Študentská vedecká aktivita

Analýza údajov z veľkých databáz s využitím nástroja SQL Server Business Intelligence Development Studio

Bc. Jaroslav Čipkala

Sekcia: kvantitatívne metódy a informatika

Konzultant: Ing. Jolana Gubalová, PhD.

Banská Bystrica
2013

Obsah

Úvod

Jednou z kľúčových podmienok úspešného riadenia organizácií, medzi ktoré patrí aj verejná vysoká škola, je rýchly prístup k relevantným informáciám, ktoré umožňujú riadiacim pracovníkom stanoviť vhodnú stratégiu a ostaným pracovníkom poskytujú znalosti na správne rozhodnutia. Aby však boli informácie na podporu rozhodovania relevantné, musia byť poskytnuté v reálnom čase a v požadovanom formáte. Úspešné riadenie firiem a organizácií nezaistí množstvo údajov vo firemných databázach, ale rýchlosť a presnosť, s akou sa manažérom darí z týchto údajov získavať kľúčové informácie. Tieto skutočnosti si vyžiadali zavedenie progresívnej technológie, ktorá prináša do súčasného sveta požadované riešenia pre presnejšie analytické, plánovacie a rozhodovacie činnosti a zohráva kľúčovú úlohu v získaní konkurenčnej výhody. Technológia, ktorá umožňuje proces transformácie dát na informácie a prevod týchto informácií na poznatky prostredníctvom objavovania, je známa pod pojmom Business Intelligence.

Náplňou predkladanej práce je ukážka využitia možností Business Inteligence na dátach extrahovaných z AiS2. Naším cieľom bude určiť profil úspešného absolventa Ekonomickej fakulty UMB. Profil bude zhotovený na základe modelom rozpoznaných kľúčových atribútov majúcich signifikantný vplyv pre úspešné zvládnutie štúdia. Pre potreby spracovania predkladanej práce sme využili databázový produkt SQL Server 2008 od spoločnosti Microsoft. Ide o modernú, dostupnú a komplexnú serverovú platformu pre ukladanie a správu dát v databázach a dátových skladoch. Tento program má implementované sady nástrojov pre Business Intelligence ako sú aplikácie SQL Server Business Intelligence Development Studio alebo MS SQL Server Mangement Studio, v ktorých budeme pracovať.

1. Architektúra Business Intelligence na platforme SQL Servera 2008

Pre MS SQL Server 2008 ako aj ďalšie databázové systémy (napr. Oracle) je charakteristické, že v sebe integrujú väčšinu potrebných nástrojov pre riešenie úloh BI. Implementácia BI na platforme MS SQL Server 2008 pozostáva zo štyroch častí a to (POUR, J.; MARYŠKA, M.; NOVOTNÝ, O., 2012):

- **Relačný databázový server** – sklad pre ukladanie dát s využitím relačných databázových technológií,
- **Integračné služby** (Integration Services) – výber, transformácia a ukladanie dát (ETL),
- **Analytické služby** (Analysis Services) – obohatenie dát o výsledky analýz a predikcií, čím sa údaje menia na cenné informácie, potrebné pre podporu rozhodovania,
- **Reportovacie služby** (Reporting Services) – sprístupnenie dát a výsledkov analýz užívateľom vo vhodnej forme, obsahu a rozsahu.

Jadro produktu pozostáva z dvoch vrstiev, do ktorých sú zaradené spomínané nástroje podľa funkcie ktorú zohrávajú pri tvorbe BI riešení. Prvá vrstva je tvorená aplikáciou SQL Server Integration Services (SSIS). Do druhej vrstvy sú postupne zaradené zvyšné komponenty a to: relačný databázový server, SQL Server Analytical Services (SSAS) a SQL Sever Reporting Services (SSRS).

2. Príprava dát - extrahovanie

Ako sme sa zmienili v úvode práce, subjektom nášho záujmu sú vybrané údaje z akademického informačného systému AiS2 a ich analýza pomocou nástrojov BI v programe SQL Server 2008. Nakoľko databázový aj aplikačný server tohto akademického systému boli poskytnuté spoločnosťou Oracle, na to, aby sme s nimi mohli pracovať v rozhraní SQL Servera 2008, museli sme najskôr uskutočniť ich konverziu. Na túto úlohu sme využili voľne dostupný nástroj od spoločnosti Microsoft, ktorým je SQL Server Migration Assistant 2008 (SSMA).

Po úspešnej konverzií dát z databázového systému Oracle do formy vhodnej pre databázový systém SQL Server 2008, máme k dispozícií vstupy potrebné pre našu analýzu. Sú nimi dáta z databázy AiS2 o celkovej veľkosti 5,33 GB, ktorá spolu obsahuje 720

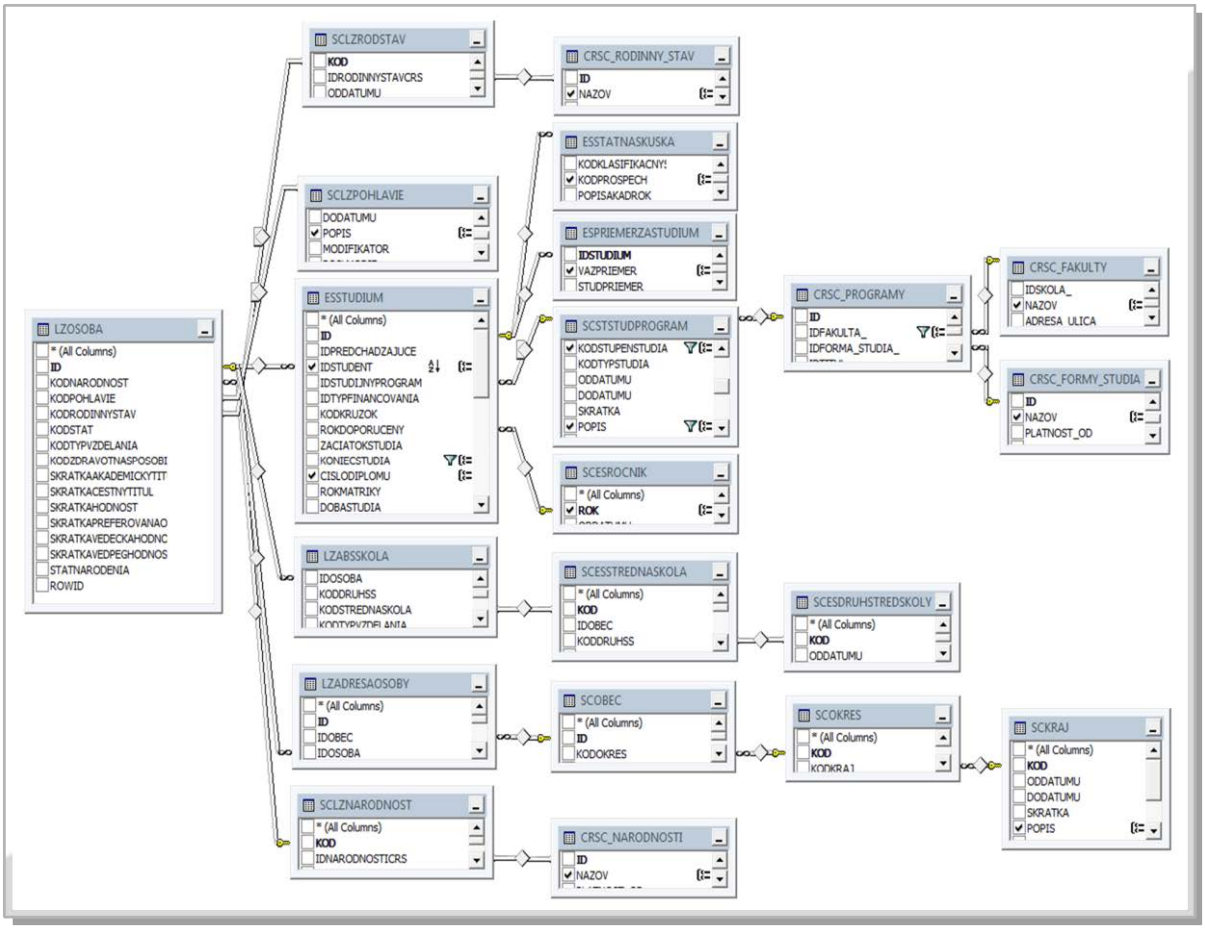
tabuliek s obsahom aktuálnym k októbru 2012. AiS2 je z pohľadu BI našou produkčnou (zdrojovou) databázou.

Akademický informačný systém AiS2 je komplexný informačný systém používaný v súčasnosti fakultami desiatich verejných a troch súkromných vysokých škôl a univerzít pôsobiacich v Slovenskej republike. Databáza AiS2 disponuje širokým spektrom údajov, poskytujúcim komplexný prehľad o všetkých činnostiach a osobách spojených s priebehom a zabezpečením štúdia na daných vzdelávacích inštitúciách, čomu zodpovedá aj jej veľkosť. Samozrejme, my budeme z tohto objemu dát potrebovať len určitú časť. Naším cieľom bude totiž určiť profil a predpoklady úspešného absolventa Ekonomickej fakulty UMB. Všetky kroky potrebné k extrahovaniu požadovaných údajov a vytvoreniu cieľového datasetu uskutočníme v na to určenom nástroji, ktorým je SQL Server Manegement Studio. Na tomto mieste je ešte dôležité uviesť, že predtým ako sme s danou databázou začali pracovať, boli z nej príslušnou inštanciou odstránené všetky osobné údaje. Osoby – študenti, učitelia tak vystupujú len pod prideleným ID_riadka ich záznamu a nie je možné ich stotožniť.

Po ujasnení a zadefinovaní nášho cieľa a pripojení databázy do servra, môžeme v prostredí SQL Server Mangement Studio pristúpiť k príprave dát potrebných na jeho naplnenie. Profil študenta môžeme definovať ako množinu vlastností (atribútov), ktoré sa viažu k jeho osobe a charakterizujú ju. Takýmito atribútmi sú napríklad demografické údaje, ako vek alebo pohlavie a údaje o jeho štúdiu, napríklad prospech a študijný odbor. Našou úlohou je spomedzi dostupných dát nájsť a získať tie, na základe ktorých bude možné takýto profil zostaviť.

Táto úloha je časovo aj prácne najnáročnejšia. Východiskovým bodom v tejto fáze bolo dôkladné oboznámenie sa s názvami a obsahom jednotlivých tabuliek v databáze a predbežné vyselektovanie tých, v ktorých by sa mohli nachádzať pre nás relevantné údaje. Po identifikovaní vhodných tabuliek, sme následne prostredníctvom dotazov a filtrov vytvorených v query editore pristúpili k samotnému extrahovaniu dát obsahujúcich požadované atribúty. Keďže sa tabuľky nepodarilo preniesť z pôvodného databázového systému aj s reláciami medzi materskými a dcérskymi tabuľkami, naše požiadavky si vynútili množstvo dodatočných relačných prepojení. Ak sme chceli napríklad priradiť k študentovi (tabuľka LZOSOBA), atribút akú strednú školu navštevoval (tabuľka SCESDRUHSTREDSKOLY), medzi ktorými neexistovalo priame prepojenie z dôvodu absencie spoločného identifikátora, museli sme nájsť a do dotazu medzi tieto dve tabuľky doplniť ďalšiu, ktorá umožnila ich prepojenie, pretože obsahovala primárny kľúč jednej aj druhej tabuľky. Takýmto spôsobom sme z databázy vyfiltrovali študentov Ekonomickej

fakulty UMB a postupne sme k nim začali priraďovať spomínané vlastnosti. Výsledný dotaz pozostával z 21 vybraných a relačne prepojených tabuliek a potrebných filtrov. Ukážku tohto dotazu v podobe diagramu vytvoreného v prostredí Design Query in Editor môžeme vidieť na nasledovnom obrázku.



Obrázok č. 1: Schéma relačných vzťahov pre identifikáciu študentov
Zdroj: vlastné spracovanie.

V dolnej časti návrhového zobrazenia vytváraného dotazu je k dispozícii podrobný popis kritérií pre zvolené atribúty, ako aj možnosti ich ďalšej úpravy. Nachádza sa tu postupne sedem záložiek:

- **Column** – obsahuje názvy vybraných stĺpcov (atribútov) z jednotlivých tabuliek s dátami, v ich pôvodnom znení,
- **Alias** – možnosť premenovať názvy stĺpcov, pod akými budú figurovať vo výstupe dotazu,
- **Table** – názov tabuľky, z ktorej daný atribút pochádza,
- **Output** – výber atribútov, ktoré sa zobrazia vo výstupe dotazu,

- **Sort Type**, **Sort Order** - možnosti pre zoradenie dát,
- **Grup by** – agregačné funkcia (COUNT, SUM, AVG, MIN a MAX),
- **Filter** – možnosti filtrácie požadovaných údajov v rámci zvoleného atribútu.

Ako vidíme na obrázku číslo 2, z daných tabuliek sme si prostredníctvom zadefinovaného dotazu vybrali spolu 17 polí, potrebných pre zostavenie nášho datasetu. Ku časti z nich sa ešte vzťahujú potrebné filtračné kritéria. Počiatočné oboznamovanie s databázou a vytypovanými tabuľkami prinieslo požadované výsledky, nakoľko sme v jednej z nich natrafili na zoznam všetkých fakúlt v databáze spolu s ich identifikačnými číslami. Ekonomickej fakulte prislúcha identifikačné číslo 714020000. Vďaka tomuto poznatku sme boli schopní z databázy vyfiltrovať študentov našej fakulty. Spoľahlivým príznakom, na základe ktorého sme boli schopní odlíšiť študenta od úspešného absolventa bolo číslo diplomu. Zamerali sme sa len na denných a externých študentov, ktorí ukončili štúdium v roku 2012 a ktorí študovali svoj študijný program priamo na EF UMB v Banskej Bystrici a nie na dištančnom pracovisku v Poprade.

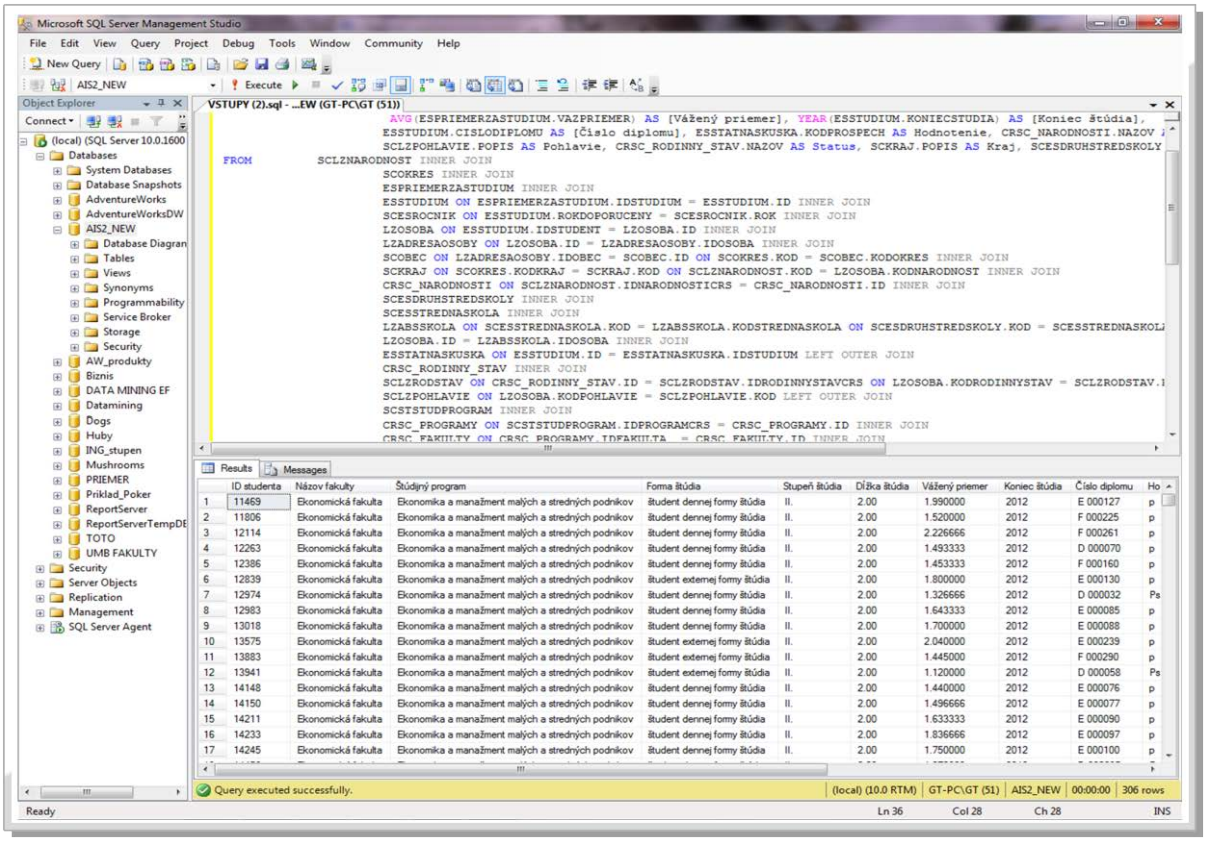
Column	Alias	Table	Output	Sort Type	Sort Order	Group By	Filter
IDSTUDENT	[ID studenta]	ESSTUDIUM	☑			Group By	
NAZOV	[Nazov fakulty]	CRSC_FAKULTY	☑			Group By	
POPIS	[Studijny program]	SCSTSTUDPROGRAM	☑	Ascending	1	Group By	NOT LIKE '%Poprad'
NAZOV	[Forma studia]	CRSC_FORMY_STUDIA	☑			Group By	
KODSTUPENSTUDIA	[Stupeň studia]	SCSTSTUDPROGRAM	☑			Group By	= 'II.'
DLZKA_STUDIA	[Dlzka studia]	CRSC_PROGRAMY	☑			Group By	
VAZPRIEMER	[Vazeny priemer]	ESPRIEMERZASTUDIUM	☑			Avg	
YEAR(ESSTUDIUM.KONIECSTUDIA)	[Koniec studia]		☑			Group By	= 2012
CISLODIPLOMU	[Cislo diplomu]	ESSTUDIUM	☑			Group By	
KODPROSPECH	Hodnotenie	ESSTATNASKUSKA	☑			Group By	
NAZOV	Narodnosť	CRSC_NARODNOSTI	☑			Group By	
POPIS	Pohlavie	SCLZPOHLAVIE	☑			Group By	
NAZOV	Status	CRSC_RODINNY_STAV	☑			Group By	
POPIS	Kraj	SCKRAJ	☑			Group By	
POPIS	Stredná	SCESDRUHSTREDSKOLY	☑			Group By	
IDFAKULTA_		CRSC_PROGRAMY	☐			Group By	= 714020000

Obrázok č. 2: Zoznam zvolených atribútov a im prislúchajúcich kritérií.

Zdroj: vlastné spracovanie.

Po prevedení týchto procedúr a ich potvrdení sa môžeme z prostredia Design Query in Editor presunúť späť do hlavného návrhové zobrazenia MS SQL Server Management Studio a prostredníctvom príkazu execute spustiť takto pripravený dotaz. Po chvíli čakania máme v dolnej časti okna k dispozícii výstup našej práce v podobe hodnôt jednotlivých atribútov,

na základe ktorých budeme schopní zostaviť zamýšľaný profil úspešného absolventa štúdia na Ekonomickej fakulte. Na obrázku číslo 3 vidíme ukážku pracovného prostredia v danej aplikácii. Tá je rozdelená na tri hlavné časti. V ľavej časti sa nachádza okno Object Explorera so zoznamom databáz. Napravo od neho vidíme v hornej časti textovú podobu nášho dotazu a pod ním samotný výsledok extrakcie dát.



Obrázok č. 3: Prostredie MS SQL Server Management Studia s vykonanými procedúrami.
Zdroj: vlastné spracovanie.

Daný výstup obsahuje spolu 306 riadkov, pričom každý riadok predstavuje jedného úspešného absolventa. Ku každému absolventovi sa vzťahuje 15 atribútov. Nie všetky atribúty však zahrnieme do výsledného datasetu. Niektoré atribúty majú totižto čisto kontrolný charakter nášho vyhľadávania. Jedným zo stĺpcov, ktorý pri nadchádzajúcich úpravach odstránime, nakoľko nemá žiadnu vypovedaciu hodnotu v procese data miningu, je napríklad stĺpček „Názov fakulty“. Tá je, tak ako to má byť, pre každý záznam rovnaká (Ekonomická fakulta) a plní len spomínaný kontrolný účel. Pravým kurzorom myši klikneme na výsledok nášho dotazu a vyberieme možnosť „Save Results As“. Získané dáta budú uložené vo formáte CSV do hárku tabuľkového kalkulátora, ktorý môžeme nazvať napríklad ABSOLVENTI.

V prostredí kancelárskeho kalkulátora MS Excel, sme vykonali finálne úpravy datasetu predtým, ako bude použitý ako zdroj pre naše analýzy.

Tak ako sme sa už zmienili, z datasetu odstránime polia, ktoré pre analýzu nemajú hlbší význam a slúžia len na pre kontrolu správnosti extrahovaných údajov. Sú nimi nasledovné stĺpce: „Názov fakulty“, „Stupeň štúdia“, „Dĺžka štúdia“, „Koniec štúdia“ a „Číslo diplomu“. Po týchto úpravách obsahuje náš dataset 10 polí. Prvý z nich – ID študenta, predstavuje jednotlivé prípady a ostatné stĺpce sú vo vzťahu k nemu jeho atribútmi. Z dôvodu prehľadnejšieho obrazu o prospechu absolventov (v tom čase ešte študentov) sme k číselnému hodnoteniu vo forme dosiahnutého váženého aritmetického priemeru za dané štúdium, priradili aj jeho slovný ekvivalent. Využili sme na to funkciu IF, ktorá v závislosti od dosiahnutého priemeru na základe nami zadaných kritérií priradila k záznamu zodpovedajúce slovné hodnotenie:

- Ak vážený priemer < 1,5 = „výborný“
- 1,51 < vážený priemer < 2 = “chvalitebný”
- 2,01 < vážený priemer < 2,5 = “dobrý” inak “uspokojivý”

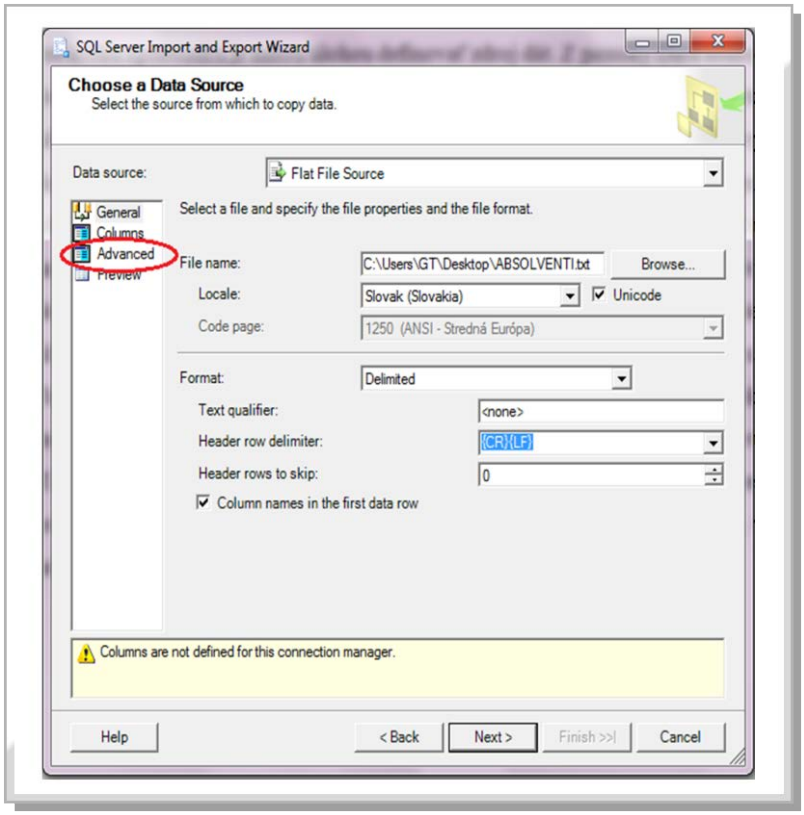
Nové pole pomenujeme „Prospech“. Celkový počet opisných atribútov je tak 11. V úpravách pokračujeme a tam kde je to potrebné, z dôvodu prehľadnosti niekoľko slovné hodnoty atribútov nahradíme vhodnou skratkou. Takýmto spôsobom upravíme postupne premenné v siedmych poliach. Ide len o kozmetické úpravy, ktoré nemajú žiadny vplyv na výsledky analýz. Napríklad namiesto celého názvu študijného odboru Financie, bankovníctvo a investovanie, bude tento teraz v tabuľke uvedený pod jeho ekvivalentom v podobe skratky FBI. Hodnoty zvyšných troch atribútov (Priemer, Prospech, Hodnotenie, Okres) sú prehľadné už v ich pôvodnom znení a preto sme ich viac neupravovali. Pre úplnosť uvedieme ešte význam hodnôt pre atribút Hodnotenie. Ten vyjadruje celkové hodnotenie štúdia študenta podľa pravidiel stanovených v sprievodcovi štúdia, v ktorom sa zohľadňuje dosiahnutý študijný priemer a výsledná známka štátnej skúšky. Môže nadobúdať nasledovné hodnoty:

- „p“ ako prospel,
- „PsV“ ako prospel s vyznamenaním,
- „v“ ako vylúčený.

Takto pripravený dataset si uložíme vo formáte txt.

3. Transformácia a zavedenie dát

Aby sme mohli pripravené dáta využiť pre data minigový príklad, potrebujeme ich v rámci prípravnej fázy preniesť do databázy. Zavedenie údajov do databázy môžeme vykonať buď prostredníctvom vytvorenia integračného projektu v prostredí SQL Server Business Intelligence Development Studio alebo na to môžeme využiť SQL Server Import and Export Wizard, ktorý je tiež súčasťou BI nástrojov MS SQL Servra 2008. Obidvoma spôsobmi sa dopracujeme k rovnakému výsledku. My využijeme druhú možnosť a spustíme SQL Server Import and Export Wizard.



Obrázok č.4: Dialógové okno v MS SQL Server Import and Export Wizard.
Zdroj: vlastné spracovanie.

V prvom kroku sprievodcu je našou úlohou definovať zdroj dát. Z ponuky Data source najskôr vyberieme možnosť Flat File Source, nakoľko súbor s našimi dátami má formát txt a následne zadefinujeme k nemu prístup. Keďže prvý riadok súboru obsahuje názvy atribútov, zaškrtneme možnosť Column names in the first row. V ľavej časti okna klikneme na záložku Advanced, kde môžeme definovať dátový typ pre jednotlivé atribúty. Máme tu k dispozícií funkciu Suggest Types, ktorú využijeme a program sám navrhne pre jednotlivé atribúty ich

dátový typ. Skontrolujeme odporúčaný typ a pokračujeme v procese integrácie. Cieľom tejto úlohy je prenos dát zo zdrojovej do cieľovej databázy. Keďže zdroj dát sme už definovali, v ďalšom kroku musíme určiť cieľovú databázu, kam skopírujeme naše dáta. Môžeme si vybrať z ponuky už vytvorených databáz alebo si vytvoríme novú. My si vytvoríme novú s názvom DATA_ABSOLVENTI. Nasledovné informatívne kroky v sprievodcovi už len cez tlačidlo next potvrdíme a ukončíme príkazom finish. Ak všetky procedúry prebehli správne, v záverečnom hlásení sprievodcu sa dočítame, že naše dáta sú úspešne zavedené do cieľovej databázy.

4. Návrh data miningového modelu

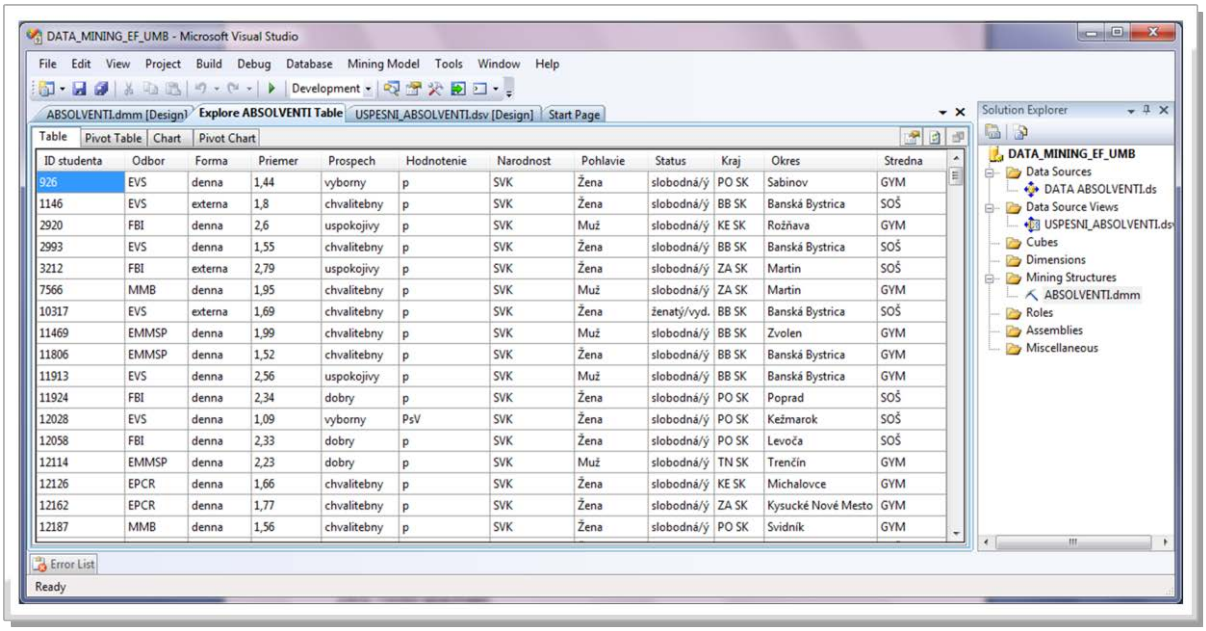
Po dlhých, ale potrebných prípravách súvisiacich s prípravou vhodných dát, môžeme pristúpiť k ich analýze. Využijeme na to techniky tzv. dolovania dát, ktoré sú spolu s OLAP analýzami a reportingom súčasťou analytických nástrojov MS SQL Servra 2008. Data minig je momentálne jednoznačne najrýchlejšie rastúcim segmentom Business Intelligence. Je principiálne založený na heuristických algoritmoch, neurónových sieťach a iných pokročilých softwerových technológiách a metódach umelej inteligencie. Pomáha sledovať a analyzovať trendy a predikovať udalosti.

V prostredí Business Intelligence Development Studio vytvoríme nový projekt typu Analysis Services Project zo zložky BI projektov a pomenujeme ho napríklad DATA_MINING_EF_UMB. Pre zrýchlenie nasadenia projektu na analytický server a zrýchlenie ladenia sa odporúča zmeniť metódu jeho nasadenia (LACKO Ľ. , 2009). V okne Solutions Explorer preto cez Properties aktivujeme zložku Deployment a parameter Deployment Mode nastavíme na hodnotu Deploy All. Postup budovania projektu je naznačený poradím zložiek v okne vývojového prostredia Solution Explorer v pravom hornom rohu. My budeme pracovať s nasledovnými zložkami:

- Data Sources,
- Data Sources Views,
- Mining Structures.

Ako dátový zdroj použijeme našu tabuľku ABSOLVENTI, ktorá sa nachádza v databáze DATA_ABSOLVENTI. Na záložke Data Source Views vyberieme tabuľku ABSOLVENTI a pohľad nazveme napríklad USPESNI_ABSOLVENTI. V prostrednom okne

vývojového prostredia sa zobrazí diagram návrhovej štruktúry vybraného pohľadu na dátové zdroje. Klikneme naň pravým tlačidlom myši a prostredníctvom položky Explore Data si môžeme prezrieť vybrané údaje.

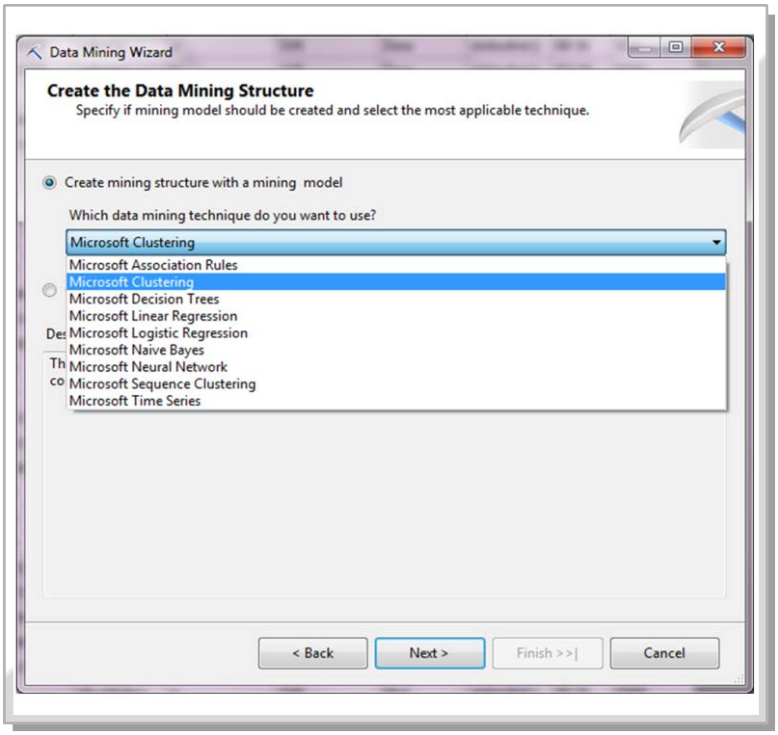


ID studenta	Odbor	Forma	Priemer	Prospech	Hodnotenie	Narodnost	Pohlavie	Status	Kraj	Okres	Stredna
926	EVS	denna	1,44	vyborny	p	SVK	Žena	slobodná/ý	PO SK	Sabinov	GYM
1146	EVS	externa	1,8	chvalitebny	p	SVK	Žena	slobodná/ý	BB SK	Banská Bystrica	SOŠ
2920	FBI	denna	2,6	uspokojivy	p	SVK	Muž	slobodná/ý	KE SK	Rožňava	GYM
2993	EVS	denna	1,55	chvalitebny	p	SVK	Žena	slobodná/ý	BB SK	Banská Bystrica	SOŠ
3212	FBI	externa	2,79	uspokojivy	p	SVK	Žena	slobodná/ý	ZA SK	Martin	SOŠ
7566	MMB	denna	1,95	chvalitebny	p	SVK	Muž	slobodná/ý	ZA SK	Martin	GYM
10317	EVS	externa	1,69	chvalitebny	p	SVK	Žena	ženatý/vyd.	BB SK	Banská Bystrica	SOŠ
11469	EMMSP	denna	1,99	chvalitebny	p	SVK	Muž	slobodná/ý	BB SK	Zvolen	GYM
11806	EMMSP	denna	1,52	chvalitebny	p	SVK	Žena	slobodná/ý	BB SK	Banská Bystrica	GYM
11913	EVS	denna	2,56	uspokojivy	p	SVK	Muž	slobodná/ý	BB SK	Banská Bystrica	GYM
11924	FBI	denna	2,34	dobry	p	SVK	Žena	slobodná/ý	PO SK	Poprad	SOŠ
12028	EVS	denna	1,09	vyborny	PsV	SVK	Žena	slobodná/ý	PO SK	Kežmarok	SOŠ
12058	FBI	denna	2,33	dobry	p	SVK	Žena	slobodná/ý	PO SK	Levoča	SOŠ
12114	EMMSP	denna	2,23	dobry	p	SVK	Muž	slobodná/ý	TN SK	Trenčín	GYM
12126	EPCR	denna	1,66	chvalitebny	p	SVK	Žena	slobodná/ý	KE SK	Michalovce	GYM
12162	EPCR	denna	1,77	chvalitebny	p	SVK	Žena	slobodná/ý	ZA SK	Kysucké Nové Mesto	GYM
12187	MMB	denna	1,56	chvalitebny	p	SVK	Žena	slobodná/ý	PO SK	Svidník	GYM

Obrázok č. 5: Pohľad na dátový zdroj.

Zdroj: vlastné spracovanie.

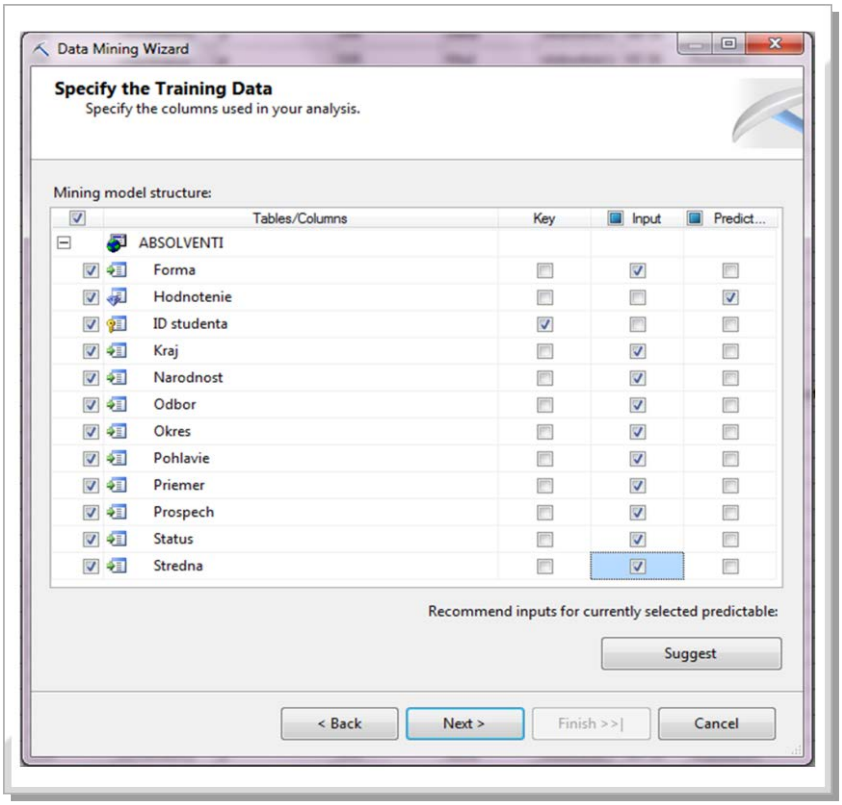
Po prípravných fázach definovania zdroja dát, môžeme pristúpiť k vytvoreniu data miningového modelu. Na záložke Mining Structures aktivujeme sprievodcu Data Mining Wizard. Tento sprievodca nám umožní vytvoriť data miningový model buď z relačnej databázy alebo z OLAP kocky. V našom prípade sa bude model vytvárať z jedinej tabuľky relačnej databázy. V ďalšom kroku nasleduje výber algoritmu. Zvolíme si algoritmus vykonávajúci zhlukovanie s označením Microsoft Clustering.



Obrázok č. 6: Výber algoritmu.

Zdroj: vlastné spracovanie.

V ďalšom kroku nasleduje spresnenie tabuľky, s ktorou bude model pracovať alebo inak povedané, na ktorej sa bude učiť. Keďže máme k dispozícii len jednu tabuľku ABSOLVENTI, učiaca tabuľka bude totožná so zdrojovou tabuľkou. Pre identifikáciu prípadov je potrebné definovať stĺpec, v databázovej terminológií primárny kľúč, pomocou ktorého bude jednoznačne identifikovaná každá entita. V našom prípade je to jednoznačne atribút ID_riadka študenta. Nasleduje špecifikácia vstupných stĺpcov a stĺpca, ktorého hodnotu chceme predikovať. Predmetom nášho záujmu je atribút, ktorý udáva to, do akej miery bol daný absolvent úspešný vo svojich štúdiách. Teda, či priebeh jeho štúdia bol bezproblémový a z pohľadu fakulty bolo s takýmto študentom minimum práce (napríklad skúšky zložil v riadnom termíne a nemuseli byť vynaložené dodatočné zdroje súvisiace s opravným termínom). Ak by sme to chceli prirovnať k bankovej terminológií, tak takýto študent by predstavoval vysoko bonitného klienta, o ktorého majú banky záujem. Predikovaným atribútom bude teda stĺpec Hodnotenie a prípady, v ktorých nadobúda hodnotu PsV, teda excelentné výsledky. Ostatné stĺpce označíme ako vstupné. Pri ich výbere nám môže opäť pomôcť program, prostredníctvom tlačidla Suggest, ktoré odporučí nevýznamné atribúty vynechať. My sa však rozhodneme všetky zostávajúce atribúty zahrnúť do modelu.



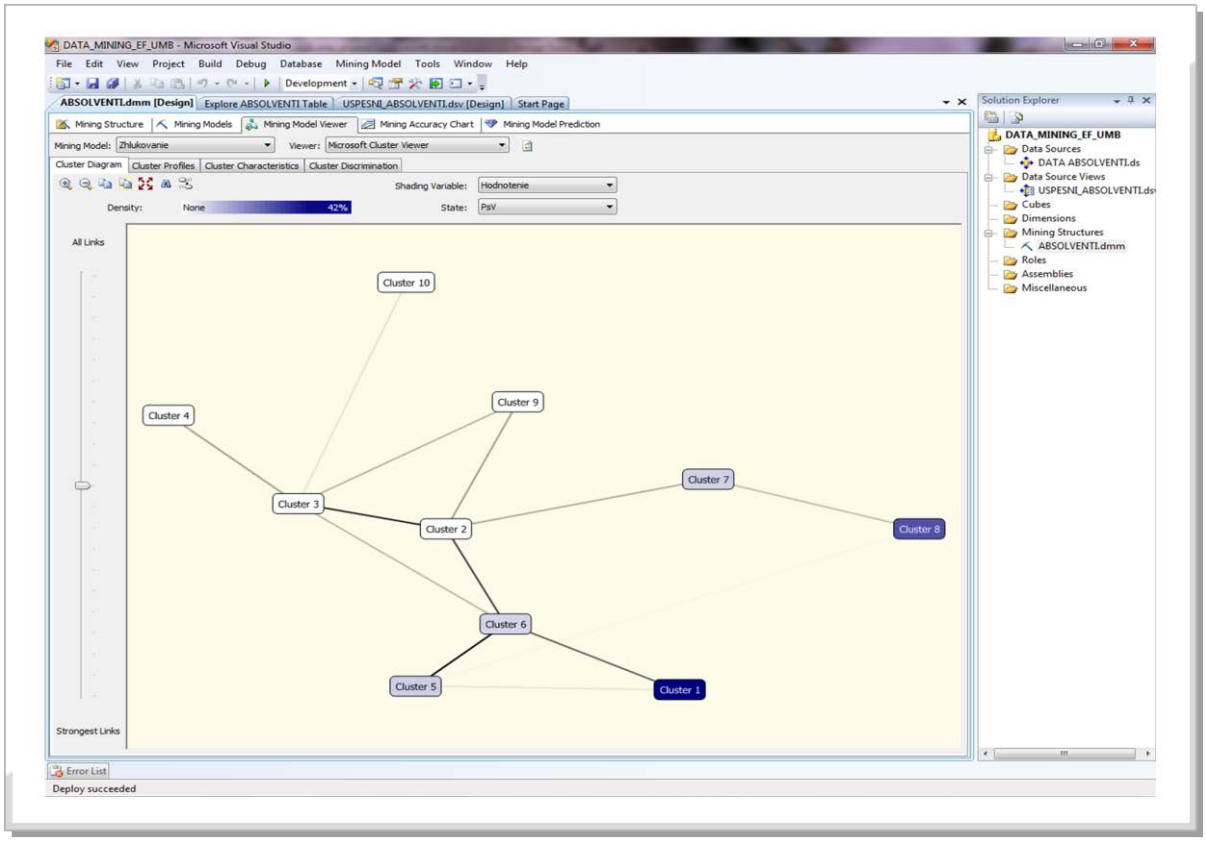
Obrázok č. 7: Výber predikovaného atribútu a atribútov

Zdroj: vlastné spracovanie.

Po výbere vstupných stĺpcov môžeme ešte špecifikovať, či obsahujú diskrétne alebo spojité hodnoty. Typickými spojitými hodnotami sú napríklad vek alebo suma obeživa, príkladom diskrétnej hodnoty môže byť napríklad počet detí. V našom prípade zmeníme pre stĺpec Priemer predurčený typ Continuous na Discrete. Ďalšia záložka v sprievodcovi nás informuje o tom, že dáta budú náhodne rozdelené na tréningovú a testovaciu množinu. Vzhľadom na veľkosť nášho datasetu, stačí, ak ponecháme veľkosť tréningovej množiny na vopred prednastavenej hodnote 30% bez obmedzenia počtu. V sumarizačnom dialógu premenujeme náš data miningový model podľa použitého algoritmu na Zhlukovanie, namiesto priradeného názvu ABSOLVENTI.

Pomaly sa blížime k cieľu nášho úsilia. Po nastavení všetkých parametrov môžeme data miningový model zostaviť a umiestniť na server cez položku záložky Build – Deploy DATA_MINING_EF_UMB. Táto časť procesu analýzy je najzaujímavejšia. Po zostavení modelu sa upriamime na záložku Mining Model Viewer. Tá je riešená univerzálne pre prehliadanie výsledkov činností všetkých algoritmov. Máme pred sebou diagram, v ktorom jednotlivé uzly znázorňujú zhluky a spojnice väzby medzi nimi. Keď nastavíme parameter Shading Variable na atribút Hodnotenie a parameter State na hodnotu PsV, farba uzla

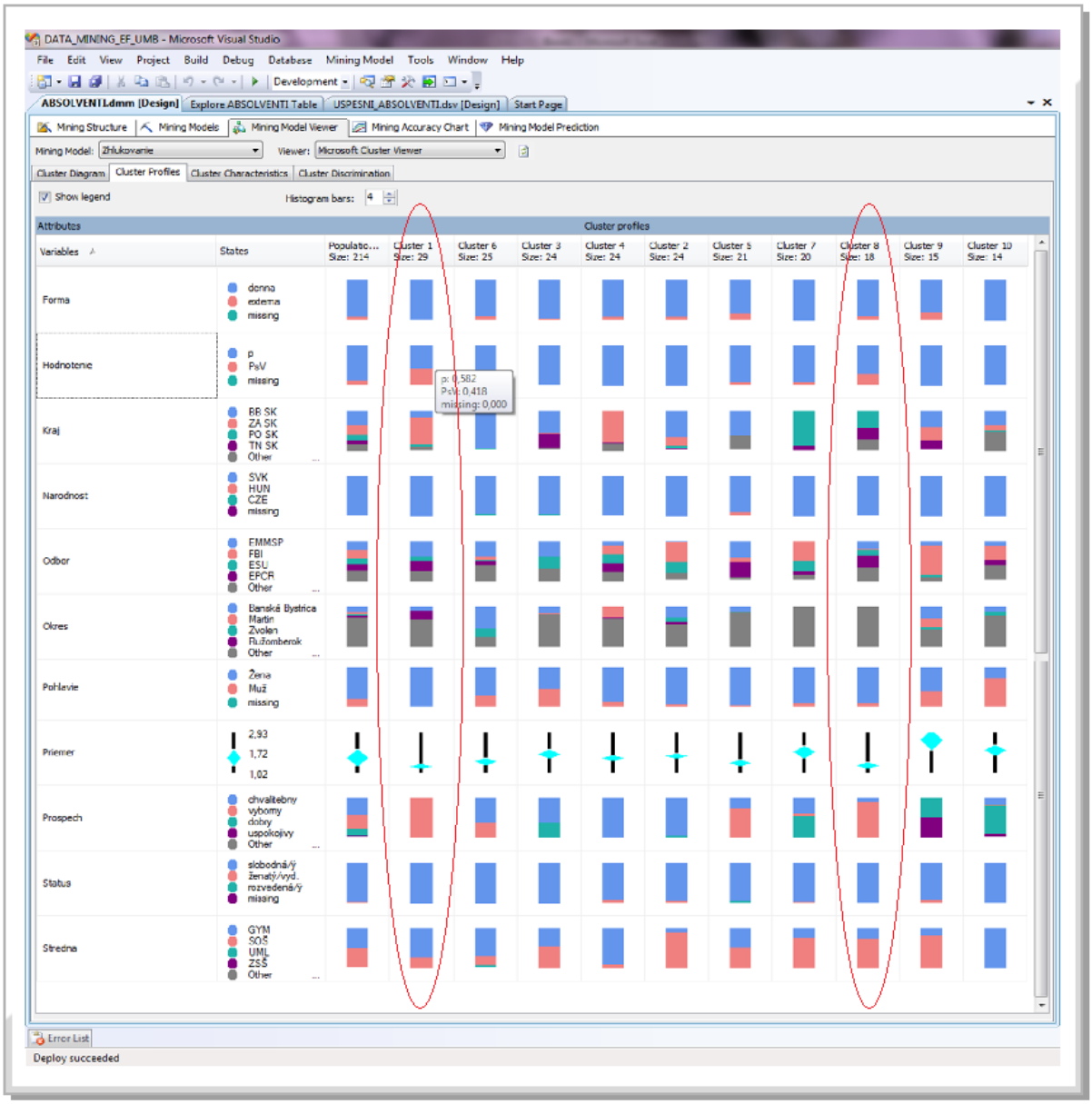
znamená frekvenciu výskytu nastavenej veličiny v zhluku. Čím je odtieň tmavší, tým vyššia je frekvencia výskytu sledovaného javu.



Obrázok č. 8: Záložka Mining Model Viewer - diagram zhlukovacieho algoritmu (Hodnotenie). Zdroj: vlastné spracovanie.

Zhlukovací algoritmus pracuje na princípe zoskupovania záznamov, pozorovaní, alebo prípadov vo forme dát, do tried na základe podobných alebo rovnakých vlastností, ktorými sa vyznačujú. Zhluk (klaster) potom predstavuje súbor záznamov, ktoré sú si jeden druhému podobné, ale zároveň rozdielne od záznamov v iných klastroch. Silu väzieb medzi jednotlivými zhlukmi môžeme zistiť veľmi jednoducho, a to pomocou posúvania „potenciometra“ v ľavej časti pracovnej obrazovky. Ako môžeme z daného výstupu vidieť, najvyššia frekvencia výskytu prípadov (absolventov) spĺňajúcich nami požadované atribúty je v klastroch 1 a 8.

Aby sme sa dozvedeli o týchto skutočnostiach viac, prepneme sa zo záložky Cluster Diagram na záložku Cluster Profiles, kde sa nám zobrazí prehľadný diagram charakterizujúci každý uzol (zhluk).



Obrázok č. 9: Diagram charakteristiky jednotlivých zhlukov.

Zdroj: vlastné spracovanie.

V tomto momente sme dosiahli náš cieľ. Z daného výstupu sme schopní stanoviť pomerne presný profil úspešného absolventa EF UMB. Vďaka výsledku data miningového zhlukovacieho modelu totiž vieme, že najviac študentov zodpovedajúcich nami určeným kritériám úspešného absolventa (Hodnotenie = PsV) sa nachádza v klastroch číslo 1 (41,8%) a číslo 8 (28,4%). K daným zhlukom stačí už len prideliť zodpovedajúce hodnoty jednotlivých atribútov, čo za nás spravil program. Tieto atribúty totiž predstavujú osobné charakteristiky daného študenta (pohlavie, druh absolvovanej strednej školy, študijný odbor..). Keďže daní úspešní absolventi boli začlenení do príslušných zhlukov práve na základe istých

spoločných čŕt a znakov, môžeme povedať, že študenti vyznačujúci sa práve danou odhalenou kombináciou jednotlivých atribútov, majú predpoklady stať sa úspešnými absolventmi Ekonomickej fakulty. Aby sme boli konkrétni, tak na základe daného výstupu môžeme povedať nasledovné:

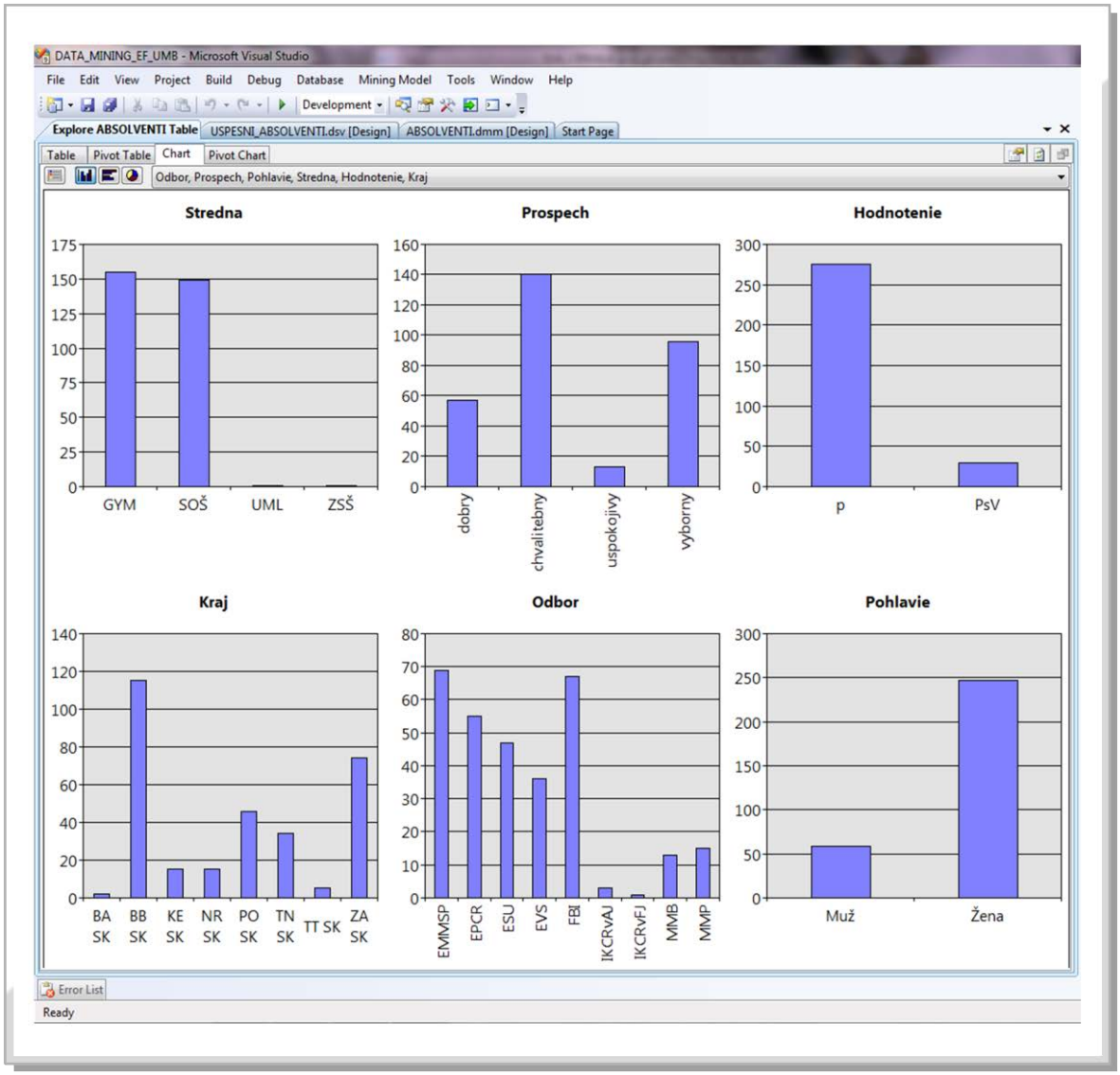
- vysoké predpoklady stať sa úspešným absolventom štúdia na EF UMB so záverečným hodnotením „prospel s vyznamenaním“ má študent, ktorým je slobodná (100%) žena (95%), Slovenka (96,2%), študujúca dennou formou (100%) druhý stupeň štúdia v študijnom odbore Ekonomika a manažment malých a stredných podnikov (39,8%), pochádzajúca zo Žilinského kraja (66,4%) v rámci ktorého majú najvyššie zastúpenie okresy Ružomberok (33,88%), Dolný Kubín (24,25%), Žilina (19,43%) a Námestovo (14,61%), kde absolvovala gymnázium (73%) a ktorá dosahovala počas svojho štúdia výborný prospech (99,7%) čomu zodpovedá aj hodnota váženého aritmetického priemeru 1,31 +/- 0,12. Percentá v zátvorkách predstavujú podiel zodpovedajúcej hodnoty atribútu na všetkých prípadoch v danom klastri.

Aj keď sme na tomto mieste dosiahli vytýčený cieľ, možnosti a výsledky, ktoré nám ponúka daný dataminingový model nie sú týmto pádom ani zďaleka vyčerpané. K ďalším zaujímavým zisteniam dospejeme po vyhodnotení poradia úspešnosti jednotlivých študijných odborov z hľadiska hodnotenia a prospechu ich študentov. Zhluk číslo 1, ktorý je z hľadiska počtu študentov s excelentnými študijnými výsledkami najpočetnejší, má napríklad nasledovné poradie úspešnosti jednotlivých študijných odborov podľa výšky podielu na ich celkovom počte:

- na prvom mieste sa nachádza študijný program Ekonomika a manažment malých a stredných podnikov (EMMSP) s podielom 39,8% študentov,
- nasleduje študijný program Ekonomika podnikov cestovného ruchu (EPCR) s 26,5% podielom,
- trojicu najúspešnejších študijných programov uzatvára Marketingový manažment podniku (MMP) so 16,4%,
- štvrté miesto patrí Ekonomike a správa území (ESU) s 9,3%,
- piate miesto obsadila Ekonomika verejných služieb (EVS) s 5,3%,

Ostané študijné odbory nemali v danej kategórií zastúpenie. Pre objektívne vyhodnotenie poradia úspešnosti jednotlivých študijných programov je však potrebné

zohľadniť aj počet študentov, ktorí daný odbor študujú. Tie môžeme vidieť na nasledujúcom obrázku na príslušnom grafe Odbor.



Obrázok č. 10: Séria grafov vzťahujúcih sa k analyzovaným absolventom.

Zdroj: vlastné spracovanie.

Ako vidíme, frekvencia výskytu najlepších študentov až na jednu výnimku korešponduje priamoúmerne s početnosťou daných študijných programov. Tou výnimkou sú študenti študijného programu Financie, bankovníctvo a investovanie. Tí majú spolu so študentmi odboru Ekonomika a manažment malých a stredných podnikov najvyššie zastúpenie medzi analyzovanými absolventmi (zhodne približne po 22%). No zatiaľ čo podiel absolventov EMMSP bolo v klastri číslo 1 (najúspešnejší študenti) až 39,8%, v prípade programu FBI je tento podiel nula. Vypovedá to pravdepodobne o rozdielnej náročnosti medzi

jednotlivými študijnými programami. Preto sa teraz zameriame na vytvorenie profilu študentov jednotlivých študijných programov.

Aj keď sme dosiahli vytýčený cieľ, možnosti a výsledky ktoré nám ponúka daný dataminingový model nie sú týmto pádom ani zďaleka vyčerpané. Ako sme už uviedli, pokúsime sa teraz z dostupných výsledkov zostaviť profil absolventov podľa ich študijných odborov. Z dôvodu získania relevantných výsledkov, si z deviatich študijných odborov zahrnutých v datasete pre túto analýzu vyberieme len tie, pre ktoré máme k dispozícií dostatok pozorovaní. Budú to študijné odbory v ktorých študovalo aspoň 10% analyzovaných absolventov. Tomu kritériu vyhovujú nasledovné študijné odbory:

- Ekonomika a manažment malých a stredných podnikov: 23,4%
- Financie, bankovníctvo a investovanie: 22,4%
- Ekonomika a správa území: 16,4%
- Ekonomika podnikov cestovného ruchu: 15,9%
- Ekonomika verejných služieb: 11,2%

V uvedených študijných odboroch študovalo spolu dovedna 89,3% všetkých nami skúmaných absolventov. Teraz môžeme na základe obdŕžaných data miningových výsledkov pristúpiť k stanoveniu profilu absolventov jednotlivých študijných programov. V našom clusterinogovom modeli nastavíme parameter Shading Variable na Odbor a parameter State postupne na FBI, ESU, EVS, EMMSP a EPCR. Týmto spôsobom zistíme, v ktorých zhlukoch je najvyššie zastúpenie študentov daných odborov, sú nimi tieto:

- pre FBI je to klaster číslo 9 s podielom študentov 74%,
- pre ESU je to klaster číslo 3 s podielom študentov 16,4%,
- pre EVS je to klaster číslo 3 s podielom študentov 20%,
- pre EMMSP je to klaster číslo 5 s podielom študentov 41%,
- a pre EPCR klaster taktiež klaster číslo 5 s podielom študentov 40%.

Na základe získaných výsledkov môžeme povedať, že 74% absolventov študijného programu Financie, bankovníctvo a investovanie ktorí tvoria daný zhluk, sú študenti, ktorými sú slobodné (91,7%) ženy (61,6%), Slovenky (100%), študujúce dennou formou štúdia (100%), z ktorých väčšina absolvovala strednú odbornú školu (82,7%) a ktorí dosahovali počas svojho štúdia najčastejšie prospech dobrý (49%) alebo uspokojivý (50,8%) čomu zodpovedá aj hodnota váženého aritmetického priemeru 2,57 +/- 0,52 a ich záverečné

hodnotenie bolo prospel (100%). Títo študenti pochádzajú z nasledujúcich samosprávnych krajov a v rámci nich najčastejšie z uvedených okresov:

- Banskobystrický kraj (41,2%): okresy Banská Bystrica (75,73%) a Zvolen (20,38%),
- Žilinský kraj (36,2%): okresy Martin (59,94%) a Liptovský Mikuláš (40,05%),
- Trenčiansky kraj (22,6%): okresy Prievidza (53,98%) a Púchov (32,3%).

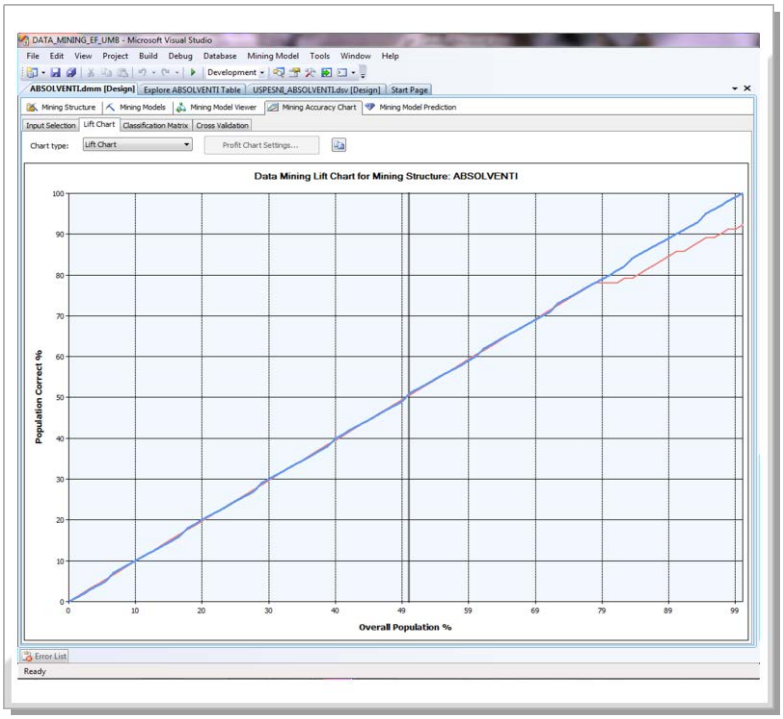
Ako vidíme, klaster číslo 5 je takmer na viac ako 81% tvorení študentmi odborov EMMSP (41,4%) a EPCR (39,8%). Znamená to, že títo študenti sú si čo sa týka zvyšných atribútov ako je ich študijný odbor veľmi podobní. Na základe získaných výsledkov môžeme povedať, že pre absolventov daných študijných odborov je charakteristické, že sú to slobodné (95,1%) ženy (95,5%), Slovenky (89,8%), študujúce dennou formou štúdia (84,1%), z ktorých viac ako polovica absolvovala strednú odbornú školu (51,1%) a ostatní gymnázium (48,9%) a počas svojho štúdia dosahovali najčastejšie prospech výborný (73,7%) alebo chválitebný (25,8%) čomu zodpovedá aj hodnota váženého aritmetického priemeru 1,48 +/- 0,16 a dosiahli záverečné hodnotenie bolo prospel (92%). Títo študenti pochádzajú najmä z nasledujúcich samosprávnych krajov (93,6%) a v rámci nich najčastejšie z príslušných okresov:

- Banskobystrický kraj (61,8%): okresy Brezno (40,78%) a Banská Bystrica (21,67%),
- Nitriansky kraj (23,5%): okresy Topoľčany (58,3%) a Levice (40%),
- Trnavský kraj (8,3%): okresy Dunajská Streda (57,83%) a Skalica (42,2%).

Obdobným spôsobom určíme aj profil absolventov študijného programu Ekonomika verejných služieb a Ekonomika a správa území. Tieto odbory nemajú samostatne dominantné zastúpenie ani v jednom klastri, preto bude výhodnejšie nájsť taký, v ktorom majú spolu väčšinový podiel. Tým je klaster číslo 3, ktorý je tvorený na 51,7% práve študentmi týchto študijných odborov (ESU 31,3%, EVS 20,4%). V týchto študijných odboroch mali muži v porovnaní s predošlými odbormi najvyššie zastúpenie (44,7%), ich status znel slobodný (100%) , boli slovenskej národnosti (96%), ich štúdium prebiehalo prevažne dennou formou štúdia (97,1%), viac ako polovica absolvovala strednú odbornú školu (53,4%) a ostatní gymnázium (46,6%) a počas svojho štúdia dosahovali najčastejšie prospech chválitebný (62,2%) alebo dobrý (37,8%) čomu zodpovedá aj hodnota váženého aritmetického priemeru 1,90 +/- 0,23 a ich záverečné hodnotenie znelo prospel (100%). Títo študenti pochádzajú väčšinou z nasledujúcich samosprávnych krajov (91,6%) a v rámci nich najčastejšie z uvedených okresov:

- Banskobystrický kraj (55,8%): okresy Banská Bystrica (33,33%) a Poltár (20,95%) a Lučenec (11,65%),
- Trenčiansky kraj (35,8%): okresy Trenčín (24,86%), Považská Bystrica (22,35%) a Nové Mesto nad Váhom (19,83%).

Presnosť daného modelu si môžeme overiť na záložke Mining Accuracy Chart, ktorá poskytuje možnosť zobrazovania prehľadných Lift Chart grafov pre testovanie data miningových modelov. Krivka pre správne navrhnutý model by samozrejme mala byť čo najbližšie ideálnej (modrej) krivke a jej sklon by sa mal plynule znižovať. V opačnom prípade je nesprávne zvolený algoritmus, alebo zle navrhnutý model.



Obrázok č. 12: Záložka Mining Accuracy Chart.
Zdroj: vlastné spracovanie

V našom prípade bol algoritmus vybraný správne, nakoľko jeho krivka kopíruje podstatnú časť krivky ideálneho modelu. Istý odklon môžeme pozorovať v poslednej štvrtine prípadov.

Záver

V predkladanej práci sme uviedli praktické ukážky možností práce s dátami, prostredníctvom riešení založených na platforme Business Intelligence. Podrobne sme sa venovali opisu celého priebehu procesu extrakcie potrebných dát z databázy prostredníctvom na to určených nástrojov a techník. Následne sme v prostredí MS SQL Server Business Intelligence Development Studio pristúpili k spusteniu nového analytického projektu, ktorý nadväzoval na predchádzajúce výsledky našej práce a spočíval vo vytvorení data miningového modelu zo získaných údajov. Po uskutočnení všetkých potrebných krokov a nastavení v procese analýzy sme sa dopracovali k želaným zisteniam, na základe ktorých sme dosiahli vytýčený cieľ a to stanovenie profilu úspešného absolventa Ekonomickej fakulty Univerzity Mateja Bela.

Veríme, že predkladaná práca bude prínosom pre každého, kto sa rozhodne oboznámiť s touto zaujímavou a veľmi perspektívnou oblasťou riešení Business Intelligence.

Zoznam použitej literatúry

ČÁBELA, M. (1. December 2009). Využite svoju BI naplno! *Infoware* , s. 30.

HEČKOVÁ, G. (1. December 2009). Výhody zavedenia riešenia BI v organizácií. *Infoware* , s. 28.

http://www.u-sluno.eu. (dátum neznámy). Cit. 26. February 2013. Dostupné na Internete: U&SLUNO: http://www.u-sluno.eu/business-intelligence.html?setLang=sk

INMON, W. (2002). *Building the Data Warehouse.* Indianopolis: John Wiley and Sons.

KADLEC, T. (27. 2 2013). *http://www.logica.cz*. Cit. 27. 2 2013. Dostupné na Internete: http://www.logica.cz: http://www.logica.cz/we-are-logica/media-centre/articles/2012/jak-optimalne-vyuzit-bi-ve-financnim-sektoru/

KACHAŇÁK, D. (January - February 2013). Nástroje na správu big data. *Infoware* , s. 15.

KRISHNNAN, M. (2013). *SQL Server Migration Assistant 2008 SSMA*. Cit. 28. 3 2013. Dostupné na Internete: MSSQLTips: http://www.mssqltips.com/sqlservertip/2230/sql-server-migration-assistant-2008-ssma/

LACKO, Ľ. (2009). Ako vydolovať z údajov čo najviac informácií na podporu rozhodovania. *Infoware , 6-7* (ISSN 1335-4787), s. 40-41.

LACKO, Ľ. (1. January - February 2013). Ako z dát získať informácie. *Infoware* , s. 14-15.

LACKO, Ľ. (2009). *Business Intelligence v SQL Servreru 2008.* Brno: Computer Press.

LACKO, Ľ. (2009). Business Intelligence: Na rozhraní ekonomických vied a mágie. *Infoware* , 22 - 23.

LACKO, Ľ. (dátum neznámy). *Microsoft SQL Server 2008.* Cit. 22. February 2013. Dostupné na Internete: Microsoft SQL Server 2008: www.microsoft.com/slovakia/sqlserver

LAROSE, D. T. (2005). *Discovering knowledge in data : an introduction to data mining.* Hoboken: John Wiley & Sons, Inc.

NOVOTNÝ, O.; POUR, J.; SLÁNSKÝ, D. (2005). *Business Intelligence: jak využít bohatství ve vašich datech.* Praha: Grada Publishing.

POUR, J.; MARYŠKA, M.; NOVOTNÝ, O. (2012). *Business Intelligence v podnikové praxi.* Praha: Professional Publishing.

ŠIMEK, R. (1. January - February 2013). Keď je dát priveľa... *Infoware* , s. 18.

URAMOVÁ, Ľ. (1. December 2009). Od skladovania údajov k inteligentnej organizácií. *Infowareq* , s. 26-27.

VERCELLIS, C. (2009). *Business Intelligence: Data Mining and Optimization for Decision Making.* West Sussex, United Kingdom: A John Wiley and Sons, Ltd., Publication.

www.dynacrongroup.com. (dátum neznámy). Cit. 23. February 2013. Dostupné na Internete: Dynacron Group: http://www.dynacrongroup.com/2011/04/bi-dw-architecture-fundamentals/