VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
BRNO UNIVERSITY OF TECHNOLOGY

FAKULTA INFORMAČNÍCH TECHNOLOGIÍ
FACULTY OF INFORMATION TECHNOLOGY

ÚSTAV INFORMAČNÍCH SYSTÉMŮ
DEPARTMENT OF INFORMATION SYSTEMS

ANALÝZA VEŘEJNĚ DOSTUPNÝCH DAT ČESKÉHO STATISTICKÉHO ÚŘADU

ANALYSIS OF PUBLIC DATA OF THE CZECH STATISTICAL OFFICE

DIPLOMOVÁ PRÁCE
MASTER'S THESIS

AUTOR PRÁCE — **Bc. ONDŘEJ POHL**
AUTHOR

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

BRNO 2017

Zadání diplomové práce/19365/2016/xpohlo01

Vysoké učení technické v Brně - Fakulta informačních technologií

Ústav informačních systémů Akademický rok 2016/2017

Zadání diplomové práce

Řešitel: **Pohl Ondřej, Bc.**
Obor: Informační systémy
Téma: **Analýza veřejně dostupných dat Českého statistického úřadu**
Analysis of Public Data of the Czech Statistical Office
Kategorie: Data mining

Pokyny:

1. Seznamte se s analytickými metodami Business Intelligence, zejména s oblastí OLAP a získávání znalostí z databází.
2. Prostudujte veřejně dostupná data Českého statistického úřadu týkající se zahraničního obchodu ČR a navrhněte aplikaci, která provede jejich stažení. Aplikaci pro stažení dat implementujte.
3. Na základě získaných dat navrhněte několik analytických úloh, které bude možné s daty provádět.
4. Vyzkoušejte navržené úlohy v dostupných nástrojích (např. nástroje MS SQL Serveru, RapidMiner, popř. jiné).
5. Po dohodě s vedoucím vyberte některou z úloh a implementujte ji. Ověřte funkčnost této úlohy.
6. Zhodnoťte dosažené výsledky a další možné pokračování v tomto projektu.

Literatura:

- Han, J., Kamber, M.: Data Mining - Concepts and Techniques, 2nd Edition. Morgan Kaufmann Publishers, 2006.
- Ponniah, P.: Data Warehousing Fundamentals. John Wiley and Sons, 2001.
- Laberge, R.: Datové sklady - Agilní metody a business intelligence, Computer Press, Brno, 2012.

Při obhajobě semestrální části projektu je požadováno:

- Body 1, 2, částečně bod 3.

Podrobné závazné pokyny pro vypracování diplomové práce naleznete na adrese http://www.fit.vutbr.cz/info/szz/

Technická zpráva diplomové práce musí obsahovat formulaci cíle, charakteristiku současného stavu, teoretická a odborná východiska řešených problémů a specifikaci etap, které byly vyřešeny v rámci dřívějších projektů (30 až 40% celkového rozsahu technické zprávy).

Student odevzdá v jednom výtisku technickou zprávu a v elektronické podobě zdrojový text technické zprávy, úplnou programovou dokumentaci a zdrojové texty programů. Informace v elektronické podobě budou uloženy na standardním nepřepisovatelném paměťovém médiu (CD-R, DVD-R, apod.), které bude vloženo do písemné zprávy tak, aby nemohlo dojít k jeho ztrátě při běžné manipulaci.

Vedoucí: **Bartík Vladimír, Ing., Ph.D.**, UIFS FIT VUT
Datum zadání: 1. listopadu 2016
Datum odevzdání: 24. května 2017

VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
Fakulta informačních technologií
Ústav informačních systémů
612 66 Brno, Božetěchova 2

L.S.

doc. Dr. Ing. Dušan Kolář
vedoucí ústavu

Abstrakt

Cílem této práce je analýza dat Českého statistického úřadu týkajících se zahraničního obchodu ČR. Čtenář se nejprve seznámí s problematikou Business Intelligence a datovými sklady. Poté jsou vysvětleny základy analýzy OLAP a dolování dat. V dalších částech se práce zabývá popisem a analýzou dat zahraničního obchodu pomocí technologie OLAP a dolování dat v MS SQL Serveru, včetně implementace vybraných analytických úloh.

Abstract

The aim of this thesis is analysis of data of the Czech Statistical Office concerning foreign trade. At first, reader familiarize with Business Intelligence and data warehousing. Further, OLAP analysis and data mining basics are explained. In next parts the thesis deal with describing and analysis of data of foreign trade by the help of OLAP technology and data mining in MS SQL Server including selected analytical tasks implementation.

Klíčová slova

BI (Business Intelligence), datový sklad, OLAP (Online Analytical Processing), dolování dat, ČSÚ (Český statistický úřad), Microsoft SQL Server, .NET, C#.

Keywords

BI (Business Intelligence), data warehouse, OLAP (Online Analytical Processing), data mining, CZSO (Czech Statistical Office), Microsoft SQL Server, .NET, C#.

Citace

POHL, Ondřej. *Analýza veřejně dostupných dat Českého statistického úřadu*. Brno, 2017. Diplomová práce. Vysoké učení technické v Brně, Fakulta informačních technologií. Vedoucí práce Bartík Vladimír.

Analýza veřejně dostupných dat Českého statistického úřadu

Prohlášení

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

.....
Ondřej Pohl
16. května 2017

Poděkování

Děkuji vedoucímu mé práce Ing. Vladimíru Bartíkovi, Ph.D. za odborné přípomínky a rady. Dále bych chtěl poděkovat Mgr. Martinovi Netolickému z ČSÚ za poskytnutí doplňujících informací.

Obsah

Kapitola 1

Úvod

Informační technologie, zejména informační systémy, jsou již dlouhou dobu neoddělitelnou součástí mnoha organizací, firem a státních institucí. S rozvojem informačních technologií se podnikové informační systémy staly pro podniky klíčovým a nezbytným prvkem pro jejich fungování. Poskytují podporu pro obchodní procesy a zefektivňují provoz podniku v mnoha směrech. Postupem času se ukázalo, že dobře implementovaný informační systém není jediný faktor ovlivňující úspěch podniku. Do popředí se dostala potřeba využít potenciál ukrývající se v datech, jež vedla ke vzniku oboru Business Intelligence.

Podnikové informační systémy pracují s daty na úrovni, která plně postačuje pro každodenní provoz podniku. Avšak řídícím pracovníkům už dávno nestačí mít k dispozici jen týdenní přehledy poskytující pohled na aktuální provoz firmy. Pro potřeby taktického a strategického řízení podniku je nutná hlubší analýza dat. Data, která podnik v průběhu své existence shromažďuje, se stávají pro firmu nejcennějším aktivem. Analýzou těchto dat mohou vedoucí pracovníci získat užitečné znalosti a podporu pro své rozhodování.

Cílem této diplomové práce je analýza dat zahraničního obchodu ČR využívající analytické možnosti Business Intelligence a implementace analytických úloh.

ČSÚ kromě zveřejňování vývoje a analýz zahraničního obchodu poskytuje webovou aplikaci umožňující získat data zahraničního obchodu a provádět jejich filtraci včetně jednoduché sumarizace (webovou aplikaci ČSÚ lze považovat za nástroj pro tvorbu tiskových sestav). Avšak není možnost podrobit data hlubší analýze, ať už jde o analýzu OLAP nebo dolování dat. Takové sofistikovanější analýzy by mohly sloužit jako doplněk k prostředkům nabízených ČSÚ a představovat, zejména pro uživatele z podnikového prostředí, užitečný nástroj pro získávání znalostí.

Druhá kapitola diplomové práce představuje oblast Business Intelligence. Součástí kapitoly je pojednání o problematice datových skladů, která s analýzou podnikových dat úzce souvisí.

Třetí kapitola se detailněji zabývá analýzou OLAP. Vysvětluje multidimenzionální datový model a jeho přínosy pro účely analýzy dat. Také se věnuje přehledu OLAP architektur a operacím nad OLAP kostkami.

Předmětem čtvrté kapitoly je získávání znalostí z databází. V této kapitole je popsán celý proces získávání znalostí spolu s postupy a metodami této pokročilé analýzy dat. Podrobněji se zde zabývám klasifikací a získáváním asociačních pravidel.

Pátá kapitola se zaměřuje na popis webové aplikace Českého statistického úřadu a dat zahraničního obchodu, která tato aplikace poskytuje.

V šesté kapitole je čtenář seznámen s nástroji Business Intelligence v Microsoft SQL Serveru. Následuje popis stažení dat a návrh analytických úloh. Zbývající část kapitoly je pak věnována analýze dat v prostředí Microsoft SQL Serveru.

Jádrem sedmé kapitoly je implementace BI aplikace pro analýzu dat zahraničního obchodu. Tato kapitola popisuje implementaci stažení dat a vybraných analytických úloh.

V osmé kapitole je funkčnost implementovaných analytických úloh prezentována na příkladech analýzy OLAP a dolování dat.

Závěrečná kapitola shrnuje výsledky dosažené během řešení této diplomové práce a uvádí možné cesty pro pokračování práce do budoucna.

Kapitola 2

Business Intelligence

V dnešní době snad nikoho nepřekvapí, že s rozvojem informačních technologií roste každým dnem množství dat a informací, které máme k dispozici. Uvážíme-li různou kvalitu zdrojů těchto informací, dojdeme k závěru, že činit rychlá a správná rozhodnutí se stává čím dál obtížnější. Pro řízení organizací a podniků jsou správná rozhodnutí klíčová a technologie Business Intelligence k tomu nabízí vhodné prostředky.

2.1 Úvod do Business Intelligence

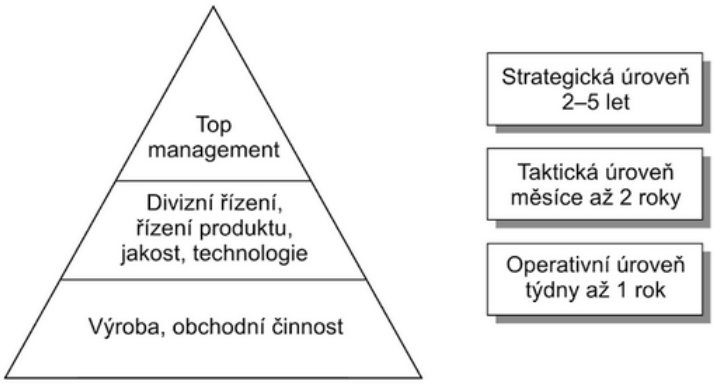
V textu věnovaném úvodu do Business Intelligence se vzhledem k rozsáhlosti této problematiky nejprve stručně zmíním o úrovních řízení podniku a podnikových informačních systémech. V dalších částech se budu zabývat charakteristikou a architekturou technologie Business Intelligence spolu s vysvětlením odlišností analytických a provozních systémů.

2.1.1 Úrovně řízení podniku a podnikové informační systémy

Řízení v podniku a s tím související rozhodování lze rozdělit do třech úrovní [23], [30]:

- **strategické řízení (rozhodování)** – Představuje nejvyšší úroveň řízení a je prováděno pracovníky vrcholového managementu nebo přímo vlastníky podniku. Rozhodnutí na strategické úrovni ovlivňují celý podnik v horizontu několika následujících let a mají značný vliv na uskutečňování dlouhodobých cílů firmy. Strategická rozhodnutí souvisejí s událostmi, které je nelehké předvídat a potýkají se s méně stukturovanými problémy.
- **taktické řízení (rozhodování)** – Řízení v rámci taktického rozhodování je realizováno v rozsahu jednotlivých oddělení nebo organizačních jednotek podniku. Na úrovni taktického řízení se nachází střední management, který plní úkoly a cíle určené v rámci strategického řízení. Z taktické úrovně jsou následně odvozovány cíle pro operativní úroveň.
- **operativní řízení (rozhodování)** – Operativní řízení pokrývá provozní aktivity firmy. Činnosti realizované na operativní úrovni jsou opakované, dají se dobře strukturovat a pohybují se v časovém rozpětí dní nebo hodin. Dosažení stanovených cílů je možné rychle hodnotit a při řešení problémů se uplatňují standardní postupy.

Úrovně řízení se liší především v časových horizontech, ve kterých dosahují zvolených cílů, a množstvím odpovědnosti během jejich plánování a samotné realizace. Rozdělení úrovní řízení z hlediska hierarchie a času ilustruje obrázek 2.1.



Obrázek 2.1: Úrovně řízení podniku [21].

Podnikové informační systémy se označují zkratkou ERP[1] a jejich základní vlastnosti a funkce zahrnují [19], [28]:

- integraci a především co nejvyšší automatizaci klíčových podnikových procesů (např. zpracování objednávek, faktur, evidence pohybu materiálu apod.),
- přístup k informacím a jejich vytváření v reálném čase (provoz ve víceuživatelském prostředí),
- standardizaci postupů při sdílení dat a dokumentů, správu kmenových dat,
- schopnost propojení a spolupráci s dalšími aplikacemi, možnost rozšíření.

Z výše uvedených vlastností ERP systémů je vidět jejich orientace spíše na provozní aktivity firmy. Některé ERP systémy se zaměřují i na okolí podniku a mohou obsahovat prvky technologie Business Intelligence [19].

Uživatelé ERP systémů jsou v prvé řadě pracovníci středního managementu. Pro pracovníky pohybující se na operativní úrovni řízení jsou podstatné základní funkce ERP systému. Vrcholový management se orientuje na práci s manažerskými informačními systémy a aplikacemi Business Intelligence [19].

Požadavky na aplikace určené pro pracovníky na strategické úrovni řízení jsou od požadavků na ERP systém v mnoha ohledech rozdílné. ERP systémy a další provozní systémy podniku mají z pohledu analýzy dat několik omezení.

Jedná se především o to, že provozní a transakční systémy podniku jsou primárně určeny k pořizování a aktualizaci dat. Jakékoliv další zatížení těchto systémů v podobě analýz dat by mělo negativní dopad na jejich výkonnost. Dalším problémem je efektivní přístup uživatelů k agregovaným datům z důvodu obrovského množství podnikových dat a způsobu jakým jsou tato data v ERP systémech uložena [26].

Vzrůstající potřeba analyzovat podniková data vytvořila prostor pro technologie a postupy Business Intelligence, které budou tématem dalších kapitol této práce.

2.1.2 Definice a charakteristika Business Intelligence

První koncepty zabývající se podporou manažerských činností a úloh začaly vznikat už na konci sedmdesátých let. První komerční počítačové aplikace zajišťující analytickou podporu

[1] Enterprise Resource Planning

manažerům byly uvedeny na trh v USA během druhé poloviny osmdesátých let. Na přelomu devadesátých let dochází k prosazování datových skladů a s rostoucím množstvím dat v těchto prostředích se objevily nástroje dolování dat umožňující provádět analýzy založené na statistických metodách a strojovém učení [26].

Aplikace pro podporu řízení na strategické a taktické úrovni procházely postupným vývojem a lze se často setkat s označením těchto aplikací pod zkratkami MIS[2], DSS[3] nebo EIS[4].

Vzhledem k tomu, že neexistuje pro termín Business Intelligence standardizovaná definice [26], dovolím si uvést více různých definic. Jako první definoval pojem Business Intelligence (BI) v roce 1989 Howard Dresner ze společnosti Gartner Group [25]:

> *Business Intelligence je množina konceptů a metodik, které zlepšují rozhodovací proces za použití metrik, nebo systémů založených na metrikách. Účelem procesu je konvertovat velké objemy dat na poznatky, které jsou potřebné pro koncové uživatele. Tyto poznatky potom můžeme efektivně použít například v procesu rozhodování a mohou tvořit velmi významnou konkurenční výhodu.*

Uvedu ještě jednu definici Business Intelligence zdůrazňující roli matematických modelů a jejich aplikaci při získávání znalostí z dat [30]:

> *Business Intelligence může být definováno jako sada matematických modelů a analytických metodologií, které využívají dostupná data k získání informací a znalostí užitečných v komplexních procesech rozhodování.*

Uvedené definice ukazují podstatu Business Intelligence, kterou je poskytování podpory pro podnikové řízení a rozhodování na základě informací a znalostí získaných z dat. Z toho plyne, že technologie Business Intelligence je orientována na oblast strategického a taktického řízení. Soustředí se na analýzu dat a nepracuje s daty způsobem typickým pro operativní úroveň řízení.

Data, informace a znalosti je nutné v kontextu Business Intelligence rozlišovat. Hierarchicky lze uspořádat tyto pojmy tak, že na nejnižší úrovni jsou data (údaje), poté informace a nejvyšší úroveň představují znalosti. Pokud provedeme zobecnění znalostí získáme navíc moudrost [25].

Čísla, znaky nebo obraz představují data. Samotná data nemají význam a například z pohledu uložení dat v databázi si je můžeme představit jako hodnoty určitého datového typu. Data se stanou informací ve chvíli, kdy jim přiřadíme význam. Informace vyjadřuje určitý vztah mezi daty a hodnota informace závisí na posouzení konkrétního uživatele. Informací je například počet prodaných produktů v určitém dni. Znalost označuje kombinaci zkušeností a informací zařazených do kontextu a vzniká během učení. Na znalost se vážou určité akce, čímž se znalost odlišuje od pouhé informace. Z pohledu řízení podniku pak znalosti ovlivňují podnikové činnosti a s tím spojenou úspěšnost těchto činností [29].

Pro správná strategická a taktická rozhodnutí je důležitá kvalita a přesnost dat. Kromě toho je nutné, aby informace a znalosti poskytované nástroji Business Intelligence byly pro jejich uživatele dostatečně užitečné a mohly být uplatňovány při rozhodování [24].

[2] Management Information Systems
[3] Decision Support Systems
[4] Executive Information Systems

Obecná architektura

Technologie BI kromě prezentačních a analytických nástrojů v sobě zahrnuje i základní datové úložiště v podobě systému datového skladu. Zajištění správné funkce celého řešení vyžaduje vzájemnou spolupráci nástrojů BI s datovým skladem a nelze opomenout ani požadavek na dostatečnou výkonnost celého systému z uživatelského hlediska.

Analytické metody Business Intelligence mohou být uplatňovány již na úrovni podnikových sestav a reportů, přes interaktivní formy analýz OLAP až po pokročilé metody dolování dat [24].

V této diplomové práci se podrobněji zabývám technologiemi OLAP a dolováním dat, kterým jsou věnovány samostatné kapitoly.

Obecná architektura prostředí Business Intelligence je znázorněna na obrázku 2.2 a sestává z několika vrstev [26], [24]:

- **Zdrojové systémy** – Tato vrstva zahrnuje produkční systémy, ERP a další zdroje dat včetně externích zdrojů.
- **Komponenty datové transformace** – Zajišťují sběr, transformaci a přenos dat ze zdrojových systémů do datových úložišť určených pro analýzu. Součástí jsou transformační ETL[5] a integrační EAI[6] nástroje. Důležitou roli hraje tato vrstva z hlediska zajištění kvality dat.
- **Databázové komponenty** – Na této vrstvě se nachází jádro podnikových dat v podobě datových skladů a datových tržišť, spolu s pomocnými databázemi využívanými během analýz (operativní datová úložiště, ODS[7]) nebo zpracování dat (Dočasná uložiště dat, DSA[8]).
- **Analytické komponenty** – Zajišťují samotné datové analýzy. K dispozici jsou nástroje pro reporting a vytváření sestav a pokročilé analýzy zahrnující OLAP a dolování dat.
- **Nástroje pro koncové uživatele** – Úkolem prezentační vrstvy je poskytnout uživatelům prostředí pro zadávání požadavků na analýzy včetně prezentace výsledků. Zde nachází uplatnění portálová BI řešení, manažerské aplikace (EIS) a vizuální prezentace založené na ovládacích panelech (dashboard) nebo přehledech výsledků (scorecard).

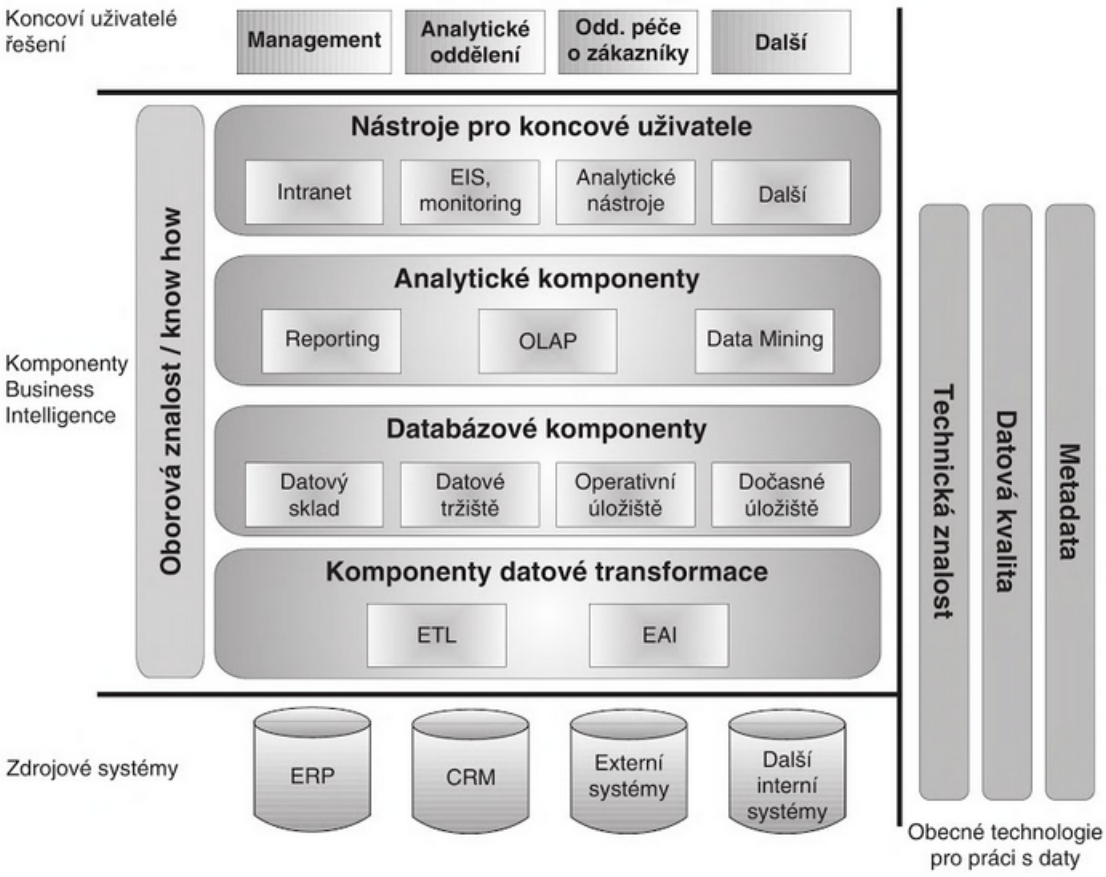
2.1.3 Rozdíly mezi OLTP a analytickými systémy

Zdrojové systémy uvedené v obecné architektuře BI patří mezi tzv. OLTP [9] systémy. Naproti tomu analytické systémy se od těchto transakčních systémů odlišují v mnoha směrech, které uvádím v následujícím souhrnu [26], [25]:

Odlišný návrh databází

Zatímco pro OLTP systémy je typické uložení dat v databázi ve třetí normální formě a snaha o co nejmenší datovou redundaci, pro analytické úlohy může být tento způsob uložení dat

[5] Extraction Transformation Loading
[6] Enterprise Application Integration
[7] Operational Data Store
[8] Data Staging Area
[9] On-line Transaction Processing



Obrázek 2.2: Obecná architektura BI [26].

v databázi značně neefektivní. Zejména u analýz založených na multidimenzionálním pohledu na data jako je OLAP je vhodnější uložení dat nenormalizovaně s využitím speciálních schémat.

Odlišné databázové modely obou systémů vychází ze skutečnosti, že OLTP systémy především zajišťují vkládání a aktualizaci dat, přičemž primárním účelem analytických systémů je podpora dotazování na data.

Různá úroveň detailu uložených dat a historická data

Během analýz dat se často pracuje s agregovanými daty nebo daty upravenými pro potřeby konkrétní analytické úlohy. OLTP systémy naopak ukládají detailní data každé transakce. Typickým rysem analytických systémů je využití historických dat pro účely komplexních analýz a predikce budoucího vývoje. Transakční systémy nemusí mít dostatečnou kapacitu na uchovávání historických dat a ani není jejich úkolem provádět časová srovnání nebo jinak zohledňovat faktor času.

Decentralizace OLTP systémů

Transakčních a produkčních systémů se může nacházet v podniku velké množství. Použití OLTP systémů pro rozsáhlou analýzu podnikových dat je z tohoto důvodu velmi kompli-

kované, jelikož jsou data uložena na různých místech. Kromě toho různé systémy mohou stejná data ukládat v rozdílných formátech.

Pro analýzy je nejvhodnější integrovat data do jednoho společného datového uložiště, nejčastěji datového skladu. Tím je zajištěno, že bude možné realizovat analytické úlohy v rozumném čase a na základě dat, která prošla procesy čištění a transformace.

Rozdílné zatížení

Transakční systémy jsou zatěžovány průběžně, vzhledem k tomu, že musí reagovat na obrovské množství požadavků v reálném čase. Pokud by se v těchto systémech současně prováděly naročné analytické operace, významně by tím byla ovlivněna odezva OLTP systémů i analytických nástrojů.

Oproti tomu zatížení analytických systémů je spíše nepravidelné. K nárazovému zatížení dochází při periodické aktualizaci a nahrávání dat do těchto systémů. Během vykonávání analýz pak zatížení závisí na složitosti analytických úloh.

2.2 Datové sklady

Důležitou součástí technologie Business Intelligence je centralizované uložiště všech podnikových dat označované jako datový sklad. Prostředí datového skladu, kde jsou data uložena v jednotné struktuře a odpovídající kvalitě, je stabilním základem pro analyzování dat.

2.2.1 Definice datového skladu

Princip datových skladů lze zjednodušeně popsat jako proces shromáždění všech dat organizace nebo podniku na jedno místo, přičemž data procházejí během tohoto procesu čištěním a transformací, aby mohla být následně použitelná pro analýzy a další rozhodování.

Datový sklad definuje Bill Inmon, který je považován za zakladatele této oblasti, následovně [27]:

> *Datový sklad je subjektově orientovaná, integrovaná, neměnná a časově variantní kolekce dat sloužící k podpoře manažerského rozhodování.*

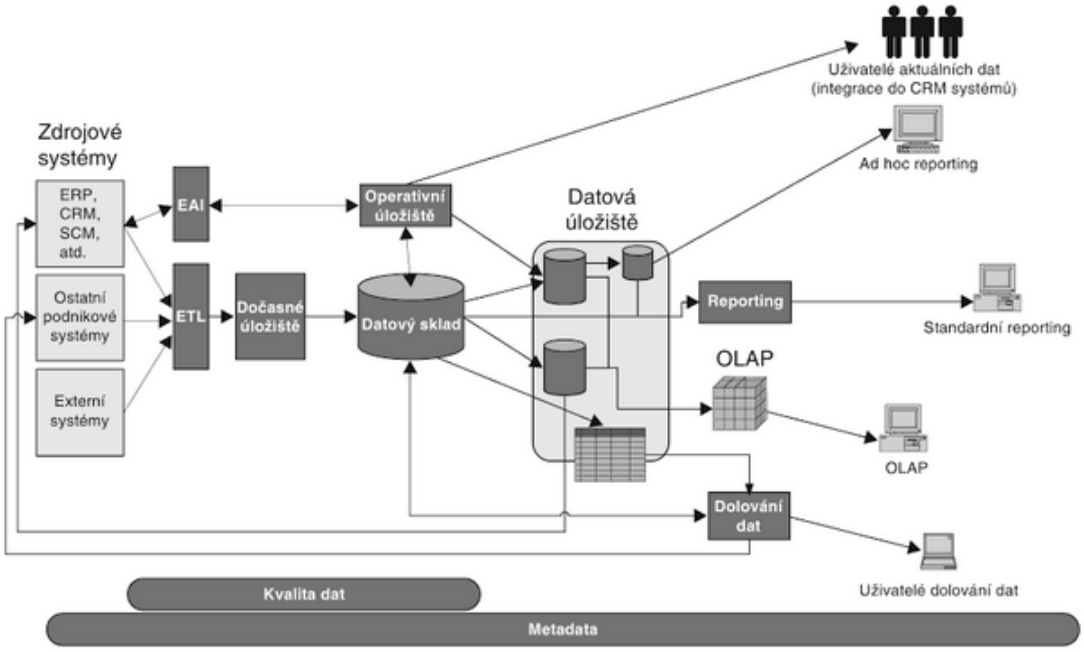
Z uvedené definice plynou čtyři základní vlastnosti datových skladů [27], [25]:

- **subjektová orientace** – Zatímco v provozních systémech probíhá ukládaní dat dle příslušné aplikace (zpracování faktur, aplikace řízení skladů, evidence zákazníků apod.), naproti tomu v datovém skladu jsou data ukládána s ohledem na subjekt. Subjekty mohou být například výrobky, dodavatelé, zaměstnanci nebo prodeje.
- **integrovanost** – Vzhledem k tomu, že se do datového skladu ukládají data z různých zdrojů, je nezbytným požadavkem konzistence dat. S tím souvisí zajištění jednotné terminologie, standardizace jednotek veličin a konvence v pojmenováních. Ke splnění těchto požadavků je využíváno nástrojů pro čištění a transformaci dat.
- **časová variabilita** – Data jsou do datového skladu ukládána ve formě časových snímků. Datový sklad uchovává historická data za delší časové období v řádu let. Naproti tomu produkční systémy pracují především s aktuálními daty.
- **neměnnost** – Do datového skladu probíhá nahrávání dat periodicky (např. jednou denně, jednou za týden) a jedná se o jediný případ, kdy nastává operace vkládání dat.

Data se dále nemění a jsou pouze přidávána. Uživatelé pak data čtou pro potřeby analýz.

Postavení datového skladu v rámci architektury BI znázorňuje obrázek 2.3. Z pohledu Business Intelligence je datový sklad úložištěm dat přicházejících ze zdrojových systémů a jak ukazuje obrázek 2.3, sám je zdrojem dat pro různé druhy analýz, zahrnující reporting, OLAP analýzu nebo dolování dat. Kromě toho datový sklad a související nástroje datové transformace zoodpovídají za zajištění kvality dat.

Pořízení a zpracování dat z provozních systémů je pravidelnou součástí provozu datových skladů a označuje se jako ETL (Extract, Transform, Load). Při budování datového skladu je ETL nejnáročnější fází, kdy je nutné namapovat data zdrojových systémů na cílový datový sklad a specifikovat aplikace, které budou extrakci a transformaci dat ze zdrojových systémů provádět [24].



Obrázek 2.3: Datový sklad v kontextu BI [26].

S datovými sklady souvisí také pojem datového tržiště (datového trhu). Datové tržiště je podmnožina datového skladu a jedná se o datový sklad pro konkrétní oddělení nebo oblast podniku. Datový sklad, který obsahuje všechna podniková data, se pak nazývá podnikovým datovým skladem a umožňuje celopodnikový pohled na data [27].

2.2.2 Přístupy k tvorbě datových skladů

Při implementaci datového skladu lze využít dvou základních přístupů. Jedná se o tzv. metodu velkého třesku a přírůstkovou metodu.

Metoda velkého třesku se dívá na implementaci datového skladu jako na jeden projekt skládající se z analýzy požadavků a vytvoření datového skladu včetně zajištění přístupu k datům. Jelikož je tvorba datového skladu náročný a dynamický proces, při němž se často mění požadavky, má metoda velkého třesku řadu nevýhod. Jednou z nich je např. dlouhá doba, než se objeví přínosy investice do datového skladu.

Vhodnějším přístupem jsou pak přírůstkové metody, kdy se nebuduje najednou celý datový sklad, ale vytváření datového skladu probíhá postupně v rámci iterativního procesu. V každé etapě je do datového skladu zahrnuta určitá oblast, například formou implementace datového tržiště. Zároveň lze během jednotlivých etap pružně reagovat na požadavky uživatelů. Výsledkem je pak úplný datový sklad s možností dalšího rozšíření [25].

Tvorbu datového skladu pomocí přírůstkové metody je možné realizovat dvěma způsoby [24], [27]:

Přístup shora dolů

Přístup shora dolů je někdy také označovaný jako centralizovaný nebo podnikový přístup a o jeho prosazení se zasloužil Bill Inmon. Základem je vytvoření centrálního úložiště, které je plněno daty ze zdrojových systémů. Uživatelé pak k tomuto úložišti přistupují prostřednictvím datových trhů. Datové trhy se v tomto případě soustředí na specifické podnikové použití a jsou odvozeny z centrálního úložiště. U přístupu shora dolů může být budování datového skladu zdlouhavé i přesto, že probíhá iterativním způsobem. Problémem může být také zajištění škálovatelného návrhu centrálního úložiště.

Přístup zdola nahoru

Princip přístupu zdola nahoru je založen na orientaci na datové trhy a je prosazován Ralphem Kimballem. Začíná se postupným budováním datových trhů, což je rychlejší a méně rizikový přístup, než vytvářet celý datový sklad už na počátku. Na druhou stranu každý datový trh má omezený pohled na podniková data a je problematické zajistit celopodnikový pohled. Jako praktické se ukázalo kombinovat přístup shora dolů spolu s přístupem zdola nahoru a využívat výhod obou metod ve formě hybridního přístupu.

Kapitola 3

OLAP

Prostředí OLTP systémů umožňuje provádět základní analýzu dat ve formě pravidelných reportů, prostřednictvím tabulkových procesorů nebo přímého dotazování na data pomocí jazyka SQL. Avšak pro komplexní analýzu je nutný multidimenzionální pohled na data zahrnující různé úrovně agregace a různé způsoby prezentace výsledků.

Cílem této kapitoly je představit multidimenzionální datový model, na kterém je postavena technologie OLAP spolu s popisem dostupných analytických operací a možnostmi implementace této technologie.

3.1 Multidimenzionální datový model

V následující podkapitole věnující se datové kostce uvádím úvodní informace k pochopení základů multidimenzionálního modelu, na které později navážu v části o implementaci multidimenzionality v relačních databázích.

3.1.1 Datová kostka

V multidimenzionálním datovém modelu jsou data modelována v podobě vícedimenzionální datové kostky. Datová kostka je definována pomocí dimenzí a faktů [22]:

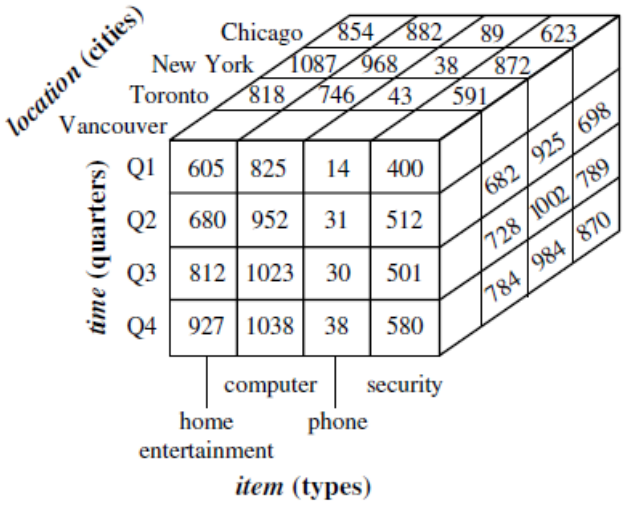
Dimenze

Dimenze jsou entity určené s ohledem na oblast, ve které se provádí analýza. Budeme-li uvažovat datový sklad sloužící pro analýzu prodejů produktů, pak mohou být pro tento účel definovány dimenze *čas*, *produkt* a *lokace*. Příklad odpovídající trojrozměrné datové kostky pro prodeje produktů je na obrázku 3.1.

Pro dimenze se definují konceptuální hierarchie umožňující pohledy na data v různých úrovních abstrakce. Konceptuální hierarchie určuje pořadí, ve kterém se mapují koncepty z nižší úrovně abstrakce na koncepty vyšších úrovní. Například pro dimenzi *lokace* je jednou z možných konceptuálních hierarchií ulice $<$ město $<$ stát.

Fakta

Fakta jsou numerické míry (metriky), které si lze představit jako veličiny sloužící k analýze vztahů mezi dimenzemi. Faktem je míra určená na základě předmětu analýzy. V případě prodejů produktů by se za fakt mohl označit počet prodaných kusů nebo celková částka za jejich prodej.



Obrázek 3.1: Příklad trojrozměrné datové kostky [22].

Na příkladu datové kostky z obrázku 3.1 ukážu nejčastěji prováděné OLAP operace:

- **Roll-up** – Umožňuje posun v rámci konceptuální hierarchie dimenze na vyšší úroveň agregace. Například pro dimenzi *čas* může být provedeno zobecnění zobrazení prodejů produktů z jednotlivých čtvrtletí na roky (za předpokladu, že jsou roky definovány v konceptuální hierarchii dimenze).
- **Drill-down** – Je opakem operace *roll-up*, kdy je prováděn posun v konceptuální hierarchii dimenze na nižší úroveň agregace (v případě dimenze *čas* např. ze čtvrtletí na měsíce).
- **Slice and dice** – Operace *slice* provádí výběr z datové kostky podle jedné dimenze (např. podle času). Výsledkem by byla dvourozměrná podkostka s dimenzemi *lokace* a *produkt*. Výsledkem operace *dice* je podkostka, který vznikla na základě výběru dvou nebo více dimenzí (např. můžeme určit podmínky pro výběr prvního a druhého čtvrtletí z dimenze *čas*, jen některých měst z dimenze *lokace* a některých produktů z dimenze *produkt*).
- **Pivot** – Někdy se nazývá *rotace* a jedná se o operaci, kdy dochází k rotaci os dimenzí. Tato operace slouží k nabídnutí jiného pohledu na data a jedná se pouze o jinou vizualizaci.

3.1.2 Systémy OLAP

Analýzy založené na multidimenzionálním pohledu na data vyžadují zavedení specializovaných nástrojů a prostředí přímo určených pro tento typ úloh. Takovým prostředím jsou systémy OLAP, které jsou definovány takto [27]:

> *OLAP (On-Line Analytical Processing) je skupina softwarových technologií, které umožňují analytikům, manažerům a řídícím pracovníkům získat pohled na data postavený na rychlém, konzistentním, interaktivním přístupu s širokým výběrem možných pohledů na informace, které byly získány transformací ze základních dat a které prezentují skutečný stav podniku formou srozumitelnou pro uživatele.*

Z výše uvedené definice plynou následující klíčové vlastnosti OLAP systémů: vícedimenzionální pohled, interaktivní přístup, konzistence a rychlost. Vícedimenzionální pohled umožňuje náhled na data z více hledisek a současně je možný pohyb na různé úrovni podrobnosti v rámci dimenzí. Odezva a konzistence OLAP systému hraje z uživatelského hlediska důležitou roli, protože uživatelé mohou vyžadovat vykonání analytických úloh různých složitostí. Z těchto důvodů využívají systémy OLAP také předpočítané agregace, které urychlují provádění komplexních analýz [27].

Příkladem takové analýzy může být hledání příčiny poklesu tržeb společnosti za posledních několik měsíců. Sledovanými dimenzemi může být např. čas, místo a produkt. Proces analýzy se pak skládá z několika kroků, kdy uživatel formuluje dotazy, při kterých analyzuje výsledky a interaktivně postupuje v analýze. Uživatel může podle potřeby měnit úrovně agregace u jednotlivých dimenzí (např. se v čase posunout z měsíců na týdny a dále sledovat tržby za jednotlivé produkty v rámci každého regionu) a pomocí dalších zmíněných operací na datové kostce nakonec identifikovat příčinu poklesu tržeb.

Systém OLAP lze charakterizovat pomocí pravidel publikovaných v roce 1993 Dr. E. F. Coddem. Celkem definoval dvanáct pravidel, která musí OLAP splňovat [27]:

1. **Multidimenzionální konceptuální pohled** – Uživatel by měl mít k dispozici jednoduše použitelný multidimenzionální datový model, který odpovídá jeho pohledu na podnikání.

2. **Transparentnost** – OLAP technologie spolu s databází a výpočetním prostředím by měly být pro uživatele transparentní a umožnit tak uživateli plnou produktivitu při používání front-end nástrojů.

3. **Dostupnost** – Systém OLAP by měl mít přístup pouze k datům, která jsou potřebná pro danou analýzu, bez ohledu na to, z jakých zdrojů pocházejí.

4. **Konzistentní výkon při vykazování** – Uživatelé by se neměli setkat se snížením výkonu při rostoucí velikosti databáze nebo zvýšení počtu dimenzí. Odezva OLAP systému by měla být stále konzistentní.

5. **Architektura klient/server** – OLAP systém musí být postaven na architektuře klient/server, aby byl zajištěn optimální výkon, flexibilita a interoperabilita.

6. **Obecná dimenzionalita** – Musí být zajištěno, že každá dimenze je stejná jak ve struktuře, tak v operačních schopnostech.

7. **Dynamické operace s řídkými maticemi** – Systém OLAP musí být schopný přizpůsobit své fyzické schéma konkrétnímu analytickému modelu, který provede optimalizaci ošetření řídkých matic při současném zachování konzistentní úrovně výkonu.

8. **Podpora více uživatelů** – Je nutné zajistit podporu pro současnou práci více uživatelů nad stejným analytickým modelem nebo při vytváření různých modelů nad stejnými daty.

9. **Neomezenené křížové dimenzionální operace** – OLAP systém musí umět rozeznávat hierarchie v dimenzích a automaticky provádět operace zvýšení nebo snížení úrovně agregace uvnitř jedné dimenze nebo napříč dimenzemi.

10. **Intuitivní manipulace s daty** – Provádění analytických operací z pohledu uživatelského rozhraní musí být intuitivní a prováděno formou akcí drag-and-drop nebo point-and-click.

11. **Flexibilní vykazování** – Uživatelé musí mít možnost uspořádání řádků, sloupců a buněk tak, aby byla analýza informací a manipulace s informacemi v sestavách co nejjednodušší.

12. **Neomezený počet dimenzí a úrovní agregace** – Systém OLAP by neměl omezovat počet dimenzí nebo úrovní agregace v analytických modelech.

3.1.3 Druhy architektur OLAP

Podle způsobu uložení dat se technologie OLAP dělí na [27]:

Relační OLAP (ROLAP)

V relačním OLAPu jsou data uložena tradičním způsobem v relačních tabulkách v datovém skladu. Pro zajištění multidimenzionálního pohledu na data se provádí mapování dimenzí na tabulky v relační databázi. K tomu slouží vrstva metadat, která skrývá před uživateli skutečnou strukturu uložení dat. Metadata umožňují podporu některých forem agregací.

Při dotazech uživatele na multidimenzionální data provádí OLAP server vytvoření multidimenzonálního pohledu dynamicky a tento pohled je pak předložen uživateli. Dotazy uživatele jsou s využitím metadat překládány do komplexních dotazů jazyka SQL a zaslány přímo pro zpracování relační databázi.

Multidimenzionální OLAP (MOLAP)

MOLAP ukládá data v multidimenzionální databázi ve formě vícedimenzionálních polí, kde jsou uložena předpočítaná data. Datový sklad zde slouží jako zdroj dat pro vytváření sumarizovaných datových kostek, které jsou ukládány do této multidimenzionální databáze. Uživatelé získávají multidimenzionální pohled na data z již vytvořených datových kostek, a tím je zajištěn vyšší výkon provádění analýz než u ROLAPu.

MOLAP je vhodný tam, kde je nutné provádět rychle komplexní analýzy. Nevýhodou můžou být požadavky na kapacitu úložiště multidimenzionální databáze, které prudce rostou při zvyšování počtu dimenzí.

Desktop OLAP (DOLAP)

DOLAP umožňuje přenést vytvořená multidimenzionální data na lokální počítač, kde je k dispozici DOLAP software pro následnou analýzu. DOLAP je variací ROLAPu.

Hybridní OLAP (HOLAP)

HOLAP kombinuje výhody ROLAPu a MOLAPu tak, že detailní data jsou uložena v relační databázi a sumarizovaná data v multidimenzionální databázi [25].

3.2 Multidimenzionalita v relační databázi

Architektura ROLAP implementuje multidimenzionalitu na úrovni relační databáze pomocí speciálních schémat a tabulek. Způsob uložení dat je v tomto případě zcela odlišný od organizace dat v relační databázi systémů OLTP, kde uložená data splňují třetí normální formu s cílem zajistit rychlé ukládání a aktualizaci dat.

3.2.1 Tabulky faktů a dimenzí

Základem datového modelu při implementaci multidimenzionality v relačních databázích jsou tabulky faktů a dimenzí, které jsou organizovány ve schématech hvězdy nebo sněhové vločky, o nichž se později zmíním.

Tabulka faktů

Tabulka faktů ukládá zvolené metriky a další ekonomické ukazatele na určité úrovni granularity, včetně klíčů do tabulek dimenzí. Primární klíč tabulky faktů je pak složen ze všech primárních klíčů tabulek dimenzí. Velikost tabulky faktů, co se týče počtu řádků, je mnohonásobně větší než v případě tabulek dimenzí, naopak počet atributů tabulky faktů bývá menší [27].

Metriky ukládané v tabulce faktů lze rozdělit na [24]:

- **aditivní** – Agregaci těchto metrik lze provádět na základě všech dostupných dimenzí. Typickým příkladem může být metrika *Počet prodejů* a dostupné dimenze *čas* a *produkt*.
- **semiaditivní** – Agregaci lze provádět jen pro některé dimenze. Příkladem takové metriky může být *Počet kusů zboží na skladě*, který má smysl sčítat za jednotlivé obchody v rámci dne, ale nikoli za několik dnů (některé zboží může zůstat na skladě i další dny).
- **neaditvní** – Není možné provádět agregaci v žádné dostupné dimenzi. Mezi neaditivní metriky patří například procenta nebo hodnoty průměrů.

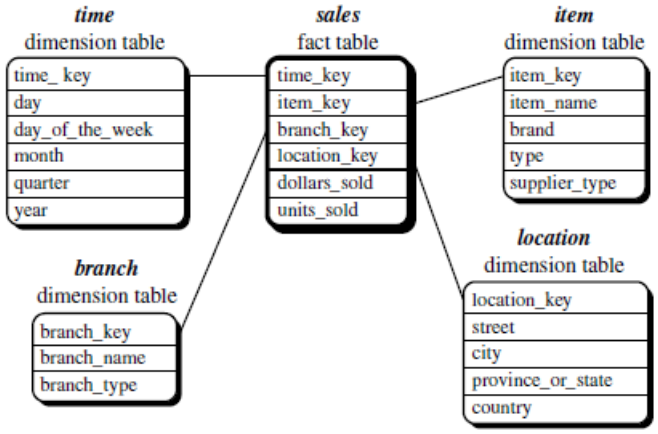
Tabulky dimenzí

Tabulky dimenzí reprezentují dimenze, v rámci kterých probíhá analýza metrik uložených v tabulce faktů. Atributy v tabulce dimenze umožňují provádění operací, při kterých se mění úroveň agregace v rámci konceptuální hierarche dimenze. Atributy obsahují textové popisy dimenze a v čase se spíše nemění. Například tabulka pro dimenzi *lokace* bude obsahovat textové atributy názvu ulic, měst a států.

Pro zajištění co nejvyššího výkonu při dotazech bývají tabulky dimenzí nenormalizované, aby byl zajištěn přímý přístup z dimenze do tabulky faktů a nebylo nutné provádět dodatečné spojování tabulek dimenzí [27].

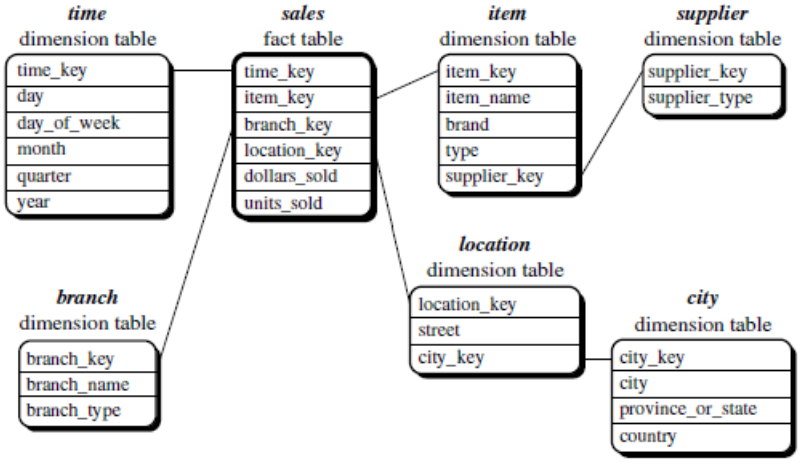
3.2.2 Schémata uložení faktů a dimenzí

Obrázky 3.2 a 3.3 znázorňují způsoby jakými je multidimenzionalita na úrovní relační databáze modelována. Tabulka faktů představuje centrální tabulku, která je napojena na sadu tabulek dimenzí. Schémata na obrázcích připomínají hvězdu nebo sněhovou vločku podle toho, jak jsou dimenze organizovány.



Obrázek 3.2: Schéma hvězdy [22].

Ve schématu hvězdy je každá dimenze reprezentována jednou tabulkou, zatímco ve schématu sněhové vločky mohou být rozděleny některé dimenze do více tabulek dimenzí. Tabulky dimenzí v případě sněhové vločky procházejí normalizací, což vede k nižší redudanci a úspoře místa. Vzhledem k velikosti tabulky faktů, je ale úspora místa zanedbatelná a navíc rozdělení tabulek dimenzí má negativní dopad na výkon při provádění analýz. Schéma hvězdy je z tohoto pohledu výhodnější a používané častěji.



Obrázek 3.3: Schéma sněhové vločky [22].

Schéma nemusí vždy obsahovat pouze jednu tabulku faktů. Schémata hvězd mohou tvořit schéma souhvězdí, kde jsou například dvě tabulky faktů, které spolu sdílí několik tabulek dimenzí [22].

Kapitola 4

Získávání znalostí z databází

Získávání znalostí z databází patří vedle analýzy OLAP mezi další pokročilé analytické metody Business Intelligence. Umožňuje objevovat nové souvislosti a získat nové poznatky, které mohou být v podobě znalostí zužitkovány v procesu rozhodování.

Tato kapitola uvádí základní informace z oblasti získávání znalostí dat a věnuje se především problematice předzpracování dat a popisu vybraných dolovacích technik.

4.1 Úvod do problematiky dolování dat

Na rozdíl od jiných analytických metod Business Intelligence se při dolování dat nepoužívá běžné dotazování v jazyce SQL (jako v případě reportování nebo analýzy v podobě přímého dotazování do databáze), ale uplatňují se postupy z oblasti statistiky a strojového učení.

V úvodu problematiky dolování dat se kromě výkladu tohoto pojmu zaměřuji na oblasti reálného použití dolování dat a dále na zdroje dat, které se využívají pro tento typ pokročilé analýzy.

4.1.1 Definice a charakteristika

Získávání nebo dobývání znalostí z databází (dat) může být definováno jako [20]:

> *Netriviální získávání implicitních, dříve neznámých a potenciálně užitečných informací z dat.*

Vlastnost netriviálního získávání informací souvisí se skutečností, že při řešení úloh v oblasti dolování dat si již nelze vystačit pouze s dotazováním do databáze, o kterém jsem mluvil v úvodu. Získávání implicitních informací znamená, že se jedná o informace, které nejsou přímo vyjádřeny, ale jsou v datech obsaženy.

Dnes se pro získávání znalostí z databází používá nejčastěji pojem *dolování dat* (data mining). Mezi méně používaná označení patří např. *datová archeologie* (data archeology) nebo *bagrování dat* (data dredging) [22].

Zajímavé vzory a vztahy v datech, které lze při dolování dat objevit, nejsou na první pohled vidět a jsou v datech skryty. Typickým příkladem aplikace dolování dat při hledání skrytých vztahů je tzv. analýza nákupního košíku, kdy se v datech o prodejích zboží hledají společně nakupované položky. Následně je možné získané znalosti využít pro vhodnější uspořádání zboží na prodejně nebo účinnější prodejní akce a reklamní kampaně.

Kromě analýzy nákupního košíku patří k dalším aplikacím dolování dat:

- identifikace pojišťovacích podvodů nebo podezřelých bankovních transakcí,
- segmentace trhu v marketingu,
- predikce vývoje cen (např. vývoje cen elektřiny, plynu na burze),
- klasifikace klientů banky při poskytování úvěrů,
- analýza biologických dat,
- analýza chování na sociálních sítí, detekce spamu.

Při získávání znalostí z dat není určen univerzální postup aplikovatelný na všechny problémy. Na základě zvoleného cíle, který má být při dolování splněn a povahy dat, je nutné provést volbu vhodné dolovací techniky. Mezi hlavní techniky používané při dolování patří klasifikace a predikce, hledání frekventovaných vzorů (dolování asociačních pravidel), detekce odlehlých objektů, shluková analýza a analýza evoluce.

4.1.2 Zdroje dat pro dolování

Dolování je možné provádět prakticky v jakýchkoliv datech, ať už se jedná o data v tradiční relační databázi nebo multimediální data. Přehled některých zdrojů dat, které lze využít pro dolování, uvádím v přehledu níže [22]:

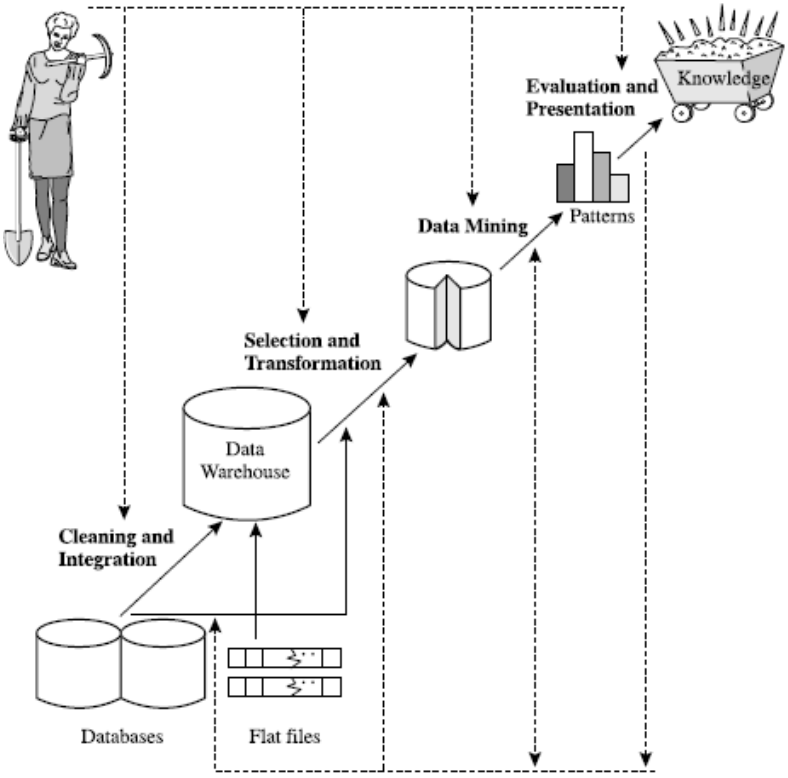
- **relační databáze a datové sklady** – Relační databáze informačních systémů obsahují velké množství dat například o zákaznících nebo prodejích, které lze pro dolování využít. V oblasti datových skladů se využívá dolování dat založené na technologii OLAP nazývané jako OLAM (On-Line Analytical Mining).
- **transakční databáze** – Jedná se o soubor obsahující záznamy o transakcích, kde se každý záznam skládá z identifikátoru a seznamu položek (např. nakupovaného zboží).
- **databáze časových řad** – Časové řady obsahují posloupnosti hodnot získané opakovaným měřením (po každé hodině, dni apod.) jako jsou burzovní data nebo data o počasí (např. teplota).
- **textové a multimediální databáze** – Zdroje multimediálních dat a textů slouží například pro vyhledávání podle obsahu nebo hledání podobností v obrazových datech.
- **datové proudy** – Zahrnuje data síťového provozu, video streamy nebo vědecká data.
- **prostředí webu** – Je vhodné pro analýzy webových stránek nebo segmentaci uživatelů sociálních sítí.

4.1.3 Předzpracování dat

Získávání znalostí z dat nezahrnuje pouze dolování dat (ve smyslu aplikace nějakého dolovacího algoritmu), ale jedná se o proces, ve kterém probíhá nejprve příprava dat, pak samotné dolování a na konci vyhodnocení a prezentace výsledků, jak ukazuje obrázek 4.1.

Vzorů dat získaných na konci procesu získávání znalostí může být obrovské množství a ne všechny jsou použitelné při potřeby dalšího rozhodování. Mezi kritéria charakterizující zajímavý vzor patří [22]: jednoduchá srozumitelnost pro uživatele, novost, potenciální užitečnost a platnost pro nová nebo testovací data s danou úrovní určitosti.

Jak je vidět z obrázku 4.1, před samotným dolováním se provádí předzpracování dat, které má za cíl během kroků čištění a integrace zajistit co nejvyšší kvalitu dat, aby výsledky dolování nebyly zkreslené a byly založeny na platných datech. Součástí předzpracování jsou



Obrázek 4.1: Proces získávání znalostí z dat [22]. Pro proces získávání znalostí z dat bylo zavedeno také několik metodik, které shrnují nejlepší postupy a doporučení pro řešení problémů pomocí dolování dat, jako je např. metodika SEMMA od firmy SAS nebo obecnější metodika CRISP-DM [20].

také transformace a redukce dat podle toho, jaká dolovací technika bude v kroku dolování dat použita.

Čištění a integrace dat

Úkolem čištění dat je vyřešit problémy s chybějícími hodnotami a zašuměnými a nekonzistentními daty. Příčinou chybějících hodnot můžou být chybějící hodnoty atributů nebo mohou chybět přímo konkrétní atributy, které jsou důležité z hlediska dolování, ale nejsou obsaženy ve zdrojových datech. Hodnota atributu nemusí být k dispozici, pokud došlo k chybě při sběru dat (chyba programu, chyba člověka) nebo pokud nebylo povinné danou hodnotu při zadávání dat uvádět (např. data z webových formulářů).

Problémy s chybějícími hodnotami lze vyřešit několika způsoby [22]:

- ignorováním záznamu,
- manuálním doplněním chybějící hodnoty,
- použitím globální konstanty, která bude nahrazovat chybějící hodnotu,
- náhradou průměrnou hodnotou atributu,
- náhradou průměrnou hodnotou atributu ze všech záznamů třídy daného záznamu,

- užitím nejvíce pravděpodobné hodnoty (nejvhodnější způsob, využívá nejvíce informací z dat pro určení chybějící hodnoty).

Zašuměná data obsahují nesprávné nebo odlehlé hodnoty, jejichž příčinou můžou být opět chyby způsobené při sběru a zadávání dat, kde dochází k neshodě mezi formáty a pojmenováními nebo chybám při měření dat. Různé konvence v pojmenování a odlišné formáty jsou také příčinou nekonzisteních dat, kdy dochází ke slučování dat z více zdrojů.

Pro ošetření zašuměných dat se používají následující metody [22]:

- **plnění** – Plnění (binning) pracuje se setříděnými daty, která jsou rozdělena do tzv. košů. Hodnoty jsou do košů rozděleny tak, že každý koš obsahuje stejný počet hodnot nebo pouze hodnoty z pevně daného intervalu. Hodnoty v koši jsou následně vyhlazeny buď průměrem, mediánem nebo pomocí hraničních hodnot koše, kdy hranice koše určují prvky koše s minimální a maximální hodnotou a hodnoty koše jsou pak nahrazovány podle toho, ke které hranici mají blíže.
- **regrese** – Vyhlazení dat probíhá dle určité regresní funkce, která může být buď lineární nebo vícenásobná lineární.
- **shlukování** – Jedná se o vhodnou metodu pro hledání odlehlých hodnot. Hodnoty s podobnými vlastnostmi tvoří shluky a každou hodnotu nacházející se mimo shluk podobných hodnot je možné považovat za odlehlou.

Integrace je prováděna za účelem sjednocení hodnot z více zdrojů a je podobná procesu integrace při vytváření datového skladu. Mezi typické problémy integrace patří sjednocení odlišností ve schématech různých zdrojů, nekonzistence mezi hodnotami nebo problémy rozdílné identifikace. Při integraci je častým problémem také redundance dat, která se detekuje například pomocí korelační analýzy.

Například identifikace zákazníka může být v každém zdroji uložena v jinak pojmenovaných atributech a v jiném formátu (číselný nebo naopak alfanumerický identifikátor). Příčinou redundace mohou být atributy, které popisují stejnou skutečnost odlišným způsobem (určení věku datem narození nebo pouze číselným údajem), nebo atributy, jejichž hodnoty lze odvodit z ostatních atributů.

Výběr, transformace a redukce dat

Při řešení konkrétního problému pomocí dolování dat se využívá určitá podmnožina dat, která se daného problému týká. Určení takové podmnožiny dat je součástí předzpracování dat, kdy je proveden výběr dat (např. z databáze datového skladu), jež budou využita při dolování.

Transformace dat zahrnuje operace, jejichž cílem je převod dat do tvaru vhodného pro aplikaci zvolené dolovací techniky. Mezi tyto operace patří vyhlazování, agregace, generalizace, normalizace a konstrukce atributů.

Operace vyhlazování slouží k odstranění šumu v datech, o kterém jsem se zmiňoval při čištění dat. Agregace dat patří mezi transformace typické při vytváření datových kostek umožňující pohled na data v různých úrovních podrobnosti. Generalizace provádí zobecnění dat na základě konceptuální hierarchie (např. zobecnění měst na země). Operace normalizace převádí hodnoty do zvoleného intervalu a operace konstrukce atributů slouží k vytvoření nových atributů za účelem zvýšení přesnosti dolování dat.

Redukce dat je prováděna za účelem zmenšení objemu dat určených pro dolování a zároveň musí tato redukce zachovat vlastnosti původního souboru dat, aby výsledky dolování na redukovaném souboru dat byly srovnatelné. K technikám redukce dat se řadí např. výběr podmnožiny atributů, kdy se detekují a odstraňují, z pohledu dané úlohy, nevýznamné nebo nesouvisejících atributy.

4.2 Typy dolovacích úloh

V následujícím textu se budu věnovat teoretickému popisu hlavních typů dolovacích úloh. Podrobněji se budu zabývat dolovacími úlohami, které budou předmětem experimentování v rámci této diplomové práce.

Dolovací úlohy lze rozdělit na deskriptivní a prediktivní [22]. Deskriptní dolovací úlohy popisují obecné vlastnosti dat zkoumaného datového souboru. Příkladem takové úlohy je zmiňovaná analýza nákupního košíku. Prediktivní úlohy provádějí predikci na základě dostupných dat. Prediktivním typem úlohy je například určení třídy spolehlivosti splácení pro nového zákazníka při poskytování bankovního úvěru.

4.2.1 Klasifikace a predikce

Základním rozdílem mezi klasifikací a predikcí je typ atributu, jehož hodnota je predikována. Klasifikace slouží k predikci hodnoty kategorického atributu (obsahuje pouze diskrétní hodnoty, např. rozdělení spolehlivosti splácení úvěru na *třídu A*, *třídu B* apod.), zatímco metody predikce slouží k předpovědi hodnoty spojitého atributu (nabývá numerických hodnot, na kterých je definováno uspořádání, např. predikce výše útraty zákazníka obchodu).

Proces klasifikace se skládá ze dvou kroků [22]:

1. **Trénování** – Na základě trénovací množiny je klasifikačním algoritmem vytvořen klasifikační model (klasifikátor). Klasifikační model může být reprezentován např. rozhodovacím stromem nebo neuronovou sítí. Trénovací množinu tvoří část záznamů ze vstupní datové sady, a je tedy známa hodnota predikovaného atributu nebo také třída (proto se klasifikace označuje za metodu založenou na učení s učitelem).

2. **Testování** – Testování vytvořeného modelu probíhá na testovací množině, tj. datech vybraných ze zbývající části vstupní datové sady, aby se mohla ověřit úspěšnost předpovědi hodnoty predikovaného atributu.

Po kroku testování je vytvořen nový klasifikační model, který má lepší schopnost klasifikace než předchozí nebo se stávající model začne používat pro klasifikaci objektů, u nichž není známa třída.

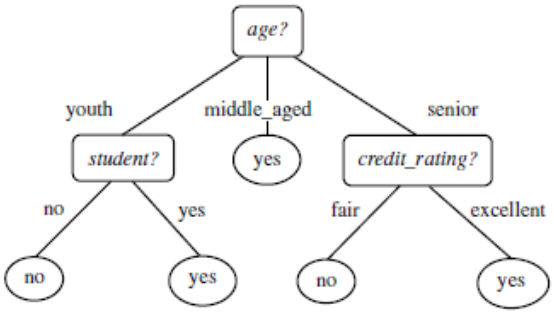
Klasifikaci lze provádět například pomocí rozhodovacího stromu, Bayesovské klasifikace, neuronových sítí nebo s využitím metody SVM [1]. Dále se budu zabývat klasifikací pomocí rozhodovacích stromů a Bayesovskou klasifikací.

Rozhodovací strom

Rozhodovací strom je z hlediska názornosti vhodným způsobem prezentace klasifikačního modelu. Příklad rozhodovací stromu pro určení, zdali si zákazník koupí nebo nekoupí určité

[1] Support Vector Machines

zboží, ukazuje obrázek 4.2. Vnitřní uzly rozhodovacího stromu označují testování hodnoty určitého atributu, větve stromu označují výsledky testu a listové uzly stromu představují třídy, do kterých je provedena výsledná klasifikace.



Obrázek 4.2: Klasifikační model v podobě rozhodovacího stromu [22].

Při vytváření rozhodovacího stromu se provádí výběr atributů podle kritéria nejlepší rozhodovací schopnosti daného atributu, přičemž atribut s nejvyšší mírou rozhodovací schopnosti je v kořeni stromu. Ideálním atributem je takový atribut, který rozdělí data na části, kde každá část bude obsahovat pouze objekty patřící do stejné třídy. K určení vhodného atributu pro rozdělování se používá informační zisk (*information gain*), poměrný informační zisk (*gain ratio*) nebo *gini index*.

Vytvořený rozhodovací strom může vlivem odlehlých hodnot a dalších anomálií v trénovací množině obsahovat větve, které snižují spolehlivost klasifikace rozhodovacího stromu. K odstranění těchto nepotřebných větví slouží následující metody provádějící tzv. prořezávání (z angl. *pruning*) klasifikačního stromu [22]:

- **Prepruning** – Nepotřebné větve jsou odstraňovány již při konstrukci stromu.
- **Postpruning** – Odstranění nepotřebných větví se provádí až po vytvoření stromu. Postpruning je výpočetně náročnější než prepruning, ale výsledkem aplikace postpruningu jsou spolehlivější rozhodovací stromy.

Bayesovská klasifikace

Klasifikace podle Bayesova klasifikátoru je založena na statistice a určuje příslušnost daného objektu do určité třídy podle vypočítané hodnoty pravděpodobnosti.

Základním klasifikátorem je *Naivní Bayesův klasifikátor* vycházející z předpokladu nezávislosti atributu na ostatních atributech při určení dané třídy a umožňuje tak při klasifikaci používat jednodušší výpočty. Avšak v praxi jsou mezi atributy závislosti, k jejichž modelování se využívají bayesovské sítě.

K výpočtu pravděpodobnosti pro určení výsledné třídy je definován Bayesův vzorec [22]:

$$P(C \mid X) = \frac{P(X \mid C)P(C)}{P(X)}$$

X označuje objekt, jehož třída má být určena. $P(C \mid X)$ označuje pravděpodobnost, že objekt X patří do třídy C. $P(X \mid C)$ označuje pravděpodobnost, že atributy náhodně

vybraného objektu ze třídy C budou stejné jako atributy objektu X. $P(C)$ označuje pravděpodobnost, že náhodně vybraný objekt patří do třídy C. $P(X)$ je konstantní.

Cílem je najít největší hodnotu $P(C \mid X)$ a určit tak třídu, pro kterou je největší pravděpodobnost, že do ní patří objekt X.

4.2.2 Dolování asociačních pravidel

Analýza nákupního košíku zmíněná v úvodu jako příklad aplikace dolování dat patří k typům dolovacích úloh, které získávají asociační pravidla z transakčních dat. Dolování asociačních pravidel se zabývá hledáním frekventovaných vzorů, tj. vzorů vyskytujících se v daném souboru dat s určitou frekvencí. V dalším popisu problematiky dolování asociačních pravidel budu pro názornost uvažovat příklad analýzy nákupního košíku.

Provedené nákupy jsou uloženy v databázi transakcí, kde každá transakce T je podmnožinou množiny všech položek I, které jsou v obchodě k dispozici. Asociační pravidlo je implikace tvaru $A \Rightarrow B$, kde A, B obsahují položky z I. Například pravidlo $koupi(chleb) \Rightarrow koupi(maslo)$ říká, že zákazník si při koupi chleba koupil také máslo.

Součástí pravidla je informace o jeho zajímavosti. Zajímavost asociačního pravidla se vyjadřuje prostřednictvím dvou základních metrik [22]:

- **podpora** – Pravidlo $A \Rightarrow B$ má podporu p, která vyjadřuje v kolika % všech analyzovaných transakcí se vyskytuje současně A a B.
- **spolehlivost** – Pravidlo $A \Rightarrow B$ má spolehlivost s, která vyjadřuje v kolika % všech transakcí, kde je přítomno A se nachází také B.

Pravidlo je označeno za zajímavé podle toho, jak jsou nastaveny hodnoty minimální podpory a spolehlivosti. Pokud pravidlo splňuje minimální podporu a spolehlivost, jedná se o tzv. silné asociační pravidlo.

Dolování asociačních pravidel probíhá ve dvou krocích [22]:

1. **Hledání frekventovaných množin** – Z databáze transakcí se získají frekventované množiny. Jedná se o množiny, které splňují stanovenou hodnotu minimální podpory.
2. **Generování asociačních pravidel z frekventovaných množin** – Z frekventovaných množin se získají pouze silná asociační pravidla.

Typy pravidel

Asociační pravidla se dělí do kategorií [22]:

- **podle úrovní abstrakce** – Dolování asociačních pravidel může probíhat na různých úrovních abstrakce, kdy výsledkem jsou pravidla reprezentující vztah položek různých úrovní. Příkladem můžou být pravidla $koupi(pocitac) \Rightarrow koupi(tiskarnu)$ a $koupi(notebook) \Rightarrow koupi(tiskarnu)$, kde položky počítač a notebook se nacházejí na různých úrovních abstrakce.
- **podle počtu dimenzí** – Pravidlo $koupi(notebook) \Rightarrow koupi(tiskarnu)$ je příkladem jednodimenzionálního pravidla, protože obsahuje pouze jednu dimenzi *koupí*. Pokud by byla přítomna další dimenze (např. *věk*) jednalo by se o vícedimenzionální pravidlo.

- **podle typu hodnot v pravidle** – Pravidlo obsahující pouze informaci o existenci nebo neexistenci položky v transakci se nazývá booleovské asociační pravidlo. Pokud pravidlo vyjadřuje vztah mezi kvantitativními atributy (např. *cena*), pak se jedná o kvantitativní asociační pravidlo.

Možnosti získání frekventovaných množin

Frekventované množiny lze získat pomocí algoritmu *Apriori* pracujícího na principu využití předchozích znalostí o frekventovaných množinách. Součástí algoritmu je tzv. apriori vlastnost, která říká, že všechny podmnožiny frekventované množiny musí být také frekventované množiny. Algoritmus *Apriori* se skládá ze spojovací a vylučovací fáze a probíhá v iteracích, kdy jsou získávány n-prvkové frekventované množiny. Ve spojovací fázi jsou generováni kandidáti na frekventované množiny, ve vylučovací fázi jsou odstraňovány množiny, které nejsou frekventované.

Nevýhodou algoritmu *Apriori* je riziko vygenerování příliš velkého množství kandidátů a časté procházení databáze. Tyto nedostatky lze odstranit získáváním frekventovaných množin bez nutnosti generování kandidátů. K tomu slouží metoda vzrůstu frekventovaných množin, která používá strukturu s názvem *FP-strom*. Tato metoda převede databázi frekventovaných položek do struktury *FP-stromu*. *FP-strom* je poté rozdělen na menší tzv. podmíněné stromy, ze kterých jsou získány frekventované množiny [22].

4.2.3 Shluková analýza

Dolovací úlohy založené na shlukování tvoří další významnou část dolovacích úloh, které nacházejí uplatnění například při rozpoznávání vzorů, průzkumu trhu nebo detekci odlehlých hodnot. O shlukové analýze se zde stručně zmíním, přestože v rámci této práce nevyužívám její metody.

Shlukování je proces, kdy jsou objekty rozdělovány do tříd nebo shluků. Rozdělování probíhá na základě hodnot atributů každého objektu, kdy bývá typicky využívána určitá vzdálenostní funkce.

Shluk lze charakterizovat jako kolekci navzájem podobných objektů v rámci tohoto shluku a zároveň odlišnou od objektů ostatních shluků.

Hlavním požadavkem kladeným na shlukovací metody je schopnost efektivního zpracování velkých objemů dat.

Shlukovací metody by proto měly mít následující vlastnosti [22]:

- **Škálovatelnost** – Je nutné zajistit, aby zpracování z pohledu velikosti dat bylo škálovatelné a shlukovací algoritmy pracovaly dobře s různě velkými objemy dat.
- **Schopnost zpracovat atributy různého typu** – Shlukování by mělo být možné nejen pro numerická data, ale i pro binární nebo kategorická data.
- **Vytvoření shluků libovolného tvaru** – Shluky mohou mít různou hustotu a tvar. Schopnost vytvářet shluky libovolného tvaru umožňuje lépe reprezentovat dané třídy.
- **Minimální požadavky na znalost problematiky při specifikaci parametrů** – Uživatel musí u některých shlukovacích metod nastavovat vstupní parametry, které ovlivňují kvalitu výsledku. Při výběru dolovací metody je nutné tuto skutečnost zohlednit.

- **Schopnost pracovat se šumem** – Vstupní data mohou obsahovat šum (např. chybějící nebo nepřesné hodnoty), což může mít vliv na kvalitu výsledných shluků.
- **Nezávislost na pořadí vstupních záznamů** – Shlukovací algoritmus by neměl být citlivý na pořadí vstupních záznamů.
- **Vysoká dimenzionalita** – Vstupní záznamy často obsahují velké množství atributů a shlukovací metoda by měla být schopna se vyrovnat s vysokou dimenzionalitou dat.
- **Použitelnost shluků** – Výstup shlukové analýzy je nutné vhodně reprezentovat a nabídnout uživatelům dobře interpretovatelné výsledky.

Existuje celá řada shlukovacích metod, které je možné využít pro shlukovou analýzu. Metody založené na rozdělování (mezi nejznámější patří *k-means* a *k-medoids*) vyžadují určení počtu tříd. Objekty jsou následně rozděleny do předem určeného počtu shluků a největší nevýhodou těchto metod je právě nutnost předem určit tento počet. Dalším problémem je schopnost hledání shluků různých tvarů a citlivost na šum. Uvedené problémy s různými tvary shluků a šumem umí lépe řešit metody založené na hustotě jako je např. *DBSCAN* nebo *DENCLUE*.

Zajímavou shlukovací metodou je metoda *WaveCluster* patřící do skupiny metod založených na mřížce. Tato metoda využívá tzv. vlnkovou transformaci, kdy dochází k transformaci původního prostoru dat. Zpracování dat metodou *WaveCluster* je velmi rychlé, metoda nevyžaduje určení vstupních parametrů, dobře hledá shluky různých tvarů, umí se vypořádat s odlehlými hodnotami a různým pořadím vstupních záznamů [22].

Kapitola 5

Data Českého statistického úřadu

Analytické úlohy využívající prostředky OLAP a dolování dat budou prováděny nad daty zahraničního obchodu ČR, která veřejně poskytuje Český statistický úřad (ČSÚ). V této kapitole nejprve představím webovou aplikaci ČSÚ zpřístupňující tato data. Poté se budu zabývat metodikami ČSÚ pro jejich zjišťování a popisem dat, která budou analyzována. Předmětem dalších částí kapitoly bude návrh a implementace aplikace provádějící automatizované stažení dat zahraničního obchodu do lokální databáze.

5.1 Webová aplikace ČSÚ

Data zahraničního obchodu jsou přístupná prostřednictvím webové aplikace Databáze zahraničního obchodu[1], kterou spravuje a vyvíjí ČSÚ. Webové rozhraní aplikace znázorňuje obrázek 5.1 a jedná se o jednu obrazovku, kde se zadávají parametry dotazu na data.

5.1.1 Funkce aplikace

Webová aplikace umožňuje uživatelům zobrazit data zahraničního obchodu za zvolené časové období. Uživatel může specifikovat konkrétní zboží, zemi a směr obchodu. Ve výsledném zobrazení má k dispozici data o celkové hodnotě a hmotnosti konkrétního zboží.

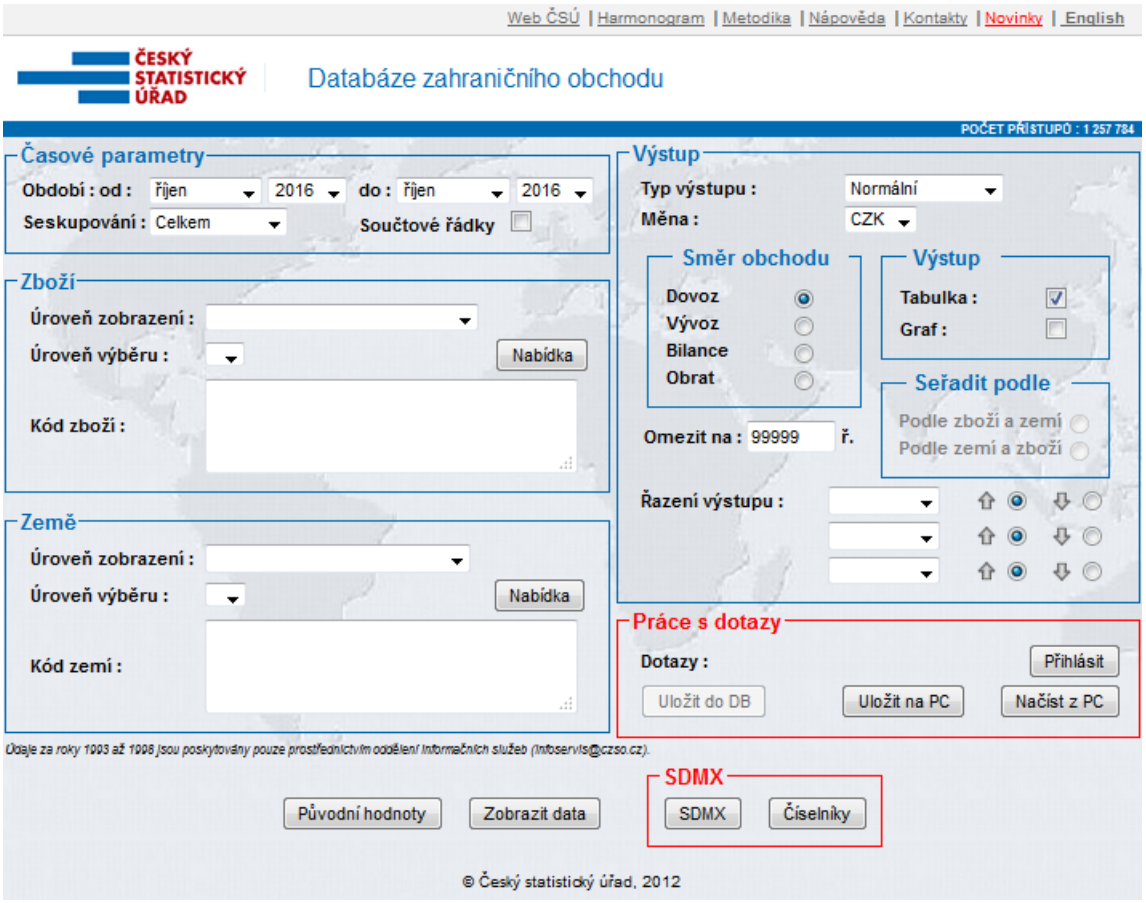
Nastavením parametrů pro získání dat zahraničního obchodu uživatel sestavuje dotaz na data, který lze uložit do vzdálené databáze nebo lokálně. Podle potřeby je možno načíst uložený dotaz na data zpět do webové aplikace a provést přímo zobrazení dat.

Funkce webové aplikace jsou popsány níže. Upřesňující popis pojmů, údajů a hodnot, se kterými aplikace pracuje, uvádím v kapitole o metodice.

Časové parametry

Pomocí časových parametrů se nastavuje začátek období, konec období a typ seskupování. Aplikace umožňuje zobrazení dat od roku 1999. Starší údaje z let 1993–1998 jsou dostupné pouze prostřednictvím Informačního servisu ČSÚ. Data lze zobrazit za celé uvedené období (parametr `Seskupování` s hodnotou *Celkem*) nebo lze celé období rozdělit podle různé úrovně detailu (po letech, po čtvrtletích nebo po měsících). Podle potřeby je možné aktivovat zobrazování součtových řádků, kdy se pro každou část období zobrazí součet hodnot.

[1] https://apl.czso.cz/pll/stazo/STAZO.STAZO



Obrázek 5.1: Webová aplikace Databáze zahraničního obchodu [1].

Zboží

Zboží je specifikováno numerickým kódem v rámci určité klasifikace zboží. Při nastavování parametrů pro zboží se rozlišuje mezi úrovní zobrazení a výběru.

Parametr `Úroveň zobrazení` určuje na jaké úrovni v rámci dané klasifikace bude zobrazeno zboží na výstupu. Parametr `Úroveň výběru` funguje podobně, a to tak, že nastavuje úroveň zobrazení zboží při výběru z nabídky ve webové aplikaci. Úroveň výběru lze nastavit vždy stejně nebo méně detailní než je úroveň zobrazení. Uživatel si po zvolení úrovně výběru pak vybírá z nabídky jeden nebo více kódů zboží.

Země

Parametry pro nastavení úrovně zobrazení a výběru jsou dostupné také při určování země. Zemi lze specifikovat podle konkrétního kódu země (hodnota *Země upravené pro statistiku*), jako skupinu zemí (hodnota *Bloky a seskupení zemí*) nebo nejobecněji pomocí kontinentu (hodnota *Kontinenty*). Mezi bloky a seskupení zemí patří části kontinentů (např. Severní Amerika) nebo ekonomická uskupení (např. země eurozóny, země EU, skupina zemí v rámci OECD).

Ostatní parametry

Parametr `Měna` nastavuje měnu, ve které bude uvedena hodnota zboží (*CZK*, *EUR* nebo *USD*). V části pro nastavení směru obchodu je možné specifikovat zobrazení zboží z dovozu nebo vývozu, případně zobrazit bilanci nebo obrat. Bilance udává rozdíl hodnot vyváženého a dováženého zboží. Obrat je součtem hodnot vyváženého a dováženého zboží. Pomocí bilance lze jednoduše zjistit, zdali se dané zboží ve zvoleném období více vyváželo než dováželo (kladná bilance) a naopak (záporná bilance).

Výstup z aplikace

Po nastavení požadovaných parametrů pro zobrazení dat zahraničního obchodu je možné data zobrazit ve formě tabulky nebo grafu, případně v obou formách v rámci jednoho okna. K dispozici je také možnost řazení hodnot na výstupu a omezení maximálního počtu hodnot. Kromě zobrazení dat v prohlížeči poskytuje aplikace export dat do formátu XLS[2] nebo SDMX[3].

Na obrázku 5.1 je ukázka výstupu s daty zahraničního obchodu pro dovoz ze zemí Evropské unie za měsíce září a říjen 2016. Přes kód zboží je možný proklik na nižší úroveň v hierarchii příslušné klasifikace zboží (pokud je nižší úroveň dostupná) a získat tak podrobnější informace.

Zahraniční obchod podle zboží a zemí

Typ výstupu :	Normální
Směr obchodu :	Dovoz
Období :	1.9.2016 – 31.10.2016
Nomenklatura zboží :	HS(2)
Data v tabulce jsou :	**s dopočty**

Období	Kód zboží	Název zboží	Kód země	Název země	Netto (kg)	Stat. hodnota CZK(tis.)
09/2016	01	Zvířata živá	AT	Rakousko	34	114
09/2016	01	Zvířata živá	BE	Belgie	15 404	1 071
10/2016	02	Maso a droby poživatelné	AT	Rakousko	2 133 717	118 926
10/2016	02	Maso a droby poživatelné	BE	Belgie	2 933 173	150 212

Obrázek 5.2: Výstup z webové aplikace *Databáze zahraničního obchodu* [1].

Parametrem `Typ výstupu` lze kromě běžného zobrazení, kdy se zobrazí hodnoty a hmotnosti pro zboží a země dle nastavených podmínek, nastavit zobrazení podílu z celku nebo meziroční index. Podíl z celku udává podíl každé položky na celkovém součtu (má smysl při zapnutém parametru `Součtové řádky`). Meziroční index udává podíl hodnot zboží pro dvě po sobě jdoucí období (uvažuje se období zadané v časových parametrech a totéž období minulého roku).

Část pro práci s dotazy umožňuje uložení a načtení dotazů pro nastavení parametrů webového rozhraní. Dotaz pro uložení na klientský počítač má podobu jednoho souboru

[2] Excel Binary File Format
[3] Statistical Data and Metadata eXchange

obsahujícího data v kódování Base64[4]. Ve skutečnosti se jedná o zakódovaný SDMX dotaz na statistická data. Uložení dotazu do databáze je možné po přihlášení uživatele přes emailovou adresu. Při pozdějším přihlášení jsou pak uživateli jeho dříve uložené dotazy nabídnuty ke spouštění.

5.1.2 Metodiky ČSÚ při zveřejňování údajů zahraničního obchodu

V následujícím přehledu uvádím způsoby, jakými jsou zjišťovány a prezentovány údaje poskytované webovou aplikací ČSÚ.

Zveřejňování a zdroj dat

Po každém měsíci, kdy jsou data sbírána, probíhá další měsíc zpracování dat a až poté jejich zveřejnění. Zveřejňování dat probíhá vždy začátkem měsíce (např. data za měsíc září jsou zveřejněna počátkem listopadu). Při každém zveřejnění dat se provádí také zpřesňování údajů za tři předcházející měsíce [1].

Zdrojem dat pro statistiku zahraničního obchodu jsou systémy Intrastat a Extrastat. Intrastat je systém statistiky zahraničního obchodu mezi státy Evropské unie, zatímco Extrastat sleduje zahraniční obchod se zeměmi, které nejsou členy Evropské unie [17].

Přeshraniční a národní pojetí zahraničního obchodu

Na zahraniční obchod ČR se lze dívat dvěma pohledy [17], [9]:

- **Přeshraniční pojetí** – Prezentuje webová aplikace ČSÚ. Zaměřuje se na fyzický přesun zboží přes hranice ČR a nerozlišuje jestli probíhá obchod mezi českým nebo zahraničním subjektem (nezabývá se změnou vlastnictví). Přeshraniční pojetí splňuje požadavky Eurostatu [5] a poskytuje mezinárodní srovnání. Přeshraniční pohled na zahraniční obchod je také vhodnější pro ekonomické subjekty, kterým může sloužit jako vhodný ukazatel vývoje zahraničního obchodu.
- **Národní pojetí** – Pohled národního pojetí poskytuje informace o výkonnosti české ekonomiky v oblasti zahraničního obchodu. Tento přístup rozlišuje, zda obchod probíhá skutečně mezi českým a zahraničním subjektem, kdy dochází ke změně vlastnictví.

Klasifikace zboží

Webová aplikace používá následující klasifikace zboží [1]:

- **Harmonizovaný systém popisu a číselného označování zboží (HS)** – Mezinárodní klasifikace zboží, která zboží značí numerickým kódem a dělí na několik úrovní: třídy, kapitoly (HS2), podkapitoly (HS4) a položky (HS6), přičemž číslo značí počet číslic v kódu.
- **Kombinovaná nomenklatura (KN8)** – Vychází z Harmonizovaného systému a je využívána všemi státy Evropské unie.
- **Klasifikace SITC**[6] – Jedná se o standardní mezinárodní klasifikaci zboží tvořenou pěti úrovněmi (SITC1 až SITC5).

[4] kódování Base64 reprezentuje binární data pomocí ASCII znaků, využívá se například v emailové komunikaci nebo při přenosu dat z HTML formulářů

[5] statistický úřad Evropské unie

[6] Standard International Trade Classification

Klasifikace zemí

Pro klasifikaci zemí využívá webová aplikace klasifikaci CZ-GEONOM, která je odvozená z evropské klasifikace GEONOM. GEONOM se skládá z geonomenklatury (názvy a dvoumístné kódy zemí světa), geografických zón (kontinenty a jejich části) a ekonomických zón (rozdělují svět podle ekonomických uskupení) [6].

Hodnota a hmotnost zboží

Hodnota zboží (statistická hodnota) je fakturovaná cena zboží spolu s dalšími náklady (např. pojistné a dopravné). Jako hmotnost zboží se uvádí čistá hmotnost. Údaje o hmotnosti zboží z let 2006-2008 webová aplikace neposkytuje, protože tyto údaje v daném období nebylo povinné vždy uvádět [1].

5.2 Data zahraničního obchodu pro analýzy

Data pro analýzu zahraničního obchodu budu získavat z více zdrojů prostřednictvím aplikace pro stažení dat, jejíž návrh a implementace bude předmětem dalšího textu této diplomové práce. Primárním zdrojem dat bude webová aplikace ČSÚ spolu s aktuálními číselníky zboží a zemí z webu ČSÚ. Dalším zdrojem budou data z portálu BusinessInfo.cz, který poskytuje informace o oborových příležitostech v zahraničním obchodu.

5.2.1 Data z webové aplikace ČSÚ

Z webové aplikace budu získávat následující údaje o zahraničním obchodu:

- rok a měsíc,
- kód země,
- kód zboží v klasifikaci HS(6),
- hodnotu zboží a hmotnost zboží,
- informaci, zdali se jednalo o vývoz nebo dovoz.

Export výstupu z webové aplikace jako souboru XLS je užitečný v případě, kdy se data dále zpracovávají například v aplikacích typu Microsoft Excel. Avšak při získávání dat pro analýzu budu využívat druhou možnost exportu z webové aplikace, a to do formátu SDMX, který je pro automatizované zpracování vhodnější, jelikož se jedná o data splňující specifikaci XML.

Formát SDMX

SDMX (Statistical Data and Metadata eXchange) je ISO standard určený k popisu statistických dat. První verze SDMX byla vydána v roce 2004, nejnovější verze 2.1 pak v roce 2011. SDMX popisuje statistická data a metadata, standardizuje jejich výměnu, umožňuje lepší sdílení dat napříč organizacemi zpracovávajícími statistické údaje a zvyšuje dostupnost statistických dat pro uživatele. Součástí standardu jsou i doporučení a postupy při zpracování a zvěřejňování statistických dat [7].

Mezi oblasti kde SDMX nachází největší uplatnění patří bankovní sektor (centrální banky) a národní statistické úřady, které nabízejí statistická data prostřednictvím svých webových stránek.

Standard SDMX specifikuje pro popis statistických dat a metadat datové struktury spolu s formáty pro výměnu těchto dat. Zabývá se také možnostmi uložení a čtení SDMX souborů z databází a způsoby dotazování na statistická data pomocí webových služeb. SDMX definuje pro výměnu statistických dat formáty SDMX-EDI a SDMX-ML. Formát SDMX-EDI je založen na formátu EDI[7], zatímco SDMX-ML je založen na XML a jedná se o formát, který používá webová aplikace ČSÚ jako jednu z možných variant exportu dat zahraničního obchodu [18].

Vzhledem k rozsáhlosti specifikace formátu SDMX se zaměřím pouze na popis struktury SDMX souboru, který poskytuje webová aplikace ČSÚ. Kompletní specifikaci SDMX a další informace je možné nalézt na oficiálních webových stránkách projektu SDMX[8].

Výsledkem exportu dat zahraničního obchodu z webové aplikace do formátu SDMX je datová sada (data set) v podobě souboru s příponou .sdmx a související metadata uložená v souboru .dsd (Data Structure Definition). Zkrácenou a zjednodušenou datovou sadu ukazuje kód 5.1.

Informace o vydavateli dat a času vydání dat poskytuje element *Header*. Data datové sady jsou obsažena v elementu *DataSet* a jsou rozdělena do dvou skupin, kde každá skupina je reprezentována elementem *Group*. Atribut `type` v elementu *Group* určuje zda hodnoty ve skupině udávají cenu (GRPZM) nebo hmotnost (GRPZH) zboží. V ukázce kódu 5.1 je pro názornost uvedena pouze skupina udávající cenu. Hodnoty dalších atributů elementu *Group* odpovídají parametrům, které byly zadávány při dotazu na data. Například informace o směru obchodu je zde specifikována atributem `FLOW` (dovoz), velikost časového intervalu, za který jsou data zobrazována, udává atribut `FREQ` (po měsících).

Element *Series* v ukázce obsahuje jednotlivé hodnoty o ceně (atribut `OBS_VALUE`). V případě skupiny GRPZH by se jednalo o hmotnosti. Kód zboží (podle zvolené klasifikace zboží a úrovně) a země jsou určeny atributy `PROD` a `COUNTRY`. Atribut `UNIT` udává jednotku, ve které se počítá hodnota atributu `OBS_VALUE`.

```
<compactData xmlns="http://www.SDMX.org/resources/SDMXML/schemas/v2_0/message">
 <Header>
  <ID>CZSO25482482</ID><Name xml:lang="cs">CZSO-FT</Name>
  <Prepared>2016-12-17T11:57:26+01:00</Prepared>
  <Sender id="CZSO"><Name xml:lang="cs">CSU</Name></Sender>
 </Header>
 <compact:DataSet dataProviderID="CZSO" dataflowID="CZSO25482482"
  reportingBeginDate="2016-09" reportingEndDate="2016-10">
  <compact:Group type="GRPZM" FREQ="M" DSET_TYPE="3" OUT_TYPE="1" FLOW="D"
   DATA_TYPE="M" PROD_NOM="CN8" COUNTRY_NOM="1" CURR="CZK" >
   <compact:Series PROD="01" COUNTRY="AT" UNIT="1000">
    <compact:Obs TIME_PERIOD="2016-09" OBS_VALUE="114" OBS_STATUS="A"/>
    <compact:Obs TIME_PERIOD="2016-10" OBS_VALUE="5503" OBS_STATUS="A"/>
   </compact:Series>
  </compact:Group>
 </compact:DataSet>
</compactData>
```

Ukázka kódu 5.1: Data zahraničního obchodu ve formátu SDMX-ML.

[7] Electronic Data Interchange - umožňuje výměnu nejčastěji obchodních dat dle národních nebo mezinárodních standardů

[8] https://sdmx.org/

5.2.2 Číselníky zboží a zemí

Kódy zemí a zboží (HS6) získané z webové aplikace budou sloužit k napojení na číselníky, ze kterých použiju další údaje o zemích a zboží.

Z číselníků klasifikace zemí CZ-GEONOM [9], které jsou dostupné na webu ČSÚ ve formátu .xlsx využiju následující informace:

- **geografické informace** - Název země, kontinent a oblast kontinentu kde země leží.
- **informace o ekonomických zónách** - Informace o tom, zda je země členem EU, eurozóny, OECD[10](Organizace pro hospodářskou spolupráci a rozvoj), ASEAN[11] (Sdružení národů jihovýchodní Asie), APEC[12](Asijsko-pacifické hospodářské společenství), OPEC[13](Organizace zemí vyvážejících ropu) nebo LDC[14](Nejméně rozvinuté země).

Ke klasifikaci zboží budu používat Harmonizovaný systém popisu a číselného označování zboží, který bude zdrojem těchto údajů:

- **úroveň 1** – Kód a název třídy zboží.
- **úroveň 2 (HS2)** – Kód a název kapitoly zboží (místo pojmu *kapitola* budu dále používat pojem *kategorie*). Odpovídá nejvyšší úrovni klasifikace HS používané ve webové aplikaci ČSÚ.
- **úroveň 3 (HS4)** – Kód a název podkapitoly zboží (místo pojmu *podkapitola* budu dále používat pojem *subkategorie*).
- **úroveň 4 (HS6)** – Název položky zboží.

Číselník klasifikace zboží je k dispozici na webu ČSÚ[15] ve formátu XML. Strukturu, ve které jsou uloženy údaje o zboží, znázorňuje ukázka XML kódu 5.2 (v ukázce byly ponechány pouze elementy důležité z hlediska dalšího zpracování).

Element *POLOZKA* specifikuje informace o zboží na určité úrovni klasifikace, která je určena hodnotou atributu `uroven`. Kód zboží na příslušné úrovni specifikuje element *CHODNOTA* a názvy zboží podle úrovně jsou určeny elementem *TEXT*.

[9] https://www.czso.cz/csu/czso/klasifikace_zemi_-cz_geonom-
[10] Organisation for Economic Co-operation and Development
[11] Association of South East Asian Nations
[12] Asia-Pacific Economic Cooperation
[13] Organization of the Petroleum Exporting Countries
[14] Least developed countries
[15] http://apl.czso.cz/iSMS/klasdata.jsp?kodcis=80034

```
<DATA>
  <POLOZKA uroven="1">
    <CHODNOTA>01</CHODNOTA>
    <TEXT>Ziva zvirata, zivocisne produkty</TEXT>
    <POLOZKA uroven="2">
      <CHODNOTA>01</CHODNOTA>
      <TEXT>Zvirata ziva</TEXT>
      <POLOZKA uroven="3">
        <CHODNOTA>0101</CHODNOTA>
        <TEXT>Kone osli muli mezci zivi</TEXT>
        <POLOZKA uroven="4">
          <CHODNOTA>010121</CHODNOTA>
          <TEXT>Plemenni cistokrevni kone, zivi</TEXT>
        </POLOZKA>
      </POLOZKA>
    </POLOZKA>
  </POLOZKA>
</DATA>
```

Ukázka kódu 5.2: Data o zboží z číselníku klasifikace zboží dle HS.

5.2.3 Doplňující data oborových příležitostí

Na oficiálním portálu pro podnikání a export BusinessInfo.cz [16] agentury CzechTrade je k dispozici v sekci zahraničního obchodu Mapa oborových příležitostí [17]. Mapu oborových příležitostí sestavuje ve spolupráci s agenturou CzechTrade také Ministerstvo zahraničních věcí a Ministerstvo průmyslu a obchodu [8].

Mapa poskytuje přehled exportních příležitostí pro české podniky na zahraničním trhu a umožňuje vyhledávání příležitostí podle země, oboru, nebo dle kódu zboží HS(4). Na obrázku 5.3 je příklad zobrazení několika oborových příležitostí v civilním leteckém průmyslu.

Perspektivní obor	Země	Konkrétní příležitosti
Civilní letecký průmysl	Francie	HS 8411 - Motory proudové, pohony turbovrtulové a ostatní
	Maroko	HS 3917 - Trouby trubky hadice příslušenství z plastů
	Spojené státy americké	HS 8409 – Části součásti pro motory pístové
	Afghánistán	HS 8411 - Motory proudové, pohony turbovrtulové a ostatní

Obrázek 5.3: Mapa oborových příležitostí pro civilní leteckých průmysl [8].

Z mapy oborových příležitostí budu využívat následující údaje:

- název oboru,
- název země,
- kód zboží v klasifikaci HS(4).

[16] http://www.businessinfo.cz/
[17] http://www.businessinfo.cz/cs/zahranicni-obchod-eu/mapa-oborovych-prilezitosti.html

Kapitola 6

Analýza dat zahraničního obchodu

Prostudování webové aplikace ČSÚ a dat zahraničního obchodu bylo základem pro další fázi této diplomové práce, kde se zabývám nástroji, které budu používat k analýze dat, stažením dat a analytickými úlohami. Uložení dat a analýza dat bude prováděna na platformě Microsoft SQL Server 2012, jež poskytuje technologii pro uložení dat a analytické nástroje Business Intelligence.

Konkrétní implementaci stažení dat a vybraných analytických úloh se budu věnovat v sedmé kapitole zaměřené na realizaci vlastní BI aplikace.

6.1 Business Intelligence v Microsoft SQL Serveru

V oblasti Business Intelligence nabízí Microsoft SQL Server 2012 podobně jako jeho předchozí vydání funkce pro integraci, analýzu a reportování dat. Platforma Microsoft SQL Server dodává pro tyto účely následující služby [14]:

- **Integrační služby (SSIS)** – Slouží pro provádění ETL procesů a další úkony spojené s integrací dat.
- **Analytické služby (SSAS)** – Poskytují prostředky pro analýzu dat, jako je OLAP nebo dolování dat. Tyto služby budu využívat během řešení této diplomové práce.
- **Reportovací služby (SSRS)** – Jsou určeny pro vytváření a distribuci podnikových reportů nad různými zdroji dat.

Instance SSAS může obsahovat více analytických databází, přičemž na daném serveru může existovat i více instancí SSAS. Jedna analytická databáze v sobě zahrnuje objekty pro analýzu OLAP (dimenze, kostky) dohromady s objekty pro dolování dat (dolovací struktury a modely). Níže uvádím základní objekty, které se při vývoji BI řešení založeného na SSAS nacházejí v každé analytické databázi bez ohledu na druh analýzy:

- **Datový zdroj** – Reprezentuje spojení s datovým zdrojem obsahujícím zdrojová data pro analýzu (takovým zdrojem může být např. relační databáze datového skladu).
- **Pohled na datový zdroj** – Představuje pohled na data datového zdroje, která budou použita k tvorbě konkrétních analytických modelů. Příkladem pohledu na datový zdroj mohou být data z tabulek datového zdroje týkající se určitého tématu, např. prodeje výrobků nebo zákazníků.

Pro tvorbu Business Intelligence řešení poskytuje společnost Microsoft vývojové prostředí SQL Server Data Tools (SSDT). Tvorba projektu v SSDT začíná definicí datových

zdrojů a pohledů na datové zdroje. Poté probíhá definice dimenzí a OLAP kostek nebo definice dolovacích struktur a modelů. Následně se provede nasazení celého řešení na SSAS a poté je možné provádět samotnou analýzu dat (prohlížení OLAP kostek, dolovacích modelů a provádění predikcí nad modely).

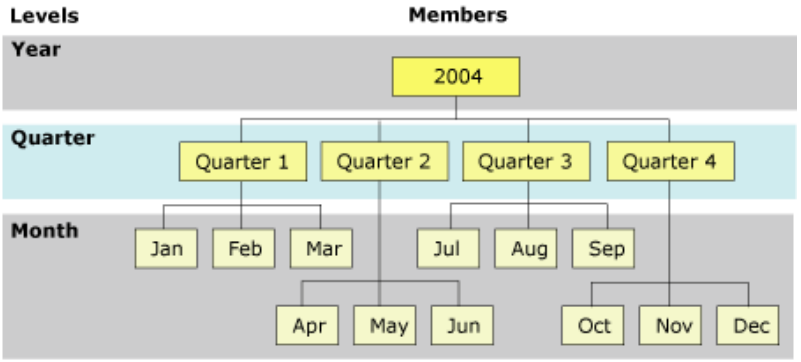
Při řešení této diplomové práce využívám kromě vývojového prostředí SSDT také nástroj SQL Server Management Studio (SSMS). SSMS umožňuje spravovat databáze SQL Serveru, provádět dotazy a nabízí i část funkcí ze SSDT, zejména prohlížení analytických modelů.

K instanci SSAS se lze připojit i pomocí aplikace Excel a využít možnosti prohlížení OLAP kostek s pomocí kontingenční tabulky (pivot table). Kromě prohlížení analytických struktur v nástrojích SSDT nebo SSMS je možné se dotazovat do analytické databáze SSAS přímo prostřednictvím speciálních jazyků MDX (pro OLAP) nebo DMX (pro dolování dat), a mít tak plnou kontrolu nad celým procesem získání dat z vytvořených analytických modelů.

Ve zbývající části této kapitoly se budu podrobněji věnovat oblastem analýzy OLAP a dolování dat, tak jak jsou implementovány v rámci platformy Microsoft SQL Server.

6.1.1 OLAP v SSAS

Platforma Microsoft SQL Server nabízí pro OLAP uložení dat založené na technologii ROLAP, MOLAP nebo HOLAP [11]. Podkladem pro tvorbu OLAP kostek jsou tabulky obsažené v datovém pohledu. OLAP kostka je v SSAS definována měřítky a dimenzemi. Měřítka jsou organizována do skupin měřítek (measure groups). Objekt dimenze je složen z atributů a hierarchií. Hodnoty atributů se nazývají členy (members). V případě, že jsou atributy organizovány do hierarchie, označují se takové atributy jako úrovně (levels). Příklad uživatelsky definované hierarchie ukazuje obrázek 6.1.



Obrázek 6.1: Uživatelsky definovaná hierarchie pro časovou dimenzi skládající se z atributů pro rok, kvartál a měsíc [16].

Poté co je OLAP kostka nasazena na server pod správu SSAS, je od této chvíle k dispozici pro provádění OLAP operací prostřednictvím nástrojů pro prohlížení OLAP kostek.

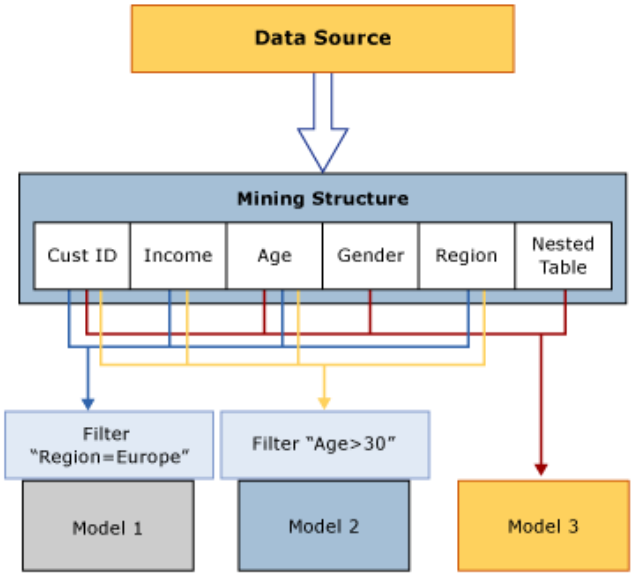
V případě, že se změní zdrojová data nebo struktura některého objektu v analytické databázi, je nutné provést tzv. zpracování (processing), jehož výsledkem jsou aktualizovaná data nebo struktura OLAP kostky. SSAS umožňuje provádět zpracování buď všech objektů v analytické databázi (např. celé kostky, včetně dimenzí) nebo jen některých objektů. První zpracování OLAP kostky se provádí při nasazení projektu vytvořeného v SSDT na instanci SSAS. Typicky je nutné provést zpracování OLAP kostky pokaždé, kdy je potřeba aktualizovat data, která kostka ukládá [12].

6.1.2 Dolování dat v SSAS

Pro dolování dat implementuje Microsoft SQL Server dolovací algoritmy pokrývající oblast klasifikace, získávání asociačních pravidel, shlukování a časových řad.

Klíčovými objekty v SSAS pro analýzu založenou na dolování dat je dolovací struktura (mining structure) a dolovací model (mining model). Nad jednou dolovací strukturou lze definovat více dolovacích modelů, které obsahují např. určitou podmnožinou sloupců dolovací struktury nebo data omezená jistou podmínkou. Vztah mezi dolovací strukturou a modelem znázorňuje obrázek 6.2.

Dolovací struktura definuje data, na jejichž základě budou vytvářeny dolovací modely. Ne všechny sloupce dolovacího modelu musí být určeny pro dolovací algoritmus. Takové sloupce mohou sloužit jako zdroj dalších informací o datech, jež byla použita pro tvorbu dolovacích modelů a jsou užitečně v případě, kdy je na dolovacím modelu aktivována funkce drillthrough. Například v případě modelu založeného na dolovacím algoritmu získávání asociační pravidel, je možné pomocí funkce drillthrough zobrazit konkrétní případy z množiny zdrojových dat, které byly použity algoritmem pro tvorbu zvoleného asociačního pravidla.



Obrázek 6.2: Ukázka vztahu mezi dolovací strukturou a modelem. V tomto případě jsou na základě dolovací struktury, která ukládá data o zákaznících, vytvořeny tři dolovací modely. Na první dva modely byla aplikovaná filtrace podle konkrétní hodnoty atributu [10].

Dolovací struktura může obsahovat tzv. nested tables. Nested table reprezentuje vztah 1:N a její použití je výhodné ve chvíli, kdy se např. data o objednávce (hlavička) a samotné položky objednávky (řádky) nacházejí v různých tabulkách. Pak je možné informace o objednávce a jednotlivých položkách spojit do jednoho případu a tento případ použít pro učení modelu [10].

Podobně jako je nutné provádět zpracování OLAP kostek, provádí se tento proces i pro dolovací struktury a modely. Zpracování je potřeba před prohlížením obsahu modelů nebo při změnách v dolovací struktuře, např. pokud je do dolovací struktury přidán další model nebo nové sloupce.

Při zpracování modelu se získají data z dolovací struktury, aplikuje se dolovací algoritmus a poté jsou nalezené vzory, pravidla a statistiky o datech k dispozici prostřednictvím dolovacího modelu. Prohlížet obsah modelu nebo analyzovat informace o přesnosti vytvořeného modelu je možné přímo v prostředí SSDT nebo SSMS.

6.2 Stažení dat ČSÚ a specifikace analytických úloh

Řešení problému stažení dat ČSÚ a jejich uložení budu nyní prezentovat nejprve z pohledu dat a analytických úloh. Popis implementace stažení dat bude předmětem sedmé kapitoly této práce.

6.2.1 Princip stažení a způsob uložení dat

Pro ukládání dat zahraničního obchodu a dalších souvisejících dat (číselníky a informace o oborových příležitostech) jsem navrhl centrální databázi *CSU_DATA* (viz obrázek 6.3), která slouží jako datový sklad.

V centrální databázi jsou data organizována okolo tabulek *Trades* a *BranchOppors*. Tabulka *Trades* obsahuje údaje o provedených obchodech (exportu a importu) v rámci zahraničního obchodu ČR. Tabulka *BranchOppors* ukládá data oborových příležitostí z webu BusinessInfo.cz. Obě tabulky jsou na sobě nezávislé a sdílejí data z tabulek uchovávajících data z číselníků zemí a zboží.

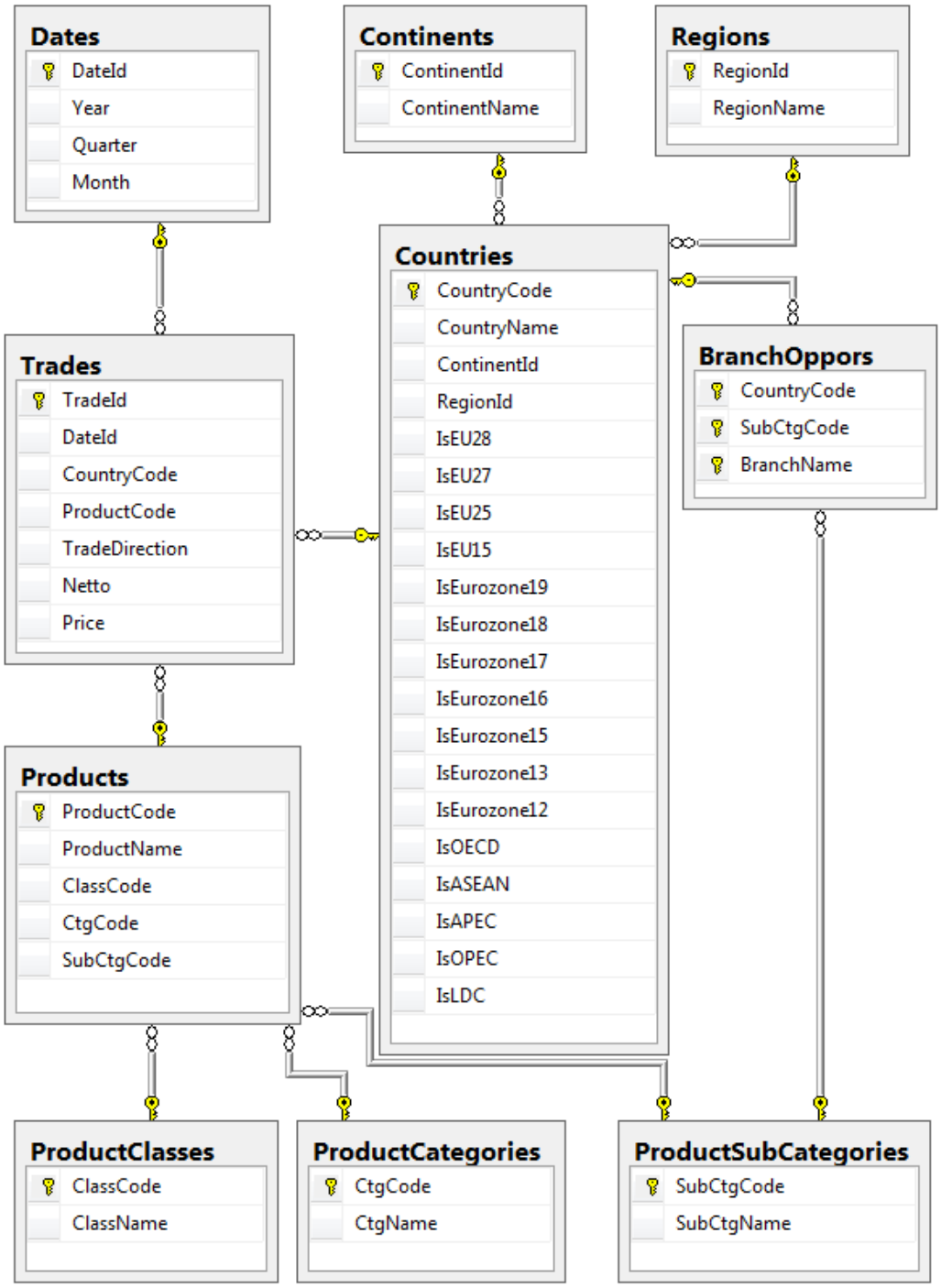
Po importu dat z číselníků zemí a zboží do centrální databáze se provádí stažení dat zahraničního obchodu z webové aplikace ČSÚ nebo stažení dat oborových příležitostí z webu BusinessInfo.cz. Po zveřejnění nových dat zahraničního obchodu je možné centrální databázi aktualizovat daty z webové aplikace ČSÚ. Údaje o oborových příležitostech bohužel neobsahují časový údaj o jejich platnosti. Pro potřeby této diplomové práce pracuji s předpokladem, že na webu BusinessInfo.cz jsou vždy aktuálně platná data.

Data v datovém skladu jsou záměrně organizována obecně bez vazby na potřeby konkrétní analytické úlohy. Datový sklad je v tomto případě výchozím bodem, který poskytuje data pro další zpracování dle požadavků a účelu dané analýzy.

6.2.2 Návrh analytických úloh

Z povahy dat poskytovaných webovou aplikací ČSÚ a dat z číselníků ČSÚ je vidět hierarchické uspořádání údajů. Data zahraničního obchodu v sobě zahrnují informace o čase, lokaci a produktu. Na základě těchto charakteristik jsem navrhl jako první analytickou úlohu analýzu OLAP. Cílem této úlohy bude definice multidimenzionálního pohledu na data ve formě OLAP kostek pro údaje o exportu a importu zboží v zahraničním obchodu ČR.

Data o oborových příležitostech pro export jsem již analyzoval v zimním semestru v rámci předmětu Získávání znalostí z databází. Konkrétně se jednalo o analýzu dolovací technikou klasifikace založenou na rozhodovacích stromech v prostředí nástroje RapidMiner. Jako druhou analytickou úlohu jsem se rozhodl zvolit tuto klasifikaci v prostředí Microsoft SQL Serveru, jelikož množství dat zpracovatelných nástrojem RapidMiner je v základní verzi z licenčních důvodů omezeno. Cílem úlohy bude určení, zdali lze data oborových příležitostí z webu BusinessInfo.cz použít jako podporu pro rozhodnutí o tom, jestli je určité zboží v dané zemi příležitostí pro export.



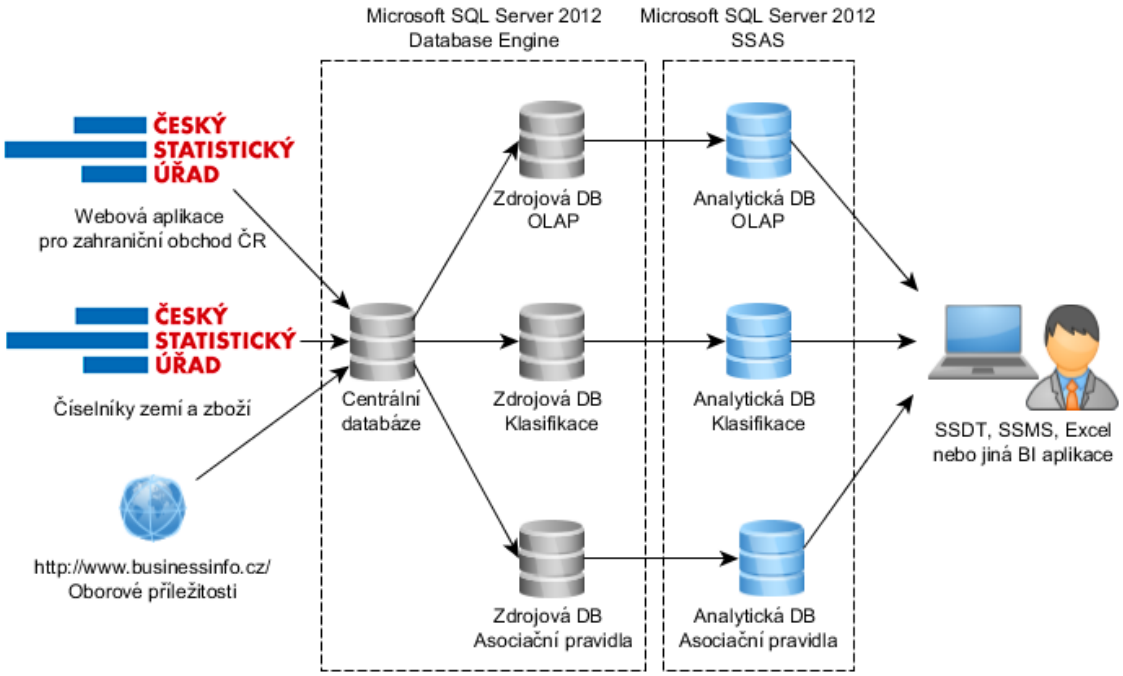
Obrázek 6.3: Logické schéma tabulek centrální databáze pro uložení dat zahraničního obchodu ČR.

Jako třetí analytickou úlohu jsem navrhl dolování dat pomocí metody získávání asociačních pravidel. Během řešení úlohy se budu snažit zjistit možnosti získání a použitelnosti asociačních pravidel z dat zahraničního obchodu ČR. Nalezená asociační pravidla mohou pomoci nalézt skryté vztahy mezi exportovaným zbožím a ulehčit potenciálním podnikovým uživatelům určení variant zboží, které se exportují často společně. Tyto znalosti o zboží by

mohly sloužit jako podpora pro rozhodnutí o rozšíření zahraniční obchodní nabídky podniku ve formě prodeje dalších skupin produktů.

Navržené analytické úlohy výrazně rozšiřují možnosti analýzy dat zahraničního obchodu, kdy uživatel není omezen jen na základní výstupy z webové aplikace ČSÚ. Dle informací od pana Mgr. Netolického z ČSÚ (ze dne 22.9. 2016) nepoužívá ČSÚ k analýze dat zahraničního obchodu žádnou pokročilou technikou BI, jako je OLAP nebo dolování dat.

Každá analytická úloha bude využívat vlastní relační databázi obsahující předzpracovaná zdrojová data a vlastní analytickou databázi, kde budou uloženy analytické objekty (OLAP kostky a dolovací modely). Analytické databáze budou spravovány v rámci jedné instance SSAS. Výslednou architekturu celého řešení ilustruje obrázek 6.4.



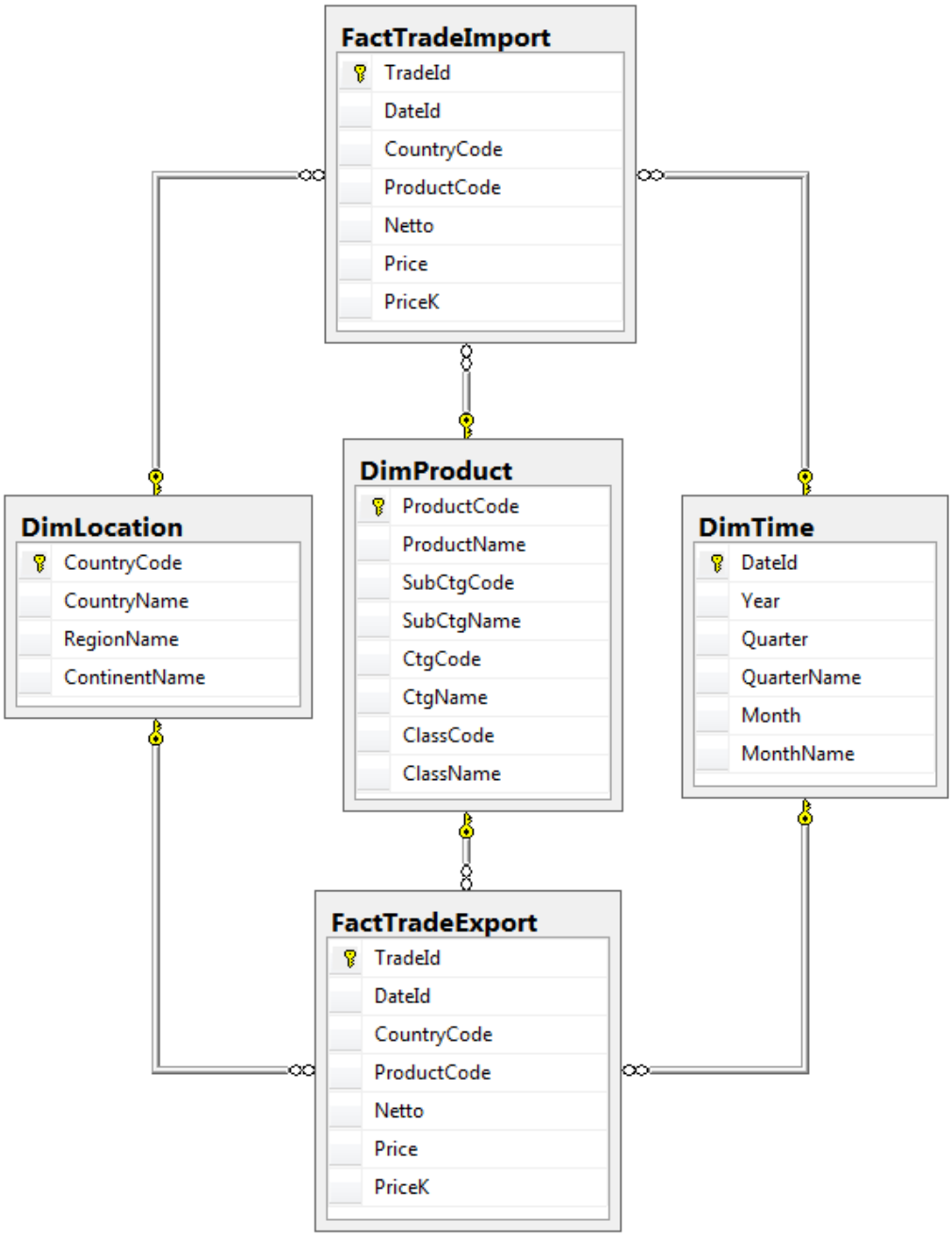
Obrázek 6.4: Architektura BI řešení pro analýzu dat zahraničního obchodu ČR na platformě Microsoft SQL Server.

Analýzu dat zahraničního obchodu jsem rozdělil do dvou fází. První fáze, která je předmětem následujícího textu, je věnována přípravě prostředí pro analytické úlohy a základnímu otestování funkčnosti vytvořených analytických struktur a navržených úloh pomocí BI nástrojů Microsoft SQL Serveru. Druhá fáze analýzy bude probíhat po dokončení implementace vlastní BI aplikace pro analýzu dat zahraničního obchodu, kdy budu ověřovat funkčnost implementovaného řešení.

Analýzu pomocí dolovací techniky klasifikace budu po dohodě s vedoucím práce provádět pouze prostřednictvím BI nástrojů platformy Microsoft SQL Server. Pro implementaci do vlastní BI aplikace jsem vybral zbývající dvě navržené analytické úlohy, tj. analýzu OLAP a dolování dat založené na získávání asociačních pravidel.

6.3 Analýza OLAP

Pro zdrojová data OLAP kostek jsem založil databázi *CSU_DATA_OLAP* obsahující tabulky faktů a dimenzí ve schématu hvězdy, tak jak ukazuje obrázek 6.5. Zdrojem dat pro databázi *CSU_DATA_OLAP* je centrální databáze *CSU_DATA*.



Obrázek 6.5: Logické schéma tabulek databáze pro tvorbu OLAP kostek.

V prostředí SSDT jsem vytvořil projekt pro OLAP analýzu, kde jsem definoval OLAP kostky, a po nasazení projektu na server analytických služeb jsou tyto OLAP kostky k dispozici pro provádění OLAP operací.

Data zahraničního obchodu se ukládají do dvou OLAP kostek podle toho, zdali se jedná o údaje o vývozu nebo dovozu zboží. Jako metodu pro uložení OLAP kostek jsem ponechal výchozí nastavení MOLAP. Plnění kostek realizuje SQL skript, který pro specifikované období provede nakopírování dat z tabulek centrální databáze *CSU_DATA* do databáze *CSU_DATA_OLAP*. Pro představu o objemu dat uvedu pro zajímavost, že pro stažená data zahraničního obchodu za roky 1999 - 2016 obsahuje tabulka *Trades* centrální databáze pro dovoz zhruba 11 mil. záznamů a pro vývoz přibližně 10,5 mil. záznamů.

OLAP kostka obsahuje celkem tři dimenze, z nichž každá obsahuje vždy dvě hierarchie. Hierarchie se liší ve formátu zobrazování hodnot atributu na dané úrovni. Podle typu atributu je možné vyjádření v číselné nebo textové formě.

- **Dimenze pro čas (DimTime)** – Pro časovovou dimenzi jsem definoval hierarchie Rok > Čtvrtletí > Měsíc. Hierarchie se liší v zobrazování údajů o měsíci, kdy je možné v jedné hierarchii zobrazit měsíc jako číselný údaj a v druhé hierarchii jeho název.
- **Dimenze pro lokaci (DimLocation)** – Dimenze pro lokaci obsahuje hierarchie Kontinent > Region > Země. K dispozici je možnost volby zobrazení kódu země nebo celého názvu země.
- **Dimenze pro produkt (DimProduct)** – Produktová dimenze obsahuje hierarchie Třída > Kategorie > Subkategorie > Produkt. Hierarchie produktové dimenze se odlišují ve všech úrovních. K dispozici je zobrazení kódů zboží (skupin zboží) nebo celých názvů.

V OLAP kostce jsou uloženy měrné jednotky pro cenu zboží v Kč (*Price*), cenu zboží v tisících Kč (*Price K*) a váhu zboží v Kg (*Netto*).

SSDT jsem v této fázi vývoje použil k definici dimenzí, OLAP kostek a k nasazení na analytický server. Funkčnost řešení jsem ověřil nejen v SSDT, ale otestoval jsem i prohlížení OLAP kostky připojené k aplikaci Excel. V rámci řešení vlastní BI aplikace budu implementovat klienta pro analýzu OLAP. OLAP klient bude umožňovat provádění OLAP operací nad vytvořenými kostkami, plnění OLAP kostek daty podle zvolených kritérií a vizualizaci výsledků multidimenzionální analýzy.

6.4 Získávání znalostí z databází

Pro dolovací úlohy klasifikace a získávání asociačních pravidel používám dvě oddělené zdrojové databáze čerpající data z centrální databáze *CSU_DATA*. Dle návrhu analytických úloh budu z algoritmů pro dolování používat konkrétně algoritmy *Microsoft Decision Trees algorithm* (klasifikace založená na rozhodovacích stromech) a *Microsoft Association algorithm* (asociační pravidla).

6.4.1 Klasifikace

Zdrojová data pro dolovací úlohu klasifikace ukládá databáze *CSU_DATA_DM_CLASS*. Tato databáze obsahuje jednu tabulku *ClassificationTrades*. Tabulka *ClassificationTrades* obsahuje všechny údaje o exportu v zahraničím obchodě (informace o zemích a zboží) za určité období a klíčový sloupec *IsBranchOppor* určený pro predikci, zdali se v případě daného zboží jedná o oborovou příležitost. Tabulka *ClassificationTrades* je plněna SQL skriptem z centrální databáze.

Pro klasifikaci jsem vytvořil samostatný projekt v SSDT s jednou dolovací strukturou a jedním dolovacím modelem (*ClassificationTradesModel*) pro klasifikaci založenou na rozhodovacích stromech. Jako zdrojová data jsem použil data o exportu zboží za rok 2016 pro všechny kontinenty (celkem přibližně 865 000 případů). Z uvedeného celkového počtu případů bylo 6,8 % případů oborovou příležitostí. Pro trénování modelu jsem vyčlenil 70 % zdrojových dat, zbývajících 30 % bylo určeno pro testování. Při tvorbě modelu určil nástroj SSDT následující atributy jako ty, které mají nejvyšší rozhodovací schopnost:

- kód a název kategorie zboží (*CtgCode*, *CtgName*),
- kód a název třídy zboží (*ClassCode*, *ClassName*),
- název regionu (*RegionName*),
- název kontinentu (*ContinentName*).

Projekt s definicí modelu jsem poté předal instanci SSAS, která jej spravuje v analytické databázi *CSU_DM_CLASSIFICATION*. Po zpracování modelu byl vytvořen velmi rozsáhlý rozhodovací strom, který zde nebudu uvádět, ale zaměřím se na vyhodnocení přesnosti vytvořeného modelu. Pro vyhodnocení přesnosti klasifikačního modelu poskytuje SSDT matici záměn (viz obrázek 6.6).

Predicted	True (Actual)	False (Actual)
True	4289	2505
False	12272	240608

Obrázek 6.6: Matice záměn klasifikačního modelu pro určení oborové příležitosti.

Klasifikační model provedl správnou predikci klasifikace zboží jako oborové příležitosti v 63 % případů. Správná predikce klasifikace zboží jako neoborové příležitosti pak proběhla v 95 % případů. Při změně počtu případů pro trénování modelu na 50 % pak zůstala přesnost modelu přibližně stejná.

Z přesnosti klasifikačního modelu je jasně vidět vliv malého počtu případů oborových příležitostí ze vstupní datové sady. Otázkou je reálná použitelnost klasifikačního modelu, která závisí na aktuální situaci v daném oboru a zemi. Vytvořený klasifikační model by mohl sloužit jako jeden ze zdrojů pro velmi hrubý přehled možností exportu zboží.

6.4.2 Asociační pravidla

Zdrojová data pro dolovací úlohu založenou na asociačních pravidlech jsou uložena v databázi *CSU_DATA_DM*. Tato databáze obsahuje dvě tabulky:

- **AssocRulesTradesHeaders** – Představuje hlavičky provedených obchodů (transakcí) při exportu zboží. Každý obchod uvedený v tabulce *AssocRulesTradesHeaders* obsahuje informace o lokaci (země, region a kontinent), kam byl export proveden, a je jednoznačně určen unikátním identifikátorem obchodu *TradeId*. *TradeId* se skládá z roku, kvartálu a měsíce, kdy byl proveden export a dvoumístného kódu země. Např. identifikátor *20020101PL* označuje export zboží do Polska v měsíci leden, v prvním kvartálu roku 2002.
- **AssocRulesTradesLines** – Ukládá řádky (položky) provedeného obchodu s informacemi o konkrétním exportovaném zboží (třída, kategorie, subkategorie, produkt).

Na získávání asociačních pravidel z dat zahraničního obchodu se lze dívat jako na dobře známou analýzu nákupního košíku ve větším měřítku. Místo hledání zboží nakupovaného zákazníky dohromady, se nyní hledá zboží, které je exportováno společně na úrovni států.

Pro asociační pravidla jsem založil projekt v SSDT, který je spravován instancí SSAS v analytické databázi *CSU_DM*. V rámci testování základní funkčnosti řešení jsem vytvořil dolovací strukturu nad daty zahraničního obchodu za rok 2015 a spustil dolovací algoritmus. Výsledný dolovací model pro testovací data obsahoval množství asociačních pravidel, která nástroj SSDT zobrazil formou kódů zboží, které byly použity pro učení modelu.

Pro analýzu je z pohledu uživatele vhodnější zobrazení pravidel ve formě názvů zboží a také pohodlnější způsob tvorby dolovacích modelů. Další experimenty s asociačními pravidly budu proto provádět s vlastním klientem pro dolování dat. Tento klient bude poskytovat uživateli jednoduché rozhraní pro analýzu bez nutnosti používat nástroj SSDT určený především k tvorbě projektů a nasazení na analytický server.

Kapitola 7

Implementace BI aplikace pro provádění analytických úloh

Nyní jsou stažená data zahraničního obchodu ČR a vytvořené analytické modely přístupné celé řadě aplikací. Zmiňoval jsem možnost analýzy v nástrojích SSDT nebo SSMS. Tyto nástroje jsou vhodné především pro vývoj a správu vytvořených řešení, přestože nabízejí prostředky pro analýzu dat. Koncový uživatel (řídící pracovník, manažer) ocení spíše nástroj, který mu umožní jednoduše provádět samotnou analýzu a interpretovat její výsledky vhodnou formou. Jednou z možností je využití analytických funkcí aplikace Excel nebo jiného BI nástroje dostupného na trhu.

V rámci této diplomové práce jsem pro analýzu dat zahraničního obchodu ČR implementoval BI aplikaci, která umožňuje na jednom místě spravovat data pro analýzy a provádět analytické úlohy založené na analýze OLAP a dolování dat.

7.1 Technologie pro vývoj

V této kapitole se stručně zmíním o vývojovém prostředí a knihovnách použitých během implementace, zbývající část kapitoly je věnována dotazování v jazycích MDX a DMX, které hrály klíčovovu roli při vývoji klientů pro analytické úlohy.

7.1.1 Aplikace

Pro implementaci desktopové BI aplikace s názvem FTA (Foreign Trade Analysis) jsem zvolil platformu .NET a programovací jazyk C#. Uživatelské rozhraní aplikace je založeno na frameworku WPF[1], který tvoří část platformy .NET. Při vývoji základní struktury aplikace jsem se částečně řídil principy návrhového vzoru MVVM.

Základní myšlenkou vzoru MVVM[2] je oddělit uživatelské rozhraní od business logiky aplikace. Takový přístup pak vede k lepší testovatelnosti aplikace a umožňuje oddělený vývoj uživatelského rozhraní. MVVM rozděluje aplikaci na tři části [5]:

- **Model** – Obsahuje třídy pro jednotlivé business entity, jejichž reprezentace je nezávislá na uživatelském rozhraní.
- **View** – Představuje uživatelské rozhraní aplikace a zajišťuje interakci aplikace s uživatelem.

[1] Windows Presentation Foundation
[2] Model-View-ViewModel

- **ViewModel** – Nachází se mezi View a Modelem. ViewModel dodává pro View data obsažená v modelech a oznamuje View změny v datech. Naopak aktualizuje obsah modelů na základě akcí provedených uživatelem prostřenictvím View.

Vývoj aplikace probíhal v prostředí Microsoft Visual Studio 2012 (edice Professional). Databázová část aplikace je postavena na platformě Microsoft SQL Server 2012 (edice Enterprise), kterou jsem při vývoji provozoval lokálně. Aplikace při správě analytických objektů a analýze komunikuje s instancí SSAS prostřednictvím knihoven ADOMD.NET[3] a AMO[4].

ADOMD.NET je .NET knihovna pro dotazování a provádění příkazů nad analytickou databází v SSAS. Jedná se o rozšíření knihovny ADO.NET pro přístup k datům v relačních databázích. ADOMD.NET zajišťuje zasílání příkazů v jazycích MDX a DMX instanci SSAS a předání výsledků dotazu zpět klientské aplikaci. Knihovna ADOMD.NET umožňuje také získat metadata o objektech analytické databáze, jako jsou informace o OLAP kostkách, dimenzích a dolovacích modelech [3].

Knihovna AMO je určena pro správu objektů instance SSAS. Poskytuje prostředky pro programové vytváření analytických objektů a jejich zpracování (processing) [4]. Knihovnu AMO využívám pro zpracování OLAP kostek při aktualizaci dat. OLAP kostky a dimenze, které jsem vytvořil v prostředí SSDT, je možné vytvořit přímo pomocí této knihovny, ale jedná se o poměrně zdlouhavý proces. Dynamické vytváření OLAP kostek a dimenzí implementovaná BI aplikace nevyžaduje.

7.1.2 Jazyk MDX

MDX[5] je dotazovací jazyk sloužící pro dotazování na data uložená v OLAP kostce. Syntaxe jazyka MDX je podobná syntaxi jazyka SQL pro dotazování nad relační databází [13].

Dotaz v jazyce MDX se skládá z následujících částí [15]:

- **SELECT** – V části příkazu `SELECT` probíhá definice os, pro které se dotaz provádí. Lze definovat až 128 os, ale používat budu vždy maximálně první dvě osy, které se označují aliasy *COLUMNS* a *ROWS* (místo aliasu os lze používat čísla os). Pro každou osu se definují členy (Members), n-tice (Tuples) nebo množiny ntic (Sets), jež mají být na ose umístěny. Member je hodnota atributu v dané hierarchii. Tuple se uzavírá do kulatých závorek a obsahuje jeden nebo více členů z různých hierarchií. Set se uzavírá do složených závorek a skládá se z jedné nebo více n-tic. Na osu je možné umístit i měrnou jednotku.
- **FROM** – Specifikuje zdroj dat pro MDX dotaz. Uvést je možné název pouze jedné OLAP kostky.
- **WHERE** – Příkaz `WHERE` plní funkci další osy, pro niž se užívá název *slicer*. Tato osa není ve výsledku dotazu zobrazena a slouží pro definici omezení hodnot atributů z dané hierarchie. Uvedení příkazu `WHERE` v MDX dotazu odpovídá realizaci OLAP operace *slice* nebo *dice*.

Pro lepší představu o dotazování v jazyce MDX uvedu dva příklady dotazování na jednu z OLAP kostek, jejichž vytvoření v SSDT jsem popisoval v předchozí kapitole o analýze

[3] ActiveX Data Objects MultiDimensional
[4] Analysis Management Objects
[5] Multidimensional Expressions

dat. Pro účely následujících příkladů budu uvažovat kostku, která obsahuje data o exportu zboží za rok 2001. Zvolenou měrnou jednotkou je cena zboží v tisících Kč (v uvedených příkladech dotazů MDX tato informace není, nastavení měrné jednotky se provádí příkazem `ALTER CUBE`).

V prvním příkladu 7.1 je specifikován MDX dotaz pro zjištění hodnoty exportovaného zboží do zemí Evropské unie za první dva kvartály roku 2001. Na ose pro sloupce je uveden konkrétní člen atributu *Region Name* z dimenze Lokace. Osa pro řádky obsahuje Set složený z jednoprvkových n-tic, kde každá n-tice obsahuje hodnotu konkrétního kvartálu.

```
SELECT [DimLocation].[Region Name].[Evropska unie] ON COLUMNS,
{[DimTime].[Quarter Name].[Q1/2001],[DimTime].[Quarter Name].[Q2/2001]} ON ROWS
FROM [CSU_DATA_EXPORT]
```

Ukázka kódu 7.1: MDX dotaz na zjištění objemu exportu do zemí Evropské unie za první dva kvartály roku 2001.

Druhý příklad 7.2 ukazuje složitější MDX dotaz, kde je definován *slicer* příkazem `WHERE`. Dotaz vrací přehled o exportu zboží pro všechny kvartály z dimenze Čas (v tomto případě všechny kvartály roku 2001) na všechny kontinenty, přičemž je sledován export konkrétní skupiny zboží uvedené pomocí hodnoty subkategorie zboží. Klíčové slovo *children* specifikuje výpis všech členů atributu.

```
SELECT
[DimTime].[Quarter Name].children ON COLUMNS,
[DimLocation].[Continent Name].children ON ROWS
FROM [CSU_DATA_EXPORT]
WHERE [DimProduct].[Sub Ctg Code].[0405]
```

Ukázka kódu 7.2: MDX dotaz s příkladem použití příkazu `WHERE`.

Na obrázku 7.1 níže je výsledek MDX dotazu 7.2, který byl proveden v SSMS. Z výsledku dotazu je vidět, že zboží z vybrané subkategorie s kódem 0405 (Máslo a jiné tuky z mléka), nebylo exportováno v žádném kvartálu roku 2001 do zemí Oceánie a polárních oblastí a na africký kontinent neproběhl export ve 4.kvartálu roku 2001.

	Q1/2001	Q2/2001	Q3/2001	Q4/2001
Afrika	8516	42566	61300	(null)
Amerika	14186	43779	26351	3231
Asie	17285	32380	28006	16342
Evropa	127340	259185	172121	228681
Oceánie a polární oblasti	(null)	(null)	(null)	(null)

Obrázek 7.1: Výstup MDX dotazu v nástroji SQL Server Management Studio.

7.1.3 Jazyk DMX

DMX[6] je jazyk pro vytváření dolovacích struktur a modelů. Umožňuje trénování dolovacích modelů a dotazování na data v modelech včetně provádění predikcí nad modely. Stejně jako jazyk MDX, používá i jazyk DMX podobnou syntaxi známou z jazyka SQL [2]. V dalším textu uvádím stručný přehled nejdůležitějších příkazů jazyka DMX spolu s příkladem.

[6] Data Mining Extensions

Pro definici dolovací struktury slouží příkaz `CREATE MINING STRUCTURE`, který vytvoří dolovací strukturu, do které se později přidávají dolovací modely příkazem `ALTER MINING STRUCTURE`. Dolovací strukturu a všechny modely v ní obsažené odstraní příkaz `DROP MINING STRUCTURE`, samotný dolovací model pak příkaz `DROP MINING MODEL`. Jakmile je vytvořena dolovací struktura a dolovací modely, je nutné provést jejich zpracování příkazem `INSERT INTO`.

Pro výpis metadat zpracovaného dolovací modelu a jeho obsahu slouží příkaz `SELECT FROM [název modelu].CONTENT`. Výsledkem dotazu je hierarchická struktura skládající se z uzlů, které obsahují v závislosti na typu dolovacího algoritmu výsledky dolování. Například pokud se jedná o model obsahující asociační pravidla, v kořeni této hierarchické struktury se nachází uzel s informacemi o modelu jako celku a uzly na nižší úrovni hierarchie obsahují data frekventovaných množin a data samotných pravidel.

Pro zobrazení případů, které byly použity k trénování modelu (funkce drillthrough) je k dispozici příkaz `SELECT FROM [název modelu].CASES`, kterým lze vypsat všechny případy, nebo při současném použití příkazu `WHERE` jen ty případy použité během učení pro identifikaci konkrétního vzoru, pravidla nebo shluku.

Následující příklad kódu v jazyce DMX 7.3 ukazuje přidání dolovacího modelu pro získávání asociačních pravidel do dolovací struktury.

První sloupec dolovacího modelu je vždy klíčový sloupec z dolovací struktury, v tomto případě idenfitikátor obchodu *TradeId*. Dále je specifikována nested table *TradesLines* obsahující položky jednotlivých obchodů a z ní sloupec *ProductCode* určený k predikci. To znamená, že asociační pravidla budou hledána nad kódy produktů. Pro nested table je definován filtr, který do modelu zahrnuje pouze produkty z kategorií 04 a 05.

Za klíčovým slovem `USING` se specifikuje dolovací algoritmus spolu s parametry modelu, pokud mají být použity jiné než výchozí. Zvlášť je zde definována minimální podpora pro pravidlo na hranici 5 % ze všech provedených obchodů určených pro učení modelu. Klíčové slovo `DRILLTHROUGH` aktivuje na modelu funkci drillthrough pro prohlížení případů použitých k učení a následuje specifikace dalšího filtru, tentokrát na základě hlaviček obchodů, omezujícího model pouze na obchody s Asií.

```
ALTER MINING STRUCTURE [AssocRulesMiningStructure]
ADD MINING MODEL [AssocRulesMiningModel]
(
    TradeId,
    [TradesLines] PREDICT (
        ProductCode
    ) WITH FILTER CtgCode = '04' OR CtgCode = '05')
)
USING Microsoft_Association_Rules (MINIMUM_SUPPORT = 0.05)
WITH DRILLTHROUGH, FILTER(ContinentName = 'Asie')
```

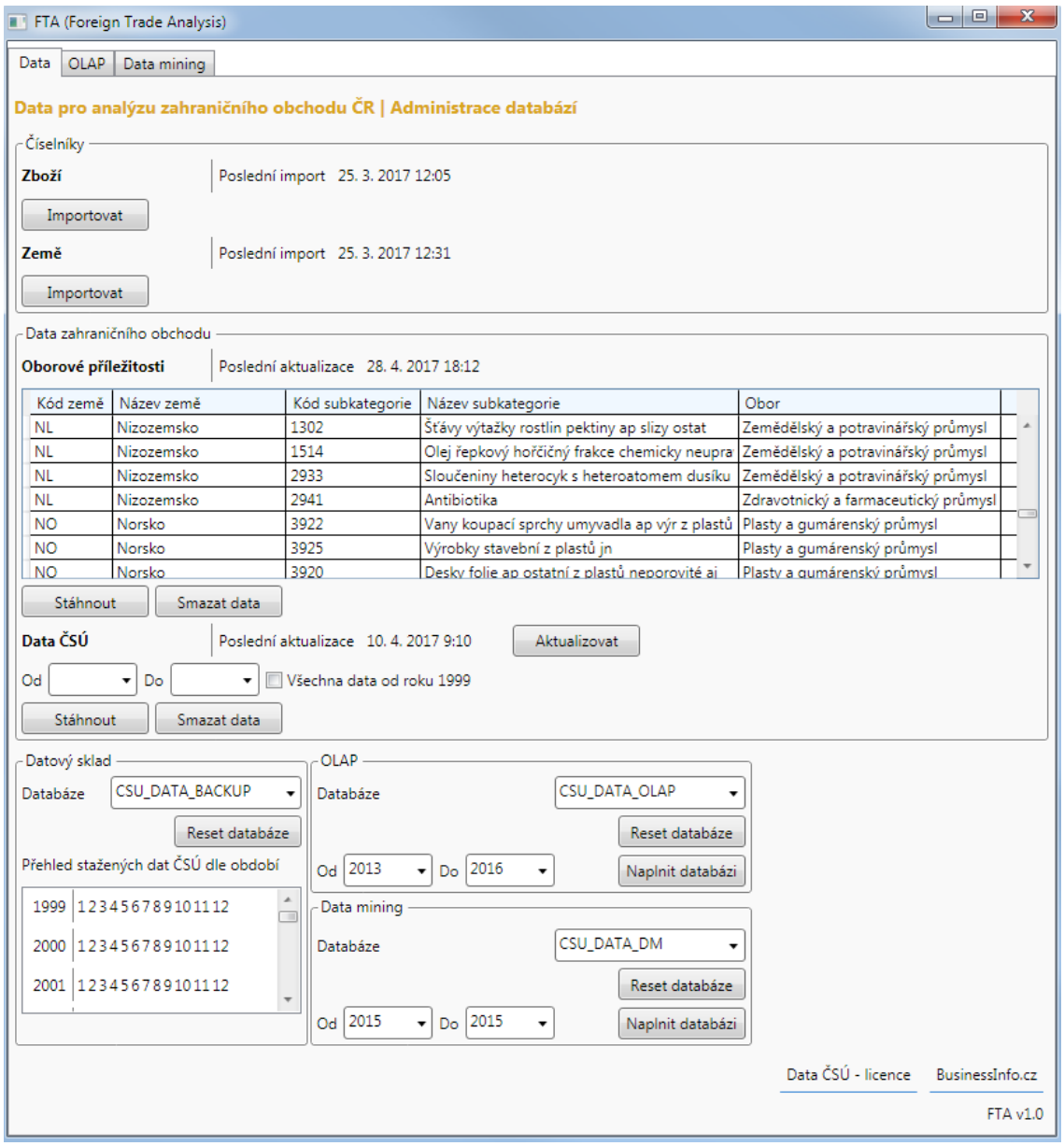
Ukázka kódu 7.3: Kód DMX pro přidání modelu do dolovací struktury.

7.2 Stažení a správa dat

Hlavní okno aplikace FTA je rozděleno do třech záložek pro správu dat, provádění analýzy OLAP a dolování dat. V následujících kapitolách se budu věnovat popisu funkcí jednotlivých záložek a detailům implementace aplikace.

7.2.1 Záložka Data

Po spuštění aplikace FTA je uživateli jako první k dispozici záložka pro správu dat (viz obrázek 7.2). Záložka Data plní dvě hlavní funkce. První je stahování a import dat do centrální databáze. Druhou je pak plnění databází určených k uložení zdrojových dat pro analýzy.



Obrázek 7.2: Aplikace FTA - Správa dat a databází.

Číselníky

V sekci pro číselníky se provádí import číselníků zboží a zemí do centrální databáze. Import nových číselníků neprobíhá často vzhledem k tomu, že číselníky se mění jednou za několik let.

Data zahraničního obchodu

V sekci dat zahraničního obchodu je zobrazen přehled informací o oborových příležitostech stažených z webu BusinessInfo.cz. Data je možné kdykoliv stáhnout znovu a získat aktuální stav oborových příležitostí.

Dolní část sekce je určena pro nastavení stahování dat z webové aplikace ČSÚ. Před stažením dat je nutné určit období, za které mají být data stažena. Při stahování dat zahraničního obchodu probíhá zpracování dat a ukládání do centrální databáze. Tento proces je časově náročný a zpracování dat za jeden rok trvá podle množství dat přibližně jednu až dvě hodiny. Data zahraničního obchodu je možné průběžně aktualizovat.

Databáze

Aplikace dovoluje uživateli zvolit libovolnou databázi běžící na Microsoft SQL Serveru, která bude plnit funkci centrální databáze. Podmínkou je, aby tato databáze obsahovala strukturu tabulek, kterou jsem definoval pro centrální databázi.

Zdrojové databáze pro analýzu OLAP a dolování dat je rovněž možné libovolně zvolit, vyhovuje-li jejich struktura tabulek. V případě analýzy OLAP uživatel určuje období pro data uložená v OLAP kostkách. Po spuštění plnění zdrojové databáze pro OLAP probíhá kopírování dat za zvolené období z centrální databáze. Poté je provedeno zpracování (processing) OLAP kostek a uživatel může pokračovat v práci na záložce pro OLAP. Před spuštěním plnění zdrojové databáze je uživatel vyzván k zadání Windows hesla, jelikož ke zpracování OLAP kostek vyžaduje SSAS oprávnění uživatele.

Při plnění zdrojové databáze pro dolování dat uživatel specifikuje období, které omezuje data pro vytváření dolovacích modelů. Během plnění databáze se stejně jako v předchozím případě kopírují příslušná data z centrální databáze. Před spuštěním plnění je opět nutné zadání Windows hesla uživatele a po dokončení procesu plnění je vytvořena a zpracována dolovací struktura. Od této chvíle je možné na záložce pro dolování dat vytvářet dolovací modely.

7.2.2 Přístup k databázi a modely dat

Spojení s databázemi SQL Serveru zpřístupňuje třída `DatabaseConn` s privátním konstruktorem. Pro celou aplikaci je k dispozici jedna instance třídy `DatabaseConn` přes statickou vlastnost `Connection`. Třída uchovává všechny připojovací řetězce k relačním i analytickým databázím. Samotná připojení k databázím poskytují příslušné vlastnosti třídy `DatabaseConn`, které vracejí instance .NET tříd `SqlConnection` nebo `AdomdConnection` podle typu zdroje dat.

Třída `DatabaseConn` obsahuje sadu statických metod pro realizaci operací spojených s administrací databáze a pro účely zpracování stažených dat. Mezi takové metody patří např. změna připojovacích řetězců (*ChangeConnString()*), reset a plnění zdrojových databází pro analýzy OLAP a dolování dat nebo metoda vracející přehled stažených dat v centrální databázi (*GetDatabaseCSUDataOverview()*).

Při zpracování stažených a importovaných dat se data ukládají do modelů, které se buď dále využívají při zpracování nebo se ukládá jejich obsah do tabulek v databázi. Všechny třídy modelů jsou seskupeny ve složce *Models* v adresářové struktuře zdrojových kódů aplikace. Každý model se skládá z třídy reprezentující konkrétní pojem (např. kategorie produktu, země) a jeho údaje, které odpovídají údajům uchovávaným v databázi. V modelu

se nachází metoda *Save()* pro uložení jeho obsahu do databáze a k danému modelu je většinou definována i třída kolekce modelů.

7.2.3 Zpracování dat

Aplikační logika záložky pro správu dat se nachází ve třídě `DataTabViewModel`. Stahování a zpracování dat probíhá v rámci dalšího vlákna aplikace, které je vytvářeno .NET třídou `BackgroundWorker`. Tím je zajištěno, že UI aplikace bude reagovat na požadavky uživatele i během zpracování dat.

Třída `BackgroundWorker` poskytuje prostředky pro běh kódu v dalším vlákně aplikace a oznamování stavu provádění operace. Kód pro vykonání v dalším vlákně je definován v handleru události *DoWork*, která je vyvolána po zavolání metody *RunWorkerAsync()* třídy `BackgroundWorker`. Kód pro zpracování informací o průběhu operace se nachází v handleru události *ProgressChanged*. Po dokončení operace je vyvolána událost *RunWorkerCompleted*, kdy se obvykle do handleru této události umísťuje kód pro zobrazení hlášky uživateli o dokončení operace.

Při stažení nebo importu dat je nejprve zavolána metoda *RunBackgroundWorker()* třídy `DataTabViewModel`, která zajišťuje vytvoření instance třídy `BackgroundWorker` a spuštění kódu pro zpracování dat v dalším vlákně. Pro zpracování dat jsem vytvořil sadu analyzátorů. Zdrojové kódy tříd analyzátorů jsou umístěny ve složce *ViewModels/Data/DataProcessing* se zdrojovými kódy aplikace. Každý analyzátor implementuje rozhraní *IAnalyzer* předepisující implementaci metody *Run()*.

V následujícím přehledu uvádím prostředky pro zpracování jednotlivých zdrojů dat:

- **Číselníky zemí** – Na webu ČSÚ jsou číselníky klasifikace zemí dle geografických a ekonomických zón k dispozici ve formátu .xlsx, pro jejichž zpracování jsem použil knihovnu ClosedXML [7].
- **Číselník zboží** – Pro parsování XML souboru číselníku zboží využívám standardní .NET knihovnu System.Xml.dll.
- **Webová aplikace ČSÚ** – Zpracování SDMX souboru s daty zahraničního obchodu provádím s pomocí .NET knihovny System.Xml.dll. SDMX soubor je poskytován webovou aplikací ČSÚ ve formátu ZIP archivu, který zpracovávám funkcemi z knihovny SharpZipLib [8].
- **Data oborových příležitostí** – Informace z webu BusinessInfo.cz stahuji ve formě HTML stránky, která je následně parsována pomocí knihovny HtmlAgilityPack [9].

Analyzátor číselníku zemí

Číselník zemí je rozdělen do dvou souborů geograficke_zony.xlsx a ekonomicke_zony.xlsx uložených v adresáři s aplikací. Z číselníku zpracovávám data ze dvou sloupců, kde jsou uvedeny dvoumístné kódy zemí a české názvy zemí, regionů, kontinentů a ekonomických zón. Data číselníku jsou uspořádána ve sloupcích hierarchicky v souvislých skupinách buněk.

Zpracování dat číselníků zemí implementuje třída `CountriesAnalyzer`. Kód pro zpracování dat číselníku zemí se nachází v metodě *GetCountriesCodeList()*. Zde je nejprve

[7] https://closedxml.codeplex.com/
[8] http://icsharpcode.github.io/SharpZipLib/
[9] https://htmlagilitypack.codeplex.com/

vytvořena instance třídy `XLWorkbook` z knihovny ClosedXML pro reprezentaci otevřeného excelovského souboru s číselníkem. Poté jsou adresovány skupiny řádků otevřeného excelovského souboru podle toho, kde se nacházejí údaje o zemích. Tyto skupiny řádků jsou uloženy v kolekcích objektů třídy `IXLRow` z knihovny ClosedXML. Následně jsou procházeny kolekce se skupinami řádků a načítány kódy a názvy. Tyto údaje ukládám postupně do kolekcí s modely kontinentů, regionů a zemí. Po naplnění modelů jsou všechna načtená data uložena do databáze.

Analyzátor číselníku zboží

Číselník zboží je jeden XML soubor, jehož strukturou jsem se zabýval v kapitole o datech ČSÚ (viz příklad 5.2). Od roku 1999 vydal ČSÚ několik číselníků zboží. Všechny číselníky se nacházejí v samostatném adresáři se soubory aplikace (*csu/codelists*).

Analyzátor pro zpracování číselníku zboží reprezentuje třída `ProductsAnalyzer`. V metodě *Run()* analyzátoru jsou postupně načítány soubory číselníků a předávány ke zpracování metodě *GetProductsCodeList()*. Pro třídy, kategorie, subkategorie a produkty jsou definovány kolekce modelů. Tyto kolekce modelů jsou plněny při procházení XML souboru číselníku. Na každé úrovni zboží probíhá kontrola, zdali se již nenachází daný kód v databázi (např. z číselníku staršího data). Kontrolu na existenci kódu zboží v databázi provádí metoda *CheckProductLevel()* třídy `DatabaseConn`. Po zpracování daného souboru s číselníkem jsou data z kolekcí modelů uložena do databáze.

Analyzátor dat ČSÚ

Analyzátor dat ČSÚ implementuje třída `CSUDataAnalyzer`. V metodě *Run()* se zjišťuje, zdali bude probíhat update dat nebo nové stahování. Při updatu je z databáze získán údaj o posledním staženém období (metoda *GetLatestCSUDataInfo* z třídy `DatabaseConn`) a nastaví se rozsah období pro stažení dat až do aktuálního období. Stahování dat podle nastaveného rozsahu období zajišťuje metoda *GetCSUDataFromTo()*. V této metodě se pro každý měsíc období volá metoda *GetCSUData()* implementující samotné stažení a zpracování dat.

Framework .NET poskytuje pro stažení souborů z webu třídu `WebClient`. Z této třídy jsem odvodil třídu `FTAWebClient`, kde jsem nastavil vlastní hodnotou timeoutu pro webový požadavek, jelikož původní hodnota timeoutu byla příliš krátká. Stažení SDMX souboru z webu ČSÚ provádí metoda *DownloadFile* třídy `WebClient`, které předávám URI zdroje. Konkrétně se jedná o URL s parametry pro stažení SDMX souboru, kde uvádím rozsah období a informaci o dovozu nebo vývozu. SDMX soubor je uložen do adresáře *csu/data* v adresáři aplikace a rozbalen.

V kapitole o datech ČSÚ jsem zmiňoval strukturu SMDX souboru (příklad 5.1), kde jsou data uložena do dvou elementů *Group* obsahujících údaje o ceně a hmotnosti zboží. Při zpracování SDMX souboru načítám nejprve všechny údaje o zboží a jeho ceně do kolekce modelů pro údaje o obchodech (třída `TradeModelCollection`). Poté procházím údaje o hmotnosti zboží a v kolekci modelů vždy vyhledám uložené zboží a doplním k němu hodnotu hmotnosti. Jakmile jsou doplněny údaje o hmotnosti u každého zboží, kolekce modelů je uložena do databáze.

Při zpracování dat ČSÚ se občas objevily nesrovnalosti v datech, které jsem vyřešil tak, že taková data nezpracovávám. Jedná se např. o případy, kdy kód země uvedený v SDMX souboru neodpovídal žádnému kódu z číselníku zemí. Nejednoznačná data se vyskytovala v jednotkách případů a nebylo nutné provádět žádná rozsáhlá ošetření dat.

Analyzátor oborových příležitostí

Data oborových příležitostí z webu BusinessInfo.cz zpracovává analyzátor reprezentovaný třídou `BranchOpporsAnalyzer`. Pro stahování souborů webových stránek používám třídu `FTAWebClient`.

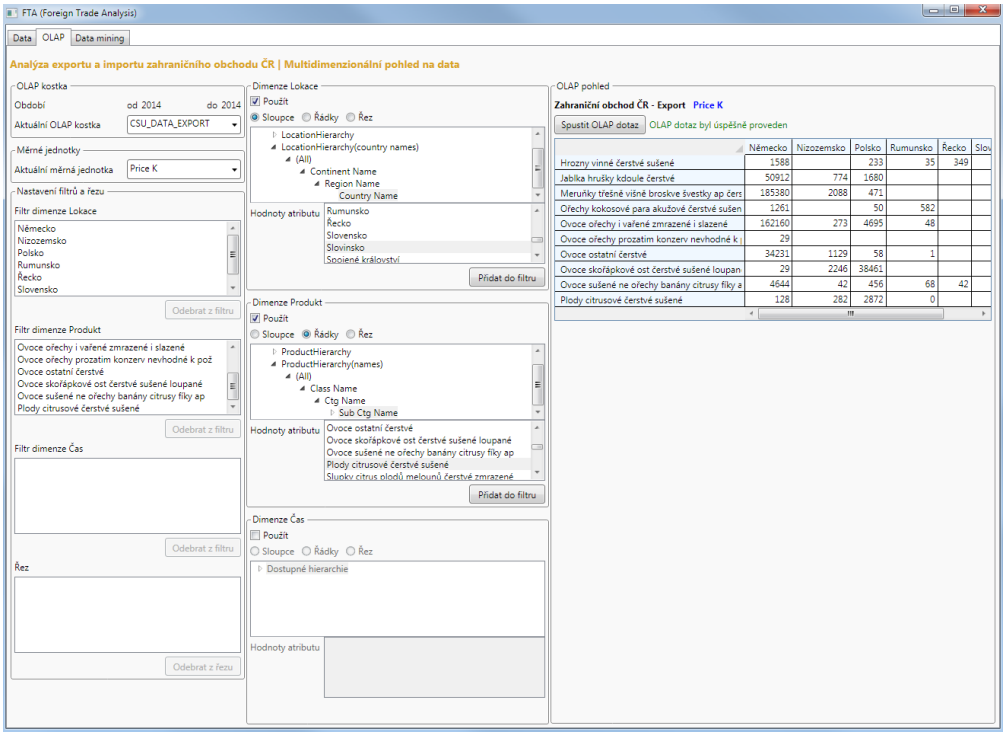
Na začátku probíhá parsování webové stránky s názvy oborů, která je stažena v metodě *GetBranches()*. Pro každý obor jsou následně stahovány jednotlivé webové stránky se seznamy oborových příležitostí pomocí metody *GetBranchOppors()*. Poté probíhá parsování názvů zemí a kódů zboží, které ukládám do kolekce s modely pro oborové příležitosti a nakonec ukládání dat do databáze.

7.3 OLAP klient

Klient pro provádění analýzy OLAP umožňuje analyzovat data zahraničního obchodu ČR za období, které bylo zvoleno na záložce pro správu dat při plnění zdrojové databáze pro OLAP. OLAP klient analyzuje data z OLAP kostek, jejichž struktura byla vytvořena dříve v nástroji SSDT.

7.3.1 Záložka OLAP

Provádění analýzy je k dispozici na záložce OLAP (viz obrázek 7.3), jejíž logika je implementována ve třídě `OLAPTabViewModel`. Před samotnou analýzou je nutné vybrat OLAP kostku a měrnou jednotku, kterou chce uživatel sledovat. Následuje výběr dimenzí a případné nastavení hodnot filtrů. Po spuštění OLAP dotazu je výsledek zobrazen v části okna nazvané OLAP pohled.



Obrázek 7.3: Aplikace FTA - Analýza OLAP.

Je-li dimenze aktivní, uživatel vybere jednu z os, kterou chce pro dimenzi použít. Jedna z dimenzí může mít funkci sliceru pro provádění řezů kostkou. Z hierarchií definovaných na dimenzi lze vybrat příslušný atribut a podle potřeby i konkrétní hodnoty atributu. Pokud je vybrán atribut bez specifikace konkrétních hodnot, v OLAP pohledu budou zobrazeny údaje pro všechny hodnoty atributu. V případě, že uživatele zajímají jen některé hodnoty atributu, může tyto hodnoty specifikovat ve filtrech.

7.3.2 Načítání informací o OLAP kostkách

Při zobrazení záložky OLAP probíhá získání informací o OLAP kostkách uložených v analytické databázi SSAS. Informace o období dat obsažených v OLAP kostkách, měrných jednotkách a dimenzích jsou poté zobrazeny v uživatelském rozhraní. Načtení těchto informací realizuje metoda *LoadOLAP()*, ve které je vytvořeno spojení s analytickou databází. Při čtení informací z analytické databáze využívám třídu `AdomdConnection` z knihovny ADOMD.NET, jež poskytuje kolekce objektů reprezentující údaje o kostkách, dimenzích a dalších objektech.

Při načítání informací o OLAP kostkách nejprve získám kolekci kostek a pak procházím kolekci dimenzí každé kostky. Pro každou dimenzi získám kolekci hierarchií a procházím všechny úrovně hierarchie, z kterých čtu hodnoty atributů.

7.3.3 Provádění OLAP dotazů

Po stisku tlačítka pro provedení OLAP dotazu je zavolána metoda *RunOLAP()*, kde nejprve probíhá ověření, že jsou dostupné všechny potřebné informace pro vykonání OLAP dotazu (kontrola výběru OLAP kostky a měrné jednotky, kontrola výběru minimálně dimenze pro osu COLUMNS apod.). Následně je vytvořena instance třídy `BackgroundWorker`, jenž vytváří instanci třídy `MDXQueryBuilder` zoodpovědné za provedení OLAP dotazu. Této třídě jsou předány údaje o uživatelem zvolené kostce, měrné jednotce a dimenzích.

Třída `MDXQueryBuilder` obsahuje metodu *RunMDXQuery()*, v níž probíhá sestavování MDX dotazu a odeslání dotazu na server SSAS. Na začátku je vytvořen a odeslán MDX příkaz `ALTER CUBE` nastavující měrnou jednotku. Vytváření MDX dotazu provádím postupnou konstrukcí dotazu podle toho, jak uživatel definoval dotaz v uživatelském rozhraní. Například při konstrukci části `SELECT` MDX dotazu kontroluji, zda uživatel zvolil kromě osy pro sloupce i osu pro řádky a jestli byly vybrány konkrétní hodnoty atributu nebo mají být zobrazeny všechny hodnoty.

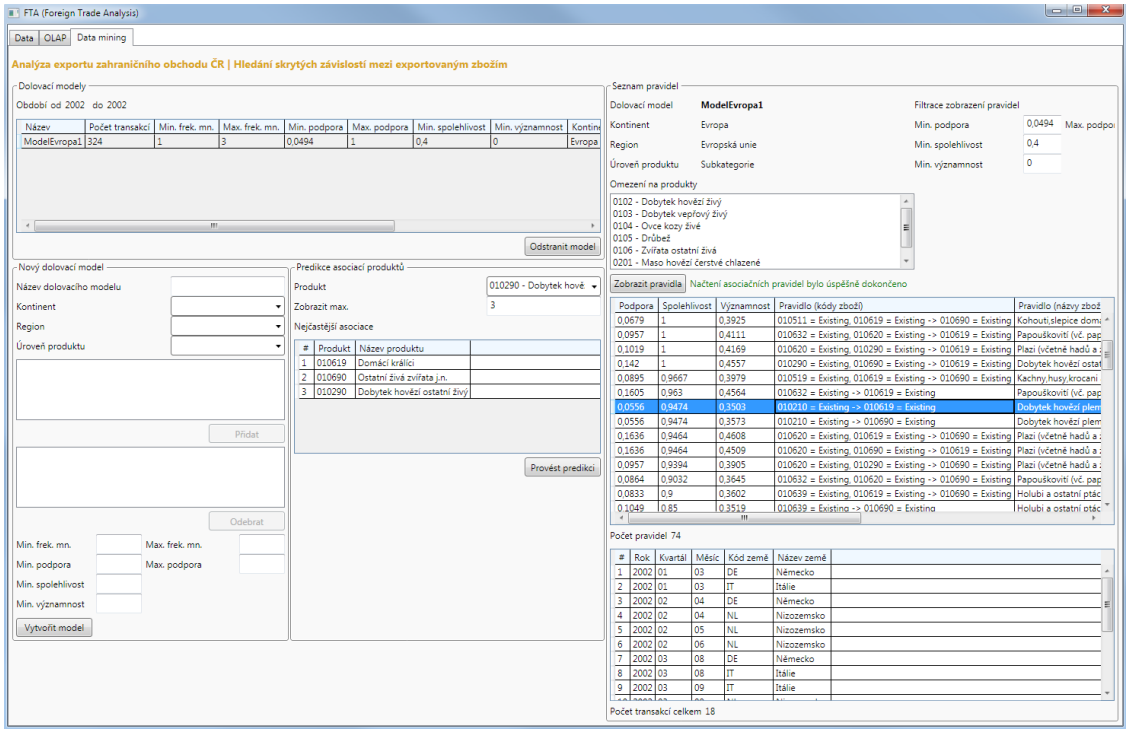
Po odeslání MDX dotazu na server je výsledek vrácen jako objekt třídy `CellSet` knihovny ADOMD.NET. Tato třída umožňuje adresovat jednotlivé osy výsledku dotazu a jednoduše přečíst konkrétní Set nebo Tuple. Údaje z výsledku MDX dotazu ukládám do objektu .NET třídy `DataTable`, která reprezentuje 2D tabulku dat v paměti. Obsah naplněné tabulky je pak předán UI komponentě `DataGrid` zobrazující výsledek uživateli.

7.4 Klient pro dolování dat

Klient pro dolování dat umožňuje uživateli provádět analýzu založenou na získávání asociačních pravidel. Data pro dolování jsou omezena dle období definovaného při plnění zdrojové databáze na záložce pro správu dat.

7.4.1 Záložka Data mining

Funkce dolování dat je k dispozici na záložce Data mining (viz obrázek 7.4) a implementována ve třídě `DataMiningTabViewModel`. Záložka je rozdělena do sekcí pro vytváření dolovacích modelů a zobrazování nalezených asociačních pravidel.



Obrázek 7.4: Aplikace FTA - Dolování dat.

Dolovací modely

V přehledu dolovacích modelů je seznam všech vytvořených modelů s jejich parametry a tlačítko pro odstranění vybraného modelu.

Přehled dolovacích modelů poskytuje uživateli následující údaje o modelu:

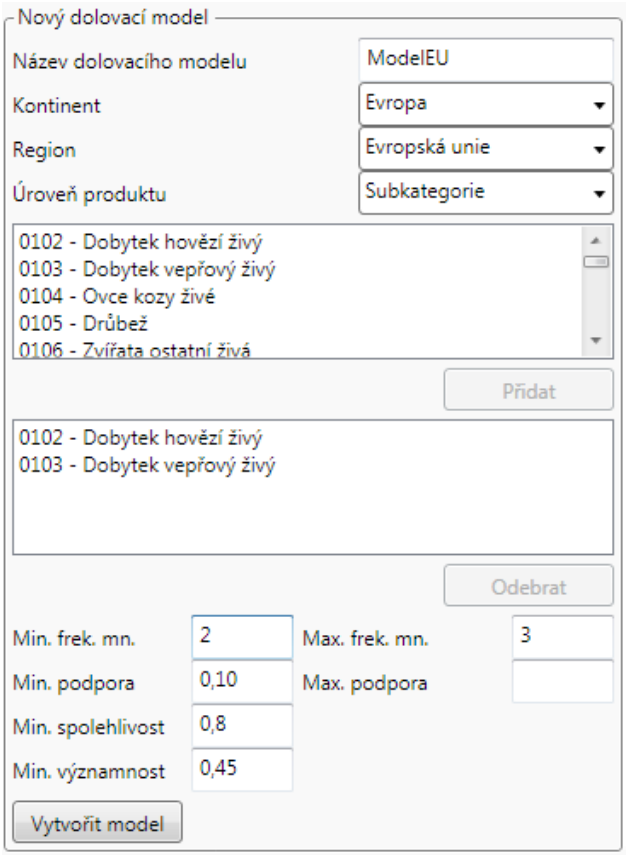
- název modelu a počet transakcí (exportů zboží) v modelu,
- minimální a maximální velikost frekventované množiny,
- minimální a maximální podpora pravidel,
- minimální spolehlivost a významnost pravidel,
- kontinent a region,
- úroveň produktu a omezení na produkty.

Na obrázku 7.5 je příklad vytvoření dolovacího modelu. Informace o kontinentu a regionu omezují data modelu pouze na transakce, kdy proběhl export v rámci Evropy, konkrétně Evropské unie. Úroveň produktu slouží k výběru zboží na úrovni třídy, kategorie, subkategorie nebo produktu, které bude v modelu zahrnuto. Uživatel tak může vytvořit několik modelů pro jednu oblast exportu a sledovat export různých skupin zboží. V příkladu jsou

vybrány dvě subkategorie zboží, mezi jejichž produkty bude dolovací algoritmus hledat asociace.

V další části vytváření dolovacího modelu uživatel specifikuje parametry pro dolovací algoritmus. Zvolená minimální a maximální velikost pro frekventované množiny významně ovlivňuje dobu následného vytváření modelu (výchozí hodnota pro maximální velikost frekventované množiny je tři prvky). Významu parametrů podpory a spolehlivosti jsem se věnoval v teoretické části této diplomové práce.

Parametr minimální významnosti (importance) udává užitečnost asociačního pravidla a jedná se o parametr specifický pro algoritmus asociačních pravidel v Microsoft SQL Serveru. Tento parametr pomáhá určit míru zajímavosti pravidla (vyšší hodnota parametru znamená vyšší užitečnost). Např. pokud všechny transakce zákazníků obchodu obsahují položku A, může přítomnost této položky zkreslovat hodnotu spolehlivosti pravidel, jelikož transakce mohly být provedeny v období, kdy ke každému nákupu byla v rámci prodejní akce přidána tato položka automaticky.



Obrázek 7.5: Vytváření dolovacího modelu.

Zobrazení asociačních pravidel

Po výběru dolovacího modelu z přehledu modelů se v sekci nazvané *Seznam pravidel* zobrazí informace o modelu včetně seznamu produktů (skupin produktů), pro které byl model vytvořen. Asociační pravidla obsažená v modelu lze vypsat po stisku tlačítka pro zobrazení pravidel a jejich zobrazení následně filtrovat. U každého pravidla je k dispozici údaj

o podpoře, spolehlivosti a významnosti. Samotné pravidlo je zobrazeno formou kódů zboží a názvů zboží.

Pokud chce uživatel zobrazit transakce (provedené exporty zboží), ze kterých bylo konkrétní pravidlo vytvořeno (funkce drillthrough), klikne na dané pravidlo v seznamu pravidel. Následně dojde k zobrazení výpisu všech transakcí, kde je možné zjistit období exportu zboží a zemi.

Sekce s názvem *Predikce asociací produktů* nabízí možnost zobrazit pro vybraný produkt nejčastěji exportované zboží společně s tímto produktem. Výchozí hodnota pro zobrazení je omezena na tři produkty a lze změnit.

7.4.2 Dolovací modely a dotazování

Informace o vytvořených modelech jsou načteny v metodě *LoadMiningModelsOverview()*, kde se vytváří spojení s analytickou databází SSAS. Při čtení informací o modelech používám třídu `AdomdConnection` z knihovny ADOMD.NET poskytující kolekce objektů reprezentující dolovací struktury a modely analytické databáze.

Dolovací struktura byla v analytické databázi vytvořena při plnění zdrojové databáze pro dolování. V metodě *LoadMiningModelsOverview()* procházím kolekci s vytvořenými modely této dolovací struktury a načítám informace o modelech do přehledu modelů. Údaje o kontinentu, regionu, úrovni produktu a omezení produktů načítám ze zdrojové databáze pro dolování dat z tabulky *AssocRulesModelsInfo*.

Zobrazení asociačních pravidel z vybraného dolovacího modelu implementuje metoda *ShowMiningModelContent()*. Třídu `AdomdConnection` zde využívám při procházení obsahu modelu, kdy z objektů reprezentujících jednotlivá pravidla čtu příslušné údaje o pravidle.

Pro vytváření dolovacích modelů a operace nad modely slouží třída `DMXQueryBuilder`, která obsahuje následující metody:

- ***RunMiningModelCreate()*** – Obsahuje konstrukci DMX příkazu `ALTER MINING STRUCTURE` pro přidání dolovacího modelu do dolovací struktury. Parametry modelu zadané uživatelem se nastavují ve filtrech modelu a v části konfigurace parametrů dolovacího algoritmu, podobně jako jsem prezentoval dříve v příkladu kódu DMX 7.3. Po odeslání DMX příkazu `ALTER MINING STRUCTURE` na analytický server SSAS je model vytvořen v analytické databázi. Následně probíhá zpracování modelu odesláním DMX příkazu `INSERT INTO MINING MODEL`.
- ***RunMiningModelDelete()*** – Odstraní dolovací model z dolovací struktury.
- ***RunRuleDrillThrough()*** – Realizuje operaci drillthrough odesláním DMX dotazu `SELECT FROM [název modelu].CASES` pro dané pravidlo.
- ***RunPrediction()*** – Provádí nalezení nejčastějších asociací pro produkt. K tomuto účelu nabízí jazyk DMX funkci *PredictAssociation()*, která se uvádí v části `SELECT` DMX dotazu.

Kapitola 8

Ukázky analytických úloh

V následující kapitole budu prezentovat příklady provádění analytických úloh ve vytvořené aplikaci FTA. Každá úloha obsahuje formulaci cíle úlohy a ukázku řešení úlohy se zhodnocením nalezených výsledků.

8.1 Příklad analýzy OLAP

Uživatel potřebuje analyzovat objem exportu zboží (v tisících Kč) s kódy subkategorie 0102 (Dobytek hovězí živý) a 0103 (Dobytek vepřový živý) z ČR do Evropské unie za rok 2016. Cílem je najít období, kdy byl exportován největší objem zboží a sousední stát, do kterého tento export proběhl.

Pro řešení této úlohy uživatel na záložce Data provede naplnění zdrojové databáze pro OLAP daty z roku 2016. Poté přejde na záložku OLAP, kde vybere vytvořenou OLAP kostku pro export a zvolí měrnou jednotku *Price K*.

Uživatel v prvním kroku analýzy zobrazí objem exportu pro celou Evropskou unii. Pro sloupce vybere z dimenze Lokace region *Evropská unie* a přidá do filtru. Pro řádky zvolí z dimenze Produkt subkategorie 0102 a 0103, které přidá do filtru. Výsledek dotazu ukazuje obrázek 8.1, kde je vidět výrazný rozdíl v objemu exportu u hovězího dobytka (asi 3,8 miliardy Kč) oproti exportu vepřového dobytka (asi 1,1 miliardy Kč).

	Evropská unie
Dobytek hovězí živý	3863855
Dobytek vepřový živý	1145097

Obrázek 8.1: Export hovězího a vepřového dobytka do Evropské unie v roce 2016.

V dalším kroku analýzy se uživatel rozhodne dál analyzovat export hovězího dobytka za jednotlivé kvartály roku 2016. Pro řádky vybere z dimenze Čas atribut s hodnotami kvartálů. Navíc přídá dimenzi pro řez (slicer), kde zvolí subkategorii zboží 0102. Na obrázku 8.2 je vidět, že objem exportu v prvních třech kvartálech se výrazně nemění a každý kvartál představuje asi 20 % z celkového objemu exportu. Avšak ve čvrtém kvartálu je objem exportu výrazně vyšší a dosahuje 35 % celkového objemu exportu hovězího dobytka za rok 2016.

Poslední krok analýzy bude zaměřen na zjištění objemu exportu v měsících čtvrtého kvartálu roku 2016 a sousedního státu, do kterého byl proveden největší export. Uživatel

	Evropská unie
Q1/2016	835586
Q2/2016	794425
Q3/2016	884534
Q4/2016	1349310

Obrázek 8.2: Export hovězího dobytka do Evropské unie v jednotlivých kvartálech roku 2016.

přidá do filtru dimenze Lokace sousední státy (tzn. Německo, Rakousko, Polsko a Slovensko). Na časové dimenzi zvolí vyšší míru detailu a vybere jednotlivé měsíce čtvrtého kvartálu (tzn. říjen, listopad a prosinec). Dimenze řezu zůstane nastavená pro zboží subkategorie hovězího dobytka. Výsledek posledního kroku analýzy ukazuje obrázek 8.3.

	Německo	Rakousko	Polsko	Slovensko
Říjen 2016	48416	220374	10682	1281
Listopad 2016	56056	310326	9678	10429
Prosinec 2016	38914	246236	18652	

Obrázek 8.3: Export hovězího dobytka do sousedních států ČR v jednotlivých měsících 4.kvartálu roku 2016.

Na první pohled je vidět, že největší export hovězího dobytka proběhl v 4.kvartále roku 2016 do Rakouska, konkrétně v listopadu. Druhý největší export proběhl do Německa, ale v porovnání s Rakouskem byl každý měsíc zhruba pětkrát menší.

Uživatel by další analýzou zjistil, že objem exportu hovězího dobytka do Rakouska tvořil v roce 2016 téměř 50 % (asi 1,8 miliardy Kč) z celkového objemu exportu hovězího dobytka z ČR do Evropské unie. Otázkou je příčina tohoto jevu, jedním z vysvětlení může být, že Rakousko má oproti ČR výrazný nedostatek hovězího dobytka a musí jej dovážet.

Chybějící hodnota exportu pro Slovensko za měsíc prosinec neznamená, že export v tomto měsíci neproběhl, ale data v době analýzy za tento měsíc nebyla dostupná.

8.2 Příklad analýzy dolování dat

Uživatel, který se začal v roce 2015 věnovat exportu živých ryb (kaprů a pstruhů) na Slovensko, uvažuje o rozšíření exportu v roce 2016 v rámci Evropské unie. V plánu má vybrat jednu zemi, kam chce směřovat své další obchodní aktivity. Cílem analýzy bude zjistit, zdali kromě Slovenska existují další země, do kterých probíhá společný export kaprů a pstruhů. V případě, že takové země existují, bude následovat zhodnocení využitelnosti nalezených výsledků.

V prvním kroku analýzy uživatel naplní zdrojovou databázi údaji pro dolování, kterými budou data za rok 2015. Na záložce pro dolování dat pak vytvoří dolovací model s následujícími vlastnostmi:

- **Kontinent a region** – Evropa a Evropská unie.
- **Úroveň produktu a omezení** – Subkategorie s omezením na kód 0301 (Ryby živé).

- **Parametry modelu** – Min. spolehlivost = 0,5 a Min. významnost = 0,5. U ostatních parametrů budou ponechány výchozí hodnoty, tzn. rozsah velikosti frekventovaných množin 1 až 3, podpora v rozsahu 0,03 až 1.

Vytvořený dolovací model obsahuje celkem 324 transakcí. Na obrázku 8.4 je výpis nalezených asociačních pravidel.

Podpora	Spolehlivost	Významnost	Pravidlo (názvy zboží)
0,071	1	0,6442	Pstruzi živí = Existing, Ryby sladkovodní okrasné, živé = Existing -> Kapři živí = Existing
0,1574	0,9623	0,832	Ryby živé ostatní = Existing, Ryby sladkovodní okrasné, živé = Existing -> Kapři živí = Existing
0,0772	0,9615	0,6388	Pstruzi živí = Existing -> Kapři živí = Existing
0,0648	0,9545	0,6126	Pstruzi živí = Existing, Ryby živé ostatní = Existing -> Kapři živí = Existing
0,1883	0,9242	0,9277	Ryby živé ostatní = Existing -> Kapři živí = Existing
0,0648	0,913	0,7632	Pstruzi živí = Existing, Ryby sladkovodní okrasné, živé = Existing -> Ryby živé ostatní = Existing
0,0679	0,8462	0,7385	Pstruzi živí = Existing -> Ryby živé ostatní = Existing
0,0648	0,84	0,7269	Pstruzi živí = Existing, Kapři živí = Existing -> Ryby živé ostatní = Existing
0,1574	0,7391	1,0706	Kapři živí = Existing, Ryby sladkovodní okrasné, živé = Existing -> Ryby živé ostatní = Existing
0,1883	0,6932	1,4366	Kapři živí = Existing -> Ryby živé ostatní = Existing

Počet pravidel 10

Obrázek 8.4: Asociační pravidla modelu pro rok 2015 a zboží s kódem 0301 (Ryby živé).

Celkem bylo nalezeno deset pravidel, z nichž čtyři pravidla obsahují kombinaci kaprů a pstruhů. Jedno z těchto čtyř pravidel obsahuje pouze kombinaci pstruhů a kaprů (na obrázku 8.4 se jedná o třetí pravidlo), které bude uživatel dále analyzovat. Pravidlo má podporu 7,7 % (tj. 25 transakcí), spolehlivost 96 % a jeho hodnota významnosti není výrazně odlišná od jiných nalezených pravidel.

Po kliknutí na pravidlo (funkce drillthrough) se provede výpis všech transakcí, na základě kterých pravidlo vzniklo, jak ukazuje obrázek 8.5 (výpis na obrázku je zkrácen).

#	Rok	Kvartál	Měsíc	Kód země	Název země
1	2015	01	01	AT	Rakousko
2	2015	01	01	HU	Maďarsko
3	2015	01	01	SK	Slovensko
4	2015	01	02	AT	Rakousko
5	2015	01	02	HU	Maďarsko
6	2015	01	02	SK	Slovensko
7	2015	01	03	AT	Rakousko
8	2015	01	03	SK	Slovensko
9	2015	02	04	AT	Rakousko

Obrázek 8.5: Drillthrough na pravidle Pstruzi > Kapři.

Ze začátku výpisu transakcí je vidět, že v lednu a únoru 2015 proběhl export kaprů a pstruhů do Rakouska, Maďarska a na Slovensko. Po zbytek roku pokračoval export každý měsíc současně do Rakouska a na Slovensko. Jako nové země pro export přicházejí v úvahu Maďarsko a Rakousko. Z pohledu uživatele a jeho dalších obchodních aktivit je výhodnější zabývat se v roce 2016 exportem ryb do Rakouska. Hlavním důvodem je skutečnost, že export do Rakouska a na Slovensko probíhal každý měsíc roku 2015 a pravděpodobně zde existuje určitá závislost, která může platit i v roce 2016.

Uživatel se rozhodne analyzovat export sledovaného zboží za rok 2014 a ověřit, zdali i v tomto roce probíhal podobným způsobem export do Rakouska a na Slovensko. Ve vytvořeném modelu pro rok 2014 je situace podobná s převahou exportu do Rakouska.

Zajímavé je zjištění, proč export do Maďarska proběhl jen v prvních dvou měsících roku 2015 a dále nepokračoval. Mohlo se jednat např. o výjimečnou situaci, anebo jde o pravidelný trend. Uživatel může v další analýze pokračovat tak, že vytvoří jiné modely pro předchozí roky a export ryb do Maďarska bude sledovat podrobněji.

Kapitola 9

Závěr

Cílem práce byla implementace stažení dat zahraničního obchodu ČR a návrh analytických úloh z oblasti analýzy OLAP a dolování dat. Dále otestování navržených úloh v prostředí Microsoft SQL Server 2012 a implementace vybraných analytických úloh ve formě BI aplikace. Na závěr byla funkčnost implementovaných analytických úloh prezentována na příkladech.

Zdrojem dat zahraničního obchodu byla veřejná webová aplikace Českého statistického úřadu a webový portál BusinessInfo.cz, který poskytuje informace o oborových příležitostech pro export. Jako dolovací úlohy jsem navrhl klasifikaci a získávání asociačních pravidel. V nástroji SQL Server Data Tools jsem provedl definici OLAP kostek a otestoval základní funkčnost dolovacích úloh. Dolovací úlohu klasifikace jsem realizoval pouze v uvedeném nástroji, kdy jsem ověřoval použitelnost klasifikace při rozhodování o tom, zdali je zboží oborovou příležitostí. Z provedeného experimentu plyne, že dolovací úloha klasifikace je pro tento účel vhodná jen jako doplňková informace k určení oborové příležitosti pro dané zboží.

Webová aplikace ČSÚ nabízí základní přehled dat zahraničního obchodu a neposkytuje pokročilou analýzu dat. BI aplikace FTA (Foreign Trade Analysis) implementovaná v rámci této práce umožňuje provádění analýzy OLAP a dolovací úlohy získávání asociačních pravidel. Významně tak rozšiřuje možnosti analýzy dat pro uživatele, kterým nedostačují výstupy webové aplikace ČSÚ.

Analýza OLAP nad daty zahraničního obchodu by našla reálné využití například jako rozšíření stávající webové aplikace ČSÚ. Výrazně by se tím zjednodušila analýza historických dat a zobrazování sumárních hodnot zboží nastavením různé úrovně detailu.

Analýza založená na získávání asociačních pravidel z dat zahraničního obchodu dokáže odkrýt zajímavé vzory v datech a závislosti mezi exportovaným zbožím, stejně jako nabídnout velké množství pravidel, která nejsou pro uživatele užitečná. Při reálném použití může implementovaná aplikace posloužit jako nástroj pro hledání potenciálních závislostí v exportovaném zboží. Zároveň by bylo nutné získané výsledky dále ověřovat v jiných zdrojích dle oboru a typu zboží.

Implementovaná BI aplikace by mohla být v budoucnu rozšířena o podobné uživatelské rozhraní a funkce jako má webová aplikace ČSÚ. Uživatelé by tak měli k dispozici data v offline režimu a zároveň funkce, na které jsou zvyklí z webové aplikace. Jako další rozšíření se nabízí implementace jiných dolovacích úloh, vizualizace zahrnující grafy a dashboardy nebo napojení na databáze jiných výrobců (Oracle apod.).

Literatura

[1] *Databáze zahraničního obchodu*. Český statistický úřad, [Online; navštíveno 15.12.2016].
URL https://apl.czso.cz/pll/stazo/STAZO.STAZO

[2] *Data Mining Extensions (DMX) Reference*. Microsoft, [Online; navštíveno 25.4.2017].
URL https://msdn.microsoft.com/en-us/library/ms132058(v=sql.105).aspx

[3] *Developing with ADOMD.NET*. Microsoft, [Online; navštíveno 24.4.2017].
URL https://msdn.microsoft.com/en-us/library/ms123483(v=sql.110).aspx

[4] *Developing with Analysis Management Objects (AMO)*. Microsoft, [Online; navštíveno 24.4.2017].
URL https://msdn.microsoft.com/en-us/library/ms124924(v=sql.110).aspx

[5] *Implementing the Model-View-ViewModel Pattern*. Microsoft, [Online; navštíveno 24.4.2017].
URL https://msdn.microsoft.com/en-us/library/ff798384.aspx

[6] *Klasifikace zemí (CZ-GEONOM)* . Český statistický úřad, [Online; navštíveno 20.12.2016].
URL https://www.czso.cz/documents/10180/39132861/metodicka_cast_2013.pdf

[7] *Learning about SDMX Basics*. SDMX, [Online; navštíveno 17.12.2016].
URL https://sdmx.org/?page_id=2555/

[8] *Mapa oborových příležitostí*. CzechTrade, [Online; navštíveno 20.12.2016].
URL http://www.businessinfo.cz/cs/zahranicni-obchod-eu/mapa-oborovych-prilezitosti.html

[9] *Metodika zahraničního obchodu se zbožím v národním pojetí (princip změny vlastnictví)*. Český statistický úřad, [Online; navštíveno 19.12.2016].
URL https://www.czso.cz/csu/czso/1-vzonu_m

[10] *Mining Structures (Analysis Services - Data Mining)*. Microsoft, [Online; navštíveno 21.4.2017].
URL https://msdn.microsoft.com/en-us/library/ms174757(v=sql.120).aspx

[11] *Multidimensional Modeling (SSAS)*. Microsoft, [Online; navštíveno 19.4.2017].
URL https://msdn.microsoft.com/en-us/library/hh230904(v=sql.110).aspx

[12] *Multidimensional Model Object Processing*. Microsoft, [Online; navštíveno 20.4.2017].
URL https://msdn.microsoft.com/en-us/library/ms174860(v=sql.110).aspx

[13] *Querying Multidimensional Data with MDX*. Microsoft, [Online; navštíveno 25.4.2017].
URL https://msdn.microsoft.com/en-us/library/e0a5dd60-35a3-4a4f-b36f-52ecea814886(v=sql.110)

[14] *SQL Server Technologies*. Microsoft, [Online; navštíveno 19.4.2017].
URL https://msdn.microsoft.com/en-us/library/ms130214(v=sql.110).aspx

[15] *The Basic MDX Query(MDX)*. Microsoft, [Online; navštíveno 25.4.2017].
URL https://msdn.microsoft.com/en-us/library/ms144785(v=sql.110)

[16] *User Hierarchies*. Microsoft, [Online; navštíveno 19.4.2017].
URL https://msdn.microsoft.com/en-us/library/ms174935(v=sql.110).aspx

[17] *Zahraniční obchod - Metodika*. Český statistický úřad, [Online; navštíveno 19.12.2016].
URL https://www.czso.cz/csu/czso/zo

[18] *SDMX 2.1 User Guide*. SDMX, Září 2012, [Online; navštíveno 18.12.2016].
URL
http://sdmx.org/wp-content/uploads/SDMX_2-1_User_Guide_draft_0-1.pdf

[19] Basl, J.; Blažíček, R.: *Podnikové informační systémy: podnik v informační společnosti*. Grada Publishing, a.s., 2008, ISBN 978-80-247-2279-5.

[20] Berka, P.: *Dobývání znalostí z databází*. Academia, 2003, ISBN 80-200-1062-9.

[21] Fotr, J.; Hájek, S.; Špaček, M.; aj.: *Tvorba strategie a strategické plánování : teorie a praxe*. Grada Publishing, a.s., 2012, ISBN 978-80-247-3985-4.

[22] Han, J.; Kamber, M.: *Data Mining: Concepts and Techniques, Second Edition*. Morgan Kaufmann Publishers, 2006, ISBN 978-1-55860-901-3.

[23] Keřkovský, M.; Vykypěl, O.: *Strategické řízení: teorie pro praxi*. C. H. Beck, 2006, ISBN 80-7179-453-8.

[24] Laberge, R.: *Datové sklady: agilní metody a business intelligence*. Computer Press, 2012, ISBN 978-80-251-3729-1.

[25] Lacko, L.: *Business Intelligence v SQL Serveru 2008: reportovací, analytické a další datové služby*. Computer Press, 2009, ISBN 978-80-251-2887-9.

[26] Novotný, O.; Pour, J.; Slánský, D.: *Business Intelligence: jak využít bohatství ve vašich datech*. Grada Publishing, a.s., 2005, ISBN 80-247-1094-3.

[27] Ponniah, P.: *Data Warehousing Fundamentals: A Comprehensive Guide for IT Professionals*. John Wiley & Sons, Inc., 2001, ISBN 0-471-22162-7.

[28] Sodomka, P.: *Informační systémy v podnikové praxi*. Computer Press, 2006, ISBN 80-251-1200-4.

[29] Truneček, J.: *Management znalostí*. C. H. Beck, 2004, ISBN 80-7179-884-3.

[30] Vercellis, C.: *Business Intelligence: Data Mining and Optimization for Decision Making*. John Wiley & Sons, Ltd., 2009, ISBN 978-0-470-51138-1.

Přílohy

Příloha A

Obsah CD

Na CD dodaném k této diplomové práci se nacházejí následující adresáře:

- `/src` – zdrojové kódy implementované aplikace
- `/prj` – zdrojové soubory projektů SQL Server Data Tools
- `/dp_pdf` – text diplomové práce ve formátu PDF
- `/dp_latex` – zdrojové kódy diplomové práce v LaTeXu

Příloha B

Konfigurace aplikace a databází

Následující návod popisuje podmínky pro zprovoznění aplikace a konfiguraci prostředí, ve kterém aplikace pracuje:

1. **Operační systém a verze .NET** - Aplikace FTA byla vyvíjena na operačním systému Windows 7 Professional a .NET frameworku verze 4.5. Novější verze operačního systému Windows nebyly testovány.
2. **Konfigurace SQL Serveru** – Aplikace vyžaduje nainstalovanou lokální instanci Microsoft SQL Serveru. Vývoj aplikace probíhal s použitím Microsoft SQL Serveru 2012 Enterprise s nainstalovanou službou SQL Server Analysis Services. Pro správný běh aplikace je nutné, aby běžely instance relačního databázového serveru a analytického serveru SSAS.
3. **Inicializace relačních databází** – Před spuštěním aplikace je nutné vytvořit v relační databázi SQL serveru alespoň tři databáze. Databázi pro datový sklad (centrální databáze), zdrojovou databázi pro OLAP a zdrojovou databázi pro dolování dat. Pro vytvoření struktury tabulek databází slouží SQL skripty, které jsou uloženy v adresáři s aplikací FTA. V centrální databázi je potřeba spustit SQL skript *db_script_dw*, ve zdrojové databázi pro OLAP SQL skript *db_script_OLAP* a ve zdrojové databázi pro dolování dat SQL skript *db_script_DM*.
4. **Inicializace analytických databází** – Ve složce se soubory projektů SSDT se nachází projekty pro vytvoření OLAP kostek, pro dolovací úlohu klasifikace a získávání asociačních pravidel. Projekty pro OLAP a získávání asociačních pravidel je nutné nasadit na instanci SSAS, aby byly vytvořeny analytické databáze a OLAP kostky. V SSDT projektu je nutné vždy nastavit datový zdroj na vytvořenou relační databázi z bodu 3. Projekt pro asociační pravidla obsahuje pouze definici datového zdroje a datového pohledu, dolovací struktura se vytváří programově v aplikaci FTA. Projekt dolovací úlohy klasifikace je možné také nasadit na instanci SSAS a prohlížet např. v SSDT nebo SSMS.
5. **První spuštění aplikace FTA** – Při prvním spuštění je nutné naplnit centrální databázi daty z číselníků zboží a zemí (přes tlačítka Import). Soubory číselníků se nachází v adresáři spolu s aplikací a názvy souborů a adresářova struktura aplikace musí zůstat zachována. Poté lze stáhnout data oborových příležitostí nebo data ČSÚ a pak lze aplikaci využívat k analýze. Centrální databázi a zdrojové databáze pro analýzy lze kdykoliv resetovat a znovu naplnit daty (není nutné ručně spouštět skripty z bodu 3).