Data Source View
ProductCateg...
CategoryKey
Category
ProductSubCategory
SubCategoryKey
CategoryKey
SubcategoryName
Product
ProductKey
SubCategoryKey
ProductName
ProductCost
ListPrice
ProductMargin


Dimension Structure
Attribute Relationships
Translations
Browser
Attributes
Dates
Calendar Quarter
Calendar Semester
Calendar Year
Date Key
English Day Name Of Week
English Month Name
Week Number Of Year
Hierarchies
MonthlyCalendar
Calendar Year
Calendar Semester
Calendar Quarter
English Month Name
English Day Name Of Week
<new level>
WeeklyCalendar
Calendar Year
Week Number Of Year
English Day Name Of Week
<new level>
To create a new hierarchy, drag an attribute here.
Data Source View
Dates
DateKey
FullDate
DayNumberOfWeek
DayNumberOfMonth
DayNumberOfYear
WeekNumberOfYear
MonthNumberOfYear
CalendarYear
CalendarSemester
CalendarQuarter
EnglishDayNameOfWeek
SpanishDayNameOfWee
FrenchDayNameOfWeek
EnglishMonthName
SpanishMonthName
FrenchMonthName



Klíčové indikátory výkonu, oddíly, perspektivy a překlady

Základní krychle sice může uspokojit podnikové požadavky s omezením pouze na dimenze, veličiny a hierarchie, ale služba SSAS poskytuje další sadu funkcí, které rozšiřují možnosti a regionální podporu krychlí.

Klíčové indikátory výkonu (KPI)

Všechny organizace mají sadu cílů, podle kterých se rozhodují při každodenním provozu, jako např.:

- Zvýšení prodejů o 10 procent oproti stejnému období předchozího roku
- Zvýšení zákaznické základny o 20 procent
- Zvýšení průměrné částky objednávky o 15 procent

Klíčové indikátory výkonu (KPI – key performance indicator) dovolují přenést podnikové cíle do krychle, aby bylo možné zpřístupnit indikátory plnění cílů.

Klíčové indikátory výkonu se definují jako sady výpočtů, které se používají k účelům porovnání. Pomocí multidimenzionálních výrazů (MDX – multidimensional expression) lze definovat výpočty, které se provedou s daty krychle, cíl porovnání, výraz indikující stav plnění cíle a trend, který ukazuje, nakolik je plnění cíle úspěšné.

Oddíly

Krychle se často vytvářejí pro zdrojová data, která obsahují desítky milionů nebo miliardy řádků dat. Navíc se do zdroje dat pravidelně přidávají nové sady dat. Chcete-li, aby se nové sady dat ve zdroji projevily i v krychli, musíte data zpracovat. Během zpracování není krychle dostupná pro koncové uživatele.

Kvůli zvýšení efektivity zpracování a zlepšení výkonu krychle můžete v rámci krychle definovat oddíly. Obdobně jako v případě funkce rozdělování tabulek umožňuje *oddíl krychle* (cube partition) určit kritéria, pomocí nichž se tabulka faktů při zpracování rozdělí na více částí. Oddíl krychle však nemá vliv na základní úložiště tabulky faktů. Místo toho definuje ekvivalent klauzule WHERE, který služba SSAS použije během zpracování.

Po vytvoření definice může služba SSAS zpracovávat všechny oddíly paralelně a poté výsledky spojit dohromady. Můžete také zaměřit zpracování na jediný oddíl a tím minimalizovat rozsah dat, která je nutné zpracovat.

Kromě toho, že poskytují další možnosti zpracování, lze pomocí oddílů využít při analýze krychle prostředky více počítačů. Oddíl se definuje pro tabulku faktů v pohledu DSV. Vzhledem k tomu, že v rámci krychle je možné definovat a využít více pohledů DSV, můžete rozdělit velkou tabulku mezi více serverů a i nadále kombinovat data dohromady do jediné krychle pomocí oddílů.

Perspektivy

Perspektivy dovolují administrátorům zpřístupnit různé pohledy na krychli. Krychle může obsahovat mnoho dimenzí a skupin veličin, které zahrnují více oblastí společnosti. Někteří uživatelé však nepotřebují přístup ke všem dimenzím skupiny veličin, aby dokázali analyzovat svou podnikovou oblast. Při definování perspektivy si můžete zvolit skupiny veličin, veličiny, dimenze, atributy, hierarchie, výpočty a klíčové indikátory výkonu.

Uživatel pak vybere konkrétní perspektivu, aby mohl omezit prvky dostupné pro analýzu. Perspektivy však nejsou navrženy jako bezpečnostní struktury. Kdokoli se znalostí tvorby výrazů MDX totiž může vždy získat přístup k libovolné části krychle bez ohledu na aktuálně platnou perspektivu. Perspektivy je tedy nutné používat pouze pro zjednodušení práce uživatelům, kteří krychli procházejí.

Překlady

Překlady (translation) dávají datům vícejazyčný aspekt. Služba SSAS není modul jazykových překladů. Pokud však poskytnete překlady, může tato služba uživatelům nabídnout volbu jazyka ze seznamu a uživatelé mohou procházet krychli ve svém upřednostňovaném jazyce.

Překlady existují na dvou různých úrovních – struktura krychle a zdrojová data. Pokud chcete zobrazit strukturní prvky krychle v daném jazyce, vložíte překlad do názvu objektu, např. dimenze, hierarchie či atributu, v definici krychle. Jestliže chcete zobrazit překlady pro členy v rámci dimenze, musíte poskytnout překlad v rámci zdrojových dat a poté namapovat přeložené sloupce na odpovídající atributy v krychli.

Dolování dat

Služba Analysis Services vyvolává představu procházení krychlí. Modul SSAS však není omezen na práci s krychlemi. Služba SSAS rovněž zahrnuje výkonný modul dolování dat.

Dolování dat (data mining) dovoluje analyzovat velké objemy dat, aby bylo možné odhalit skryté vzory. Pomocí dolování dat lze sice analyzovat historické záznamy o klimatu a zjistit, že slunce vychází každé ráno a zapadá každý večer, ale taková analýza by neposkytla mnoho podnikových znalostí. Bez dolování dat by však bylo mnohem těžší najít optimální rozmístění více než 8000 produktů v regálech typického supermarketu na základě analýzy nákupních vzorů zákazníků.

S algoritmy dolování dat se ve svém životě setkáváme všichni, i když si to neuvědomujeme. Jestliže jste někdy navštívili supermarket, interagujete s dolováním dat od chvíle, kdy zastavíte na parkovišti. Při vstupu do dveří najdete čerstvé ovoce a zeleninu po pravé straně za předpokladu, že se obchod nachází v USA. Pokud jste v Evropě, pravděpodobně uvidíte čerstvé ovoce a zeleninu vlevo, protože Američané po příchodu do obchodu obvykle míří doprava, zatímco Evropané zpravidla směřují doleva. Možná jste se domnívali, že vedoucí obchodu měl dobrý nápad a vyšel vstříc zákazníkům, když umístil slané pochoutky typu čipsů na každý konec regálu s pivem. Poloha žádné položky ve velkém obchodě není náhodná a nesouvisí s ohledem na potřeby zákazníků. Všechny předměty v supermarketu jsou umístěny tak, aby se maximalizovala útrata zákazníků, kterým se kolem zboží, které potřebují, neustále nabízejí „impulzivní" položky. Každá z „impulzních" položek je určena aplikací technik dolování dat na miliony nákupních vozíků, které jsou každý den zpracovány. Slané pochoutky s vysokým rabatem jsou na konci regálu s pivem proto, že u osoby kupující pivo je současný nákup čipsů mnohem pravděpodobnější než u zákazníka, který kupuje láhev vína.

Cílem dolování dat v rámci podniku je dospět k podnikovému rozhodnutí, jehož pravděpodobnost je mnohem vyšší. Pokud se Podnik A vždy rozhoduje na základě náhodné volby s pravděpodobností 50 procent, bude pravděpodobnost příznivého výsledku právě 50 procent. Jestliže se jeho konkurent, Podnik B, dokáže rozhodovat s lepším než náhodným výsledkem (pravděpodobností vyšší než 50 procent), bude mít rozhodování Podniku B větší šanci na úspěch. Pokud podnik trvale přijímá rozhodnutí, která v průměru přinášejí vyšší zisky, dokáže mnohem snáze překonat a porazit své konkurenty.

Algoritmy

V centru dolování dat leží sada algoritmů, které služba SSAS aplikuje na podniková data. Algoritmus dolování dat je matematická rovnice, která po aplikaci na data dokáže určit pravděpodobnost požadovaného výsledku.

Poznámka: Když učitel matematiky tvrdil, že „vše na světě lze popsat matematicky", měl pravdu, i když jsme jej všichni považovali za blázna.

Kategorie algoritmů

Algoritmy dolování dat lze rozdělit do šesti základních kategorií:

- Klasifikace
- Asociace
- Regrese
- Předpovídání
- Analýza sekvencí
- Analýza odchylek

Algoritmy *klasifikace* umožňují předvídat výsledek na základě vstupních atributů. Atributy dostanou stejnou váhu a poté procházejí více iteracemi dat, aby se objevily přirozené skupiny.

Algoritmus *asociace* dovoluje předvídat korelaci, kde nezáleží na pořadí uvažovaných položek. Výsledkem algoritmu asociace je pravděpodobnost, že nastane konkrétní výsledek, v závislosti na aktuálních podmínkách. Pokud například zákazník již do svého nákupního košíku vložil basu piv, může algoritmus asociace určit pravděpodobnost, že zákazník rovněž koupí velké balení čipsů.

Regresní algoritmy umožňují předpovědět budoucí výkony na základě minulých výsledků. Klíčem k regresním algoritmům je plynulá změna předpovídaného atributu.

Atributy lze třídit takto:

- Diskrétní – sada konkrétních hodnot, které spolu nesouvisejí a nepřekrývají se
- Plynulé – data jsou tvořena řadami, které tvoří postupně rostoucí sekvenci
- Uspořádané – vstupní data jsou seřazena
- Cyclické – data obsahují opakující se vzory
- Diskretizované – plynulá data jsou rozdělena na diskrétní sady rozsahů

Můžete mít také diskretizovaný atribut. *Diskretizované atributy* obsahují plynulé řady dat, které jsou pro analýzu rozděleny do intervalů, např. při rozdělení atributu příjmu či věku do nepřekrývajících se sad hodnot. Můžete například rozdělit sloupec příjmu do třech kategorií, které pak pro statistické účely označíte jako nízký, střední a vysoký.

Algoritmy *předpovídání* se pokoušejí odhadnout budoucí výkon na základě minulých výsledků stejně jako regresní algoritmy. Předpověď však na vstupu vyžaduje časové řady a snaží se předpovědět budoucí hodnotu nikoli pouze na základě minulého výkonu, ale také s ohledem na cyklické trendy v datech, např. sezónnost nákupu oděvů.

Analýza *sekvencí* předpovídá pravděpodobnost, že v budoucnu dojde k určité události, na základě událostí, které nastaly v minulosti. U analýzy sekvencí je při předvídání pravděpodobnosti budoucí události stejně důležité pořadí minulých událostí jako vlastní událost, ke které došlo.

Analýza *odchylek* dovoluje najít data, která se liší od normálního stavu. Mezi aplikace algoritmu odchylek patří vyhledávání podezřelých osob na letišti podle toho, že se jejich chování výrazně odchyluje od normálu.

Algoritmy dolování dat služby SSAS

Systém SQL Server se dodává se sedmi algoritmy dolování dat, které pokrývají všechny kategorie dolování dat kromě analýzy odchylek. K dispozici jsou tyto algoritmy dolování dat:

- Naive Bayes
- Microsoft Decision Trees
- Microsoft Linear Regression
- Microsoft Regression Trees
- Microsoft Clustering
- Sequence Clustering
- Association Rules
- Neural Networks
- Time Series

Modely a struktury dolování

Proces dolování začíná definováním struktury dolování. Model dolování je založen na zdroji dat, což může být buď relační databáze, nebo krychle OLAP. Po definování dochází ke zpracování modelu dolování. Zpracování modelu dolování se označuje jako „trénování" modelu. V podstatě se jedná pouze o to, že služba SSAS pomocí příkazu *OPENQUERY* načte data z definovaného zdroje dat a zpracuje všechna data vybraným algoritmem dolování.

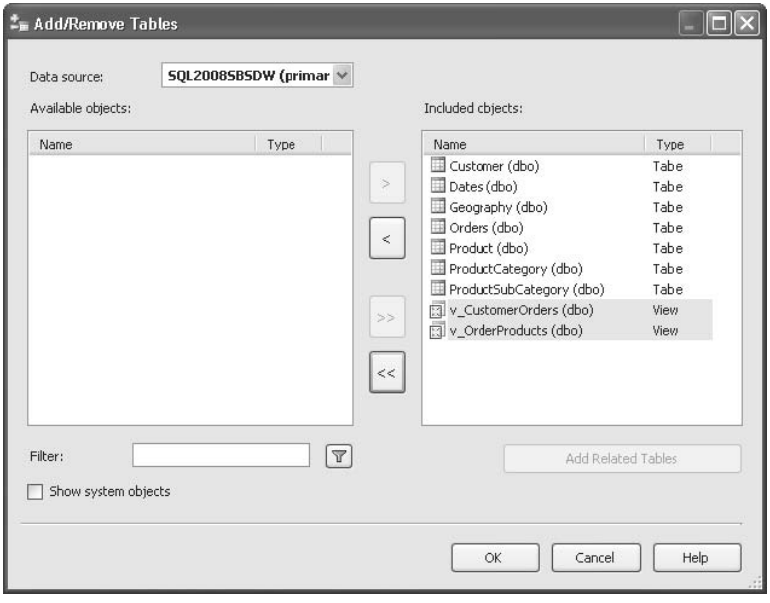
Poznámka: Každý algoritmus provádí výpočty odlišným způsobem a algoritmy dolování služby SSAS lze aplikovat na více problémů. Struktura dolování proto obvykle zahrnuje více algoritmů dolování. Každý algoritmus nastavený pro strukturu dolování se nazývá model dolování.

Každý řádek, který je načítán ze zdroje dat ke zpracování, se označuje jako *případ* (case). Zpracovávané sloupce či atributy fungují jako proměnné algoritmu. Po dokončení trénování je struktura dolování připravena k analýze a bude obsahovat definici každého modelu dolování a také data, pro která ve struktuře proběhlo trénování.

V následujícím cvičení vytvoříte model dolování, který umožní určit produkty, které se budou s nejvyšší pravděpodobností nacházet ve stejném nákupním košíku.

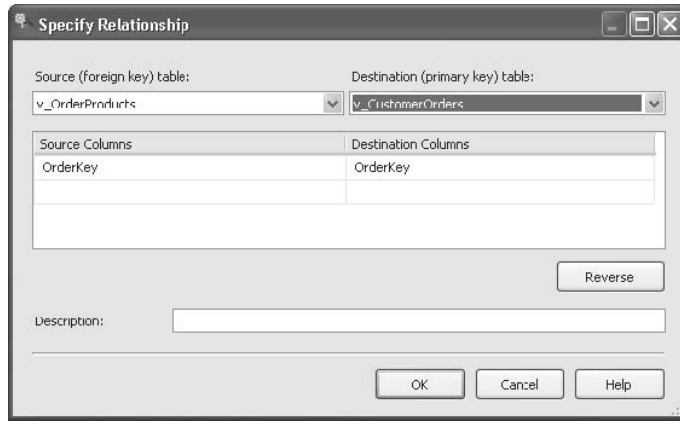
Sestavení modelu a struktury dolování

1. Poklepáním na pohled DSV s názvem SQL2008SBSDW spusťte editor pohledů DSV.
2. Klepněte na možnost Add/Remove Objects, přidejte do pohledu DSV pohledy v_CustomerOrders a v_OrderProducts a klepněte na tlačítko OK.

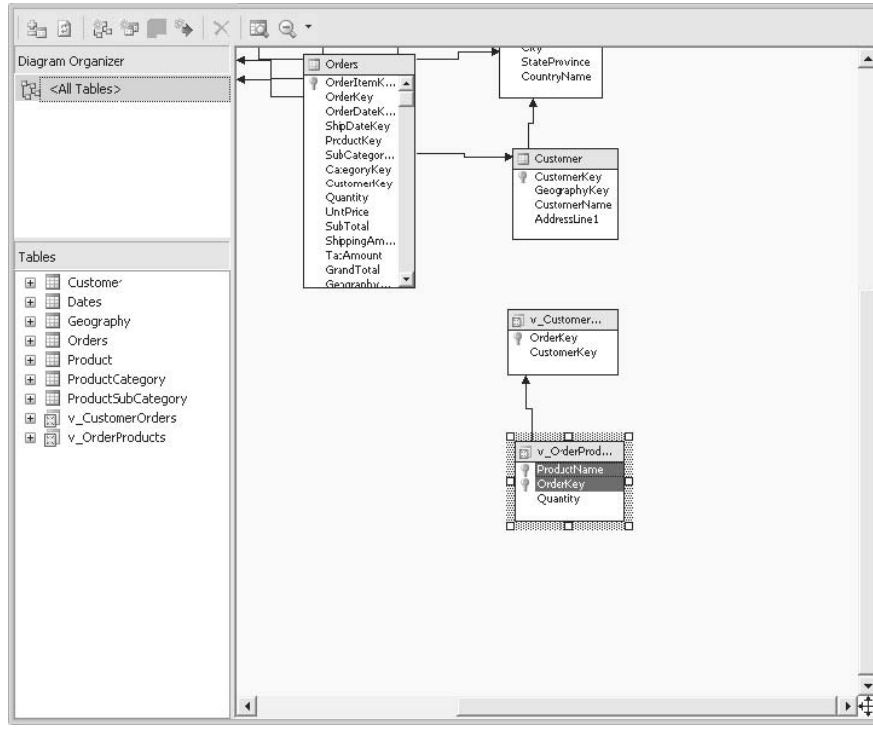


3. Klepněte pravým tlačítkem myši na sloupec OrderKey v pohledu v_OrderProducts a vyberte příkaz New Relationship. Spojte sloupce OrderKey mezi pohledy v_OrderProducts

a v_CustomerOrders, klepněte na tlačítko OK a vytvořte logický primární klíč klepnutím na tlačítko Yes.



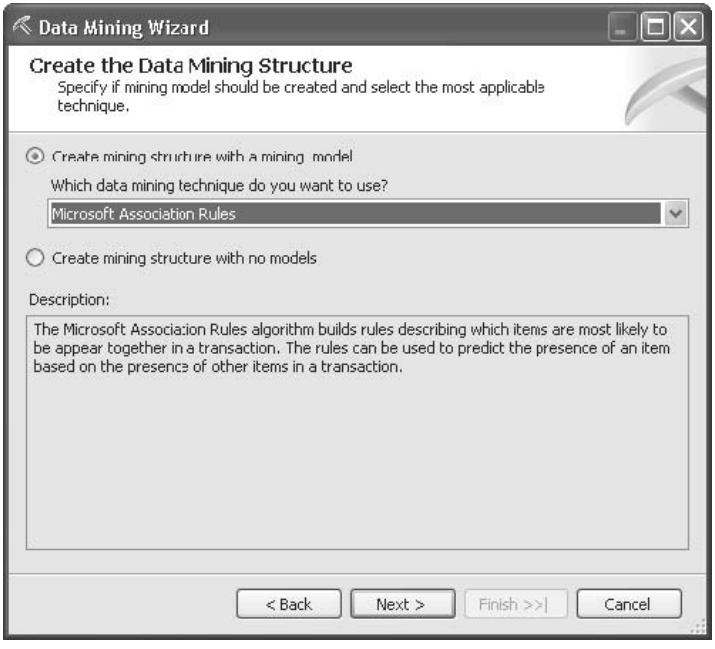
4. Vyberte sloupce OrderKey a ProductName v pohledu v_OrderProducts, klepněte pravým tlačítkem a vyberte položku Set Logical Primary Key a uložte a zavřete pohled DSV.



5. V okně Solution Explorer klepněte pravým tlačítkem myši na položku Mining Structures, vyberte příkaz New Mining Structure a klepněte na tlačítko Next.

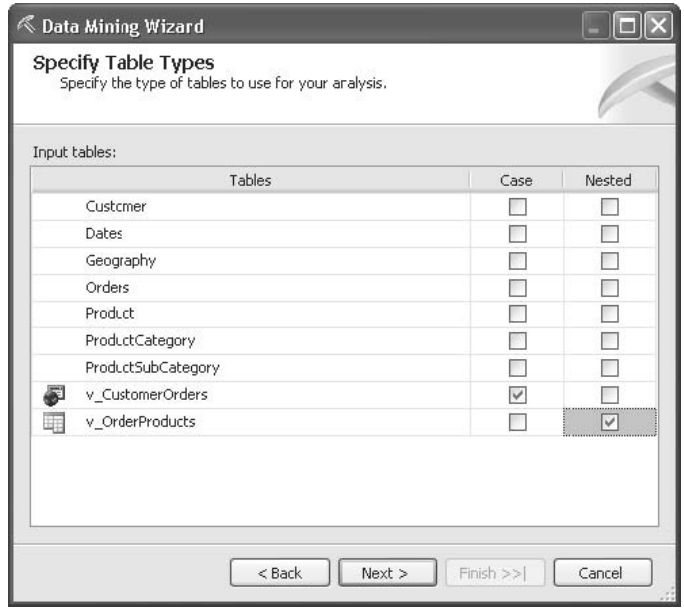
6. Vyberte přepínač From Existing Relational Database... a klepněte na tlačítko Next.

7. Z rozevírací nabídky zvolte položku Microsoft Association Rules a klepněte na tlačítko Next.

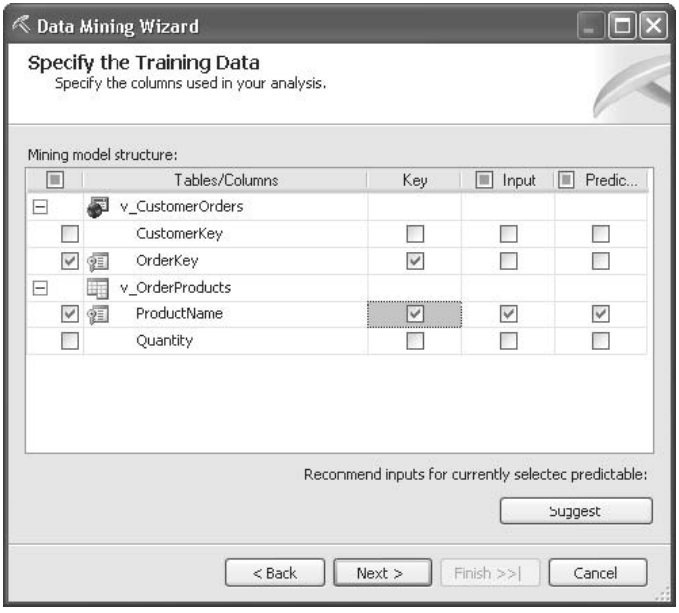


8. Vyberte pohled zdroje dat SQl2008SBSDW, který jste ve svém projektu již definovali, a klepněte na tlačítko Next.

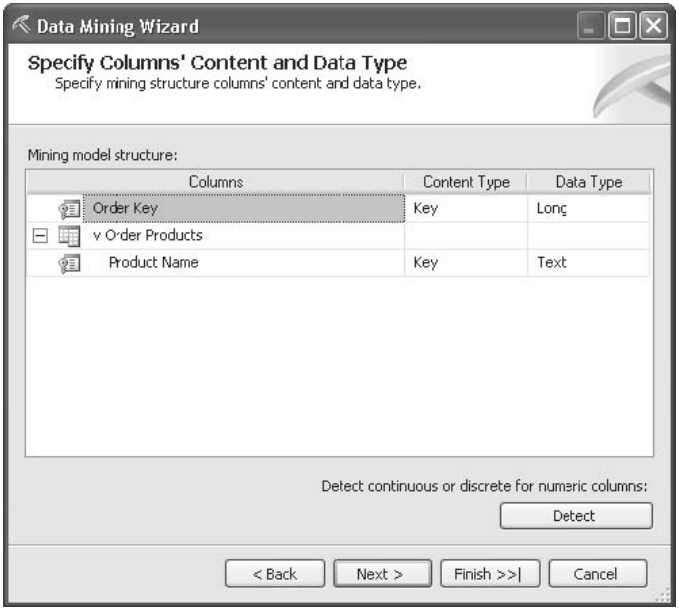
9. Jako tabulku Case vyberte pohled v_CustomerOrders a jako tabulku Nested pohled v_OrderProducts a klepněte na tlačítko Next.



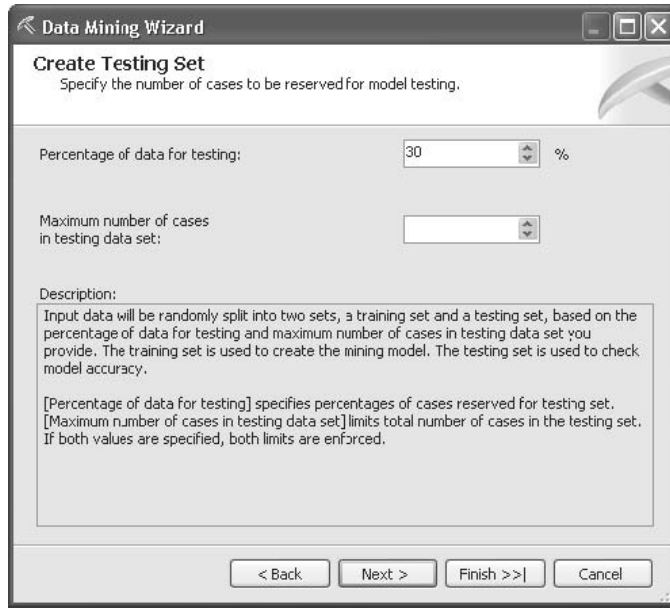
10. Vyberte položku OrderKey jako sloupec Key pro pohled v_CustomerOrders a položku ProductName jako sloupce Key, Input a Predictable pro vnořenou tabulku a klepněte na tlačítko Next.



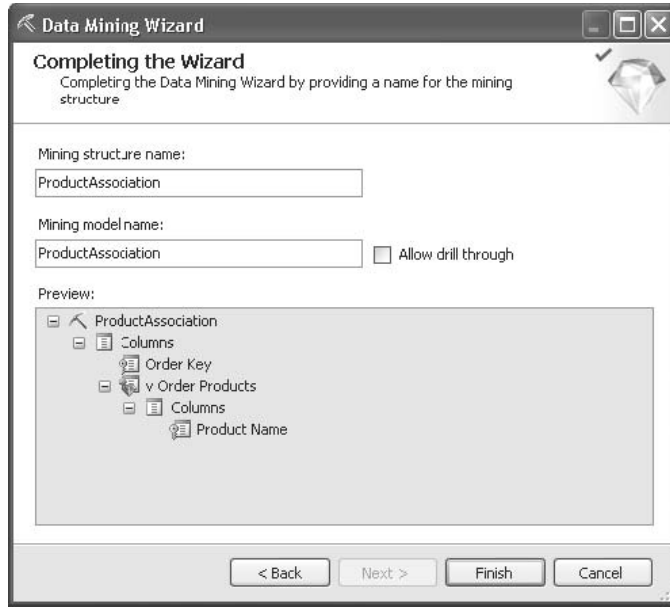
11. Ponechejte nastavené výchozí hodnoty obsahu a datových typů a klepněte na tlačítko Next.



12. Nastavte v poli Percentage Of Data For Testing hodnotu 30 % a klepněte na tlačítko Next.



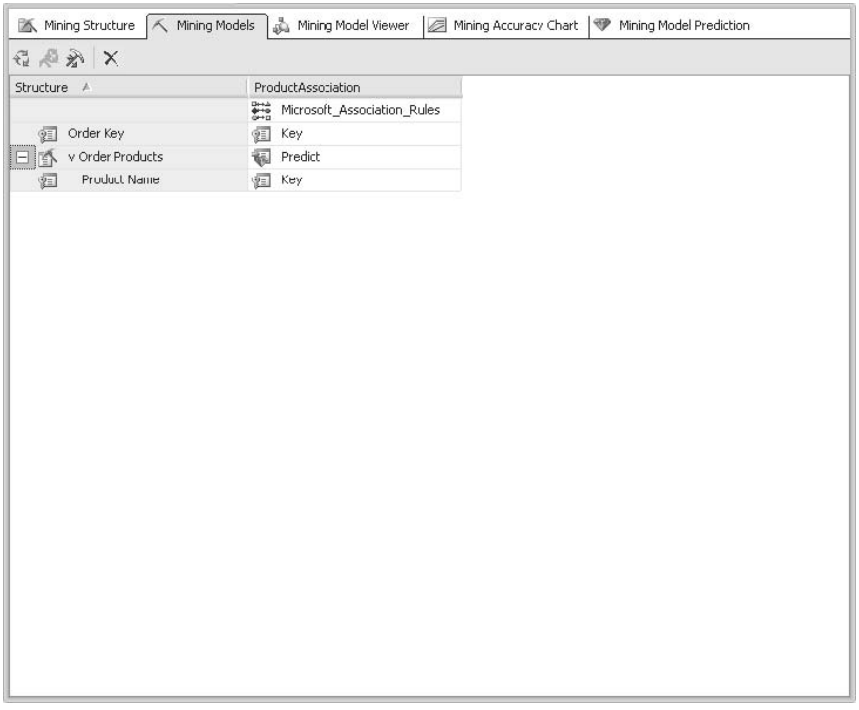
13. Pojmenujte model a strukturu dolování jako Product Association a klepněte na tlačítko Finish.

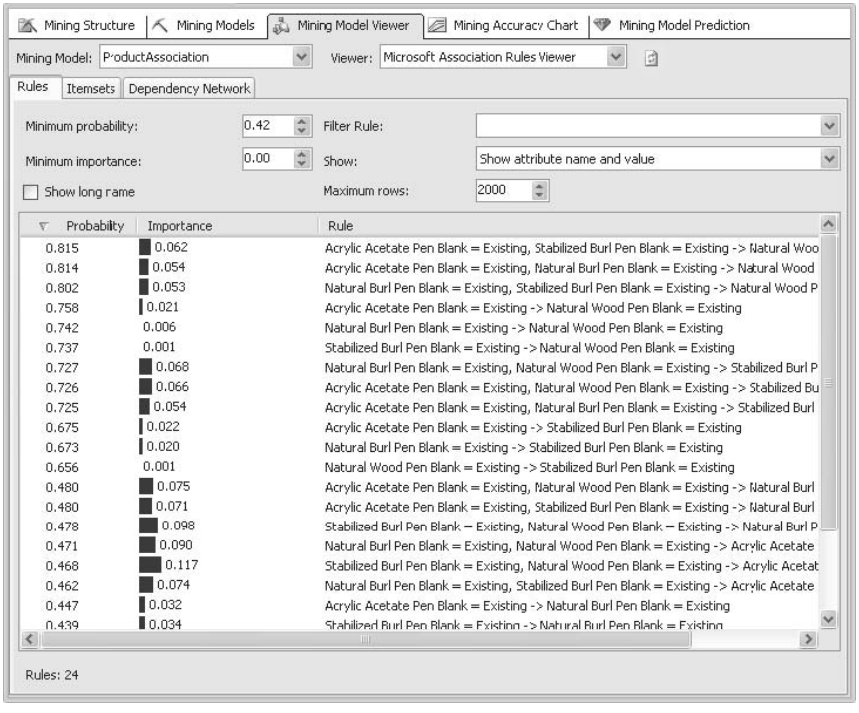


14. Klepněte na kartu Mining Models a zkontrolujte nastavení modelu ProductAssociation.

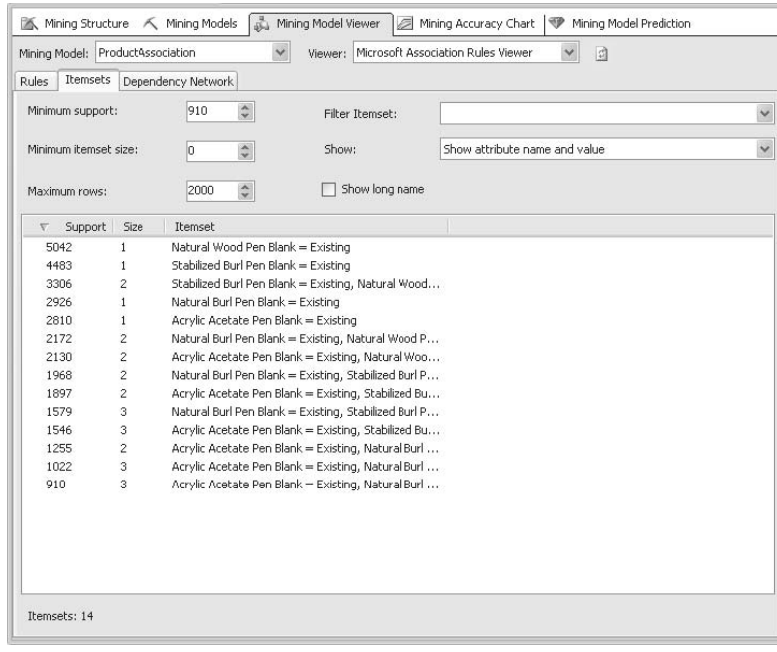
15. Zaveďte svůj projekt do služby SSAS a zpracujte strukturu dolování.

16. Vyberte pohled modelu dolování, abyste zkontrolovali pravidla nalezená v datech.

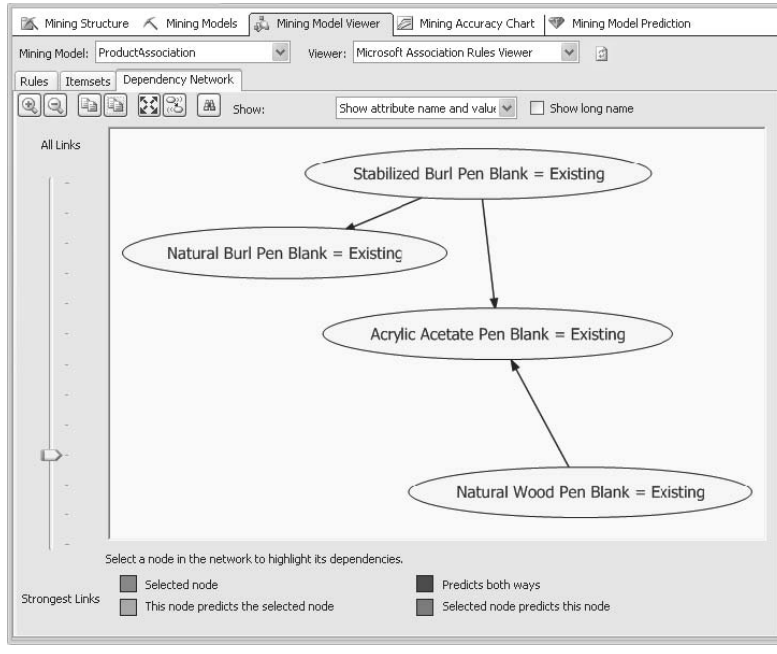




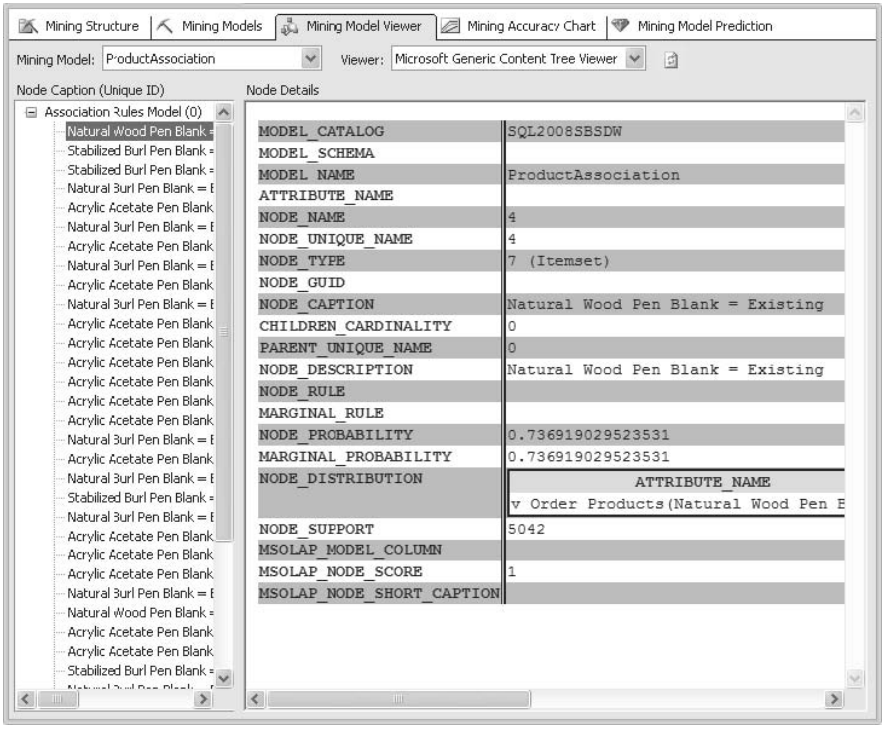
17. Po klepnutí na kartu Itemsets načtěte sady, které byly nalezeny v datech.



18. Vyberte kartu Dependency Network a prohlédněte si vztahy mezi různými produkty.



19. Vyberte prohlížeč Generic Content a zkontrolujte výsledky.



Praktické dolování dat

Dolování dat se považuje za doménu „odborníků v bílých pláštích s 18 doktoráty v aplikované matematice" a za něco nedostupného pro „normální" lidi. Dolování dat není žádné tajné umění ani čarodějnictví. Jedná se pouze o aplikaci matematických (přesněji řečeno statistických) metod na sady dat. V posledních letech nabídlo několik dodavatelů včetně společnosti Microsoft sady nástrojů, které „pouhým smrtelníkům" umožňují definovat a využívat dolování dat, aniž by museli rozumět základní teorii.

Pokud věnujete čas na seznámení se s matematickými principy, zjistíte, že dolování dat se příliš neliší od jiných operací se systémem SQL Server. Kvůli odhalení závoje tajemství, který obestírá dolování dat, si v následujícím scénáři ukážeme aplikaci algoritmu Naive Bayes spolu se základními výpočty. Příklad dokumentuje odvození výsledků z modelu dolování dat. Poté byste mohli vytvořit sadu dat odpovídající scénáři, zpracovat ji algoritmem Naive Bayes a zobrazit stejné výsledky ve službě SSAS.

Algoritmus Naive Bayes má tento tvar:

$$P(H|E) = \frac{P(E|H) * P(H)}{P(E)}$$

P(H) definuje pravděpodobnost hypotézy. P(E) určuje pravděpodobnost důkazu. P(E|H) udává pravděpodobnost hypotézy s ohledem na získané důkazy.

Cílem průzkumu s 1 277 účastníky bylo určit nákupní preference. Každá osoba dostala řadu otázek, které měly zjistit, jaké faktory ovlivňují rozhodování o nákupu určitého vozidla. Skupina zahrnovala 652 mužů a 625 žen. Úkolem je na základě odpovědí na otázky průzkumu zjistit, zda je kupcem auta muž či žena.

Pokud byste náhodně odhadli, že kupující je muž, měli byste 51,1 procent pravděpodobnosti správné odpovědi a 48,9 procent, že odpověď nebude správná – což je statistická remíza. Aplikujete-li na odpovědi průzkumu algoritmus Naive Bayes, měli byste svůj odhad vylepšit a zvýšit pravděpodobnost správné odpovědi na výše položenou otázku.

Tabulka 26.1 shrnuje pět uvažovaných faktorů a také počet členů každé skupiny, která na pět konkrétních faktorů odpovídá kladně či záporně.

Tabulka 26.1: Faktory ovlivňující nákup vozidla

Odpověď	Prestiž značky		Typ karosérie		Záruka		Hodnocení zákazníky		Cena	
	M	Ž	M	Ž	M	Ž	M	Ž	M	Ž
Ano	126	482	238	522	631	479	492	603	141	463
Ne	518	85	397	58	20	143	53	6	492	161

Na základě výsledků průzkumu můžete vypočítat pravděpodobnosti každé odpovědi pro každý faktor. Tabulka 26.2 ukazuje pravděpodobnost každé odpovědi na průzkum.

Tabulka 26.2: Pravděpodobnost jednotlivých odpovědí na průzkum

Odpověď	Prestiž značky		Typ karosérie		Záruka		Hodnocení zákazníky		Cena	
	M	Ž	M	Ž	M	Ž	M	Ž	M	Ž
Ano	19,6 %	85,0 %	37,5 %	90,0 %	96,9 %	77,0 %	91,8 %	99,0 %	22,3 %	74,2 %
Ne	80,4 %	15,0 %	62,5 %	10,0 %	3,1 %	23,0 %	8,2 %	1,0 %	77,7 %	25,8 %

Pokud byste se dozvěděli, že pro respondenta průzkumu není důležitý typ karosérie a jako důležité faktory uvedl značku, záruku, hodnocení zákazníky a cenu, mohli byste nyní mnohem lépe předpovědět, zda se jedná o muže či o ženu.

Chcete-li vypočítat, že respondent je muž, vynásobíte pravděpodobnosti odpovědi pro každý faktor: 19,6 % * 62,5 % * 96,9 % * 91,8 % * 22,3 %. Jestliže chcete vypočítat, že respondent je žena, vynásobíte pravděpodobnosti odpovědi pro každý faktor: 85,0 % * 10,0 % * 77,0 % * 99,0 % * 74,2 %. V této fázi dosahuje pravděpodobnost, že respondentem je muž, hodnoty 0,01238, a pravděpodobnost, že se jedná o ženu, má hodnotu 0,02354.

Chcete-li výpočet dokončit, vydělte číslo pravděpodobnosti pro muže součtem obou pravděpodobností. Totéž proveďte s hodnotu pro ženu. Konečný výsledek udává, že v závislosti na poskytnutých odpovědích průzkumu je kupcem s pravděpodobností 65,54 procenta žena, zatímco pravděpodobnost, že se jedná o muže, činí pouhých 34,46 procent.

Stručný přehled kapitoly 26

Požadovaná akce	Postup
Definování pohledu na zdroj dat	■ Vytvořte zdroj dat. ■ Vytvořte pohled na zdroj dat založený na zdroji dat. ■ Vyberte tabulky či pohledy, které chcete zahrnout do pohledu DSV.
Úpravy vlastností dimenze	Vyberte dimenzi a změňte přidružené vlastnosti v podokně Properties.
Změna hodnot zobrazených pro člena dimenze	Určete sloupec NameColumn atributu.
Změna metody agregace veličiny	Vyberte veličinu a změňte vlastnost AggregationFunction.
Nastavení formátování zobrazení	Konfigurujte vlastnost *FormatString*.
Přidání atributu do dimenze	■ Poklepejte na dimenzi v okně Solution Explorer. ■ Přetáhněte sloupec z podokna DSV do seznamu Attributes.
Přidání hierarchie do dimenze	■ Přetáhněte atributy do podokna Hierarchies. ■ Uspořádejte atributy v pořadí, ve kterém chcete hierarchii procházet.
Definice pořadí řazení atributu	Vyberte atribut a nastavte vlastnost *OrderBy*.
Přidání vlastního výpočtu do krychle	Vyberte kartu Calculations návrháře krychle, přidejte nový výpočet a uveďte použitý výraz MDX.
Vytvoření struktury dolování dat	■ V okně Solution Explorer klepněte pravým tlačítkem myši na položku Mining Structures a vyberte příkaz New Mining Structure. ■ Vyberte metodu dolování. ■ Nastavte použitý pohled DSV. ■ Nastavte tabulku či tabulky a pohled či pohledy, které se použijí pro tabulku případů a volitelně pro vnořenou tabulku. ■ Vyberte sloupec či sloupce Key, Input a Predictable. ■ Nakonfigurujte typ obsahu.
Trénování modelu dolování	Zpracujte strukturu dolování načtením zdrojových dat a výpočtem pomocí vybraného algoritmu dolování.