VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
BRNO UNIVERSITY OF TECHNOLOGY

FAKULTA INFORMAČNÍCH TECHNOLOGIÍ
ÚSTAV INFORMAČNÍCH SYSTÉMŮ

FACULTY OF INFORMATION TECHNOLOGY
DEPARTMENT OF INFORMATION SYSTEMS

DATOVÉ SKLADY A OLAP V PROSTŘEDÍ MS SQL SERVERU

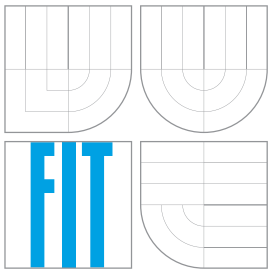
BAKALÁŘSKÁ PRÁCE
BACHELOR'S THESIS

AUTOR PRÁCE
AUTHOR

PAVEL RUDOL

BRNO 2010

VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
BRNO UNIVERSITY OF TECHNOLOGY



FAKULTA INFORMAČNÍCH TECHNOLOGIÍ
ÚSTAV INFORMAČNÍCH SYSTÉMŮ

FACULTY OF INFORMATION TECHNOLOGY
DEPARTMENT OF INFORMATION SYSTEMS

DATOVÉ SKLADY A OLAP V PROSTŘEDÍ MS SQL SERVERU

DATA WAREHOUSING AND OLAP IN MS SQL SERVER ENVIRONMENT

BAKALÁŘSKÁ PRÁCE
BACHELOR'S THESIS

AUTOR PRÁCE — PAVEL RUDOL
AUTHOR

VEDOUCÍ PRÁCE — Ing. VLADIMÍR BARTÍK, Ph.D.
SUPERVISOR

BRNO 2010

Abstrakt

Tato práce se zabývá problematikou datových skladů a OLAP analýzy. Je zde uvedena jejich definice a využití. Dále je zde popsáno prostředí MS SQL Serveru pro podporu vývoje datových skladů. Zde popsané technologie jsou využity v ukázkové aplikaci.

Abstract

This thesis deals with the issue of data warehouses and the OLAP. It contains their definitions and usage. The MS SQL server as a tool for supporting the data storage development and OLAP is also described. The technologies are used in a sample application

Klíčová slova

Datové sklady, OLAP (online analytical processing), OLTP (online transaction processing), Bussiness Intelligence (BI), datová kostka, MS SQL Server, SSIS (integrační služby MS SQL serveru), SSAS (analytické služby MS SQL serveru), SSRS (reportovací služby MS SQL serveru)

Keywords

Data warehouses, OLAP (online analytical processing), OLTP (online transaction processing), Bussiness Intelligence (BI), data cube, MS SQL Server, SSIS (SQL Server Integration Services), SSAS (SQL Server Analytical Services), SSRS (SQL Server Reporting Services)

Citace

Pavel Rudol: Datové sklady a OLAP v prostředí MS SQL Serveru, bakalářská práce, Brno, FIT VUT v Brně, 2010

Datové sklady a OLAP v prostředí MS SQL Serveru

Prohlášení

Prohlašuji, že jsem tuto bakalářskou práci vypracoval samostatně pod vedením pana Ing. Vladimíra Bartíka, Ph.D. Uvedl jsem všechny literární prameny a publikace, ze kterých jsem čerpal.

.....
Pavel Rudol
17. května 2010

Poděkování

Děkuji za odbornou pomoc a vedení Ing. Vladimíru Bartíkovi, Ph.D. a také Ing. Lukáši Strykovi.

Obsah

Kapitola 1

Úvod

Informační technologie jsou dnes využívány ve všech oblastech lidské činnosti. Ve většině použitých případů jde o datově orientované aplikace. Jedná se o různé informační systémy firem, bank, institucí apod.

Tato data jsou většinou využívána pouze v aktuální podobě. To znamená, že jakákoliv předešlá data jsou nenavrátně ztracena. Ovšem tato data lze využít i k jiným účelům. Své uplatnění mají například v obchodní sféře, kde analytici mohou na základě analýzy chování zákazníků nabízet nové produkty. Uchováváním a využitím dat za širší časové období se zabývá Business Inteligence.

Tato práce je zaměřena na problematiku datových skladů a OLAP. Tyto pojmy jsou vysvětleny v počátečních kapitolách. Je zde také vysvětlen jejich význam v rámci procesů pro podporu rozhodování, označovaných jako Business Intelligence.

Dále je zde popsáno prostředí MS SQL Serveru pro podporu vývoje datových skladů a OLAP analýzy.

V předposlední kapitole je prezentována ukázková aplikace využívající zde zmíněné technologie.

Kapitola 2

Business Intelligence

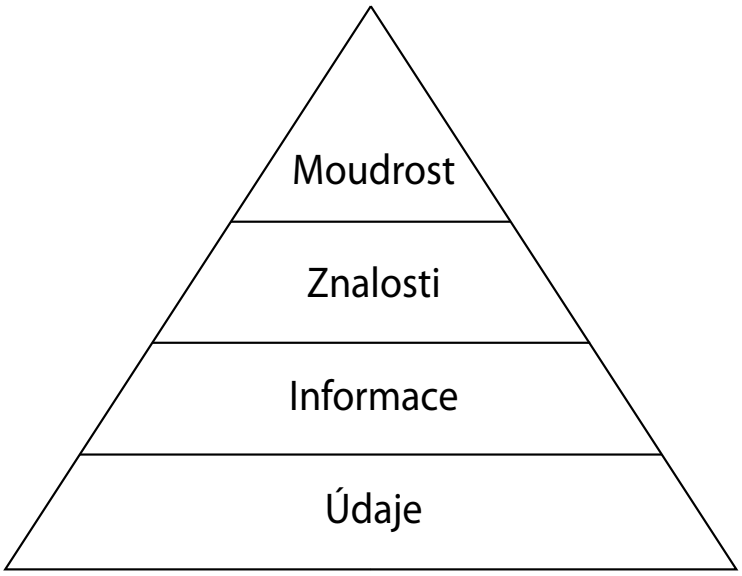
Práce se zabývá datovými sklady a analýzou OLAP. Na úvod je ale nutné zmínit pojem Business Intelligence, který je v tomto tématu velmi důležitý.

2.1 Definice

Business Intelligence je proces transformace dat na informace a převod těchto informací na poznatky prostřednictvím objevování.[6] Jeho účelem je tedy konverze velkých objemů dat na poznatky, které jsou důležité pro koncové uživatele. Tyto poznatky jsou potom využitelné například v procesu rozhodování.

2.2 Hierarchie informací

Proměnu údajů na informace lze zobrazit pomocí hierarchické pyramidy informačních úrovní. [5]



Obrázek 2.1: Hierarchie informací

Základem jsou **údaje**, ty ale obsahují jen jednoduchá fakta. Přidáním souvislostí získáváme **informace**. Jejich zpracováním získáme **znalosti**. Zobecněním znalostí získáme **moudrost**, nebo-li schopnost správného zhodnocení znalostí a jejich využití v praxi.

2.3 Význam

Přínos Business Intelligence je ve zlepšení procesu analýzy. Jedná se ovšem o velmi nákladnou investici, proto se jeho použití pečlivě zvažuje. Business Inteligence nebývá zaváděn jako jediný projekt, ale jako řada menších. To umožňuje společnosti zhodnotit jeho přínos a postupně jej začít nasazovat ve větší šíři. Analýzu a plánování je nutné provést na globální podnikové úrovni, ale provedení je možné po částech odpovídajících nějakému výrobnímu procesu.

Celkový úspěch je pak dán mírou spokojenosti zakázníků a zvýšením schopnosti rozhodování, kterou systém Business Intelligence přinese.

V rámci Bussiness Intelligence jsou data z informačních systémů firmy ve fázi ETL převáděna do datového skladu a dále transformována do multidimenzionálních databází. Tyto zdroje slouží k vytváření reportů a publikování dat v aplikacích na podporu rozhodování.

Za srdce celého systému lze pak označit datový sklad, který prezentuje jediný bod datové pravdy pro všechny uživatele. Mezi další důležité komponenty patří:

- aplikační server (OLAP server)
- reportingový server, uživatelský portál
- dataminingové nástroje
- uživatelské nástroje (klientské aplikace pro přístup k datům)
- vývojové nástroje (tvorba reportů, aplikací)

Kapitola 3

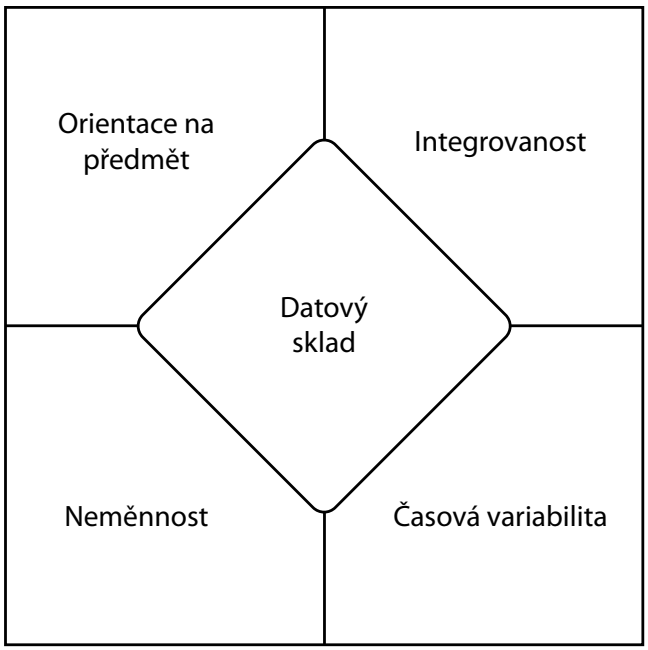
Datové sklady

Nyní se dostáváme k pojmu, na který je zaměřeno téma práce. V této kapitole popíšeme datový sklad, jeho tvorbu a provoz.

3.1 Definice

Autorem nejznámější definice datového skladu je Bill Inmon. Jeho definice říká, že datový sklad je podnikově strukturovaný depozitář subjektově orientovaných, integrovaných, časově proměnlivých, historických dat použitých na získávání informací a podporu rozhodování. V datovém skladu jsou uložena atomická a sumární data.[6]

Tato definice je graficky znázorněna na 3.1.



Obrázek 3.1: Schéma definice datového skladu

Znázorněné pojmy lze definovat takto:

- **Subjektová orientace**: Údaje jsou do datového skladu zapisovány spíše podle předmětu zájmu než podle aplikace, ze které pochází. Z toho vyplývá, že data jsou v datovém skladu uložena v kategoriích zaměřených na nějaký předmět, např.: dodavatel, výrobek, zaměstnanec apod.
- **Integrovanost**: Pro datovou integrovanost je důležité, aby data v datovém skladu týkající se konkrétního předmětu zde byla jen jednou. Z tohoto důvodu je nutné zavést jednotnou terminologii, jednotné a konzistentní jednotky veličin. Jelikož data pochází z různých zdrojů je nutné je před zavedením upravit vyčistit a sjednotit.
- **Časová variabilita**: Znamená, že datový sklad obsahuje data z různých časových období. Narozdíl od operačního databázového prostředí se údaje ukládají za delší časové období, typicky několik roků. Klíčové atributy v datovém skladě obsahují čas, který v operačních databázích nemusí být uváděn.
- **Neměnnost**: Představuje důležitý atribut datového skladu, jelikož údaje v něm nebývají nijak měněny, ani odstraňovány, pouze se v pravidelných intervalech přidávají nové. To znamená, že s datovým skladem jsou prováděny jen dva typy operací: zavedení údajů do skladu a přístup k nim.

3.2 Budování datového skladu

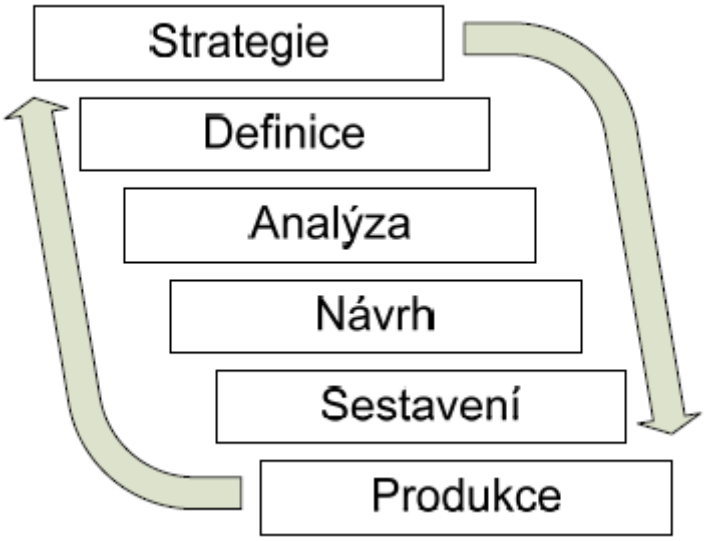
Než lze datový sklad použít je nutné jej vytvořit, k tomu existují dvě metody:

1. **Metoda „velkého třesku“**.
   Tato metoda má tři etapy:
   - Analýza požadavků podniku
   - Vytvoření podnikového datového skladu
   - Vytvoření přístupu buď přímo, nebo přes datové trhy

   Jedinou výhodou této metody je možnost celý projekt kompletně vypracovat ještě před začátkem jeho realizace. Ovšem i toto může představovat problém, jelikož během tvorby se mohou změnit technologie nebo požadavky uživatelů. Hlavní nevýhodou je dlouhá doba realizace znemožňující použití datového skladu a jeho výhod.
2. **Přírůstková metoda**.
   Cílem této metody je budování datového skladu po částech, které lze začít hned využívat. Přináší tedy rychleji výsledky a odezvu uživatelů.

   Metoda se skládá z několika fází:
   - **Strategie**: V této fázi je nutné definovat cíle. Jedná se o především o cíl podnikání a účel řešení datového skladu. Dlouhodobý cíl budování datového skladu definuje i strategii jeho správy, dokumentaci a zaškolení uživatelů.
   - **Definice**: Během této fáze dochází k definování rozsahu a cíle přírůstkového vývoje. Vytváří se počáteční přírůstek, konceptuální modely, dokumentují se zdroje dat a dochází k vymezení rozsahu kvality těchto údajů. Také je zde navržena architektura datového skladu a technických prostředků.

- **Analýza**: Cílem této fáze je zaměřit se na informace o uživatelích, získávání dat a požadavků na přístup k datům. Také se dokončí výběr nástroje pro komponenty datového skladu, řeší se problém kvality dat a stanovují požadavky na metadata.
- **Návrh**: V této fázi dochází k transformaci požadavků získaných během analýzy do detailních podmínek návrhu. Také se dokončuje instalace technické architektury.
- **Sestavení**: Cílem této fáze je vytvoření a otestování navržené databázové struktury, modulů získávání dat, správy datového skladu, metadat, přístupu k datům.
- **Produkce**: Během této fáze dochází k nainstalování datového skladu a spuštění jeho provozu. Také začíná řízení růstu a údržby datového skladu.



Obrázek 3.2: Schéma jedné iterace přírustkové metody

3.3 ETL

Jedná se o nástroje a postupy sloužící k zavedení dat do datového skladu. Tato data většinou získáváme z různých OLTP systémů, takže jsou často různého typu. Proto je potřeba specifikovat jednotné kategorie a provést „čištění“ dat.

Proces ETL se zkládá z několika etap, tvořících tuto zkratku:

- **Extrakce**: Jedná se o proces výběru dat z různých, mnohdy nehomogenních systémů (operačních, databázových). Jako základ datového skladu často slouží archivní data, tato data se pak již, ale nevyužívají k obnově údajů v datovém skladu. K extrakci jsou dostupné různé postupy, nástroje a technologie. Je možné k tomuto účelu vytvořit aplikace ve vyšších procedurálních programovacích jazycích, C++, C# nebo v procedurálních nadstavbách jazyka SQL (T-SQL, PL/SQL, ad.) Někdy lze využít výstupy z vlastních podnikových systémů, které umožňují konverzi a vyčištění údajů.
- **Transformace**: Během této etapy je data potřeba ověřit, transformovat a časově označit. Ověřuje se zda data mají kvalitu postačující pro jejich zavedení do datového

skladu. Může nastat i případ, že systém OLTP sice obsahuje kvalitní údaje, ale tyto údaje nemusí být zárukou kvalitního datového skladu. OLTP systémy totiž neobsahují historické údaje. Transformace samotná je soubor úloh a úkonů, které vedou ke zvýšení kvality údajů, hlavně k odstranění anomálií. Během čištění dat se sjednocuje formátování údajů, přiřazení datových typů, jednotek míry a peněžních měn. Složené primární klíče se rozkládají na atomické hodnoty. Během tansformace dochází nejčastěji k problémům s:

  - **nejednoznačností údajů**: Například různé uložené údaje o pohlaví M, F a Male, Female.
  - **chybějícími hodnotami a duplicitními záznamy**: Duplicity lze odstranit, u chybějících údajů je možné tyto záznamy vypustit.
  - **konvencí názvů pojmů a objektů**: Zavádění z různých zdrojů, kde jsou stejné entity jinak pojmenovány.
  - **různé peněžní měny**: Suma 29,5 má rozdílnou hodnotu v eurech a českých korunách.
  - **formáty čísel a textových řetězců**: Čísla mohou být v databázích uložena do numerického nebo řetězcového datového formátu, např. rodné číslo ve tvaru bez lomítka, naproti tomu s lomítkem už může být uloženo jen jako řetězec.
  - **referenční integrita**: Databáze obsahují mimo hodnot i vztahy master - detail, organizační strukturu firmy apod. Problém vzniká, když dojde například ke zrušení oddělení firmy bez odpovídající změny v databázi.
  - **chybějící datum**: V mnohých transakčních systémech se údaje neoznačují časem, jinde čas tvoří důležitou veličinu např. datum objednávek. Časový údaj musí být v datech přítomem před jejich zavedením do datového skladu, nebo se musí určit a přidat při zavádění dat.

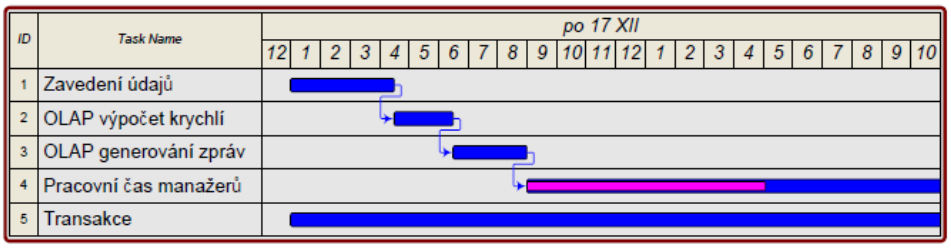
  Transformaci je možné provádět sériově nebo paralelně se zaváděním údajů. U sériového způsobu se transformace vykoná před zavedením dat. U paralelní metody se provádí současně se zaváděním.

- **Loading**: Přenos dat z paměti zdrojových dat nebo přechodné vynášecí oblasti do datového skladu. Přenos spočívá v přesunu údajů a jejich uložení do databázových tabulek, tento proces by měl být plánovaný a automatizovaný. Při prvotním naplnění datového skladu se jedná většinou o obrovské množství dat. Poté se údaje již přenáší v pravidelných intervalech např. denně.

Volně převzato z [5]. Proces ETL je komplexní a časově náročný a jeho hlavním cílem je centralizace údajů.

3.4 Provoz

Při návrhu koncepce řešení je nutné vycházet z reálného provozu a zabezpečit, aby manažeři a analytici měli údaje k dispozici skoro v reálném čase. Data je nutné doplňovat v pravidelných intervalech. Ty mohou být denní, týdenní nebo měsíční.



Obrázek 3.3: Provoz datového skladu

Kapitola 4

OLAP

Tato kapitola se zabývá analýzou OLAP sloužící k podpoře rozhodování. K analýze využívá data uložená v datových skladech.

4.1 Definice

Termín OLAP zavedl Dr. E. F. Codd na popsání technologie překlenující mezery mezi využitím osobních počítačů a řízením podnikových dat. Jedna z definic zní: OLAP je volně definovaný řád principů, které poskytují dimenzionální rámec pro podporu rozhodování.[6]

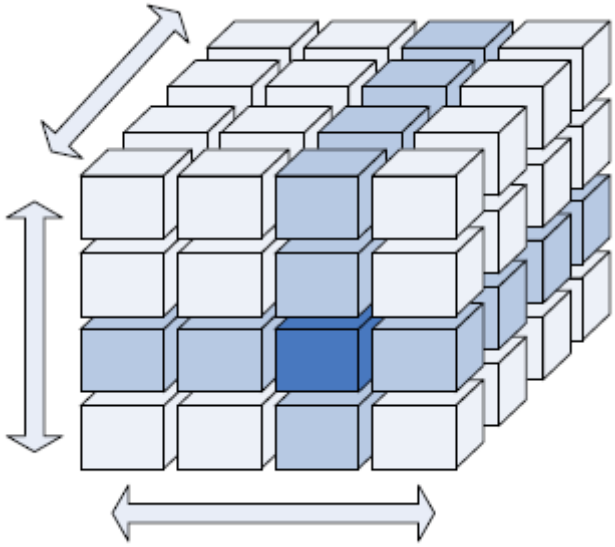
4.2 Porovnání OLTP a OLAP

OLTP (On-line transaction processing) je zaměřené, jak z názvu vyplývá, na transakce. To znamená je optimalizované pro operace jako je vkládání, změna a mazání dat. Data jsou uložena v relačních databázích a normalizována.

OLAP (On-line analytical processing) je orientované na analytické služby. Probíbají zde jen operace ukládání a čtení dat. Za tímto účelem jsou jinak organizována i data. Ta jsou ukládána do datových skladů. Je zde kladen důraz na rychlé zpracování dat pro různé analýzy.

4.3 Multidimenzionální datový model

Narozdíl od relačního databázového modelu, kdy jsou data v dvoudimenzionálních tabulkách, je multidimenzionální model možné zobrazit jako kostku. Ta představuje ekvivalent tabulky v relační databázi. Kostka se skládá z několika dimenzí, jež představují ekvivalent indexových polí v relačních tabulkách. Narozdíl od geometrické krychle není počet dimenzí omezen. To umožňuje uchovávat velké množství údajů. Na průsečíku dimenzí se pak nachází konkrétní údaje v tabulce faktů.



Obrázek 4.1: Data na průniku dimenzí v datové kostce

Při ukládání multidimenizionálních databází se používají následující technologie sloužící ke kompresi objemu využitého diskového prostoru.

4.3.1 Multidimenzionální OLAP (MOLAP)

Data v tomto úložišti jsou ukládáná jako dopředu vypočítaná pole. Databáze je organizována pro co nejrychlejší získávání dat z dimenzí. Část dat může být zavedena ke klientovi, což zrychlí analýzy a sníží síťový provoz. Hlavní výhodu představuje maximální výkon vzhledem na dotazy uživatele, nevýhodu pak redundance dat z důvodu jejich uložení v relační i multidimenzionální databázi.

4.3.2 Relační databázový OLAP (ROLAP)

Data jsou získávána z relačního datového skladu a po zpracování jsou uživateli předkládána jako multidimenzionální pohled. Data a metadata jsou v ROLAP úložišti uloženy jako záznamy v relační databázi. OLAP server pak dynamicky využívá tato metadata ke generování SQL příkazů potřebných k získání dat vyžadovaných uživatelem. Díky tomu, že data zůstávají v relačních databázích, nevzniká problém s redundancí.

4.3.3 Hybridní OLAP (HOLAP)

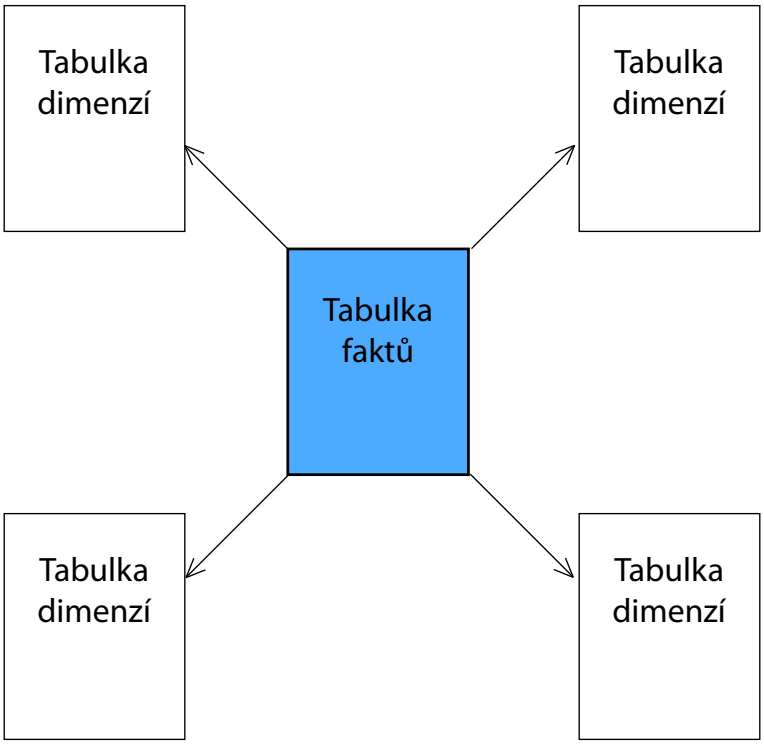
Hybridní OLAP je tvořen kombinací MOLAP a ROLAP za účelem využití výhod těchto typů úložišť. Data jsou uložena v relačních databázích a vypočítané agregace se ukládají do multidimenzionálních struktur. Při dotazování jsou pak data ukládána do multidimenzionální paměti cache.

4.4 Fakta a dimenze

K vytvoření OLAP kostky jsou potřeba fakta a dimenze. Fakta jsou numerické měrné jednotky obchodování (typicky cena, množství, apod.). Dimenze obsahují logicky nebo or-

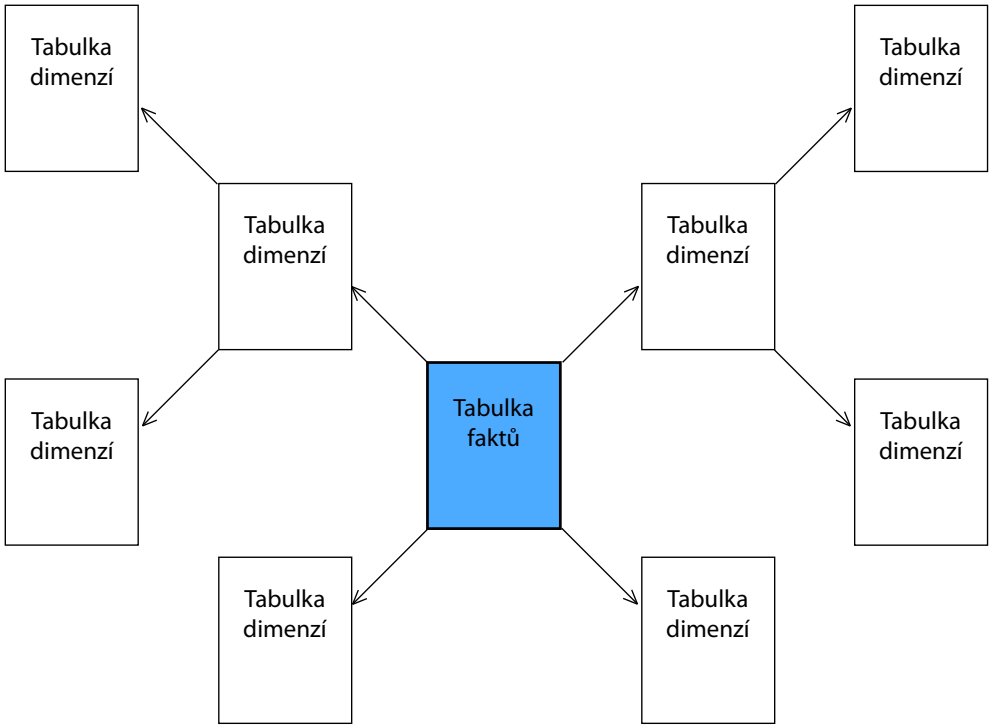
ganizačně hierarchicky uspořádaná data. Jsou to vlastně textové popisy obchodování (čas, region, produkt, apod.). Nejčastěji používáná schémata jsou [1]:

- **Schéma hvězdy** se skládá z rozsáhlé centrální tabulky s hodnotami (tzv. tabulka faktů) a řadou malých doprovodných tabulek pro každou dimenzi. Grafické vyjádření schématu připomíná hvězdu, s tabulkami dimenzí zobrazenými v paprskovité struktuře okolo centrální tabulky faktů. Ve hvězdicovém schématu je každá dimenze reprezentována právě jednou tabulkou. Toto schéma nemá normalizované dimenze ani relační propojení mezi tabulkami dimenzí. Vytvoření tohoto modelu je relativně pomalé, ale poskytuje vysoký „dotazovací výkon“.



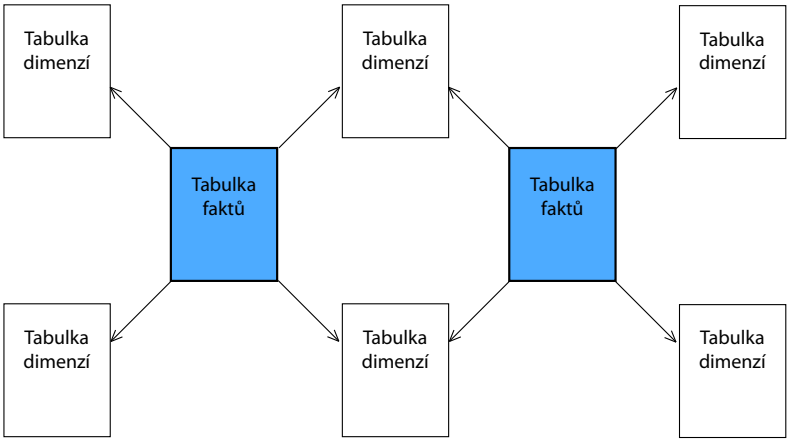
Obrázek 4.2: Schéma hvězdy

- **Schéma „sněhové vločky“** je určitým druhem hvězdicového schéma, ve kterém jsou tabulky dimenzí normalizovány, čímž se data rozdělují do dalších tabulek. Výsledné grafické schéma pak vytváří tvar podobný sněhové vločce. Tento model umožňuje rychlejší zavedení údajů do normalizovaných tabulek, ale jeho dotazovací výkon je nižší z důvodu většího množství spojení tabulek.



Obrázek 4.3: Schéma sněhové vločky

- **Schéma souhvězdí** obsahuje narozdíl od hvězdicového schématu více tabulek faktů, které mohou mít mezi sebou sdílené dimenze.

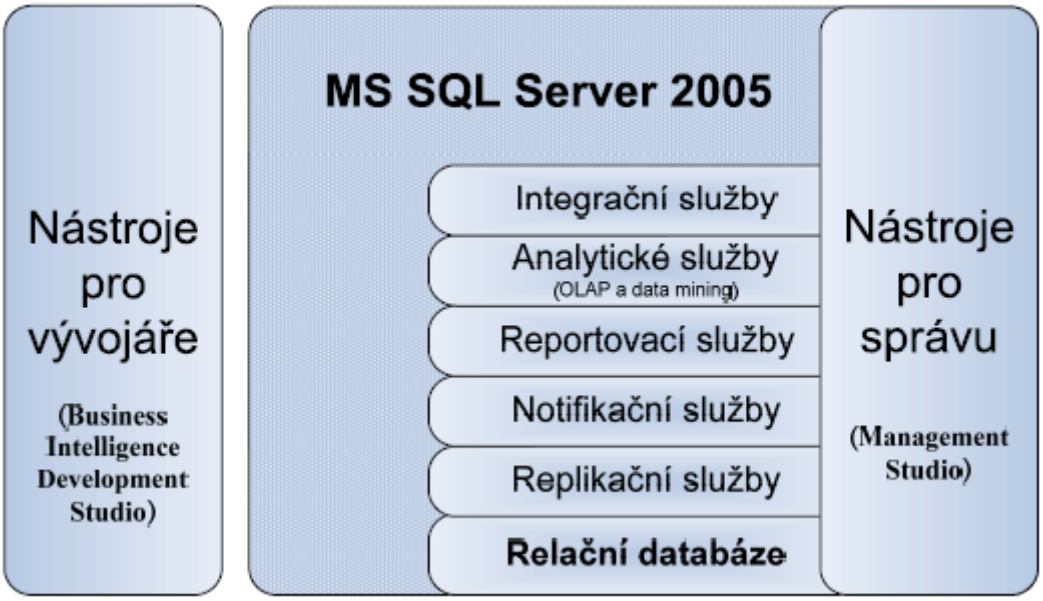


Obrázek 4.4: Schéma souhvězdí

Kapitola 5

MS SQL Server

Databázový systém MS SQL Server poskytuje podporu pro implementaci datových skladů, vytváření OLAP analýz a reportů. V této kapitole popíši MS SQL Server ve verzi 2005.



Obrázek 5.1: Schéma MS SQL Serveru

Datová platforma MS SQL Server 2005 obsahuje následující nástroje [2]:

- Relační databáze: Relační databázový stroj.
- Služby Replication Services: Replikace dat pro aplikace zpracovávající distribuovaná či mobilní data.
- Služby Notification Services: Funkce zasílání upozornění pro vývoj a nasazení škálovatelných aplikací. Umožňuje zasílat aktuální informace přizpůsobené individuálním požadavkům na nejrůznější připojená i mobilní zařízení.
- Služby Integration Services: Funkce extrakce, transformace a načítání dat (ETL) pro datové sklady a integraci dat v celém podniku.

- Služby Analysis Services: Funkce OLAP pro analýzu velkých a složitých datových sad s využitím vícedimenzionálních úložišť.
- Služby Reporting Services: Komplexní řešení pro vytváření, správu a zasílání statických i interaktivních webových sestav.
- Nástroje pro správu (MS SQL Server Management Studio): SQL Server zahrnuje integrované nástroje pro pokročilou správu a ladění.
- Nástroje pro vývojáře (Business Intelligence Development Studio): SQL Server nabízí integrované nástroje pro vývojáře, určené pro databázový stroj, extrakci, transformaci a načítání dat (ETL), dolování dat, funkce OLAP a vytváření sestav, které jsou úzce integrovány se sadou Microsoft Visual Studio a poskytují komplexní funkce pro vývoj aplikací.

5.1 Integrační služby SSIS

Integrační služby (SSIS – SQL Server Integration Services) slouží především k zavádění dat do datových skladů (fáze ETL). Kromě toho však umožňují vytvářet aplikace, které mohou spravovat databáze či systémové prostředky, reagovat na databázové a systémové události a dokonce interagovat s uživateli. SSIS zahrnuje různé úkoly, díky nimž mohou jím vytvořené balíčky odesílat nebo stahovat soubory ze serverů pomocí protokolu FTP, manipulovat se soubory v adresářích, importovat soubory do databází nebo exportovat data do souborů. Integrační balíček vytvořený pomocí služby SSIS se skládá ze dvou částí: workflow a dataflow. Workflow umožňuje provádět různé sekvence, smyčky úloh, provádí odchytávání a zpracování událostí a chyb. Dataflow obsahuje úkoly spojené s výběrem, transformací a uložením dat do databáze.

5.2 Analytické služby SSAS

Analytické služby (SSAS – SQL Server Analysis Services) tvoří dvě komponenty: OLAP a Data Mining (dolování dat). Modul OLAP umožňuje zavádění, dotazování a správu kostek, které byly vytvořeny v aplikaci Bussiness Intelligence Development Studio. Je možné zahrnout více hierarchií v rámci dimenze a zvolit různé možnosti, např. atributy dostupné pro zobrazení a způsob řazení členů. Veličiny lze navrhnout jako jednoduché aditivní prvky nebo nasadit složitá uživatelsky definovaná schémata agregace.

5.3 Reportovací služby SSRS

Reportovací služby (SSRS – SQL Server Reporting Services) poskytují pružnou platformu pro návrh sestav i distribuci dat. SSRS se skládá ze serveru sestav a nástroje Report Designer. Server sestav odpovídá za hostování všech sestav a zajišťuje jejich zabezpečení. Integrované komponenty Report Designeru umožňují uživatelům vytvářet širokou škálu sestav od jednoduchých tabulek až po sestavy s více úrovněmi dílčích sestav, vnořených sestav, grafů, propojených sestav a odkazů na externí prostředky.

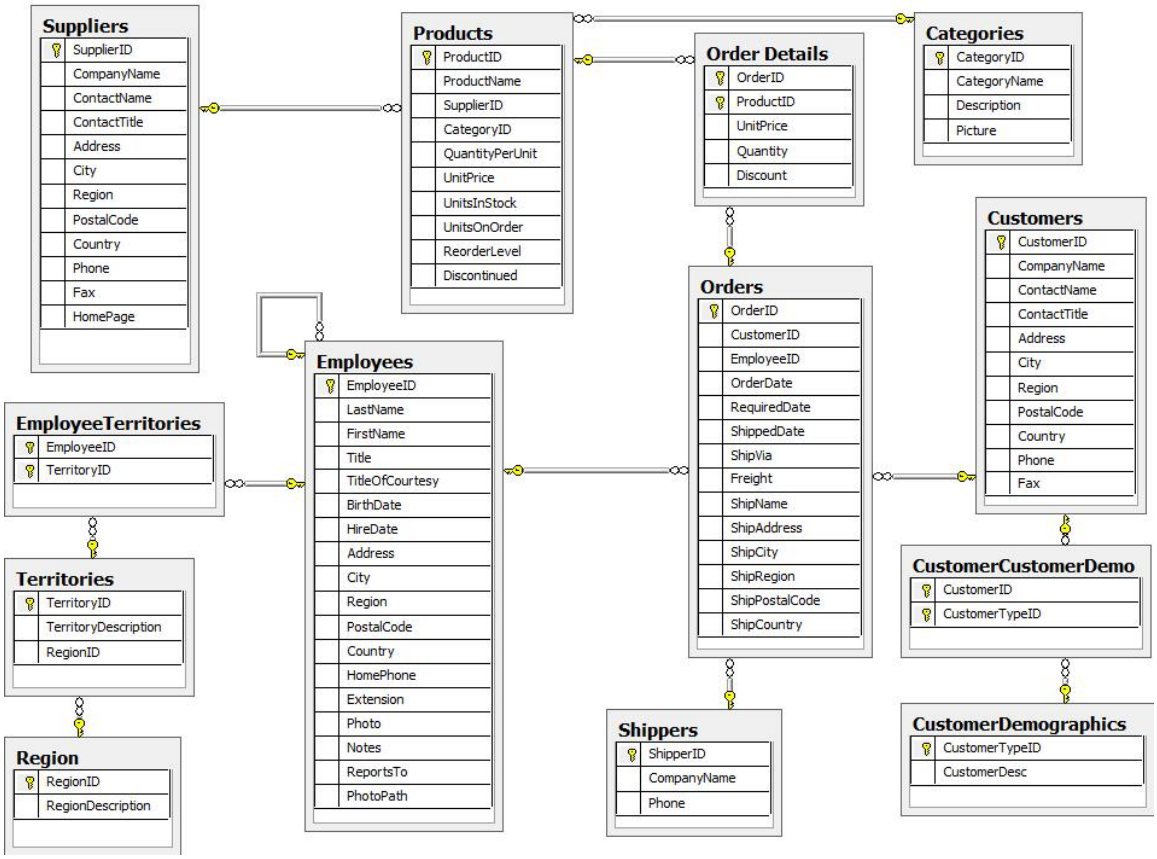
Kapitola je volně převzata z [4].

Kapitola 6

Ukázková aplikace

V této kapitole je popsána ukázková aplikace vytvořená v prostředí MS SQL Serveru 2005 se Service packem 3.

Aplikace je rozčleněna na menší projekty demonstrující možnosti MS SQL Serveru. Jako zdrojová databáze datového skladu je použita OLTP databáze Northwind dodávaná s předchozí verzí MS SQL Serveru (2000). Tato databáze obsahuje 13 tabulek a její schéma je následující (převzato z [3]):



Obrázek 6.1: Schéma databáze Northwind

Projekty představují způsob integrace datového skladu, vytvoření OLAP analýzy a reportů a jejich prezentaci pomocí aplikace pro Windows.

6.1 Vytvoření datového skladu

6.1.1 Strategie

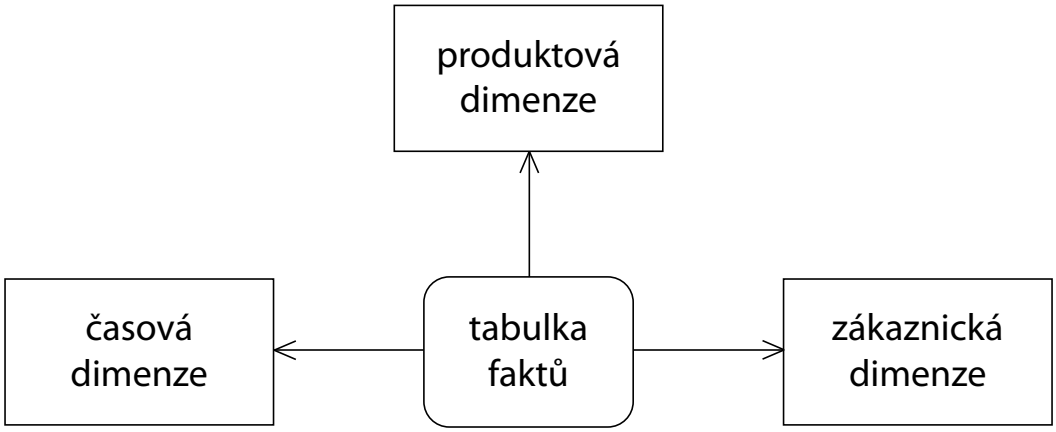
Vytvářený datový sklad má za úkol posloužit firemním manažerům a analytikům k rychlejším a kvalitnějším rozhodnutím. K tomu poslouží hlavně reporty (provozní i strategické), ad-hoc analýzy. Z dlouhodobého hlediska se využije ke sledování plnění plánů a trendů vývoje. Projekt vybudování datového skladu je založen na přírůstkové metodě.

6.1.2 Definice

Na základě prostudování schématu zdrojové databáze fiktivní firmy Northwind jsem vybral tabulky vhodné k vytvoření datového skladu. Díky využití jediného zdroje nehrozí problém s integritou dat. V rámci čištění dat jsem odstranil řádky, ve kterých sloupec s datem doručení obsahoval hodnotu NULL.

6.1.3 Analýza

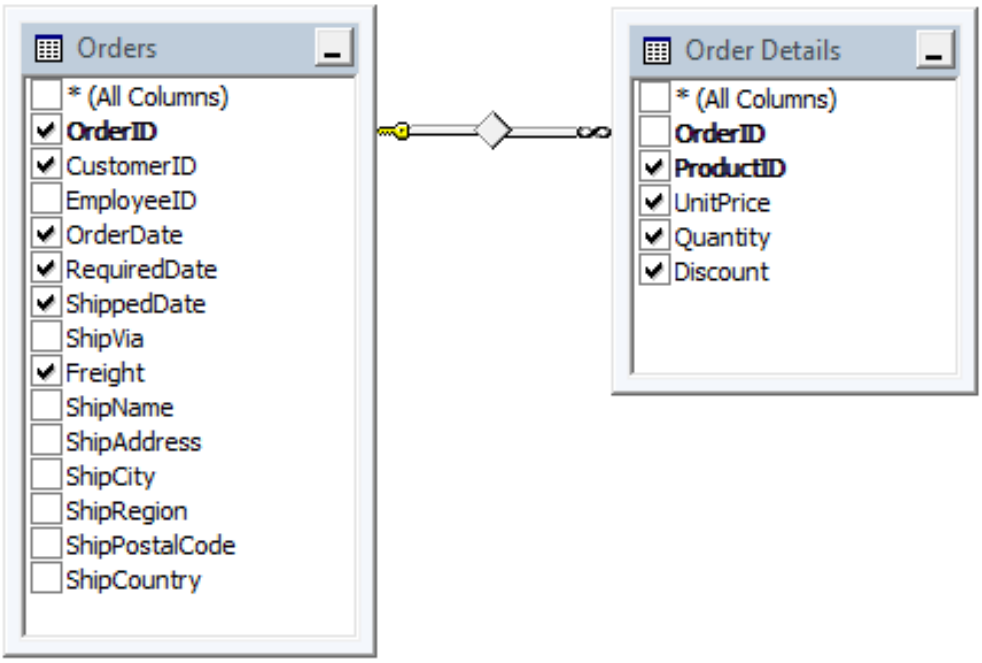
Uživateli datového trhu budou pracovníci obchodního oddělení a jim nadřazení zaměstnanci. Řešení bude založeno na produktu MS SQL Server 2005. Na základě zdrojových dat jsem pro konceptuální model skladu vybral schéma hvězdy, ve kterém centrální tabulka faktů obsahuje počet a cenu prodaných produktů. K tabulce faktů se vztahují 3 dimenze - produktová, zákaznická a časová (viz obrázek 6.2).



Obrázek 6.2: Schéma datového skladu

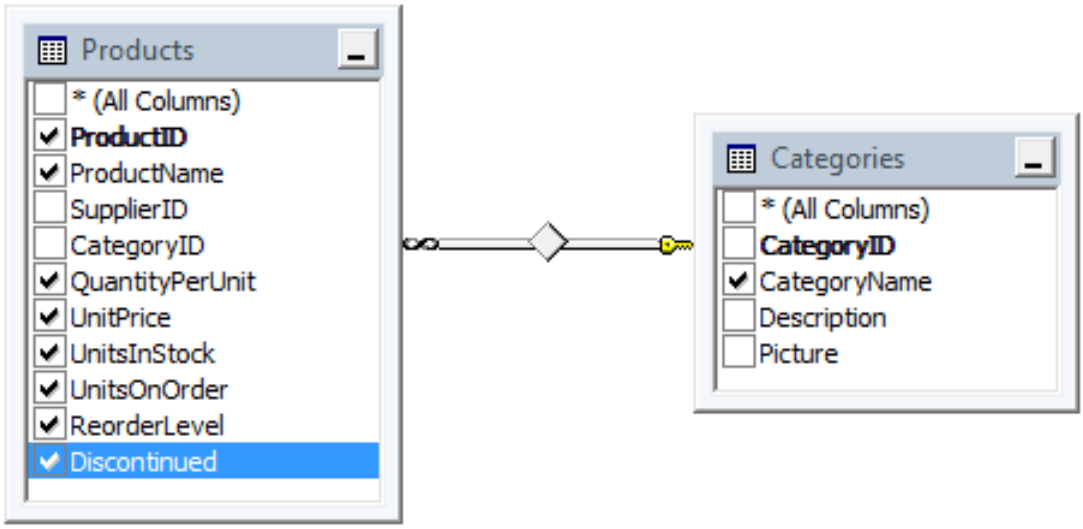
6.1.4 Návrh

Tato fáze transformuje výsledky analýzy do detailních podmínek návrhu. Pomocí datově řízeného přístupu [8] jsem jako základ pro tabulku faktů použil tabulky Orders a Order Details obsahující informace o objednávkách. Na obrázku jsou zaškrtnuty sloupce, které jsou přenášeny do datového skladu.



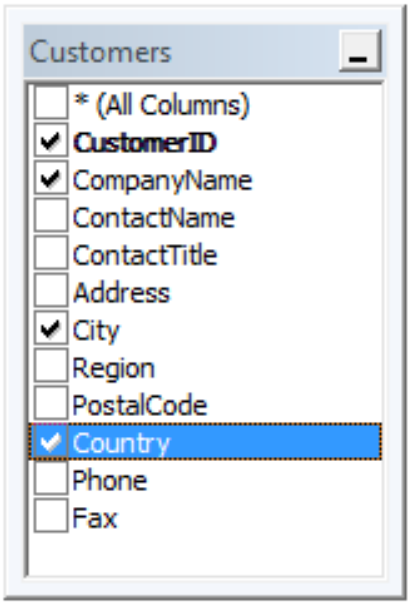
Obrázek 6.3: Zdroj pro tabulku faktů

Produktová dimenze vznikla z tabulek Products a Categories, které obsahují informace o produktech firmy a jejich zařazení do kategorií. Obrázek opět ukazuje přenášené sloupce.



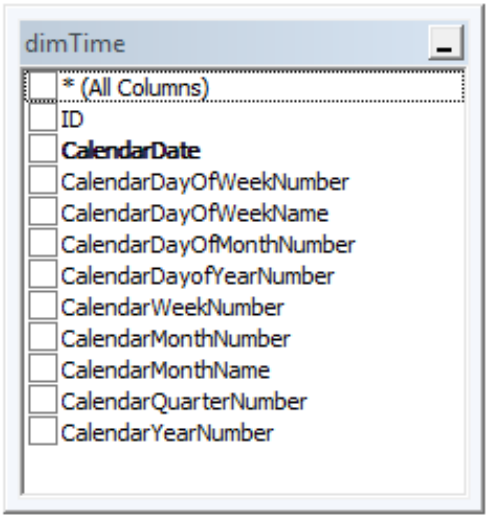
Obrázek 6.4: Zdrojová tabulka Products

Pro zákaznickou dimenzi jsem využil tabulky Customers a sloupců se jménem firmy, zemí a městem, kde sídlí:



Obrázek 6.5: Zdrojová tabulka Customers

Pro časovou dimenzi nebyla využita existující data, ale byla vytvořena tak, aby pokryla časové období, pro které v databázi Northwind existují objednávky. Časová dimenze má tedy tuto strukturu:



Obrázek 6.6: Struktura časové dimenze

K jejímu vytvoření byla využita procedura z [7].

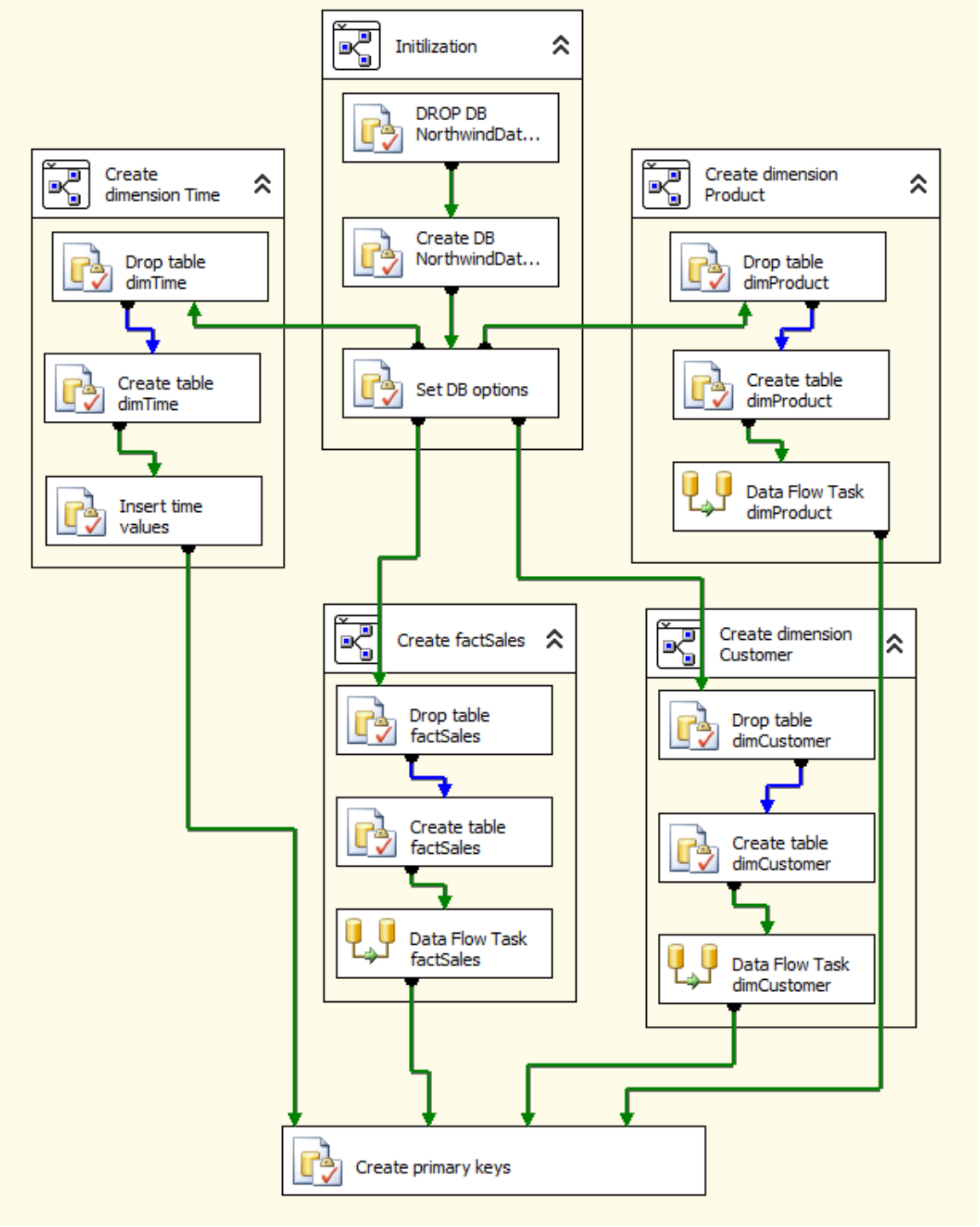
6.1.5 Sestavení

Pomocí služby SSIS MS SQL Serveru jsem sestavil balíček sloužící k vytvoření databáze datového skladu, jejího nastavení a vytvoření a naplnění tabulky faktů a tabulek dimenzí. V Bussiness Intelligence Development Studiu (BIDS) nadefinujeme potřebná připojení:

- k databázi master pro vytvoření databáze datového skladu,
- ke zdrojové databázi Northwind,
- k cílové (vytvářené) databázi datového skladu.

Projekt na záložce Control Flow je na obrázku 6.7. Jednotlivé bloky představují úlohy vedoucí k vytvoření databáze datového skladu. Ve fázi inicializace je odstraněna případná databáze a znova vytvořena, také jsou jí nastaveny parametry, velikost 5 MB s růstem 10 %. V dalších blocích jsou odstraněny a znovu vytvořeny tabulky dimenzí a faktů. V závěrečné fázi jsou nastaveny primární klíče.

Pro Data Flow úlohy je v BIDS vyčleněna samostatná záložka, zde se zpracovávají data, dochází k jejich transformaci a vyčištění. V projektu se nachází Data Flow úlohy sloužící k výběru dat ze zdrojové databáze a jejich převedení do tabulek dimenzí a tabulky faktů, u které bylo nutné vynechat řádky s hodnotou NULL u sloupce Shipped Date. Také jsem se setkal s problémem při validaci projektu, kdy kontrola hlásila neexistující tabulky dimenzí, resp. nemožnost připojit se k vytvářené databázi k ověření existence tabulek. Z tohoto důvodu bylo nutné v Control Flow nastavit pro Data Flow úlohy zpoždění validace.



Obrázek 6.7: Projekt v BIDS

6.1.6 Produkce

Po vytvoření a odladění balíčku je možné jej nasadit. U projektu je proto potřeba nastavit parametr CreateDeploymentUtility na hodnotu True a provede se build. V podadresáři bin/Deployment tím vzniknou soubory „nazev_projektu.dtsx“

a „nazev_projektu.SSISDeploymentManifest“. Tyto soubory přeneseme na cílový počítač, kde spuštěním souboru „nazev_projektu.SSISDeploymentManifest“ vyvoláme průvodce Package Installation Wizard. Ten nám umožní vybrat zda balíček uložíme do souborového systému nebo úložiště MS SQL Serveru.

Nyní můžeme balíček spustit. To provedeme příkazem Run package u daného balíčku v MS SQL Server Management Studiu (po připojení k serveru Integration Services). Vytvořený balíček integračních služeb slouží k iniciálnímu naplnění datového skladu. To se obvykle provádí pouze jednou a lze jej provést také jednoduše spuštěním balíčku příkazem Execute Package v Solution Exploreru vývojového prostředí BIDS.

6.2 Vytvoření datové kostky

Po zavedení datového skladu s tabulkami dimenzí a tabulkou faktů máme vytváření datové kostky pomocí služby SSAS usnadněné.

6.2.1 Datové zdroje

Jako datový zdroj použijeme datový sklad vytvořený v předchozím kroku. K definování připojení je možné využít průvodce, který umožňuje i nastavení přihlašovacích údajů.

6.2.2 Pohledy na datové zdroje

Pokud zdroj obsahuje více tabulek dimenzí a tabulek faktů, umožňuje pohled na datový zdroj vybrat tabulky dimenzí a tabulku faktů, které chceme použít pro vytvoření datové kostky. Opět je možno použít služeb průvodce, kde vybereme tabulky, které chceme použít a vztahy mezi nimi, ty je průvodce schopný sám detekovat na základě shody jmen sloupců s primárními klíči. Další vztahy je možné dodefinovat pomocí designeru pohledů, což bylo nutné pro časovou dimenzi, jelikož v tabulce faktů jsou rozdílné časové údaje pro datum objednávky, požadované datum doručení a datum odeslání.

6.2.3 Datová kostka

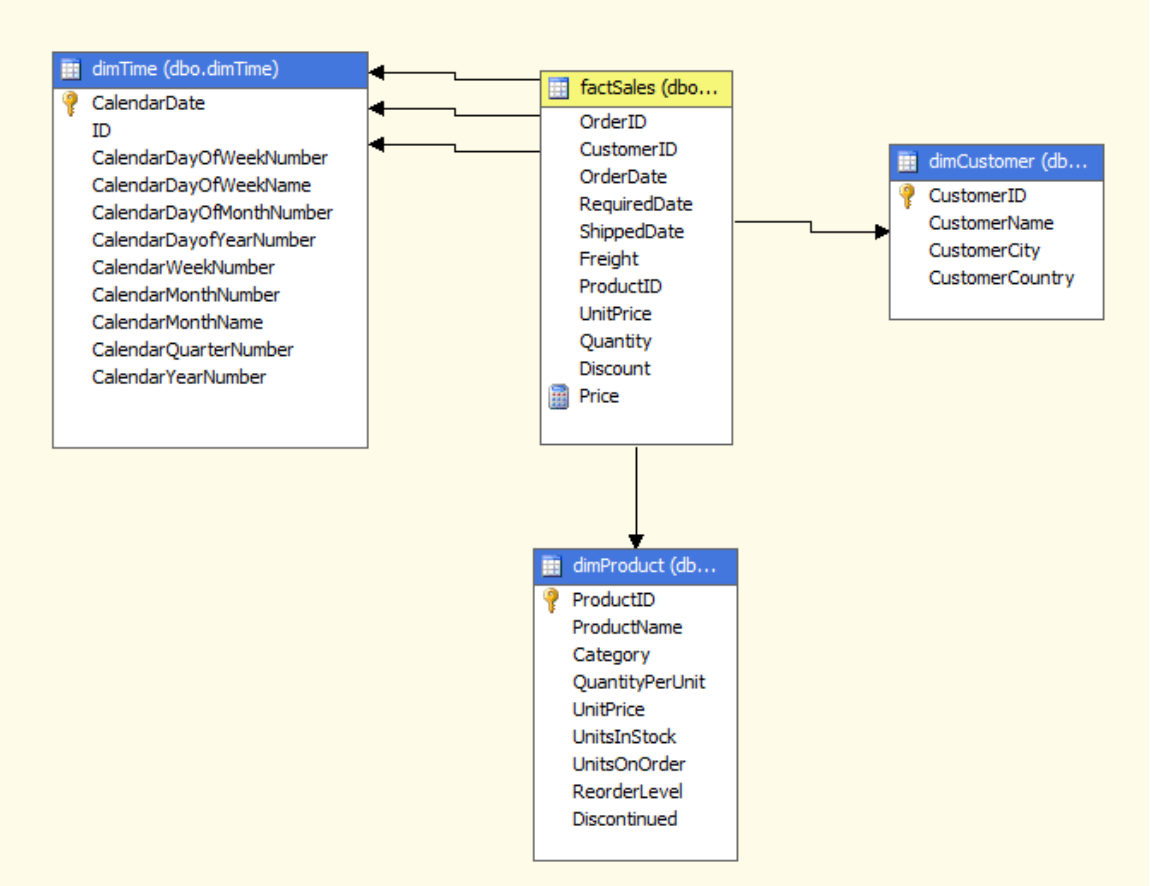
K vytvoření datové kostky slouží průvodce Cube Wizard. Ten nabízí možnost Auto build k automatické detekci dimenzí a faktů. V následujícím kroku je možné toto nastavení upravit a také nastavit tabulku pro časovou dimenzi. Průvodce se poté pokusí detekovat hierarchie dimenzí, což je možné ještě poupravit v příštím kroku. Tímto máme vytvořenou kostku, jejíž schéma je na obrázku 6.8.

6.2.4 Hierarchie dimenzí

U datové kostky jsem u dimenzí definoval následující hierarchie:

- Časová dimenze — rok – kvartál – měsíc – datum
- Produktová dimenze — kategorie – název produktu
- Zákaznická dimenze — země – město – jméno zákazníka

Po vytvoření kostky ji zavedeme příkazem Deploy z menu Build.



Obrázek 6.8: Stuktura datové kostky

6.3 Vytvoření a doručování reportů

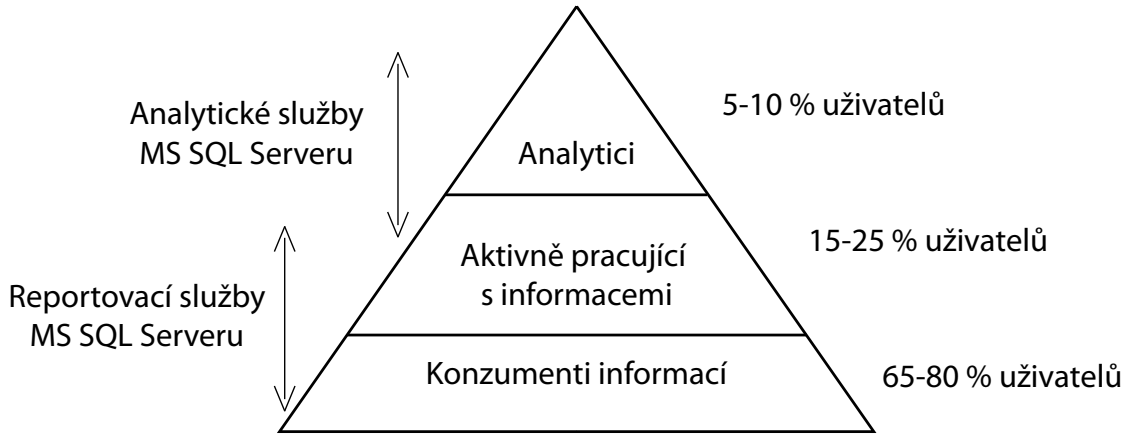
Reportovací služby MS SQL Serveru představují prezentační vrstvu projektů Bussiness Intelligence. Tato podkapitola popíše tvorbu a doručení reportů uživatelům a je převzata z [6].

6.3.1 Filosofie reportovacích služeb

Reporting Services v SQL Serveru 2005 slouží k vývoji, správě a zobrazování datových sestav. Obsahuje nástroje pro návrh sestav, jednorázové dotazy i administraci. Definice sestav jsou ukládány v jazyce XML, publikují se do databáze v SQL a ovládají přes webové rozhraní. Služba ve Windows ovládá zpracování sestav, jež odděluje získávání dat od zobrazování sestav. Jednu sestavu tak lze zobrazit mnoha uživatelům v různých výstupních formátech.

Uživatelé

Pro objasnění pozice reportovacích služeb je důležité si uvědomit, kteří uživatelé budou tyto služby využívat. V kontextu celého BI řešení existují tři základní typy uživatelů (viz 6.9):



Obrázek 6.9: Uživatelé reportů

- **Analytici**: Jejich úlohou je vybírat data, která jsou vhodná pro analýzy, analyzovat je a na základě výsledků analýz poskytovat informace pro podporu rozhodování.
- **Aktivně pracující s informacemi**: Pracovníci této skupiny jednak data analyzují, zpracovávají a potřebují je v různé formě zobrazovat.
- **Konzumenti informací**: Tvoří většinu uživatelů. Dostávají výpisy dat ve formě dvourozměrných sestav. Přístup k datům ve formě reportů zabezpečují právě reportovací služby.

Ještě zásadnější rozdělení reportů vychází ze skutečnosti, zda bude přístup k datům umožněn pouze pracovníkům firmy nebo například obchodním partnerům. Z tohoto hlediska je možné rozlišovat následující dvě kategorie:

- **Enterprise a Embedded Reporting**: Prezentace dat v rámci podniku pomocí podnikových portálů, intranetu nebo informačních systémů.
- **B2B Reporting**: Prezentace dat obchodním partnerům, například přes firemní webový portál.

Komponenty

Reportovací služby se skládají z těchto komponent:

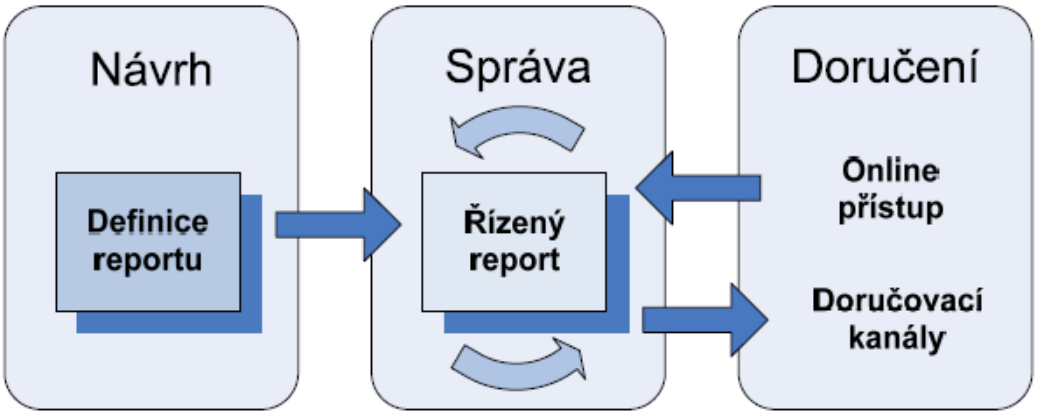
- **Report Server**: Serverová komponenta, zahrnující webové a reportovací služby. (Webové služby představují množinu programátorských rozhraní, která využívají klientské aplikace pro přístup k reportům.)
- **Report Model Designer**: Nástroj pro návrh modelů reportů.
- **Report Designer**: Vizuální nástroj pro vytváření reportů. Je součástí produktu MS Visual Studio.
- **Report Manager**: Webová aplikace pro správu a prohlížení reportů.

- **Report Builder**: Umožňuje vytvářet a nasazovat sestavy pomocí modelu vytvořeného v Report Model Designeru.

Koncepce reportovacích služeb v SQL Serveru je označována jako „managed reporting“. Její hlavní výhodou je skutečnost, že metadata, na jejichž základě se generují reporty, jsou uložena centrálně. Mohou tedy být i centrálně spravována. To znamená, že je možné reporty organizovat do různých sestav, adresářů a zpřístupňovat je na různé úrovni uživatelům, popř. skupinám uživatelů. Přístup k reportům je pak možné řídit také přes přístup k adresářovým strukturám.

Životní cyklus reportu

Pomocí reportů můžeme zobrazovat data z relačních i analytických databází. V případě, že data čerpáme z datového skladu, je nutné při návrhu reportu brát v úvahu i životní cyklus datového skladu. Cílem je, aby plynule navazovaly kroky naplnění datového skladu, vypočítání OLAP kostek a následně generování reportů z těchto zdrojů. Z jiného hlediska je možné rozdělit životní cyklus reportu na tři základní fáze: návrh, správa a doručení (viz obrázek 6.10).



Obrázek 6.10: Životní cyklus reportu

6.3.2 Návrh reportu

Výsledkem návrhu reportu je kód v jazyce RDL, který se zapisuje ve formě XML. Návrh je proto usnadněn přítomností vizuálního návrhového prostředí Report Designeru. Reporty mohou data zobrazovat formou tabulek, grafů, vnořených reportů. Variabilitu rozšiřují dynamické a hierarchické parametry, možnosti třídění, filtrování, seskupování dat a výpočet částečných výsledků pomocí agregačních funkcí.

Vytvořené reporty

Po vytvoření datového skladu a OLAP kostky z něho vycházející, využijeme nyní kostku k vytvoření několika reportů, ukazujících možnosti reportovacích služeb v MS SQL Serveru.

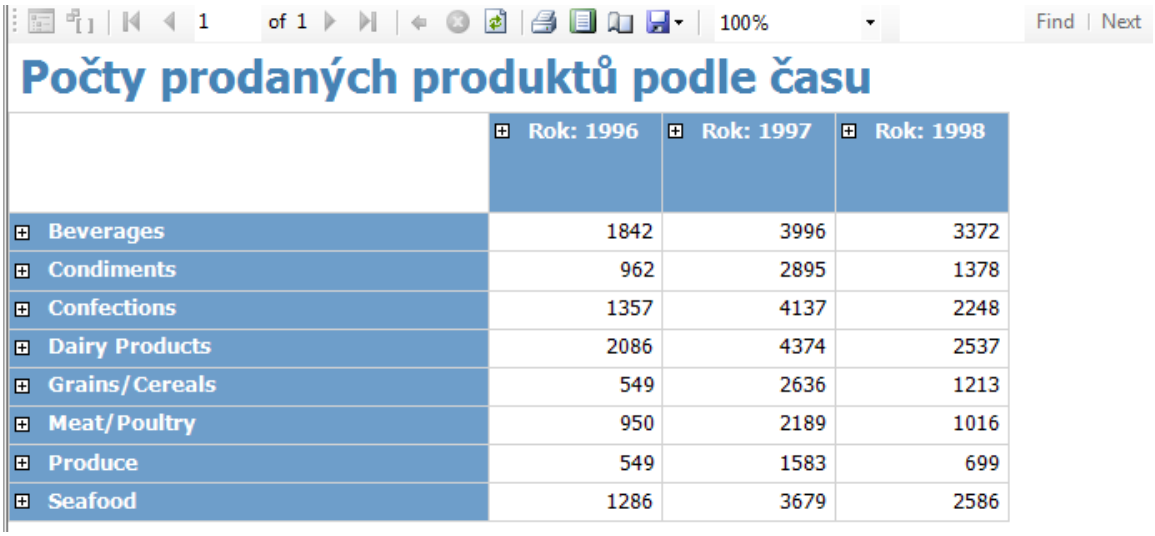
Nejprve je potřeba nastavit datový zdroj, který budou sdílet všechny vytvářené reporty. Na výběr je několik typů zdrojů dat (MS SQL Server, Oracle, ODBC, XML, ad.). K vy-

tvořené kostce se dostaneme vybráním připojení MS SQL Server Analysis Services. Pro připojení je ještě potřeba nastavit název serveru a databáze.

Po nastavení datových zdrojů můžeme začít vytvářet reporty. K tomu je možné využít průvodce Report Wizard a jehou součást Query Builder, usnadňující výběr dimenzí a faktů z datové kostky a nastavení filtrů vizuálním prostředím. Následující krok průvodce nabízí výběr ze dvou typů rozvržení dat – v běžné tabulce nebo matici, způsob uspořádání jednotlivých dimenzí a také šablonu designu, která se má použít. Tímto je usnadněna tvorba základních reportů.

K vytvoření složitějších reportů slouží návrhové prostředí Report Designeru. Je rozděleno na tři části – na záložce Data je možné upravovat výběr dimenzí a faktů použitých v reportu, v záložce Layout se vytváří vizuální vzhled a rozmístění jednotlivých prvků reportu (jako jsou grafy, tabulky ad.), na poslední záložce Preview je možné zobrazit výslednou podobu reportu.

Po vytvoření reportů je potřeba je nahrát na server, aby byly dostupné pro další uživatele. K tomu slouží příkaz Deploy z menu Build. Reporty jsou poté přístupny přes prohlížeč na URL adrese http://localhost/ReportServer (v případě že server běží na lokálním počítači). Po zadání této adresy do prohlížeče se objeví stránka obsahující seznam datových zdrojů a projektů s reporty. Po vybrání požadovaného projektu se nám zobrazí jednotlivé reporty (ukázka reportu viz obrázek 6.11).



Obrázek 6.11: Ukázka reportu

6.4 Klientská aplikace

Při vývoji aplikací, prezentujících reporty, určených pro Windows je možné využít komponenty Report Viewer. U této komponenty je potřeba nastavit adresu Report Serveru a konkrétního reportu. Komponenta pak poskytne stejné ovládací prvky jako při přístupu přes web. Aplikace představuje možnosti prezentace reportů. Toto řešení by pak bylo možné zabudovat i do firemních aplikací, tak aby uživatelé měli přístup k výsledkům BI řešení z jednoho místa.

Kapitola 7

Závěr

I přes rozsáhlost oblasti datových skladů a OLAP analýz jsem se pokusil ji věcně zmapovat v rámci kontextu Bussiness Intelligence.

Získané teoretické znalosti jsem využil k vytvoření ukázkové aplikace v prostředí MS SQL Serveru, který představuje jeden z několika obsáhlých databázových nástrojů.

Klientská aplikace a reportovací služby poskytují možnosti rozšíření. Například integraci BI řešení do informačního systému firmy.

Literatura

[1] Datové kostky [online]. http://datamining.xf.cz/view.php?cisloclanku=2002102811, 2002-10-28 [cit. 2010-05-10].

[2] Přehled produktu SQL Server 2005 [online]. http://www.microsoft.com/cze/windowsserversystem/sql/prodinfo/overview/default.mspx, 2005-11-07 [cit. 2010-05-10.

[3] Northwind sample database [online]. http://www.wilsonmar.com/northwind.htm, [cit. 2010-05-10.

[4] Hotek, M.: *Microsoft SQL Server 2008 krok za krokem*. Brno: Computer Press, 2009, 488 s. ISBN 978-80-251-2466-6.

[5] Lacko, L.: *Databáze: datové sklady, OLAP a dolování dat s příklady v Microsoft SQL Serveru a Oracle*. Brno: Computer Press, 2003, 488 s. ISBN 80-7226-969-0.

[6] Lacko, L.: *Business Intelligence v SQL Serveru 2005*. Brno: Computer Press, 2006, 388 s., 1 DVD-ROM. ISBN 80-251-1110-5.

[7] Pavliashvili, B.: Case Study of Building a Data Warehouse with Analysis Services [online]. http://www.scribd.com/doc/6618422/Case-Study-of-Building-a-Data-Warehouse-With-Analysis-Services, 2006-02-10 [cit. 2010-05-09].

[8] Wrembel, R.; Koncilia, C.: *Data Warehouses and OLAP: Concepts, Architectures and Solutions*. Hershey: IRM Press, 2007, 332 s. ISBN 1-59904-364-5.

Dodatek A

Manuál

Ukázkovou aplikaci představují 4 projekty vytvořené s pomocí MS SQL Serveru 2005 SP3 a MS Visual Studia 2005. Pro spuštění projektů je nutné mít nainstalován webový server IIS a MS SQL Server. Všechny projekty lze spustit, po otevření ve vývojovém prostředí MS Visual Studio 2005, kliknutím na zelenou šipku na pruhu nástrojů. Popis jednotlivých projektů je následující:

1. **NorthwindDataMart**:
   Integrační balíček sloužící k vytvoření datového skladu.

2. **NorthwindCube**:
   Vytváří OLAP kostku nad datovým skladem.

3. **NorthwindReport**:
   Projekt s reporty vytvořenými nad OLAP kostkou.

4. **NorthwindWinApp**:
   Windows aplikace obsahující integrované reporty.