Sem vložte zadání Vaší práce.

České vysoké učení technické v Praze
Fakulta informačních technologií
Katedra teoretické informatiky

Diplomová práce

Analýza výsledků absolventů středních škol na VŠ

Bc. Eliška Hrubá

Vedoucí práce: Ing. Pavel Kordík, Ph.D.

7. května 2014

Poděkování

Tímto bych ráda poděkovala vedoucímu své diplomové práce Ing. Pavlu Kordíkovi, Ph.D. za ochotu a věcné připomínky, Ing. Stanislavu Kuznetsovi za pomoc při integraci dat do datového skladu a Ivaně Dolejšové za vysvětlení významu dat v přihláškách. V neposlední řadě bych pak chtěla poděkovat svému příteli za trpělivost a podporu během psaní této práce.

Prohlášení

Prohlašuji, že jsem předloženou práci vypracoval(a) samostatně a že jsem uvedl(a) veškeré použité informační zdroje v souladu s Metodickým pokynem o etické přípravě vysokoškolských závěrečných prací.

Beru na vědomí, že se na moji práci vztahují práva a povinnosti vyplývající ze zákona č. 121/2000 Sb., autorského zákona, ve znění pozdějších předpisů, zejména skutečnost, že České vysoké učení technické v Praze má právo na uzavření licenční smlouvy o užití této práce jako školního díla podle § 60 odst. 1 autorského zákona.

V Praze dne 7. května 2014 .....

Odkaz na tuto práci

Hrubá, Eliška. *Analýza výsledků absolventů středních škol na VŠ.* Diplomová práce. Praha: České vysoké učení technické v Praze, Fakulta informačních technologií, 2014.

Abstract

This work deals with the processing of data from the system Přihláška ČVUT, which is used for submitting application forms to the study. In the first part we will focus on uploading the data into the data warehouse in which there is cleaning and integration with other data. Cleaned data then will be used for analysis with a focus on high schools and for creating dashboards

Keywords data, data warehouse, integration, Kimball, Inmon, data mining, business intelligence, analysis, dashboards

Abstrakt

Tato práce se zabývá zpracováním dat ze systému Přihláška ČVUT, který slouží pro podávání přihlášek ke studiu. V první části se zaměříme na nahrání těchto dat do datového skladu, během kterého dojde k jejich vyčištění a integraci s dalšími daty. Vyčištěná data poté využijeme pro analýzy se zaměřením na střední školy a pro tvorbu dashboardů.

Klíčová slova data, datový sklad, integrace, Kimball, Inmon, dolování dat, business intelligence, analýzy, dashboardy

Obsah

Seznam obrázků

Seznam tabulek

Úvod

Využívání dat z procesních systémů, které primárně slouží pro zajištění chodu firmy, k dolování dat (ang. Data Mining) a získávání informací (typicky o chování zákazníků), je dnes velmi populární téma. Skryté závislosti v datech mohou výrazně ovlivnit rozhodování vedení, neboť se může nalézt lepší strategie pro celkový záměr, nalézt problémy, které je nutné řešit, nebo odhalit nejpřínosnější odvětví.

Speciálním případem firem jsou vzdělávací instituce, kde se dolování dat také velmi rozšířilo, a dokonce dalo vzniknout úplně novému odvětví anglicky nazvanému Educational Data Mining. Nejrozšířenější je v oblasti e-learningových kurzů, kde je k dispozici velké množství dat o studentech, ale čím dál oblíbenější je i u klasických škol. Takový případ je i v případě Fakulty informačních technologií ČVUT, která má poměrně vysokou úmrtnost studentů po prvním semestru a pomocí analýz je možné zjistit, zdali jsou správně nastaveny podmínky přijímacího řízení, hodnocení jednotlivých předmětů, apod.

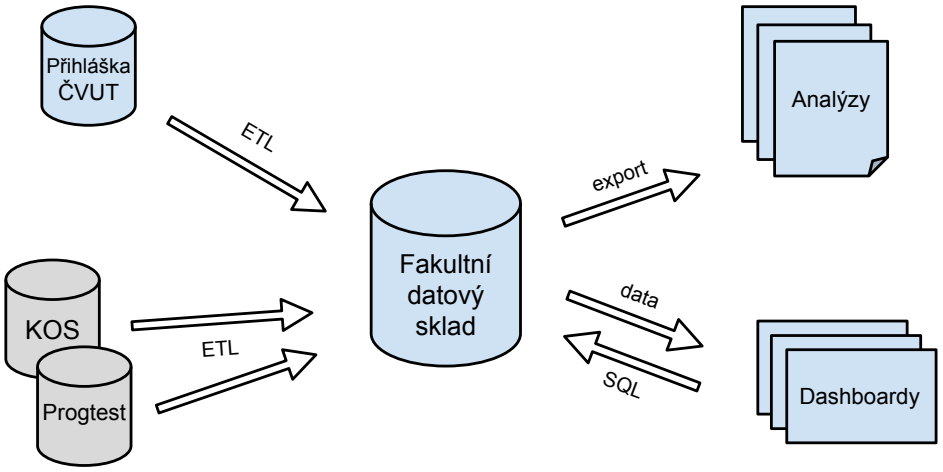
Cíl práce

Na univerzitě existuje mnoho systémů, ze kterých je možné data získat, ty jsou nestejnorodé a často může trvat dlouho, než by se analytik k těmto datům dostal. Proto vzniknul fakultní datový sklad [9], který má za úkol sjednocovat data z různých zdrojů a integrovat je do jednoho celku. Tato data jsou již předzpracovaná a hlavně vyčištěná, proto jsou vhodná i pro analýzy.

Ve skladu však ještě nejsou zakomponované všechny systémy, které fakulta používá, proto je cílem této práce se zaměřit na další z nich, Přihláška ČVUT. Ta slouží k podávání přihlášek ke studiu a uchovávání informací z přijímacího řízení. Tato data jsou velmi důležitá, protože u studentů známe pouze jejich studijní výsledky na fakultě, ale nic před tím. Přihláška ČVUT nám umožní

nahlédnout i do studentovy historie – jak si vedl na střední škole, odkud na fakultu vlastně přišel a můžeme se podívat, jak tato data ovlivňují jeho další studium na fakultě.

Střední školy jsou obzvlášť důležité, protože se může zjistit, že většina studentů určité střední školy má problémy s konkrétním předmětem, zřejmě jim tedy chybí nějaké znalosti, a fakulta jim může nabídnout pomoc ve formě doučování nebo přípravných kurzů, případně střední škola může látku přidat do výuky. V opačném případě, pokud budou studenti nějaké střední školy excelovat, je to vhodná střední škola pro propagaci a marketing, neboť fakulta má samozřejmě zájem o kvalitní studenty.



Obrázek 0.1: Souhrnné schéma práce

Tato práce je členěna do tří hlavních oblastí (viz obrázek 0.1): datový sklad, analýzy a dashboardy. Základem všeho je datový sklad, protože nahrání dat do skladu zároveň obsahuje seznámení se s daty a jejich předzpracování. Poté můžeme ze skladu získat např. export dat pro analýzy. Data jsou předzpracovaná, odpadá tedy nejnáročnější část dolování dat. Doteď bylo nutné data předzpracovat před každou analýzou a ručně je zintegrovat s dalšími požadovanými daty a tím samozřejmě narůstala náročnost práce.

Analýzy ale nejsou jediná možnost, jak z dat získat nějaké informace. Poslední částí této práce je vytvoření dashboardů, které zobrazí data ze skladu ve vizuální a snadno čitelné podobě. Vedení fakulty pak může sledovat vývoj studentů za různé semestry, porovnávat známky z jednotlivých předmětů nebo se zaměřit na konkrétní střední školy.

Kapitola 1

Rešerše

Data ze vzdělávacího prostředí (školy, kurzy, apod.) se stávají velmi populárními pro data mining. Jedná se totiž o poměrně přesná a velmi obsáhlá data, která mohou být důležitým vodítkem pro předpovědi výkonů studentů, doporučení pro nastupující studenty nebo mohou jednoduše poskytovat zpětnou vazbu pro instituci (např. univerzitu). Souhrnně se tato oblast nazývá Educational Data Mining (EDM) [18].

Problém predikce výkonu studentů je poměrně složitý z toho důvodu, že je zde mnoho rizikových faktorů, které ovlivňují studentův výkon – demografický, kulturní, sociální, psychologický, apod. Důležité je také, jak kvalitního se mu dostalo vzdělání doposud (a jistě se všichni shodneme na tom, že škola je velmi důležitou částí života každého z nás) a jaké má rodinné zázemí.

V posledních letech se tento výzkum rozvíjí a bylo vytvořeno několik studií, které se EDM zabývají. Velká část těchto prací se zabývá daty z e-learningových kurzů, tudíž jsou k dispozici například i průběžné výsledky studenta nebo jak dlouho mu trvalo splnit daný test. Jsou ale i takové práce, které se zabývají studenty základních, středních nebo vysokých škol.

1.1 Studie z jiných škol/institucí

1.1.1 Predikce neúspěšných studentů

První ze zkoumaných studií [12] se zabývá nalezením faktorů, které vedou k neúspěchu (čili ukončení) studia a tedy vedou k rozpoznání studentů, kteří budou mít sklony studium nedokončit. Těm může být nabídnuta pomoc v průběhu studia.

Pro tuto studii byla využita data o studentech střední školy, kteří navštěvují přípravný kurz Univerzity v Zacatecas v Mexiku. Zde je střední škola tříleté

všeobecné vzdělání, které studenty připravuje na studium na vysoké škole. Zaměření je pouze na studenty prvního ročníku (tedy 15–16 let), protože právě v tomto roce je největší úmrtnost studentů, a to zhruba 10 %. Celkový počet záznamů je 670, avšak zkoumaná data obsahují velké množství atributů (77), které obsahují informace o studentovi, střední škole a získaných známkách, tak i o rodinném zázemí (např. zaměstnání a vzdělání rodičů).

Tato data byla rozdělena do 10 skupin, které byly využité pro křížovou validaci. Vzhledem k velkému množství příznaků byla data rozdělena na následující skupiny:

- 10 trénovacích a testovacích skupin se 77 příznaky
- 10 trénovacích a testovacích skupin s nejlepšími 15 příznaky (feature selection)
- 10 trénovacích a testovacích skupin s nejlepšími 15 příznaky po použití techniky SMOTE pro vybalancování dat

V práci je zkoumáno celkem 10 algoritmů programu Weka DM Software [5], z toho 5 rozhodovacích stromů a 5 algoritmů využívajících pravidla a následně evoluční algoritmus využívající gramatiku ICRM (Interpretable Classification Rule Mining) vyvinutý autory, který má celkem tři verze.

	Algoritmus	**Přesnost postupivší**	**Přesnost nepostupivší**	**Celková přesnost**
Všechny	ADTree[1]	**99,7 %**	76,7 %	**97,6 %**
příznaky	ICRM v3	84,4 %	**93,3 %**	85,2 %
Nejlepší	ADTree	**99,2 %**	78,3 %	**97,3 %**
příznaky	ICRM v1	92,0 %	**93,3 %**	92,1 %
Nejlepší	ADTree	**98,2 %**	86,7 %	**97,2 %**
příznaky,	ICRM v2	98,0 %	96,1 %	**97,2 %**
SMOTE	ICRM v3	86,7 %	**98,7 %**	92,7 %

Tabulka 1.1: Výsledky měření

Pro predikci studentů, kteří postoupí, se nejlépe osvědčil rozhodovací strom ADTree a pro predikci těch, kteří nepostoupí, byl nejvhodnější algoritmus ICRM. Po vybalancování dat mají oba algoritmy zhruba stejnou přesnost, čili autoři dosahují vynikajících výsledků jak pro predikci studentů, kteří nepostoupí (což byl jejich původní záměr), tak pro predikci těch, kteří postoupí.

[1] Alternating Decision Tree

Při porovnání výsledků všech algoritmů se obecně více osvědčily rozhodovací stromy než jiné metody.

1.1.2 Atributy pro detekci rizikových studentů

Na Technické univerzitě v Eindhovenu v Nizozemí [4] se také zabývali podporou jejich studentů v prvním roce, zejména se snažili detekovat rizikovou skupinu studentů, kteří mohou být úspěšní, ale potřebují pomocnou ruku, bez které by skončili hned po prvním semestru. Úmrtnost studentů v jejich bakalářském programu je kolem 40 %.

Autor v úvodu studie uvádí několik klíčových atributů, které se ve výzkumech prokázaly jako nejvýznamnější při studiích:

- pohlaví (u technických oborů),
- věk při nástupu do studia,
- výsledek ze zkoušky před nástupem na školu (tedy maturita),
- typ předchozího vzdělání,
- zdali byl student přijat na preferovaný obor,
- typ finanční podpory,
- vzdělání otce,
- bydliště v okolí univerzity.

Tyto atributy opět potvrzují to, že úspěch závisí na mnoha různorodých faktorech, včetně zázemí studenta. Naneštěstí na výše zmíněné univerzitě neměli všechna data k dispozici (zejména informace o rodinném zázemí, jak daleko musí student dojíždět a u některých nemají ani známku ze středoškolského studia).

Celkový počet záznamů je 648 a tato data byla rozdělena do svou skupin. V první jsou studenti, kteří studovali přípravnou školu pro univerzitu a u nich jsou známé veškeré známky a jejich studijní historie. Druhá skupina obsahuje studenty a jejich známky na univerzitě, celkově 74 atributů.

Pro klasifikaci je použito několik algoritmů: rozhodovací stromy, Bayesův klasifikátor, logistický model, Random Forest a další, opět pomocí pomocí Weka DM Software a také používají křížovou validaci, protože dataset je příliš malý na rozdělení na trénovací a testovací data.

Výsledky na prvním datasetu (tedy historie studenta) dosáhly přesnosti kolem 70 %, přičemž nejvíce měl Bayesův klasifikátor (71,1 %). U druhého datasetu jsou výsledky o něco lepší, průměrně kolem 75 %, nejlepší byl rozhodovací strom (80,8 %), který měl jako kořen známku z lineární algebry a hranici úspěšnosti absolvování kurzu (55 %). Při dolování na celém datasetu opět mírně vítězí rozhodovací stromy s nejlepším výsledkem 79,9 %.

Podle výsledků bylo také zjištěno, že univerzitní známky jsou lepším prediktorem než známky z předchozích studií a také to, že nemá cenu mít datasety oddělené. Dále jsou ve studii popsány podrobněji právě rozhodovací stromy a sledováno více metrik, než pouze přesnost, a informace o možných příčinách chybné klasifikace (např. to, že 0 ve výsledku nemusí znamenat 0 bodů, ale jiný způsob zkoušky, apod.). To je ale velmi konkrétní pro jejich způsob hodnocení, proto se tím zde nebudeme již zabývat.

1.1.3 Faktory ovlivňující studijní výsledky

Velmi zajímavé na této práci [11] jsou data, která mají autoři k dispozici. Kromě klasických studijních výsledků mají také dovednosti jako je asertivita, schopnost řídit tým, odolávat stresu, apod. Data byla sesbírána formou dotazníků od různých institucí spolupracující s GGSIP[2], celkem bylo k dispozici 250 záznamů s 25 atributy.

Autoři zvolili jako nástroje dva rozhodovací stromy (J48 a Random Tree), na základě rešerše, která prokázala, že rozhodovací stromy jsou v této oblasti přesnější nebo minimálně stejně tak vhodnými klasifikátory jako bayesovské klasifikátory či neuronové sítě. Dosáhli vynikajících výsledků, a to 88,4 % správně klasifikovaných vzorů pro J48 a 94,5 % pro Random Tree. Došli také k zajímavým výsledkům: ze všech zkoumaných vlastností má největší vliv právě schopnost vést ostatní lidi a naopak sociálně ekonomické podmínky mají minimální dopad na výkony studentů.

1.1.4 Úspěšní studenti přírodních věd

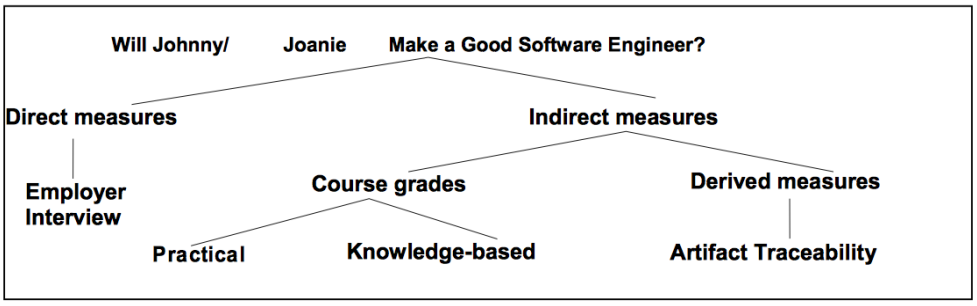
Až do nedávna byla politika univerzit na Novém Zélandu taková, že univerzity nabíraly všechny studenty, kteří splňovali základní podmínky, a ti po prvním roce museli získat alespoň známku B (odpovídá 5 bodům z 9, kdy 9 je rovno A+). To se však změnilo a studenti jsou přijímáni na základě výsledků NCEA[3] testů, kde jsou výkony v jednotlivých předmětech ze střední školy. Tyto testy jsou obecné a nemají technické zaměření. Cílem této studie [2] bylo nalézt vhodné podmínky pro přijetí na přírodní vědy na Viktoriině univerzitě ve Wellingtonu.

[2] Guru Gobind Singh Indraprastha University, Nové Dillí, Indie
[3] National Certificate of Educational Achievement

Pro klasifikaci byl použit rozhodovací strom J48 pomocí Weka DM Software. Autoři se zaměřili na přírodovědné předměty (matematiky, biologie, chemie) a ty rozdělili do několika skupin, které použili jako datasety. Nejdříve zkoušeli binární klasifikátor (uspěl/neuspěl po prvním roce), ale nakonec bylo zvoleno jemnější členění pro dosažení lepších výsledků. Po zhodnocení výsledků jednotlivých předmětů je opět zmínka o tom, že mimo školní výsledky mají vliv na výkony studentů také sociální a kulturní metriky, čili nastává případ, kdy studenti, o kterých se předpokládal výborný studijní výkon, ho nedosahují např. z rodinných či jiných osobních důvodů.

1.1.5 Jsou známky z předmětů vhodným měřítkem?

Na Univerzitě v Kentucky vznikl zajímavý průzkum [6] o tom, jak moc reflektuje známka znalosti studenta, resp. absolventa, softwarového inženýrství. Autoři zde zmiňují dvě základní metody, jak odhadnout studentův potenciál: přímou a nepřímou. První z nich zahrnuje hodnocení zaměstnavatele po pár měsících, co u něj student pracuje. Druhá z nich zahrnuje dvě kategorie: známky z předmětu a metriky, kdy známky z předmětu opět sestávají ze dvou částí: teoretická (midterm, závěrečný test) a praktická (známka z projektu). Mezi další metriky, na které se ve studii zaměřili, je sledovatelnost projektu (traceability of a project), což mimo jiné ukazuje schopnosti studenta dokončit životní cyklus projektu a může tedy sloužit jako vhodný ukazatel pro budoucí uplatnění studenta jako softwarového inženýra. Cílem je zjistit, jak moc je tato metrika korelovaná se známkami z předmětu.



Obrázek 1.1: Dělení metrik pro metody.

Testování proběhlo na 22 projektových skupinách s 85 studenty z Univerzity Waterloo a nebyla nalezena žádný významná vazba mezi známkou a sledovatelností projektu. Známky velmi dobře ukáží, jaké má student programovací schopnosti v určitém jazyce, ale nelze je dobře aplikovat na predikci studentova úspěchu v pracovním prostředí (tedy jestli bude nebo nebude dobrý softwarový inženýr). Na závěr však zmiňují, že je vhodné provést detailnější studie

s lepšími daty (neměli k dispozici informace o jednotlivcích v týmu, tedy kdo měl co na starosti a kolik toho udělal) a zaměřit se také na otázku, zdali je nepřímá metoda vhodná pro predikci úspěchu absolventa.

1.1.6 Shrnutí

V této oblasti existuje opravdu velké množství studií, protože téměř každá z nich musí brát v úvahu jiný způsob hodnocení studenta. V některých zemích mají tu výhodu, že mají jednotné hodnocení středoškoláků a je tedy možné tato hodnocení brát v potaz, snad k tomu jednou dospějeme i my – možná se státními maturitami. Na čem se shodly všechny studie ale je, že známka z předmětu je ovlivněna mnoha faktory a její vypovídající schopnost se také různí od předmětů a zaměření. Je vhodné mít více osobních informací o studentovi, zejména o jeho zázemí a velmi užitečné by bylo i nějaké hodnocení osobnostních dovedností, např. prezentační dovednosti, schopnost vést tým, asertivita, odolnost vůči stresu apod., protože právě ty velmi ovlivňují potenciál studenta. Co se známkování týče, může být velmi nepřesné zejména v oblasti studia informačních technologií, kdy je nezbytné hodnotit jak teoretické, tak praktické znalosti, v některých oblastech (např. softwarové inženýrství) také další metriky, které mohou prokázat studentovy dovednosti.

1.2 Situace na Fakultě informačních technologií ČVUT

Narozdíl od výše uvedených studií, naše fakulta nemá k dispozici tak rozsáhlé množství atributů, úplně chybí informace o psychologickém, sociálním, a tudíž o rodinném zázemí. Můžeme vycházet pouze z informací o střední škole (v lepším případě i z prospěchu) a známek získaných na fakultě.

1.2.1 Hodnocení studenta

ČVUT používá pro hodnocení studenta mezinárodní kreditový systém ECTS[4], kdy je každý předmět ohodnocen počtem kreditů podle zátěže. Obecně tedy platí, že čím více kreditů, tím náročnější je předmět. Na univerzitě se spolu s aritmetickým průměrem používá průměr vážený, kde je počet kreditů zohledněn, avšak tím způsobem, že čím více kreditový předmět, tím větší má známka váhu. Je nutné brát v úvahu také to, že větší počet absolvovaných předmětů za semestr pravděpodobně povede k horším známkám, protože časová náročnost bude vyšší a čas na jeden předmět menší. To je velmi důležité si uvědomit, protože student s více kredity a horšími známkami může být stále lepší než student s méně kredity a lepšími známkami.

[4]European Credit Transfer System

Získané kredity se sčítají a jejich počet určuje postup ve studiu (viz tabulka 1.2) nebo nutnou podmínku pro ukončení studia. Pro možnost ukončení bakalářského studia je nutné získat nejméně 180 kreditů, u magisterského je to adekvátně k počtu semestrů doporučeného studijního plánu 120.

Doba studia	**Bc. studijní program**	**Mgr. studijní program**
za první semestr studia	15	20
za první akademický rok studia	30	40
za každý další ak. rok studia	40	40

Tabulka 1.2: Minimální počet kreditů nutný pro pokračování ve studiu [19]

Klasifikace jednotlivých předmětů opět odpovídá standardu ECTS (viz tabulka 1.3), používá se stupnice A až F, kdy A znamená nejlepší známku a F nejhorší (tedy neprospěl). Podmínka získání známky E, a tedy úspěšného zakončení předmětu, je získat aspoň 50 % bodů z daného předmětu. Některé školy mají tuto hranici vyšší (55 %, 60 % i 70 %), vždy ale záleží na nastavení celkových podmínek.

	A	**B**	**C**	**D**	**E**	**F**
Bodové hodnocení	100–90	89–80	79–70	69–60	59–50	< 50
Číselná klasifikace	1	1,5	2	2,5	3	4
Česky	výborně	velmi dobře	dobře	uspokojivě	dostatečně	nedostatečně
Anglicky	excellent	very good	good	satisfactory	sufficient	failed

Tabulka 1.3: Klasifikační stupnice [19]

Bohužel žádná tato stupnice nedokáže objektivně zachytit skutečný výkon studenta, ke kterému je potřeba mnohem více ukazatelů – např. body ze semestrální práce, kurzy a certifikáty, mimoškolní aktivity studenta (v daném oboru). Proč jsou tyto věci důležité si popíšeme v následující kapitole.

1.2.2 Typy studentů

Na fakultě se vyskytuje několik typů studentů, které můžeme shrnout do několika základních skupin. Nejedná se nutně o rozdělení na dobré nebo špatné studenty, ale každá ze skupin má určité charakteristiky a silné i slabé stránky.

Důležité jsou dva ukazatele – praktické znalosti (semestrální práce, zaměstnání, apod.) a teoretické znalosti (většinou známka z předmětu). Ti nejlepší studenti nemusí být ti, kteří mají nejlepší prospěch.

Studenti s excelentním prospěchem mají některé předpoklady, které jiné skupiny postrádají. Zpravidla je to rodinné zázemí či finanční jistota. Pak mohou soustředit veškerou energii na studia. Spadají sem ale také nadšenci nebo studenti se zaměstnáním v oboru, kdy jim pak škola rozšiřuje obzory. Pokud se jedná o ty první, je na fakultě, aby se na tyto studenty zaměřila a dala jim možnost vyzkoušet si teoretické poznatky v praxi (fakultní projekty, reálné úlohy, apod.), jinak je možné, že jejich potenciál zůstane nevyužit.

Průměrní studenti (z pohledu známek) se najdou všude, proto je důležitý druhý z ukazatelů. Sem totiž mohou spadat právě ti studenti, jejichž semestrální práce budou nadprůměrné, budou aktivní, zúčastní se soutěží, akcí a konferencí. Zpravidla jsou již zaměstnaní, tudíž studium berou pouze jako rozšíření znalostí, případně kvůli titulu, ale nejde jim o vynikající prospěch. Pokud se o něco zajímají, získávají výborné známky, pokud ne, stačí jim, že „prolezou“. Tito studenti jsou z pohledu uplatnění na pracovním trhu ti nejlepší, které fakulta má, protože budou mít dostatečné znalosti (nepodložené známkami) a vynikající praxi.

Poslední kategorií jsou studenti, kteří jakž takž procházejí studiem, známky mají nevalné a s velkou pravděpodobností studium nedokončí. Ty, kteří mají zájem o studium a mohou být úspěšní, je nutné rozpoznat a podat jim pomocnou ruku. Opět mohou mít velmi dobré výsledky ze semestrálních prací, případně vynikat v konkrétních předmětech nebo naopak mají problém s tím či oním předmětem. Pokud se jim v těžkých začátcích pomůže, mohou se stát velmi dobrými studenty.

Dále je také nutné myslet na to, že výsledná známka z předmětu je ovlivněna mnoha faktory – zdravotní stav studenta, celkové množství zkoušek, varianta testu, časová náročnost zaměstnání, apod. Mnoho vynikajících studentů získá špatnou známku jen proto, že jim nesedne konkrétní test a kvůli velkému množství ostatních zkoušek nejdou na opravný termín. Získané body ze semestru by mohly být vhodným ukazatelem pro detekování těchto studentů, kteří jsou dobří i přes špatnou výslednou známku.

Z výše uvedeného tedy plyne, že by bylo vhodné sledovat tyto další ukazatele:

- body ze semestru,
- vynikající semestrální práce (velké bodové ohodnocení, plusové body, upozornění cvičícího, apod.),

- vynikající závěrečné práce (ty pak získávají Cenu děkana FIT za vynikající závěrečnou práci),
- aktivita studentů (zapojení se do fakultních projektů, doučování spolužáků, apod.),
- absolvování kurzů, konferencí (tito studenti prahnou po vzdělání a sami ho vyhledávají i mimo školu, ve svém volném čase se věnují prohlubování znalostí),
- práce v oboru (nejlepší uplatnění vědomostí nabytých na škole a prokázání hodnoty samotného studia).

1.2.3 Vztah fakulty k okolí

Fakulta potřebuje znát tyto faktory úspěchu a neúspěchu, protože je vázána na své okolí. Z jedné strany jsou to střední školy a z druhé strany je to trh práce, tedy zaměstnavatelé.

Pokud vezmeme vazbu na střední školy, může se zjistit, že navstupivší studenti z nějaké konkrétní školy jsou lepší nebo horší než jiní. To se pak může promítnout do spolupráce fakulty s danou školou, jednak kvůli získání jejich absolventů nebo naopak se detekce problému může promítnout do změny studia na střední škole. Další možností je podpora studentů v začátcích z horších středních škol.

Fakulta má zájem získat kvalitní studenty, je tedy nutné vhodně nastavit podmínky přijímacího řízení a detekovat rizikové skupiny. Pokud přes přijímací řízení projde velké množství studentů, kteří skončí v prvním roce studia, není asi něco v pořádku. Snižování nároků fakulty na studenty (a tedy snižování kvality vzdělání) by mělo být až velmi krajním řešením.

Firmy, které mají zájem o absolventy, se naopak mohou obracet na univerzitu, potažmo fakultu, a klást na ni stejné nároky jako klade fakulta na střední školy, ale impuls může být i obrácený, tedy fakulta sleduje pohyb svých absolventů a podle toho může upravovat studijní plány. Toto je spíš vize do budoucna, protože prozatím má fakulta vzhledem k roku svého založení (1. 7. 2009) ještě malé množství absolventů.

Fakulta má tedy zájem o kvalitní studenty, aby mohla produkovat kvalitní absolventy. Proto je pro ni klíčové:

- získat kvalitní studenty
  - tedy nastavit vhodné podmínky přijímacího řízení
  - zaměřit se na střední školy

- detekovat rizikové skupiny
  - studenti konkrétních středních škol mohou mít opakovaný problém s určitými předměty
  - podobné skupiny studentů mohou mít stejné problémy
- vzdělávat studenty tak, aby byli uplatnitelní na trhu práce
  - kontakt s absolventy
  - zpětná vazba od firem

V prvním případě je nutné začít s tím, co vlastně kvalitní student znamená, což se nejlépe zjistí podle uplatnění absolventů. Pak je možné zjistit, jaké vykazovali podobné vlastnosti či známky a podle toho lze určit i vhodné skupiny středoškoláků, na které se lze zaměřit. Jelikož ale prozatím nemáme dostatečné informace o absolventech, musíme se spoléhat na známkování na fakultě jako ukazatel dobrých a špatných studentů.

1.2.4 Systémy na ČVUT

ČVUT, resp. FIT, má k dispozici poměrně velké množství dat z různých informačních systémů. Bohužel většina těchto dat podle výše uvedených studií není úplně vhodným ukazatelem pro měření výkonu studentů, některé ale uchovávají velmi zajímavé informace.

Jelikož se stále pohybujeme v rámci fakulty, budeme se zabývat systémy, které jsou přínosné pro fakultu, nikoli ČVUT jako celek. Tam se to komplikuje tím, že různé fakulty používají jiné systémy, mají jiné nároky na studenty (např. talentové zkoušky u architektů) a není cílem této práce se jimi zabývat.

KOS
Celoškolní systém (Komponenta studia ČVUT) má jednu velmi důležitou funkci, a tou je uchovávání hodnocení studenta. Všechny zapsané předměty, zkoušky, pokusy, známky i jednorázové akce jsou zachyceny právě v tomto systému. Jak si tedy student vedl v průběhu studia a jak studium ukončil, najdeme právě v tomto systému.

Přihláška ČVUT
Další celoškolský systém, který slouží pro podávání přihlášek ke studiu. Potenciál tohoto systému je obrovský, můžeme tak totiž získávat informace o historii studenta, případně o sociálním či psychologickém zázemí zájemce o studium. Většina ale zůstává nevyužita, protože se soustředí pouze na sběr informací nezbytných k přijímacímu řízení daný rok. Pokud se tedy nepřijímá na základě prospěchu, nejsou požadovány známky ze střední školy a ztrácíme tak cenné informace pro analýzy.

Edux
Narozdíl od systému KOS je Edux pouze fakultní záležitost a slouží jako základ známkování právě pro KOS. Na Eduxu má student v každém z předmětů určité hodnocení, získává body za testy a semestrální práce, po splnění podmínek pak získá známku, kterou vyučující zapíše do KOSu. Tyto informace jsou velmi důležité, protože mohou označovat právě ty vynikající studenty, kteří byli aktivní přes semestr, měli excelentní semestrální práce či nadprůměrné výsledky z průběžných testů.

Progtest
Další z fakultních systémů slouží pro automatické opravování programovacích úloh. Tento systém využívá hned několik předmětů jak pro automatické opravy (student nahraje úlohu a vidí okamžitě výsledek), tak pro opravy ruční (tedy jako úložiště pro úlohy a učitel sem pak může zadat výsledek), případně kombinace obojího. Na konkrétní úlohu může mít student více pokusů a je vidět veškerá historie, jak se student snažil odevzdat úlohy a jak si celkově vedl.

Moodle
Podobný systém předchozímu, avšak slouží pouze pro testy, nikoli pro programovací úlohy. Ve využíván prozatím jen v několika málo předmětech, většina testů (i zkouškových) stále probíhá klasickou písemnou formou. Výsledek zkoušky (testu) je ale zapsán do Eduxu, takže informace o klasifikaci není uložena pouze papírově.

Portál SSP
Portál pro spolupráci studentů s průmyslem prozatím běžel v pilotním provozu, ale skrývá v sobě velký potenciál. Jednak se snaží hodnotit výkony studentů na základě jejich známek, ale co je důležitější, studenti se pak mohou subjektivně ohodnotit. Lze tedy porovnat jejich subjektivní cítění se známkami z předmětu, a pokud na závěr přidáme ještě hodnocení průmyslového partnera, získáme představu o tom, jak moc známky odpovídají znalostem a schopnostem studentů. Při vytváření osobních profilů můžeme také získat cenné informace o mimoškolních aktivitách studentů – certifikace, zaměstnání, jinde získané dovednosti.

Systémů, které fakulta využívá je celé řada, výše byly shrnuty ty nejvýznamnější z nich. Nově by šlo sledovat aktivitu studentů přes Marast (Matematika radostně), dále přes fakultní fórum či sociální sítě, zapojení do fakultních projektů, apod. Mnoho důležitých informací se skrývá zejména v Eduxu, kdy můžeme sledovat výkon studenta přes semestr a nejsme odkázáni jen na závěrečnou známku, a v Portálu pro spolupráci studentů z průmyslem, na základě kterého můžeme zjistit relevanci známek vůči skutečným dovednostem studentů.

Kapitola 2

Datový sklad

2.1 Úvod do problematiky

S rozvojem informačních technologií, ať už po stránce hardwaru nebo softwaru, narůstají také nároky firem na jejich systémy[5]. Podstatné je, že došlo ke zvýšeným požadavkům na systémy ve společnostech a stávající řešení zajišťující provozní chod firmy přestávaly uspokojovat potřebu rozhodovací, tedy podporu dlouhodobého plánování. Sem můžeme zařadit například analýzy prodeje a predikce chování zákazníka, což je možné odvodit z nasbíraných dat, které má většinou firma k dispozici.

2.1.1 Business Intelligence

Business Intelligence (BI) je velmi široké téma, které ale velmi úzce souvisí s problematikou datových skladů. Nebudeme se zde zabývat definicí nebo podrobnou analýzou tohoto tématu, ale zkusíme si stručně a jednoduše popsat, co to BI vlastně je a jak s datovými sklady souvisí.

Mnoho velkých firem má obrovské databáze a v nich uložená data. Ta sama o sobě zajišťují chod firmy, ale mají také důležitou vypovídající hodnotu pro rozhodování. Uveďme si jednoduchý příklad. Do banky přijde zákazník a žádá o půjčku. Pracovník v bance potřebuje okamžitě vidět informace o klientovi, včetně toho, jestli je dobré mu půjčku poskytnout nebo ne. Tato data mohou být v odlišných systémech, může se jednat dokonce o externí informace (např. registr dlužníků). Pokud tento příklad zobecníme, jedná se o rozhodování managementu na základě historických informací (čili dat) – kteří zákazníci mají zájem o které půjčky a kam jako firma směřovat v dalších letech.

BI jsou tedy nástroje a postupy sloužící k tomu, jak z uložených dat získat nějaké znalosti, které slouží k podnikovému rozhodování či k určení podnikových

[5]Existuje také opačný názor, tedy že nejdříve byly nároky, ale to pro nás není důležité.

cílů. Ukládat data dnes není žádný problém, ale získání cenných informací vyžaduje mnohem více práce.

První požadavky na nové technologie přišly právě z business sféry. Z tohoto pohledu se jednalo o několik klíčových prvků:

- Rozhodnutí musí být rychlá a správná za použití veškerých dat, která jsou k dispozici.
- Množství dat narůstá, což má za následek prodloužení reakční doby.
- Uživatelé systémů nejsou počítačoví experti.

Problémem je, že klasická úložiště dat mohou obsahovat chybné údaje (např. překlepy, různě vyplněná data), bývají vytížené, nejsou konzistentní a v neposlední řadě většinou neuchovávají historii (pokud se změní údaj o klientovi, nejčastěji se nahradí za údaj nový). To ale není vhodné prostředí a často ani struktura pro provádění analýz a právě v tuto chvíli přichází na řadu datový sklad.

2.1.2 Co je datový sklad

Datové sklady se datují už do 70. let 20. století, ale jejich prudký rozvoj zažíváme až v posledních letech. Na počátku jsou dva pánové, William H. Inmon a Ralph Kimball. Oba dva definují datové sklady, každý z nich ale používá odlišný způsob. Obecně nemůžeme říci, že jeden přístup je špatný a druhý dobrý, jednoduše jsou to přístupy odlišné a jsou zaměřeny na různé potřeby a požadavky na datový sklad. Při budování skladu je tedy vhodné seznámit se s oběma přístupy a vybrat si ten lepší pro naše potřeby.

Abychom ale příliš neodbíhali, vraťme se k otázce, co je vlastně datový sklad. S první definicí tohoto pojmu přišel v roce 1991 právě William H. Inmon, také často označován za „otce datových skladů", který datový sklad definuje takto: „Datový sklad je integrovaný, subjektově orientovaný, stálý a časově rozlišený souhrn dat, uspořádaný pro podporu potřeb managementu." [15]

- Integrovaný – datový sklad má mnoho zdrojových systémů a jeho cílem je tato data vzít a zintegrovat je do jednoho celku.
- Subjektově orientovaný – v klasických systémech jsou data uložena podle aplikací, které je používají nebo kde vznikla. V datovém skladu jsou data uložena podle subjektů (např. zaměstnanec).
- Stálý – jakmile se do datového skladu nahrají data, nikdy se nemění. Přidávají se nová data, stará zde zůstávají po celou dobu životnosti datového skladu.

- Časově rozlišený – data mají přidanou dimenzi času, je tedy možné jednoznačně rozlišit, z jakého časového období pochází. Veškerá data se do skladu nahrají v určitou dobu, jedná se tedy o jakýsi snímek v určitém čase. Pokud se data změní, vytvoří se snímek nový.

Druhá z definic je od Ralpha Kimballa, také přezdívaného „otec business intelligence“, a zní takto: „A data warehouse is a copy of transaction data specifically structured for query and analysis“ [17]. Zde je vidět už první z rozdílů, W. Inmon se zaměřuje na komplexní sklad, R. Kimball má na prvním místě použitelnost pro koncové uživatele.

Ve své knize R. Kimball popisuje metaforu, kde datový sklad přirovnává k restauraci. V kuchyni proběhne příprava jídla, které pak číšníci roznesou mezi hosty, zatímco na vše dohlíží management. Kuchyně je velmi dobře naplánovaná a využívá nejlepší možné suroviny. Jídelna bývá od kuchyně striktně oddělena, zákazníci si jen tak nemohou po kuchyni procházet. Je to dané zejména bezpečností a hygienou (nechceme, aby zákazníci strkali prsty do omáček nebo na sebe převrhli pánev plnou vroucího oleje). V kuchyni tedy probíhá úprava surových dat na integrovaná, kvalitní, vyčištěná a konzistentní data, ke kterým se může dostat koncový uživatel. Není efektivní, aby získával data přímo ze zdrojových systémů, tedy aby mu na stůl přinesli suroviny, aby si uvařil polévku.

Jídlo ale není všechno, co dělá restauraci úspěšnou. Důležitými faktory jsou také interiér, služby (obsluha, donáška, zabalení, apod.) a cena. Pokud uživatel zadá nějaký dotaz, chce získat správnou odpověď (jinými slovy chce jídlo, které si objednal) a v co nejkratším čase.

Pokud tedy vezmeme poznatky uvedené výše, můžeme odvodit obecnou definici. Datovým skladem můžeme nazývat systém, který shromažďuje, sjednocuje a organizuje velké množství dat z více zdrojů a uchovává je napříč časem.

2.1.3 Vlastnosti datových skladů

Požadavky kladené na datové sklady daly formulaci vlastnostem datových skladů, které nám také zpravidla ukazují rozdíl oproti klasickým systémům, které nesplňují jednu nebo více z vlastností.

Jedním z nejdůležitějších prvků datových skladů je, že data jsou integrována z různých datových zdrojů. Je zde jen jeden a tentýž zákazník, o kterém máme veškeré dostupné informace. Ve zdrojových systémech figuruje několikrát podle potřeb oddělení využívajících dané zdrojové systémy a data mohou mít různou podobu. Datový sklad nám zajišťuje stejný formát dat a jejich jednotnou

podobu (např. pohlaví je označováno „M“ a „F“, barvy jsou malými písmeny anglicky, jména začínají velkým písmenem a ostatní jsou malá).

Rozdělení ve zdrojových systémech má také za následek zpomalení dotazů nad daty. Typický příklad, kdy potřebujeme všechna data rychle, jsou nabídky šité na míru zákazníka. Pokud přijde do firmy a chce si zakoupit nějaký produkt (pojištění, investice, apod.), musíme mu dát nějakou nabídku. Pokud bychom ho požádali, aby počkal 14 dní, pravděpodobně odejde ke konkurenci. Datový sklad má tu výhodu, že jsou data na jednom místě a získáme je tedy zpravidla v mnohem rychlejším čase, než kdybychom je získávali přímo z různých zdrojových systémů a poté je dávali dohromady.

Zdrojové systémy mohou mít různou podobu. Mohou být elektronické či papírové, data mohou být ukládána strojově nebo ručně. To je poměrně velký prostor na vytvoření chyby. Pokud manažer firmy potřebuje získat data (např. pro poradu vrcholového managementu) z různých poboček firmy, dva týdny bude data sbírat a poté je na schůzi předloží. Problém ale nastane, pokud jiný manažer bude pracovat se stejnými daty, ale dojde k jinému závěru, protože čísla byla jiná. Důvodů může být několik: data získali v rozdílném čase, z různých poboček nebo se někde stala chyba. Při použití datového skladu toto nestane, protože než se k datům uživatel dostane, data jsou konzistentní a kvalitní, jednotná a je jednoznačně prokazatelné, ze kterého data informace pochází.

Další z klíčovou vlastností datových skladů je uložení velkého množství dat. Existuje několik důvodů, proč je jich takové množství:

- Datový sklad obsahuje historická data

  Předchozí metody používaly pouze aktuální data a staré informace se jednoduše přepisovaly. Datové sklady toto omezení nemají a sbírají veškerá data, takže je vidět, jak se data měnila v průběhu času, což je důležité pro analýzy. Během několika let přirozeně musí, resp. muselo, dojít k obrovskému nárůstu dat.

- Neznámé požadavky

  Je nutné mít data pro splnění známých, tak neznámých požadavků, protože nikdy nemůžeme dopředu vědět, která data budou potřeba za několik měsíců nebo let. Proto se ukládají i data, která nyní nevyužíváme, ale v budoucnu bychom mohli. Neznamená to však, že vezmeme všechna data a bez rozmyslu je uložíme. Pokud u některých dat nevidíme využití, není nutné je nahrávat nyní a můžeme je třeba nahrát později.

- Detailní data

  Souvisí s předchozím bodem, data se snažíme uchovat co nejdetailnější pro potřeby stávajících a budoucích analýz.

- Externí data

  Data z externích zdrojů jsou velmi důležitá pro potřeby rozhodování. Může se jednat o demografická či psychologická data, která mohou složit k předpovědi, kdo bude dobrý nebo špatný zákazník. Například banka při žádosti o půjčku potřebuje znát i jiná fakta než jen historii zákazníka u své banky – například, jestli je dotyčný v registru dlužníků nebo exekučním řízení.

Přidání časové dimenze s sebou nese další možnosti, a to pravidelnou aktualizaci (nahrávání) dat. Pomocí procesů (např. ETL) se může jednat o automatickou činnost. Tedy každý den ve 2 hodiny ráno se budou nahrávat nová data do skladu. Tato časová perioda se může lišit, princip zůstává stejný. Odpadá namáhavá ruční práce, kdy se řeší stále stejné postupy, a vše je řešeno automaticky.

Pokud si tedy shrneme výše uvedené, nejdůležitější vlastnosti datového skladu jsou následující:

- integrace dat,
- rychlost,
- důvěryhodnost dat,
- uložení velkého množství dat včetně jejich historie,
- automatizace.

2.1.4 Datové sklady ano či ne

Nenechme se zmást myšlenkou, že datové sklady jsou spásné a bez nich nemůžeme existovat. Mnoho podniků datové sklady nepoužívá, protože je nevyužije nebo má jiné postupy. A pokud jim to vyhovuje, proč ne. Většina funkcí datového skladu lze nasimulovat i v jiném prostředí, např. přidáním časové dimenze. Datové sklady to ale většinou zjednodušují a urychlují.

Nejčastější otázka, se kterou se setkáme při budování datového skladu, říká, proč vytvářet datový sklad, když můžeme přistupovat přímo ke zdrojovým systémům [10]? A to je vskutku ta správná otázka, se kterou musíme začít při rozhodování, zdali se nám vyplatí investovat čas a peníze do datového skladu. Odpověď není jednoznačná a velmi závisí na zdrojových systémech.

Struktura dat
Pokud zdrojové systémy nedokáží udržovat historii (data přepisují), datový sklad nenahradí. Jednou z jeho předností (a často i požadavkem při jeho budování) je právě zachovávání dat v průběhu času. Pokud to dokáží, přichází v úvahu další aspekt, a to optimalizace a struktura dat z hlediska použití.

Zdrojové systémy jsou většinou navrženy za jiným účelem než je BI, a proto mají zpravidla nevhodné struktury. Navíc mohou být zcela uzavřené a přístup je možný pouze přes API a bez jakékoli dokumentace.

Vytížení zdrojových (provozních) serverů
Druhým důležitým argumentem ke zvážení je zatížení provozních serverů. Ty zajišťují fungování chodu firmy a je nutné předem rozmyslet, jestli je možné je zatížit složitými a časově náročnými dotazy. Pokud výrazným způsobem prodloužíme reakční dobu zdrojových systémů, bude to mít samozřejmě velmi negativní dopad na chod firmy.

Terminologie
Jak jsou na tom zdrojové systémy s terminologií? Pokud mají podobné nebo zcela odlišné termíny pro stejnou věc, pro analýzy bude nezbytné tyto termíny sjednotit. V tom případě je možný zásah do původních systémů nebo vytvoření centrálního prostředí, tedy sjednocení pojmů v jednotném datovém skladu.

Kvalita dat
Důležitým prvkem k zamyšlení jsou také samotná data, která máme k dispozici. Pro podporu rozhodování jsou nutná historická data, čím méně dat máme, tím nepřesnější rozhodování může být. Pokud historická data nemáme vůbec, je nutné zajistit nějaké mechanismy pro uchovávání dat do budoucna. Jestliže je máme, musíme se podívat na jejich kvalitu. Jakou mají vypovídající hodnotu? Je k nim případná dokumentace nebo se jedná o data, u kterých už nikdo neví, co znamenají? Pokud je kvalita natolik nízká, že činí data nepoužitelnými, je vhodné zvážit, zda má cenu budovat datový sklad, který by tato data obsahoval. Je ovšem možné přijmout opatření vedoucí ke zvýšení kvality budoucích dat a poté má návrh datového skladu opět smysl.

Takže při rozhodování si můžeme položit několik otázek ohledně zdrojových systémů:

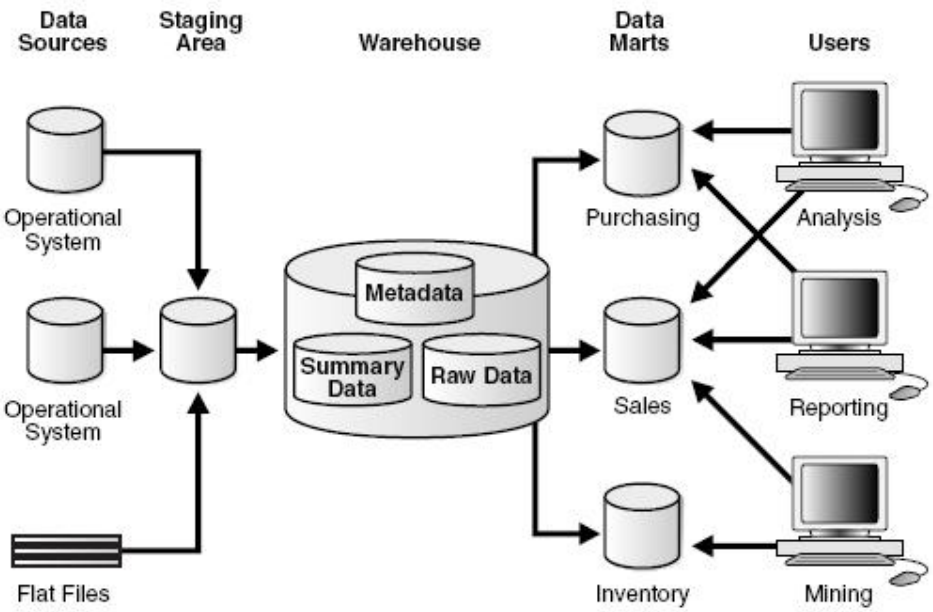
- Umí uchovat historii dat?
- Jsou zde data optimalizována a vhodně strukturována?
- Mám do zdrojových systémů přístup, příp. mám potřebnou dokumentaci?
- Jak moc dotazy zatíží provozní servery?
- Liší se terminologie v jednotlivých systémech?
- Jaká je kvalita dat?

Podle těchto otázek je možné se orientačně rozhodnout, zdali je datový sklad vhodné řešení nebo ne. Málokteré zdrojové systémy jsou ovšem navrženy vhodným způsobem, a proto se většinou řeší kopírování dat do datového skladu

a jejich následná úprava. V některých případech se ale opravdu může stát, že je datový sklad zbytečný a stačí pouze drobná úprava stávajících systémů nebo jsou na to již dokonce připraveny. Pak samozřejmě nemá cenu vymýšlet nové komplexní řešení, když stávající systémy splní stejnou službu.

2.1.5 Architektura datových skladů

Pokud chceme navrhovat datový sklad, musíme znát jeho základní architekturu, která sestává z několika vrstev. Ačkoli architektur již dnes existuje celá řada, popíšeme si zde pouze tu základní, protože i u dalších architektur se tyto vrstvy stále opakují. Jedná se konkrétně o vrstvu se zdrojovými systémy, stage vrstvu, úložiště datového skladu a datová tržiště. Zde je vidět, proč je termín „datový sklad“ občas zavádějící. Správně označuje všechny tyto vrstvy dohromady, ale často pro zjednodušení používá pouze v kontextu úložiště datového skladu. Pokud je v této práci uvedeno, že se data nahrávají do datového skladu, je tím myšleno pouze úložiště, pokud není výslovně uvedeno jinak.



Obrázek 2.1: Vrstvy datového skladu. Zdroj: `http://oracle-datawarehousing.blogspot.cz` [13]

Data pochází z několika zdrojových systémů (Data Sources), mohou to být relační databáze, exporty, apod. My chceme tato data zpracovat s co nejmenším zatížením těchto produkčních systémů, proto je uložíme do dočasného úložiště

(Staging Area), kde jsou neagregovaná, nekonzistentní, bez časové dimenze. Po jejich zpracování se nahrají do úložiště datového skladu (Warehouse) a ze Stage vrstvy jsou odstraněny, čili jsou zde pouze aktuální data a jenom po dobu nezbytnou pro jejich zpracování (často jsou ještě po určitou dobu zálohovány, někdy je to řešeno tak, že jsou ponechány v této vrstvě).

Pokud chceme přizpůsobit data pro jednotlivé části organizace, můžeme použít datová tržiště (Data Marts). Ta narozdíl od samotného datového skladu, který obsahuje velké množství dat, obsahují pouze jejich podmnožinu. Často se vytváří jedno datové tržiště pro jedno oddělení nebo pobočku.

Existují také architektury, kterým chybí datová tržiště nebo Staging Area, ale ty dnes již bývají spíše výjimkami nebo staršími implementacemi datových skladů.

2.1.6 Metody implementace

Existují dva základní přístupy pro implementaci datových skladů, které byly vymyšleny W. H. Imnonem a R. Kimballem. Oba pánové se datovými sklady začali zabývat přibližně ve stejnou dobu a každý z nich se na datové sklady díval jinak. Oba jsou jednoznačně velkými průkopníky v této oblasti a jejich odlišné přístupy nám umožňují široké použití datových skladů.

Metoda shora dolů (W. H. Inmon)
Cílem je uspořádat všechna data v rámci celé organizace najednou. Začíná se obecnou analýzou činností organizace a následnému dělení na nižší úrovně. Tato analýza je komplexní, po ní přichází souhrnný návrh a nakonec přichází implementace. Vše je vytvořeno naráz.

Tato metoda využívá jednoho centrálního a normalizovaného úložiště. Data ze zdrojových systémů jsou převedena do požadovaných formátů a nahrána do tohoto centrálního úložiště. Teprve po vytvoření tohoto prvku pro celý podnik mohou vzniknout jednotlivé datové trhy přizpůsobené pro potřeby jednotlivých oddělení. Datové trhy jsou tedy podmnožina skladu, ale již nemusí být ve stejných strukturách, protože jsou optimalizovány konkrétně pro určitou aplikaci.

Mezi výhody tohoto přístupu patří velmi jednoduché vytvoření datových tržišť, která jsou konzistentní, protože čerpají ze stejného jednotného centrálního prvku. Další výhodou je, že již od začátku máme jasnou datovou strukturu celého skladu a jsou minimalizovány duplicity. Hlavní nevýhodou tohoto přístupu je jeho náročnost, časová i finanční. Vytvoření datového skladu pomocí této metody může trvat i několik let, během kterých se mohou změnit požadavky nebo zdrojové systémy.

Vzhledem ke komplexitě tohoto řešení je tato metoda hojně využívána pro enterprise datové sklady.

Metoda zdola nahoru (R. Kimball)
Začátek budování datového skladu probíhá přesně v obráceném pořadí než u předchozí metody, čili začíná se od datových tržišť. Ta typicky vzniknou na základě požadavků jednotlivých oddělení, která potřebují přistupovat k datům ze zdrojových systémů. Vznikne pak několik etap vývoje datového skladu, které se seřadí podle priorit a postupně tak vznikají jednotlivé části datového skladu (datová tržiště).

Výhody a nevýhody jsou patrné. Největší výhodou je rychlý výsledek, možnost inkrementálního růstu skladu a nízká počáteční investice. Tento přístup se také dokáže poměrně jednoduše přizpůsobit případným změnám požadavků nebo zdrojových systémů. Nevýhodou je pak komplikace při změně do stávající struktury a redundantní data v jednotlivých datových tržištích. Mohou také vzniknout vyšší náklady na provoz.

Hybridní metoda
Existuje také kombinace obou výše uvedených přístupů, kde obě metody spolupracují dle potřeby. Jinými slovy tento hybridní model slučuje výhody obou modelů.

Jak již bylo zmíněno výše, nelze tvrdit, že jeden z přístupů je dobrý nebo špatný, jsou jednoduše odlišné a jejich oblasti použití jsou různé (dáno zejména vysokými počátečními náklady na datový sklad podle W. Inmona).

Zjednodušeně můžeme říci, že pokud se jedná o velkou společnost s velkými přírůstky dat v malém časovém úseku (dny, hodiny), vyplatí se vyšší počáteční investice pro vytvoření komplexního datového skladu podle W. Inmona. Naopak, pokud se jedná o menší společnost a není třeba datový sklad takového rozsahu, je vhodnější použít přístup R. Kimballa. To je příklad i fakultního datového skladu, neboť fakulta nemá tak obrovský přísun dat v krátkých intervalech. Typická délka intervalu je 1 semestr a tím pádem data nenarůstají tak rychle jako u velkých společností, kde je denní nebo hodinový přísun dat. Spolu s nižšími pořizovacími náklady a jednoduchou možností rozšiřování, se stala právě druhá možnost jasnou volbou.

2.1.7 Dimenzionální modelování

Datový sklad využívá dimenzionálního modelování, které spočívá v denormalizaci struktur za účelem vytvoření schémat vhodných pro podporu rozhodování. Jsou využívány dva typy tabulek: faktové a dimenzionální.

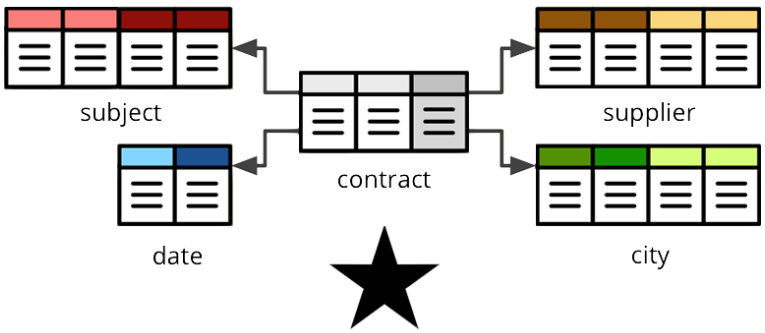
Faktové tabulky
Tyto tabulky slouží pro ukládání zejména číselných záznamů, která lze sumarizovat a analyzovat. Narozdíl od tabulek dimenzionálních zabírají většinu prostoru datového skladu (uvádí se až 90 % [15]) a používají se jak v normalizované, tak denormalizované podobě.

Typicky se jedná o objem prodeje, v případě studenta jsou to pak jednotlivá studia.

Dimenzionální tabulky
Tabulky dimenzí obsahují atributy popisující jednotlivá fakta a mohou být opět v denormalizované podobě, což umožňuje poskytovat všechny relevantní hodnoty k faktům v jedné tabulce. Dimenze uchovávají atributy, které mají vypovídající schopnost, jsou obvykle textové a nezkracují se. Jejich název by také měl být jednoznačný.

Pokud použijeme předchozí příklad, dimenzionální tabulky pro prodej budou obsahovat informace o zákazníkovi, produktu a firmě; pro studium pak informace o studentovi, předmětech a škole. Další důležitou dimenzí je také čas (semestr studia, čtvrtletí prodeje, apod.).

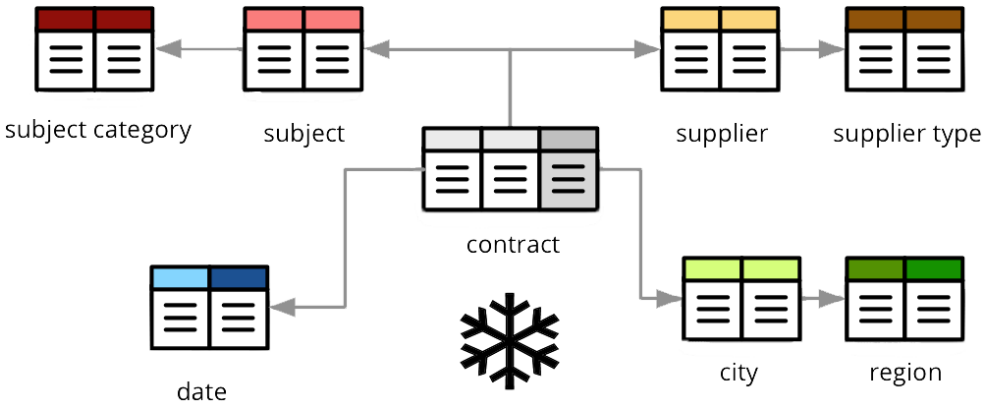
Fakta + dimenze = schéma
Schématem pak můžeme označit tabulky faktů a tabulky dimenzí. Jsou dvě nejčastější schémata pro dimenzionální modelování, a to hvězdicové a sněhové vločky.



Obrázek 2.2: Hvězdicové schéma. Zdroj: `http://databrewery.org` [14]

Hvězdicové schéma (Star Schema)
Obsahuje poměrně vysokou redundanci dat, protože v základní dimenzionální tabulce jsou zahrnuty i nadřazené dimenze (např. dimenze Produkt obsahuje také Kategorii produktu). Výhodou ovšem je, že jsou mnohem rychlejší odezvy, neboť odpadají operace sjednocení (join) mezi dimenzionálními tabulkami. Neefektivní je to v případě, kdy potřebujeme změnu v hierarchii dimenzí, protože tato jedna změna se promítne do mnoha tabulek (kvůli redundanci dat).

Schéma sněhové vločky (Snowflake Schema)
Vychází z předchozího schématu, avšak obsahuje více dimenzionálních tabulek provázaných kardinalitou 1:N, což schéma normalizuje a dochází ke snížení redundance dat. To odstraňuje nevýhodu hvězdicového schématu s aktualizacemi dimenzí. Výhody přechozího schématu se ale stávají nevýhodami: dojde k nárůstu reakční doby kvůli spojování tabulek (join) a samotné schéma je méně přehledné.



Obrázek 2.3: Schéma sněhové vločky. Zdroj: `http://databrewery.org` [14]

V praxi se často používá kombinace obou přístupů v jednom datovém skladu.

2.1.8 Historizace dat

Jednou z důležitých vlastností datového skladu je uchovávání historie dat, kdy v ideálním případě mám u každé dimenzionální tabulky všechny hodnoty po celou dobu historie. Podle R. Kimballa se přístupy dělí na několik kategorií, které souhrnně nazývá SCD (Slowly Changing Dimension), tedy pomalu se měnící dimenze. Data se zpravidla nemění podle předem daného časového rozvrhu, ale mění se pomalu v průběhu času (např. pokud si zákazník změní

jméno, požádá o novou půjčku, apod.) a datový sklad si s tím musí nějak poradit. Způsobů, jak se s tím může datový sklad vypořádat je hned několik.

Typ 0: Ponechání originálu
Jedná se o pasivní způsob, kdy se po změně atributu nic nezmění a jednoduše se ponechá originální hodnota. Tento typ je využíván v případě, kdy potřebuji zachovat původní hodnotu, což je většina dat (například datum první objednávky).

Typ 1: Přepsání hodnoty
Druhá metoda pasivního způsobu je, že původní hodnotu nahradíme za novou. Nevýhoda je, že tímto způsobem nedojde k uchování historie. Tento typ je ovšem užitečný pro případy, kdy historii není nutné zachovávat. Například změna atributu Titul nebo Rodinný stav není většinou údaj, jehož sledování historie by přineslo nějaký užitek. Vždy ale záleží na konkrétní situaci a potřebách firmy, která datový sklad buduje (třeba pro banku by změna rodinného stavu mít vliv mohla).

Typ 2: Přidání nové řádky
Pro změněný údaj vytvořím celý nový záznam (řádek). Tato metoda uchovává kompletní historii, proto se používá pro atributy, kdy potřebuji zaznamenávat jakoukoli změnu. Vznikne mi více řádků, které budou obsahovat stejný záznam, proto musím přidat atributy, které mi určí platnost záznamu (tedy platnost od a do). U platného záznamu je koncová platnost nastavena na nekonečno (může být např. hodnota 2999 či 9999), která se změní s příchodem dalšího záznamu (nastaví se na datum nového záznamu, protože ten den platnost skončila).

Typ 3: Přidání nového atributu
Při použití tohoto typu nezachovávám celou historii, ale pouze nejnovější hodnotu a hodnotu původní nebo předchozí (podle potřeb). Nejnovější hodnotu přepisuji, jako je tomu u Typu 1, a do dalšího atributu buď nakopíruji hodnotu původní nebo ponechám originální. Výhoda je, že nemusím vytvářet nové řádky, ale vždy aktualizuji ten stávající.

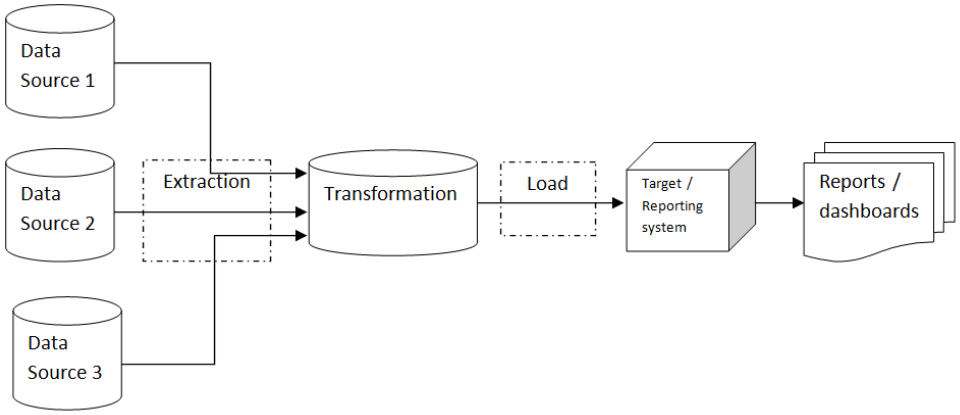
Typ 4: Přidání nové dimenze
Pokud dochází k velmi častým změnám, je vhodné tyto záznamy uchovávat v přidružené dimenzi. V původní dimenzi se uchovává pouze nejnovější údaj.

Tyto typy jsou těmi základními, ale v praxi se ukázalo, že v některých případech nedostačují. Proto byly vytvořeny další, které vznikly kombinací těch předchozích, konkrétně se jedná o: Typ 5 (kombinace 4 a 1), Typ 6 (kombinace 1, 2 a 3) a Typ 7 (kombinace 6 a 1).

2.1.9 ETL procesy

Nedílnou součástí datového skladu jsou ETL procesy, které mají na starosti mechanismus plnění skladu daty ze zdrojových systémů, někdy jsou také označovány pojmem „datová pumpa“. Zkratka „ETL“ popisuje celý průběh manipulace s daty, tedy jejich získání ze zdrojových systémů (Extract), následné úpravy a uspořádání (Transform) a nahrání do předpřipravených struktur (Load).

Existuje několik variant těchto procesů, které se odvíjí od pořadí jednotlivých částí a architektury datového skladu. Druhým nejčastějším typem jsou ELT procesy, které jsou typicky používané ve chvíli, kdy máme přímý přístup do zdrojových systémů a data si nejdříve uložíme do dočasné vrstvy datového skladu a teprve poté provedeme zpracování dat. Podle potřeb firem (datového skladu) vznikají i komplikovanější, např. ETTL nebo ELTETLELT. V této práci budou nadále popsány ETL procesy, ale základní prvky jsou pro všechny stejné.



Obrázek 2.4: Průběh ETL procesu. Zdroj: `http://blog.performancearchitects.com` [16]

Extract
Tato část se zaměřuje na získání dat ze zdrojových systémů, které bez jakýchkoli úprav uložíme do úložiště (Stage Area). Zdrojových systémů bývá zpravidla více, proto se v datovém skladu není možné spoléhat na stávající identifikační klíče, ale je nutné vytvořit nové.

Transform
Cílem transformace je upravit data pro naše podmínky, tedy organizovat je do požadovaných struktur a získat dostatečnou kvalitu, což obnáší několik kroků.

Čištění dat
Tento krok provádíme za jediným účelem, a tím je zvýšení kvality dat. Ta často obsahují nějaké chyby, které je nutné odstranit. Jedná se o velmi rozsáhlou oblast, která vyžaduje velké množství času. V rámci čištění dat se snažíme docílit toho, aby data dodržovala následující podmínky:

- Jednotná forma dat
  - stejný formát telefonních čísel, PSČ, webových adres, apod.
  - jednotné identifikátory: např. převod M/Ž, M/Z, M/F na Muž nebo Žena
- Sjednocení chybějících hodnot
  - prázdné hodnoty integeru vyplním 0 (-1 či nekonečnem, záleží na datech), string „N/A“, apod.
- Validace dat
  - v rámci jednoho atributu: správný formát e-mailu nebo tel. čísla
  - v rámci více políček: rodné číslo dělitelné 11, PSČ by mělo odpovídat městu, resp. okresu a zemi, apod.

Konsolidace dat
Pokud má jedna firma více systémů, předpokládá se, že jedna osoba se může vyskytovat ve více zdrojích pod různými klíči. Je nutné najít tyto duplicity a sjednotit je pod jeden záznam, což se provádí zpravidla porovnáním jména, příjmením, rodným číslem; názvu firmy a IČ, apod.

Vytvoření nových klíčů
Z důvodů zmíněných výše je vhodné vytvářet nové klíče a nespoléhat na ty stávající. Jednak dojde k překryvům ve více zdrojových systémech, čili jeden subjekt může mít více klíčů, a při vytváření nových struktur si nikdy nemůžeme být jisti unikátností záznamů, které původem nejsou klíči.

Vhodné strukturování
Na konci této fáze bychom měli mít připravená data pro naše nové struktury, což může obnášet:

- Vytvoření odvozených sloupců – např. spočítání BMI z váhy a výšky
- Agregace – součet prodejů v daném období
- Rozdělení/sloučení dat – podle potřeb mohu rozdělit data z jednoho do více sloupců či naopak
- Transpozice dat – záměna sloupců za řádky a naopak

Load
Uložení dat do požadovaných struktur datového skladu. Neznamená to, že až nyní jsou data nahrána do samotného skladu, protože tyto procesy jsou součástí skladu jako takového a již by měly být prováděny v rámci datového skladu.

2.1.10 Metadata

Datový sklad obsahuje kromě samotných dat také metadata, která se často označují jako data o datech. Jedná se o strukturovaný popis dat jako takových, jakým způsobem jsou získávána, jak se ukládají, zpracovávají, mohou zahrnovat také informace o síti nebo technických prostředcích. Pomáhají nám orientovat se v datech a lépe jim porozumět. Existují celkem tři skupiny metadat: business, technická a procesní.

Business metadata
Popis dat jednoduchou a přívětivou formou – nejedná se o technické specifikace, ale o vymezení jednotlivých pojmů (např. co znamená „Klient“ či „Zákazník“), odkud data pochází a jakým procesem projdou, než se uloží ve skladu.

Technická metadata
Tato metadata popisují technický pohled na datový sklad, typicky sem spadá popis datových modelů, tedy popis tabulek (dimenze, fakta, apod.), atributů, indexů, klíčů, dále pak technická realizace ETL procesů, definice jejich zdrojů a cílů, apod.

Procesní metadata
Popisují výsledky jednotlivých operací v datovém skladu, včetně jejich délky trvání, výsledků, počtu operací čtení a zápisu. Tato metadata jsou vhodná pro detekování a opravy případných chyb či monitorování provozu.

2.2 Návrh a implementace

Jedním z cílů této práce je získat data z přihlášky ČVUT[6] a po jejich zpracování je pomocí ETL procesů nahrát do fakultního datového skladu. V první řadě bylo nutné data získat, což spolu s interpretací těchto dat zabralo nejvíce času. Po tomto kroku přišly na řadu ELT procesy. Ty se postaraly o nahrání dat do datového skladu, který bylo nutné předem rozšířit o nové tabulky. Stav,

[6]https://prihlaska.cvut.cz, systém používaný na ČVUT pro správu elektronických přihlášek

který bude nadále popsán (tedy aktuální stav) samozřejmě není stavem ideálním. Ten se v první a poslední řadě odvíjí od zdrojových dat, jejichž problémy a možná řešení jsou popsány v kapitole 2.2.1.

2.2.1 Popis dat

Tato data jsou používaná pro potřeby přijímacího řízení na fakulty na ČVUT. Jedním zdrojem těchto dat je aplikace prihlaska.cvut.cz, kde si sám zájemce údaje vyplní, druhým zdrojem je KOS, který dodává informace o předchozích studiích dotyčného na ČVUT (to je potřebné zejména, pokud se hlásí bakalář na magisterské studium na stejné fakultě a na základě průměru je přijat bez přijímací zkoušky). Další údaje, které jsou k dispozici jsou z průběhu přijímacího řízení, tedy jak si zájemce vedl a na základě čeho byl případně přijat.

Jelikož je aplikace Přihláška jednotná pro všechny fakulty a všechny typy přihlášek, obsahují data velké množství sloupců. Některé se využívají pro přihlášky do programu bakalářského, magisterského či doktorského, další jsou zde pro potřeby jednotlivých fakult. Aby se se změnou podmínek přijímacího řízení nemusely pokaždé měnit sloupce, existují univerzální atributy s názvy „H#“ či „HODN#“, kde „#“ označuje číslo od 1 do 11. Je jen na dohodě studijního oddělení, co se bude do těchto sloupců ukládat.

Akademický rok	Počet sloupců	Počet řádek (tzn. přihlášek)
2009/2010	177	520
2010/2011	177	1775
2011/2012	207	2108
2012/2013	—	2144
2013/2014	224	2519

Tabulka 2.1: Exporty dat v průběhu let

Každý rok se data trochu liší. První dva roky (tedy 2009/2010 a 2010/2011) se používal jiný formát dat než roky následující a i poté se ještě nějaké sloupce přidávaly. V tabulce 2.1 je pro jednotlivé ročníky vidět, kolik bylo celkově přihlášek (do BSP a MSP) a kolik obsahovaly sloupců.

Všímavý čtenář si na první pohled všimne, že u akademického roku 2012/2013 není vyplněný počet sloupců. Je tomu tak proto, že jediný dostupný export s těmito daty neobsahoval sloupce s rodnými čísly, podle kterého je možné rozlišit studenty se stejnými jmény. Z toho důvodu tento export nebyl použit.

Výsledný model obsahuje 55 atributů. Ne všechny roky mají všechny atributy, což bude mít samozřejmě vliv na použitelnost těchto dat. V tabulce 2.2

je vyznačeno, které roky používají které sloupce. Slovem používají zde mám na mysli, že tyto sloupce nejen obsahují, ale mají v nich také smysluplná data.

2.2.1.1 Problémy s daty

Tato data s sebou bohužel přinesla několik problémů, z nich některé jsou pro případné analýzy kritické, jiné naopak řešitelné.

Není přístup ke zdroji
Fakulta pro získávání dat používá rozhraní, které ale není zdrojovým systémem. To s sebou může přinést problémy v podobě chybějících sloupců a nemožnosti zjistit, která data jsou ta původní bez dalších úprav.

Různé formáty dat
První dva roky byla data jakž takž podobná, poté došlo ke změně sloupců, včetně jejich názvů. To by samo o sobě takový problém nebyl, pokud by zůstal stejný obsah. To se naštěstí podařilo částečně vyřešit tím, že byly dohledány odpovídající číselné a slovní hodnoty např. u země, odkud zájemce pochází, nebo u sloupce označujícího stupeň předchozího vzdělání, kdy v prvních letech byla v datech uvedena pouze kódová jména (K, L, M, apod.).

Povinné sloupce
Problém, který ale bohužel vyřešit nemůžeme, je s povinnými údaji. Samozřejmě to souvisí s podmínkami pro přijetí daný rok, ale bohužel odebráním povinnosti vyplnit konkrétní sloupec pak ztrácíme určitou hodnotu. Jedná se zejména o výsledky ze středních škol. V prvních letech bylo povinné vyplnit známky z matematiky, studijní průměr i známky z maturity. V dalších letech už to vyžadováno nebylo a tyto informace nadále nemáme.

Univerzální sloupce
Se změnou podmínek pro přijetí se mění také potřeby ukládat určité hodnoty. Pokud se nadále nepřijímá podle průměru z matematiky, sloupec není potřeba, pokud se začne brát v potaz výsledek se státní maturity, potřebujeme nový sloupec. Toto bylo vyřešeno univerzálními sloupci, kdy si na studijním oddělení zvolí, co do kterého sloupce budou ukládat. Vznikne pak jednoduchá mapovací tabulka, která se každý rok trochu mění. Bohužel jsou tyto tabulky v papírové podobě a nejsou nikde archivované. Zpětně je není možno nikde dohledat a pokud nějaký rok tato tabulka chybí, dojde ke ztrátě důležitých informací. Jelikož ne všechny tabulky byly k dispozici, bylo nutné vytvořit odpovídající mapování podle paměti studentů.

IZO střední školy vs. předchozí vysoká škola
Pokud se zájemce hlásí do magisterského studia, musí vyplnit údaje o předchozím studiu na vysoké škole. V tom případě si v nabídce vyhledá svoji fakultu

a my máme k dispozici veškeré potřebné údaje. Naopak, pokud se zájemce hlásí do bakalářského studia, musí vyplnit IZO (kód) své střední školy. To už není možné výběrem ze seznamu, ale dostane k dispozici odkaz, aby si IZO dohledal sám[7] a vyplnil ho do políčka a s tím samozřejmě přibývá pravděpodobnost, že dojde k chybě (někteří zájemci např. vyplnili IČ místo IZO).

Chybějící sloupec
Ze změnou formátu dat (od roku 2011/2012) zmizel z dat sloupec Předchozí vysoká škola (i když ho zájemci vyplnili). Již bylo zažádáno o jeho opětovné přidání, ale zpětně tato data bohužel není možné získat.

Na základě čeho byl student přijat
V datech se dozvíme, jaké měl zájemce výsledky z přijímacího řízení, ale již není možné dohledat, na základě čeho byl skutečně přijat. Rozlišeno je pouze přijetí na základě přijímací zkoušky či mimo ni.

Originální sloupce
Pokud si zkusíme vyplnit přihlášku sami, zdá se celkem jednoduchá a jasná. Export dat ovšem tak jednoznačný není, jsou zde přidány sloupce pro další účely, kdo došlo ke kopírování či agregaci nějakých dat. Např. jméno zájemce je zde ve třech sloupcích, rodné číslo ve dvou a v jednom upravené, apod. Jelikož k datům neexistuje dokumentace, je možné se orientovat pouze pomocí názvů sloupců, kde se bohužel nedá zjistit, které sloupce jsou ty původní.

Předchozí studia
Každý zájemce je požádán o vyplnění svých předchozích studií. Jelikož není daný žádný formát (narozdíl od papírové přihlášky), může do pole vyplnit cokoli. Někteří studenti vyplní pole zkratkami, které mohou a nemusí být jednoduše interpretované, jiní z úsporných důvodů vynechají název fakulty, programu či oboru. Data v takovém formátu bohužel není možné jednoduše zpracovávat a se zmizením sloupce Předchozí vysoká škola dochází ke ztrátě informace, odkud přichází noví studenti (zejména zájemci o magisterský studijní program).

Pro bližší představu si uvedeme pár příkladů:

ČVUT-FEL
1999-2001
(nedokončeno)

ZČU FAV - 2007/2008 nedokončeno
VŠE - 2009/2010 nedokončeno

[7]Podle registru MŠMT ČR na adrese http://stistko.uiv.cz/registr/vybskolr.asp

ČVUT,strojní,prezenční bakalářské,30.6.05-28.2.06 - 1 semestr
ČVUT,dopravní,kombinované bakalářské, technika a technologie v dopravě a spojích, management a ekonomika dopravy a telekomunikací 26.8.11-29.2.12 - 1 semestr

1999-2002 FEL CVUT 2002-2003 FIS VSE 2010-2012 FIT CVUT

2.2.1.2 Vhodné změny

Pro jednodušší práci s daty či předejití některým problémům popsaným výše, by bylo vhodné zavést následující změny:

- Získat přístup do zdrojových systémů

  Tedy do aplikace Přihláška, příp. do PřiŘíz na FIT, který slouží právě pro agendu kolem přijímacího řízení. Tímto přístupem by se kromě exportu mohly dohledat původní sloupce, které vyplnil student a případně sloupce chybějící, díky metadatům v aplikaci PřiŘíz by pak odpadly problémy s univerzálními sloupci.

- IZO střední školy jako seznam

  Zájemci by si pouze dohledali střední školu v předem definovaném seznamu a nemuseli by vyplňovat ručně IZO.

- Strukturovat předchozí studia

  Například podle papírové přihlášky vytvořit kolonky: Škola, Fakulta, Program, Obor, Datum (od–do) a příznak, jestli bylo studium ukončené, neukončené či stále probíhá.

2.2.1.3 Vypovídající hodnota dat

Vzhledem ke změnám v datech každý rok (vyplňované sloupce, relevantnost údajů v nich, změna formátů) byla vytvořena tabulka, kde je vidět, které atributy jsou pro který rok k dispozici. Tato tabulka je velmi důležitá pro další práci s daty, neboť je zde vidět, které sloupce jsou irelevantní v celkových datech (a tudíž mohou být zavádějící, ale mohou být použity v rámci konkrétního ročníku). Kvůli sjednocení názvů sloupců jsou zde již použity jednotné názvy z datového modelu (viz. kapitola 2.2.2).

Např. pokud budeme chtít použít sloupec s výsledky z maturity, je nutné brát v úvahu pouze roky, kdy byl sloupec vyplněn. Pokud použijeme všechna data, výsledky budou velmi zkreslené, ale pro jeden či dva konkrétní ročníky můžeme základní analýzy provést.

	09/10	10/11	11/12	12/13	13/14
d_student					
a_academic_degree_prefix	Ano	Ano	Ano	Ne	Ano
a_academic_degree_suffix	Ano	Ano	Ano	Ne	Ano
a_sex	Ano	Ano	Ano	Ne	Ano
a_citizenship	Ano	Ano	Ano	Ne	Ano
a_pernament_residency_CZ	Ano	Ano	Ano	Ne	Ano
d_address					
a_country	Ano	Ano	Ano	Ano	Ano
a_district	Ano	Ano	Ano	Ne	Ano
a_city	Ano	Ano	Ano	Ano	Ano
a_postal_code	Ano	Ano	Ano	Ano	Ano
d_high_school					
a_IZO	Ano	Ano	Ano	Ne	Ano
a_high_school_type	Ano	Ano	Ano	Ne	Ano
d_high_school_field					
a_field_code	Ano	Ano	Ano	Ne	Ano
d_high_school_study					
a_high_school_leaving_exam_year	Ano	Ano	Ano	Ne	Ano
a_high_school_leaving_exam_results	Ano	Ano	Ano	Ne	Ne
a_high_school_leaving_exam_average	Ano	Ano	Ano	Ne	Ne
a_high_school_average	Ano	Ano [8]	Ne	Ne	Ne
a_high_school_math_results	Ano	Ano	Ano	Ne	Ne
d_application					
a_registration_date	Ano	Ano	Ano	Ne	Ano
a_form_of_study	Ano	Ano	Ano	Ano	Ano
a_study_programme	Ano	Ano	Ano	Ano	Ano
a_prev_study_degree	Ano	Ano	Ano	Ne	Ano
a_enrollment_to_study	Ano	Ano	Ano	Ano	Ano
a_bachelor_CTU_success	Ne	Ne	Ne	Ne	Ano

[8] Pouze pro cca 100 záznamů

	09/10	**10/11**	**11/12**	**12/13**	**13/14**
a_bachelor_FIT_success	Ne	Ne	Ne	Ne	Ano
a_from_school	Ano	Ano	Ano	Ne	Ano
a_last_university	Ano	Ano	Ne	Ne	Ne
a_application_num	Ano	Ano	Ano	Ano	Ano
a_applicant_num	Ano	Ano	Ano	Ano	Ano
f_application_results					
a_decision_date	Ano	Ano	Ano	Ano	Ano
a_admission_decision	Ano	Ano	Ano	Ano	Ano
a_admission_exam_num	Ne	Ne	Ano	Ne	Ano
a_admission_exam_result	Ne	Ne	Ano	Ne	Ano
a_SCIO_math	Ne	Ne	Ano	Ne	Ano
a_competitions	Ne	Ne	Ano	Ne	Ano
a_school_leaving_exam_math	Ne	Ne	Ano	Ne	Ne
a_bachelor_weighted_average	Ne	Ne	Ne	Ne	Ano
a_bachelor_arithmetic_average	Ne	Ne	Ne	Ne	Ano
d_prev_study					
a_description	Ne	Ne	Ne	Ne	Ano
export KOS (string)	Ne	Ne	Ano	Ne	Ano

Tabulka 2.2: Odpovídající atributy pro různé roky

2.2.2 Datový model

Datový model byl vytvořen pomocí programu Enterprise Architect[9] a slouží ke strukturalizaci původních dat do odpovídajících tabulek, které jsou dvojího typu: faktové a dimenzionální. Podle prefixu je možné jednoduše poznat typ tabulky, tedy „d_“ pro dimenzionální a „f_“ pro faktové tabulky. Podobnou konvenci mají také jednotlivé atributy, u kterých prefix znamená původ dat, tedy „k_“ pro KOS, „p_“ pro Progtest a „a_“ pro export z přihlášek. Ty bez prefixu jsou atributy odvozené, pomocné nebo bez jednoznačné identifikace původu dat.

Grafickou podobu modelu naleznete v příloze (obrázek A.1 na straně 92). Jedná se o poslední, tedy implementovanou podobu. Není to ovšem první verze tohoto modelu, neboť s nově získávanými daty a informacemi o nich, bylo nutné model několikrát upravit, aby vyhovoval nejnovějším podmínkám. I přes tento inkrementální vývoj není dokonalý, velké nedostatky má zejména

[9] http://www.sparxsystems.com.au

v části, která uchovává informace o předchozích studiích. Ta jsou v ideálním případě potřeba změnit již na úrovni aplikace Přihláška ČVUT (viz kapitola 2.2.1.2 na straně 33).

2.2.2.1 d_student

Dimenzionální tabulka, která uchovává informace o studentech, resp. o zájemcích o studium, na naší fakultě. Na tuto tabulku jsou navázány všechny ostatní, které vlastně jen rozšiřují informace o jednotlivých studentech. Tato tabulka již existuje ve stávajícím datovém skladu v souvislosti s jednotlivými studii studenta, čili se jedná o pojící prvek nové části z přihlášek a původního řešení. Níže si popíšeme všechny atributy této tabulky, v původním řešení byly pouze *k_name*, *k_surname*, *k_p_username* a *k_rc*, ostatní jsou nově přidané.

Název atributu	**Datový typ**	**Popis**
ID_student	integer	Primární klíč tabulky d_student
k_name	text	Jméno
k_surname	text	Příjmení
k_p_username	text	Username v rámci ČVUT
k_rc	text	Rodné číslo (pozor, cizinci mohou mít i alfanumerické znaky)
begin_study	boolean	Tento atribut slouží k rozlišení zájemců o studium, kteří nastoupili a těch, kteří nenastoupili.
a_academic_degree_prefix	varchar(10)	Titul před jménem
a_academic_degree_suffix	varchar(30)	Titul za jménem
a_sex	varchar(1)	Pohlaví (M, Z)
a_citizenship	varchar(60)	Občanství
a_pernament_residency_CZ	boolean	Trvalé bydliště v ČR (true, false)

Tabulka 2.3: Popis atributů d_student

2.2.2.2 d_address

V této dimenzionální tabulce jsou uloženy údaje o bydlišti studentů. Vzhledem k účelu těchto dat není nutné zde mít uloženou konkrétní adresu (tedy ulici a číslo popisné).

Název atributu	Datový typ	Popis
ID_address	integer	Primární klíč tabulky d_address
a_country	varchar(60)	Země
a_district	varchar(60)	Okres
a_city	varchar(50)	Město
a_postal_code	varchar(15)	Poštovní směrovací číslo (cizí PSČ mohou obsahovat i alfanumerické znaky)

Tabulka 2.4: Popis atributů d_address

2.2.2.3 d_student_d_address

Pomocná tabulka, která řeší M:N vztah mezi tabulkami *d_student* a *d_address*. Kromě klíčů obsahuje také atribut *type*, který je typu varchar a označuje typ adresy (kontaktní, trvalé bydliště, apod.). V těchto datech je vyskytuje prozatím pouze trvalá adresa.

Název atributu	Datový typ	Popis
ID_student_address	integer	Primární klíč tabulky d_student _d_address
ID_student	integer	Cizí klíč tabulky d_student
ID_address	integer	Cizí klíč tabulky d_address
type	varchar(20)	Typ adresy (kontaktní, trvalé bydliště, apod.)

Tabulka 2.5: Popis atributů d_student_d_address

2.2.2.4 d_high_school

Informace o konkrétní střední škole, kterou student vystudoval. V přihlášce musí vyplnit IZO (identifikátor střední školy) a typ střední školy. Ostatní atributy jsou doplněné podle identifikátoru.

Název atributu	Datový typ	Popis
ID_high_school	integer	Primární klíč tabulky d_high _school
a_IZO	varchar(20)	Identifikační číslo střední školy

Název atributu	Datový typ	Popis
a_high_school_type	varchar(30)	Typ střední školy
a_high_school_name	varchar(50)	Název střední školy
a_high_school_name_full	text	Plný název střední školy (může obsahovat i město, apod.)
a_high_school_city	varchar(50)	Město, ve kterém je střední škola

Tabulka 2.6: Popis atributů d_high_school

Atribut *a_high_school_type* může nabývat těchto hodnot:

- střední odborné učiliště
- gymnázium
- střední odborná škola
- konzervatoř
- ostatní

Atribut *a_IZO* může obsahovat speciální hodnoty:

- 999999999 – zahraniční škola
- 888888888 – škola Ministerstva vnitra
- 777777777 – škola Ministerstva obrany
- 666666666 – střední odborná škola (rok maturity 2005 a starší)
- 555555555 – gymnázium (rok maturity 2005 a starší)
- 444444444 – střední odborné učiliště (rok maturity 2005 a starší)
- 333333333 – integrovaná střední škola (rok maturity 1998 a starší)

2.2.2.5 d_high_school_field

Obor střední školy, např. gymnázia mají čtyřletý a osmiletý obor, které pak mají kódy 7941K41 a 7941K81.

Název atributu	Datový typ	Popis
ID_high_school_field	integer	Primární klíč tabulky d_high_school_field
ID_high_school	integer	Cizí klíč tabulky d_high_school
a_field_name	text	Název oboru

Název atributu	Datový typ	Popis
a_field_code	varchar(20)	Kód oboru

Tabulka 2.7: Popis atributů d_high_school_field

2.2.2.6 d_high_school_study

Konkrétní studium na střední škole daného studenta. Vyplněné atributy se liší podle akademického roku, ve kterém se student hlásil. Obecně můžeme říct, že čím starší ročník, tím více informací o jeho středoškolském studiu máme k dispozici. Čím mladší, tím méně důrazu se klade na prospěch na střední škole, a proto tato pole již nejsou v přihlášce vyžadována.

Název atributu	Datový typ	Popis
ID_high_school_study	integer	Primární klíč tabulky d_high _school_study
ID_student	integer	Cizí klíč tabulky d_student
ID_high_school	integer	Cizí klíč tabulky d_high_school
ID_high_school_field	integer	Cizí klíč tabulky d_high_school _field
a_high_school _leaving_exam_year	integer	Rok maturity
a_high_school _leaving_exam_results	varchar(15)	Známky z maturity
a_high_school _leaving_exam_average	integer	Průměr z maturity * 100 (průměr 1,25 je uložen jako 125)
a_high_school _average	integer	Průměr ze střední školy * 100
a_high_school_math _results	varchar(15)	Známky z matematiky

Tabulka 2.8: Popis atributů d_high_school_study

2.2.2.7 d_application

Student při vyplňování přihlášky vyplní informace nejen o své historii jako studenta, ale také informace o tom, odkud se aktuálně hlásí a kam se hlásí. Tyto informace jsou časově omezené (a svázané právě s přihláškou), proto jsou uvedeny právě zde. Například to, že se hlásí ze střední školy je aktuální pouze pro tuto přihlášku, za rok může být tato informace irelevantní.

Název atributu	Datový typ	Popis
ID_application	integer	Primární klíč tabulky d_application
ID_student	integer	Cizí klíč tabulky d_student
ID_time	integer	Cizí klíč tabulky d_time
a_registration_date	timestamp	Datum registrace
a_form_of_study	varchar(20)	Forma studia, do které se uchazeč hlásí
a_study_programme	varchar(60)	Studijní program, do kterého se hlásí
a_prev_study_degree	varchar(65)	Předchozí stupeň vzdělání
a_enrollment_to_study	boolean	Zápis do studia
a_bachelor_CTU_success	boolean	Úspěšný bakalář z ČVUT
a_bachelor_FIT_success	boolean	Úspěšný bakalář z FIT
a_from_school	varchar(50)	Ze které školy se uchazeč hlásí (pokud se nehlásí ze školy, uvede např. zaměstnání či domácnost)
a_last_university	text	Univerzita/fakulta, ze které se student hlásí
a_application_num	integer	Číslo přihlášky
a_applicant_num	integer	Číslo uchazeče

Tabulka 2.9: Popis atributů d_application

Atribut *a_form_of_study* může nabývat těchto hodnot:

- prezenční
- kombinované

Atribut *a_study_programme* může nabývat těchto hodnot:

- Informatika (bakalářská)
- Informatika (magisterský)
- Informatics
- Informatics (master)

Atribut *a_prev_study_degree* může nabývat těchto hodnot:

- Úplné střední vzdělání – gymnázium

- Úplné střední odborné vzdělání s vyučením i maturitou – SOU
- Úplné střední odborné vzdělání s maturitou (bez vyučení) – SOŠ
- Vyšší odborné vzdělání
- Vysokoškolské bakalářské vzdělání
- Vysokoškolské magisterské vzdělání
- Vysokoškolské doktorské vzdělání

Atribut *a_from_school* může nabývat těchto hodnot:

- střední škola
- zaměstnání
- domácnost
- vojenská služba
- vyšší odborná škola
- vysoká škola
- přichází přes Dům zahraničních služeb
- jiné

2.2.2.8 f_application_results

V této jediné faktové tabulce (v rámci dat z přihlášek) jsou uloženy informace o průběhu samotného přijímacího řízení, tedy výsledky přijímací zkoušky, body ze Scio testů, rozhodnutí o přijetí, apod. Data z této tabulky nepochází všechny od samotného studenta, ale také od studijního oddělení či z KOS.

Název atributu	**Datový typ**	**Popis**
ID_application	integer	Cizí klíč tabulky d_application
ID_time	integer	Cizí klíč tabulky d_time
a_decision_date	timestamp	Datum rozhodnutí
a_admission_decision	integer	Rozhodnutí o přijetí
a_admission_exam_num	integer	Varianta přijímací zkoušky
a_admission_exam_result	integer	Výsledek přijímací zkoušky (body 0 až 100)
a_SCIO_math	integer	Percentil ze Scio testu z matematiky (0 až 100)
a_competitions	boolean	Úspěšný řešitel olympiád

Název atributu	Datový typ	Popis
a_school_leaving _exam_math	integer	Body z vyšší státní maturity z matematiky
a_bachelor_weighted _average	numeric	Vážený průměr z bakalářského studia (studenti FIT ČVUT)
a_bachelor _arithmetic_average	numeric	Aritmetický průměr z bakalářského studia (studenti FIT ČVUT)

Tabulka 2.10: Popis atributů f_application_results

Atribut *a_admission_decision* může nabývat těchto hodnot:

- 10 přijat na základě přijímací zkoušky
- 11 přijat bez přijímací zkoušky (Scio, olympiády, průměr, apod.)
- 12 přijat na odvolání (nedodal včas maturitní vysvědčení – např. kvůli opravnému termínu)
- 13 přijat mimo přijímací řízení (např. přestupující)
- 21 nepřijat pro nesplnění podmínek přijímacího řízení

2.2.2.9 d_faculty

Informace o fakultách, na kterých zájemce studoval. Jedná se o vnitřní kódy fakult v rámci ČVUT. Pokud fakulta působí ve více městech, má dva různé kódy (například Fakulta dopravní, která má součást v Děčíně).

Název atributu	Datový typ	Popis
ID_faculty	integer	Primární klíč tabulky d_faculty
Code	integer	Kód fakulty v rámci ČVUT
Name	text	Název fakulty

Tabulka 2.11: Popis atributů d_faculty

2.2.2.10 d_study_field

Opět souvisí s předchozím studiem, jedná se o obory v rámci fakult na ČVUT.

Název atributu	Datový typ	Popis
ID_study_field	integer	Primární klíč tabulky d_study_field
ID_faculty	integer	Cizí klíč tabulky d_faculty
Code	varchar(30)	Název oboru v rámci ČVUT
Name	text	Název oboru

Tabulka 2.12: Popis atributů d_study_field

2.2.2.11 d_prev_study

Podstatnou informací, kterou nám zájemce o studium sděluje, jsou jeho předchozí studia. Je několik typů:

- Kód předchozí školy (fakulty) – nejpodstatnější informace, kterou nám zájemce může poskytnout. Bohužel po změně exportů došlo ke ztrátě této informace a je pouze u prvních dvou ročníků. Již bylo zažádáno o přidání tohoto atributu pro další ročníky, ale zpětně bohužel informaci nezískáme.
- Studium na ČVUT – u každého zájemce je zjištěno, jestli má historii na ČVUT. Pokud ano, jsou vrácena jeho studia (fakulta, obor, průměr a datum zápisu).
- Poslední možností je, že student není z ČVUT. V tom případě může do pole Předchozí studia napsat jakýkoli text. Bohužel tento text není možné v rozumné míře zpracovávat, protože má omezenou délku a zájemci šetří místo například tím, že vynechají název oboru, fakulty nebo školy. Výsledné informace nedávají mnohdy příliš velký smysl (viz také kapitola 2.2.1).

Název atributu	Datový typ	Popis
ID_prev_study	integer	Primární klíč tabulky d_prev_study
ID_student	integer	Cizí klíč d_student
ID_faculty	integer	Cizí klíč d_faculty
ID_study_field	integer	Cizí klíč d_study_field
a_at_CTU	boolean	Předchozí studium na ČVUT
a_enroll_date	timestamp	Datum zápisu
a_description	text	Text vyplněný uchazečem, žádná struktura nebo stejné členění

Název atributu	Datový typ	Popis
a_weighted_average	numeric	Vážený průměr studia
a_arithmetic_average	numeric	Aritmetický průměr studia

Tabulka 2.13: Popis atributů d_prev_study

Vzhledem k tomu, že v atributu *a_description* může být opravdu cokoli, mohou být vytvořena dvě totožná studia (jedno podle dat z KOS a druhé právě kvůli tomuto popisu). Atribut *a_at_CTU* tedy dokáže jednoznačně identifikovat studia, která jsou původem z ČVUT, obráceně to ale neplatí.

2.2.2.12 code_names

Jelikož se měnily exporty dat v průběhu let, bylo nutné data v datovém skladu zintegrovat. V prvních letech se používalo převážně číselné označení, v dalších jak číselné, tak slovní. Bylo nutné vytvořit mapování těchto čísel na jejich slovní atributy, k čemuž slouží právě tato tabulka. Aby nebylo nutné vytvářet tyto tabulky pro každý atribut, jsou nahrány v tabulce jediné a rozlišeny pomocí atributu *description*. Tato tabulka může sloužit jako určitá podoba dokumentace (k dispozici jsou samozřejmě i textové soubory), ale její hlavní účel je pro integraci dat v rámci ETL procesů.

Název atributu	Datový typ	Popis
ID_code_name	integer	Primární klíč tabulky d_code_names
code	varchar(15)	Kód (většinou vnitřní kódování na ČVUT nebo číselníky ze samotné Přihlášky ČVUT)
name	text	Název (význam)
description	text	Slouží pro rozlišení jednotlivých typů kódů (např. Studijní program, Forma studia nebo Stupeň předchozího vzdělání)

Tabulka 2.14: Popis atributů code_names

2.2.2.13 d_time

Dimenzionální tabulka ze stávajícího datového skladu určuje jednotlivé semestry. Použití ve vztahu k přihláškám je jejich zařazení do určitého akademického roku, konkrétně do toho, do kterého se zájemce hlásí. Např. přihláška

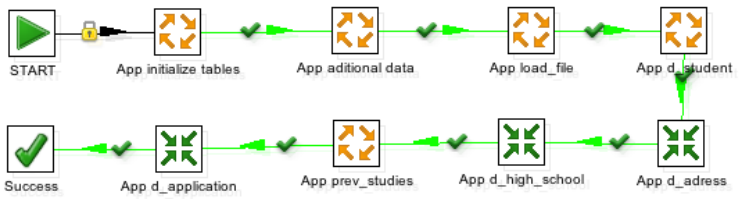
podaná v březnu 2009 bude mít zařazení do zimního semestru 2009/2010. Tabulka obsahuje několik atributů, pro nás budou důležité pouze tyto:

Název atributu	**Datový typ**	**Popis**
id_time	integer	Primární klíč tabulky d_time
k_semester_code	varying(50)	Kód semestru (např. B131 označuje zimní semestr 2013/2014)
k_content_name	text	Název semestru
k_content_startdate	date	Datum začátku semestru
k_content_enddate	date	Datum konce semestru

Tabulka 2.15: Popis atributů d_time

2.2.3 ETL procesy

Pro nahrávání dat do datového skladu se používají ETL procesy, které byly vytvořeny pomocí programu Pentaho Data Integration (PDI)[10]. Procesy se dělí na úlohy (jobs) a transformace (transformations), kde úloha je nadmnožina transformací. Úlohy jsou vždy prováděny sériově (následující začíná po dokončení první), transformace mohou být prováděny pseudoparalelně (vnitřní metodika PDI). Proto jsou některé transformace uvnitř samostatné úlohy, aby byla jistota pořadí prováděných kroků.



Obrázek 2.5: Úloha Applications Data

Celý proces nahrání přihlášek je zpracován v úloze Applications_data (viz obrázek 2.5), která data přečte, zpracuje a nahraje do koncové databáze. Pro správné fungování je potřeba před samotným spuštěním úlohy provést:

- Nastavit správné připojení k databázi (Database connection) s názvem *Application*. Zde je předpokládáno schéma *public* a již existující tabulky

[10] http://community.pentaho.com/projects/data-integration/, program pro vytváření ETL procesů a datové integrace

vytvořené podle přiloženého SQL skriptu (což by mělo být vytvořeno z předchozích let).

- Nastavit následující proměnné vedoucí na 3 různé složky:
  - *preprocessed-data* ke složce, kam se budou ukládat dočasné soubory,
  - *original-data* ke složce, která obsahuje data ke zpracování,
  - *additional-data* ke složce, která obsahuje dodatečná data (stačí pouze při prvním spuštění, data se jednou nahrají do datového skladu a není nutné nahrání opakovat).
- Mít aktuální tabulku *d_time*, která obsahuje kódy semestrů. Po ukončení přijímacího řízení, které se koná v roce 2014, musím mít v této tabulce před zpracováním dat již nahraný semestr B141, tedy zimní semestr 2014/2015. To by mělo být vždy, ale je vhodné data zkontrolovat.

Poznámka k původním datům: ETL procesy jsou nastaveny na konkrétní soubory, protože každý rok se data lišila a je nutné je zpracovat do stejné podoby. V dalších letech je předpokládané využití pouze posledního procesu (tedy pro rok 2013/2014) za podmínky, že se struktura a interpretace dat již nebude lišit. V opačném případě bude nutné vytvořit novou transformaci a nastavit správné mapování sloupců (a případné další změny).

V této kapitole jsou popsány základy toho, co které transformace nebo úlohy dělají. Podrobný popis (včetně jednotlivých kroků transformací) najdete v příloze A.1 na straně 91.

2.2.3.1 Úloha App initialize tables

Některá data se nemusí nahrávat pokaždé, ale pouze při prvním spuštění. To je případ i této úlohy, která má na starosti vytvoření záznamů pro fakulty ČVUT (tranformace *App d_faculty*) a středních škol (transformace *App create high schools*).

2.2.3.2 Úloha App additional data

Tato úloha obsahuje jedinou transformaci *Load additional data*, která slouží pro nahrání dodatečných dat (mapování kódů a jejich názvů pro tabulku *code_names*). Vezme všechny soubory, které budou nalezeny ve složce, ke které vede proměnná *additional-data*.

Tyto soubory musí být stejného typu (**.xls*) a obsahovat sloupce v tomto pořadí: *Kód*, *Název* a *Popis* (tedy klíčové slovo, podle kterého se rozliší typ kódů).

Kód	Název	Popis
1	střední odborné učiliště	Typ střední školy
2	gymnázium	Typ střední školy
3	střední odborná škola	Typ střední školy
4	konzervatoř	Typ střední školy
5	ostatní	Typ střední školy

Tabulka 2.16: Příklad souboru TypSS.xls, který slouží pro sjednocení hodnot atributu *a_high_school_type*

2.2.3.3 Úloha App load_file

Nahrání jednotlivých souborů s daty se děje v této úloze. Úpravy dat specifické pro konkrétní ročníky jsou provedeny v těchto transformacích, úpravy jednotné pro všechna či většinu dat jsou provedeny později.

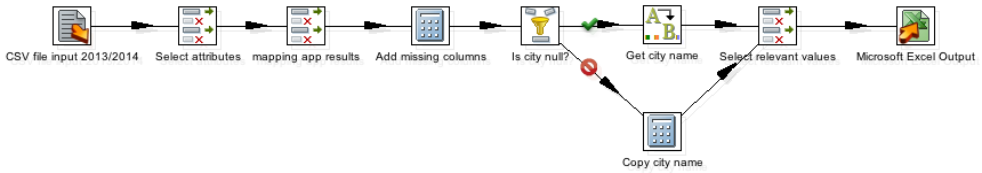
Jak je uvedeno výše, každý ročník má svoji vlastní transformaci, protože se lišil význam sloupců. Všechny transformace mají společné jádro: načtení relevantních sloupců, jejich přejmenování podle jmenných konvencí a doplnění neexistujících sloupců. Výsledkem je tedy jednotný formát dat. Níže jsou popsána specifika jednotlivých ročníků.

Pozn: V každém roce je zvlášť krok *mapping app results*, který páruje hodnoty H1 až H10 (případně HODN1 až HODN10) s jejich významovou hodnotou (body ze Scio, varianta přijímací zkoušky, apod.). Je tomu tak z toho důvodu, že pokud se v dalších letech nezmění formát dat, ale pouze tyto hodnoty, je možné použít transformaci z posledního roku a upravit pouze tento krok.

App load_file 2009_10, App load_file 2010_11
V těchto datech začínají některá IZA nulami, v dalších letech tomu tak již není. Z toho důvodu jsou nuly odstraněny.

App load_file 2011_12
Od tohoto roku došlo ke změně formátu dat, některé sloupce byly odstraněny, jiné přidány, změnily se jejich názvy. Další změna je, že někteří cizinci mají uvedeno město v jiném sloupci než Češi, je tedy nutné vzít tuto informaci z různých sloupců. A co se týče přijímacího řízení, maximální hranice bodového zisku z přijímací zkoušky byla 70, v dalších letech je to 100. Aby nedošlo ke zmatkům, všechny body jsou přeškálovány na maximum 100.

App load_file 2013_14
Kromě přeškálování totožné jako *App load_file 2011_12*.

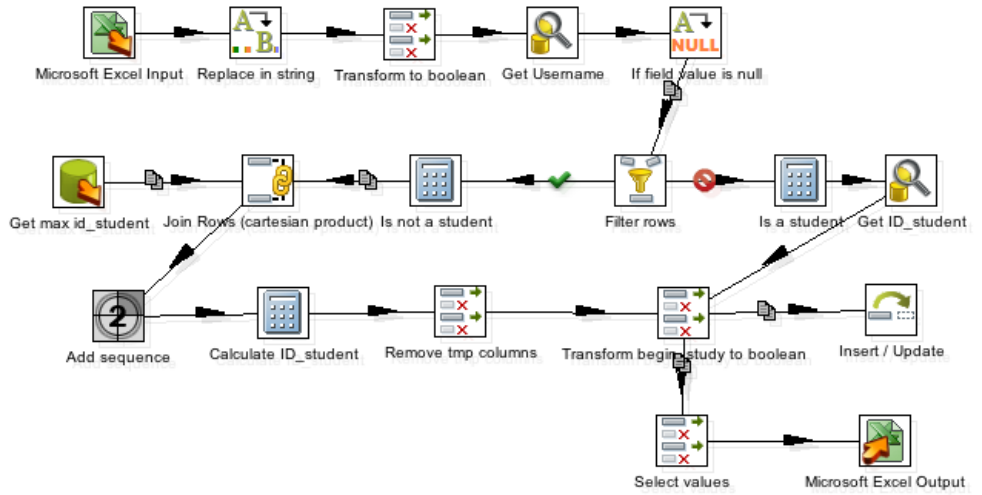


Obrázek 2.6: Transformace App load_file 2013_14

2.2.3.4 Úloha App d_student

Opět obsahuje jedinou transformaci, která aktualizuje dimenzionální tabulku *d_student*.

Nejdříve dojde ke sjednocení terminologie pohlaví („Ž“ a „Z“ na „Z“), podle username je vyplněna boolovská hodnota atributu *begin_study* a jsou nahrány nové atributy. Nikdy nedochází k úpravě již stávajících hodnot atributů označujících jméno, příjmení, username a rodné číslo. Zájemci, kteří v databázi ještě neexistují (tzn. nenastupivší) jsou anonymizováni a nadále vystupují pouze pod *id_student*, což je umělý klíč, přes který není možné dohledat konkrétní osobu. Tato data slouží pouze pro analytické a výzkumné účely, proto není nutné (a ani možné) uchovávat citlivé osobní údaje.



Obrázek 2.7: Transformace App d_student

Název	**Popis**
Microsoft Excel Input	Nahrání všech **.xls* souborů začínajících na „App_“ ze složky *preprocessed-data*
Replace in string	Sjednocení Z a Ž u pohlaví a změna hodnot u trvalého bydliště „A“ na „Y” (kvůli převodu na boolean)
Transform to boolean	Převedení *a_pernament_residency_cz* ze stringu na boolean
Get username	Podle jména, příjmení a rodného čísla je vrácen username z datového skladu
If field value is null	Nahrazení null hodnot odpovídajícími hodnotami (0, N/A)
Filter rows	Rozdělení dat na ty, které mají a nemají username
Is a student	Nastavení atributu *begin_study* na hodnotu „Y“
Get ID_student	Vrácení ID u stávajících studentů
Is not a student	Nastavení atributu *begin_study* na hodnotu „N“
Get max id_student	Získání maximálního ID ve skladu
Join rows	Spojení s daty nestudentů
Add sequence	Vytvoření posloupnosti čísel od 1 do počtu řádek (základ pro výpočet ID)
Calculate ID_student	Přičtení posloupnosti k maximálnímu ID a uložení jako nové ID
Remove tmp columns	Odstranění sloupců, které dočasně sloužily pro výpočet nových ID
Transform begin_study to boolean	Převod atributu *begin_study* na boolean
Insert / Update	Nahrání nových atributů ke stávajícím studentům a vytvoření anonymizovaných záznamů o nestudentech
Select values	Vybrání sloupců pro uložení
Microsoft Excel Output	Uložení zpracovaných dat do dočasného souboru *application-all.xls*

Tabulka 2.17: Popis jednotlivých kroků transformace d_student

2.2.3.5 Transformace App d_address

Tato transformace vytváří jednak tabulku *d_address*, tak tabulku *d_student_d_address*. Je to kvůli společným úpravám dat, které by bylo jinak nutné provádět dvakrát nebo upravená data ukládat do dočasného souboru. Tyto

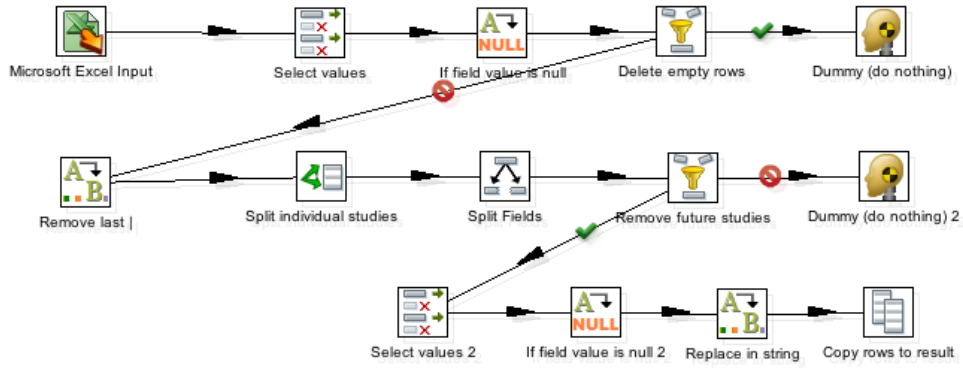
úpravy zahrnují vyhledání chybějících hodnot země a okresu, které jsou dohledány pomocí kódu v tabulce *code_names*.

2.2.3.6 Transformace App d_high_school

Podobně jako v předchozím případě má také tato tranformace za úkol vytvořit více tabulek najednou, konkrétně *d_high_school*, *d_high_school_field* a *d_high_school_study*. Jelikož informace o střední škole vyplňují pouze zájemci o bakalářské studium, dochází na začátku k vyfiltrování pouze těchto přihlášek.

2.2.3.7 Úloha App prev_studies

Tato úloha obsahuje několik transformací, které mají za úkol zpracovat informace o předchozích studiích. Součástí dat z přihlášky jsou také exportovaná data z KOS o studiích na ČVUT, bohužel se ale jedná o jeden dlouhý řetězec znaků. Ten je nutné rozparsovat na logické celky: fakulta, obor, datum zápisu, aritmetický a vážený studijní průměr. Také zde jsou všechna studia, takže pokud je export až po zápise studentů, mají zde jako předchozí studium to, do kterého se hlásili. To je nutné pomocí data zápisu vyfiltrovat. Ti, kteří nejsou z ČVUT mají vytvořené předchozí studium pouze s popisem, který vyplnili do přihlášky. Strojově tento údaj nelze jednoduše zpracovat, slouží pouze jako doplňující údaj, kdyby někoho zajímal konkrétní student, resp. konkrétní studium.



Obrázek 2.8: Transformace App parse_prev_studies

2.2.3.8 Tranformace App d_application

Jedná se o nejrozsáhlejší transformaci, která je použitá pro data z přihlášek. Jednak je to proto, že přihláška se páruje se semestrem, do kterého se studenti hlásí. Porovnává se rok, ve kterém byla přihláška vytvořena s rokem počátku

zimního semestru. Dále se provádí doplňování významů zkratek či vnitřních číselníků (studijní program, stupeň předchozího vzdělání, apod.) a vytváření boolenovských hodnot z číselných nebo znakových (zápis do studia, bakalář z FIT nebo ČVUT).

2.2.4 Automatizace celého procesu

Jednou z předností datových skladů je možnost automatizace procesů a jednoduchého nahrávání. To byl i jeden z požadavků na nahrávání dat z přihlášek. Vzhledem k častým změnám ve zdrojových datech byla automatizace velmi těžko proveditelná, i přesto jsou ale procesy maximálně automatizované. Data z přihlášek jsou nejdříve předzpracována do jednotného formátu a zbytek lze provádět automaticky (požadované soubory na zpracování jsou nahrány do určité složky a všechny se zpracují). Současný stav tedy vyžaduje ruční zásahy:

- získání exportu dat a nahrání do příslušné složky, odkud se bude zpracovávat
- je nutné porovnat formát dat s předchozím rokem a pokud se liší, je nutné vytvořit či upravit odpovídající ETL proces

Aby odpadl ruční zásah a proces byl zcela automatický, je nutné:

- dodržet jednotný formát dat,
- získávat export dat automaticky či získat přímý přístup do zdrojových systémů.

Bohužel hned první podmínka je velmi obtížná splnit. Většina automatizovaných procesů provádí denní nebo týdenní nahrávání dat do skladu. Tady se tak děje jen jednou ročně, tudíž je mnohem větší pravděpodobnost změny na úrovni zdrojových systémů.

Kapitola 3

Analýzy

V této kapitole si popíšeme, co zajímavého můžeme z dat získat, jakými metodami a k čemu nám to bude dobré. Získávání těchto znalostí se nazývá Data mining (tedy dolování dat), přesnější definice říká, že se jedná o hledání zajímavých vztahů, které jsou netriviální a mohou být užitečné pro majitele dat.

Jedním z nejčastějších přístupů dolování dat je CRISP-DM[11] metodika. Ta má celkem 6 částí:

- Porozumění problematice (Business Understanding)
- Porozumění datům (Data Understanding)
- Příprava dat (Data Preparation)
- Volba vhodného modelu (Modelling)
- Získání výsledků (Evaluation)
- Využití výsledků (Deployment)

Porozumění problematiky je odlišné od porozumění vlastním datům. V tomto konkrétním případě problematika obsahuje způsob přijímacího řízení a pochopení, co je důležité pro přijetí, nepřijetí, absolvování a neabsolvování. Jako studentka mám výhodu, že většinu věcí znám, protože se v prostředí každodenně pohybuji. To je rozdíl od bodu dvě, porozumění datům. Do toho spadá rozkódování způsobu uložení dat v aktuální databázi, resp. exportu, a odhalení, co tím bylo vlastně myšleno (obzvlášť, pokud chybí dokumentace).

Příprava dat proběhla z největší části při nahrávání do datového skladu, kde jsou data vyčištěna a připravena pro analýzy. To je další z velkých výhod datového skladu (či centrálního úložiště). Doteď, pokud bylo potřeba udělat analýzy, musela se data pokaždé zpracovat znovu neboť zdrojem byl export

[11] CRoss-Industry Standard Proces for Data Mining

(většinou z KOS), tedy soubor s maticí dat. Nyní už bude stačit pouze uložit data do požadovaného formátu a předzpracování dat se omezí jen na úpravy pro konkrétní model.

Nyní přejdeme k vytvoření prediktivního modelu a poté k analytickým otázkám, které nám umožní hledat závislosti v datech. U nich začneme s technikou asociačních pravidel a podle zjištěných skutečností se pak zaměříme podrobněji na jednotlivé podmnožiny dat.

3.1 Prediktivní model

Podobně jako u jiných škol jsme se i my pokoušeli vytvořit model, který by na základě atributů o studentech dokázal určit, zdali po prvním semestru získají či nezískají 15 kreditů, tedy uspějí nebo neuspějí.

Pro vytváření modelů budeme používat program Weka.

3.1.1 Data

Použili jsme pouze atributy, které známe o studentovi před samotným studiem, tedy:

- Osobní údaje – Pohlaví, Občanství, Trvalé bydliště v ČR, Bydliště (Země, Okres, Město)
- Údaje přijímacího řízení – Forma studia, Stupeň předchozího vzdělání, Odkud se uchazeč hlásí, Rozhodnutí o přijetí
- Střední škola – Průměr z maturity, Průměr z matematiky, IZO, Typ střední školy

Vzhledem k použitým atributům jsme vzali v úvahu pouze studenty, kteří mají vyplněné studijní průměry. Celkem se jedná o 479 studentů, z toho 159 po prvním semestru neuspělo a 320 uspělo.

V první řadě se musíme zamyslet nad relevancí daných atributů – např. mezi občanstvím a zemí trvalého bydliště je logicky vysoká korelace a data jsou poměrně nevyvážená. Použijeme tedy metodu Feature ranking pro ohodnocení jednotlivých atributů (viz tabulka 3.1). Přibližně polovina z nich má minimální přínos. V prvním datasetu ponecháme atributy všechny, ve druhém ponecháme pouze 6 nejlepších atributů: Bydliště – město, IZO, Bydliště – okres, Stupeň předchozího vzdělání, Typ střední školy a Průměr z maturity.

Další věcí, kterou budeme řešit, je nevyváženost dat. Pro třetí dataset tedy zvolíme pouze nejlepší atributy a použijeme techniku SMOTE pro balancování dat.

Hodnota	Název atributu
0,444285	Bydliště – město
0,412772	IZO
0,1107	Bydliště – okres
0,038125	Stupeň předchozího vzdělání
0,037324	Typ střední školy
0,03538	Průměr z maturity
0,018231	Průměr z matematiky
0,013932	Občanství
0,012599	Bydliště – země
0,010057	Odkud se uchazeč hlásí
0,00768	Forma studia
0,000506	Trvalé bydliště v ČR
0,000393	Pohlaví
0	Rozhodnutí o přijetí

Tabulka 3.1: Ohodnocení atributů

3.1.2 Modelování a testování

Pro vytvoření našeho modelu použijeme více klasifikátorů, konkrétně Naivní bayesovský klasifikátor (NaiveBayes) a rozhodovací stromy ADTree, J48 (neprořezávaný), RandomTree a RandomForest. Pro měření výsledků si popíšeme matici záměn (Confusion matrix), která říká, kolik klasifikací bylo správných a kolik špatných. Mějme dvě třídy, pozitivní (P) a negativní (N). V matici jsou pak uvedeny všechny možné typy klasifikace:

- TP (True Positive) – počet správně klasifikovaných vzorů, pokud patří do třídy P
- TN (True Negative) – počet správně klasifikovaných vzorů, pokud patří do třídy N
- FP (False Positive) – počet vzorů klasifikovaných do třídy P, pokud patří do třídy N
- FN (False Negative) – počet vzorů klasifikovaných do třídy N, pokud patří do třídy P

Jako míru pro ohodnocení klasifikátoru budeme používat celkovou přesnost (Accuracy), která se vypočítá následovně:

$$ACC = \frac{TP + TN}{TP + FP + TN + FN}$$

Pro testování a trénování dat budeme používat desetinásobnou křížovou validaci.

Klasifikátor	**Dataset 1**	**Dataset 2**	**Dataset 3**
NaiveBayes	63,88 %	63,04 %	70,21 %
ADTree	**66,38 %**	**64,30 %**	71,31 %
J48	63,25 %	63,04 %	**73,04 %**
RandomTree	62,21 %	64,09 %	71,94 %
RandomForest	62,63 %	62,21 %	71,47 %

Tabulka 3.2: Výsledky měření

Výsledky měření vidíme v tabulce 3.2. Pro žádný dataset ani klasifikátor nebyly výsledky nijak ohromující, přesnost kolem 65 % je vskutku poměrně malá. Vzhledem k malé různorodosti použitých atributů, jejich nevyváženosti se to ale dalo očekávat (hodně napovědělo i ohodnocení atributů). Z toho vyplývá, že použité atributy nejsou bohužel pro prediktivní model postačující.

Pro řešení tohoto problému bylo by vhodné datasety doplnit o další atributy, například schopnosti studentů jako jsou soft skills (viz studie [11]). Takové informace ale bohužel k dispozici nemáme a ani techniky pro ohodnocení takových schopností. Můžeme ovšem použít data z Progtestu, tedy jak si studenti vedou během semestru v programování, neboť úlohy se odevzdávají online a opravují automaticky. Je možné tak sledovat aktivitu studentů, zdali si úlohu vyzvednou, nevyzvednou a jak dlouho trvá případné vypracování.

3.2 Asociační pravidla

Asociační pravidla hledají zajímavé vztahy mezi dvojcemi booleovských atributů, které jsou odvozeny ze sloupců analyzované matice dat. Mezi booleovskými atributy je 4ft kvantifikátor, kterému je přiřazena podmínka týkající se čtyřpolní tabulky. Skládá se z antecedentu, sukcedentu, příp. podmínky a popisuje kritérium pro míru zajímavosti. Výsledek asociačního pravidla může být true nebo false.

Čtyřpolní tabulka označuje vztah dvou booleovských atributů, má 4 políčka (proto čtyřpolní) a v každém z nich zobrazuje četnost, obecně označovanou písmeny a, b, c a d. Platnost antedentu označují a a b, platnost sukcedentu pak a a c.

3.2.1 Kvantifikátory

Existuje několik typů kvantifikátorů, my si zde popíšeme pouze 4 základní, a to fundovanou implikaci, dvojtou fundovanou implikaci, fundovanou ekvivalenci a Above Average.

Každý kvantifikátor má přiřazenou funkci, pomocí které provede zobrazení na $\{0, 1\}$, tedy na true nebo false. Ve funkcích se používají prvky čtyřpolní tabulky, antecedent ϕ a sukcedent ψ, podpora $Base$ a spolehlivost p. Podpora (support) určuje podíl počtu položek, kdy platí antecedent i sukcedent vůči všem položkám, jedná se tedy o frekvenci výskytu daného pravidla. Spolehlivost (confidence) určuje podíl počtu položek, kdy platí antecedent i sukcedent vůči položkám, kdy platí antecedent, jedná se tedy o sílu pravidla.

4ft kvantifikátor fundované implikace

$$\varphi \Rightarrow_{p,Base} \psi : \frac{a}{a+b} \geq p \wedge a \geq Base$$

Zajímá nás, zda platnost nějaké kombinace znamená s vysokou pravděpodobností i platnost nějaké jiné kombinace.

Platnost $\varphi \Rightarrow_{p,Base} \psi$ znamená, že nejméně $100 \cdot p$ % řádků splňujících φ splňuje také ψ a nejméně $Base$ řádků splňuje jak φ, tak ψ.

4ft kvantifikátor fundované ekvivalence

$$\varphi \equiv_{p,Base} \psi : \frac{a+d}{a+b+c+d} \geq p \wedge a \geq Base$$

Zajímá nás, zda platnost nějaké kombinace je téměř ekvivalentní platnosti nějaké jiné kombinace.

Platnost $\varphi \equiv_{p,Base} \psi$ znamená, že pro nejméně $100 \cdot p$ % řádků mají φ a ψ stejnou hodnotu a nejméně $Base$ řádků splňuje jak φ, tak ψ.

4ft kvantifikátor dvojité fundované implikace

$$\varphi \Leftrightarrow_{p,Base} \psi : \frac{a}{a+b+c} \geq p \wedge a \geq Base$$

Zajímá nás, zda pro nějaké dvě kombinace platí, že když je splněna alespoň jedna z nich, tak jsou velmi často splněny obě.

Platnost $\varphi \Leftrightarrow_{p,Base} \psi$ znamená, že pro nejméně $100 \cdot p$ % řádků splňujících φ nebo ψ splňuje φ i ψ a nejméně $Base$ řádků splňuje jak φ, tak ψ.

4ft kvantifikátor Above Average

$$\varphi \Rightarrow^{+}_{p,Base} \psi : \frac{a+c}{a+b+c+d}(1+p) \leq \frac{a}{a+b} \wedge a \geq Base$$

Zajímá nás, zda platnost nějaké kombinace znamená výrazné zvýšení relativní četnosti nějaké jiné kombinace.

Platnost $\varphi \Rightarrow^{+}_{p,Base} \psi$ znamená, že relativní četnost řádků splňujících ψ mezi řádky splňujícími φ je o $100 \cdot p$ % vyšší, než relativní četnost řádků splňujících ψ v celé matici M a nejméně $Base$ řádků splňuje jak φ, tak ψ.

Pro získávání hypotéz pomocí asociačních pravidel je použit program LISp-Miner[12], který je vyvíjen na Vysoké škole ekonomické pod vedením pana profesora Jana Raucha a docenta Milana Šimůnka.

3.2.2 Příprava dat v programu LISp-Miner

Cílem této práce je nalézt zajímavé vztahy mezi výsledky ze střední školy a studii na naší fakultě. Z toho důvodu byla zohledněna pouze první studia studentů, protože ta další jsou již ovlivněna předchozím studiem na fakultě.

Předzpracování dat z LISp-Mineru nebyla složitá záležitost, protože data jsou z datového skladu v dobrém stavu. Bylo ale nutné zvolit význam NULL hodnot a rozdělit data na ekvidistanční nebo ekvifrekvenční intervaly. Ve většině případů je používáno něco mezi, tedy snaha o ekvifrekvenční intervaly s logickými krajními hodnotami.

Dále byly vytvořeny skupiny atributů:

- Osobní údaje
- Střední škola
- Přihláška
- Studium
- Předměty

Osobní údaje
Atributy: Občanství, Město a okres bydliště a Pohlaví.

Většina těchto dat je nevyvážených, na což je třeba brát ohled při formulaci analytických otázek. České občanství má 71 % studentů, 5 % jsou občané Slovenské republiky a 3 % Ruské republiky, ostatní jsou zanedbatelné. U trvalého bydliště je 24 % studentů z Prahy, 2,3 % z okresu Praha-západ a 2,2 %

[12] http://lispminer.vse.cz

z okresu Kladno. Co se pohlaví týče, na fakultě je 92 % studentů a 8 % studentek.

Střední škola
Atributy: Průměr z matematiky, Průměr z maturity, Rok maturity, IZO střední školy, Město střední školy a Typ střední školy.

Většina těchto atributů byla povinná pouze v prvních letech fakulty, proto je poměrně málo záznamů, které má tyto údaje vyplněné. Průměr z matematiky nemá vyplněno 61 % studentů, průměr z maturity dokonce 82 %. Naopak u mnoha studií známe IZO střední školy, ale kvůli příliš velké rozmělněnosti studentů není pro tento typ analýz příliš vhodné. Typ střední školy ukazuje, že gymnazisté a studenti středních odborných škol jsou vyvážené skupiny (33 % vs 34 %), ostatní skupiny jsou zanedbatelné.

Přihláška
Atributy: Odkud se uchazeč hlásí, Rozhodnutí o přijetí, Scio, Olympiády, Varianta přijímací zkoušky a Výsledek přijímací zkoušky.

Bez přijímací zkoušky bylo přijato 54 % (2225) uchazečů, na jejím základě pak 16 % (679). Tento nepoměr je dán také tím, že v prvním roce byli přijati všichni uchazeči (cca 500). U 26 % studentů tento atribut není známý.

Studium
Atributy: Součet kreditů za první semestr studia, Obor, Stav studia, Studijní program a Semestr začátku studia.

Vzhledem k malé historii fakulty je absolventů stále poměrně málo (8 % v BSP a 6 % v MSP), 43 % jsou aktivní studenti a 40 % ukončilo studia (20 % vyloučeno a 20 % zanechalo). Za první semestr mají studenti nejčastěji 30 kreditů (27 %) nebo 0 (17 %). Obory jsou opět nevyvážené, nejvíce studentů BSP je na Softwarovém inženýrství (8 %), v MSP na Webovém a softwarovém inženýrství (10 %). Tato čísla jsou ale směrodatná jen u absolventů, protože ostatní studenti nemají povinnost si obor volit dříve než s volbou bakalářské práce a u většiny je tedy obor neznámý (60 %).

Předměty
Atributy: povinné předměty BSP a MSP.

V této skupině jsou všechny povinné předměty obou programů Informatika a jejich známky. Rozložení známek se liší podle předmětů. Pokud si odmyslíme známky F (ty většinou převažují ostatní), můžeme předměty rozdělit do tří kategorií: lehké, střední a těžké. Lehké předměty mají nejlepší známky A

a četnosti klesají. Ty těžké mají rozložení přesně opačné, nejvíce je známek E, A je nejméně. Středně těžké předměty mají normální rozdělení: známek A a E je zhruba stejně, nejčastější známka je C. Existují také speciální případy s rovnoměrným rozdělením.

- Lehké předměty
  - BSP: ČAO, SAP, TED
  - MSP: TES
- Středně těžké předměty
  - BSP: AAG, BEZ, DBS, MLO, PAI, PPR, PSI, PST
  - MSP: MPI, PAA, SPI
- Těžké předměty
  - BSP: LIN, PA2, UOS, ZDM, ZMA
  - MSP: PAR (téměř žádná E)
- Speciální případy (rovnoměrné rozdělení)
  - BSP: OSY, PA1

3.2.3 Analytické otázky

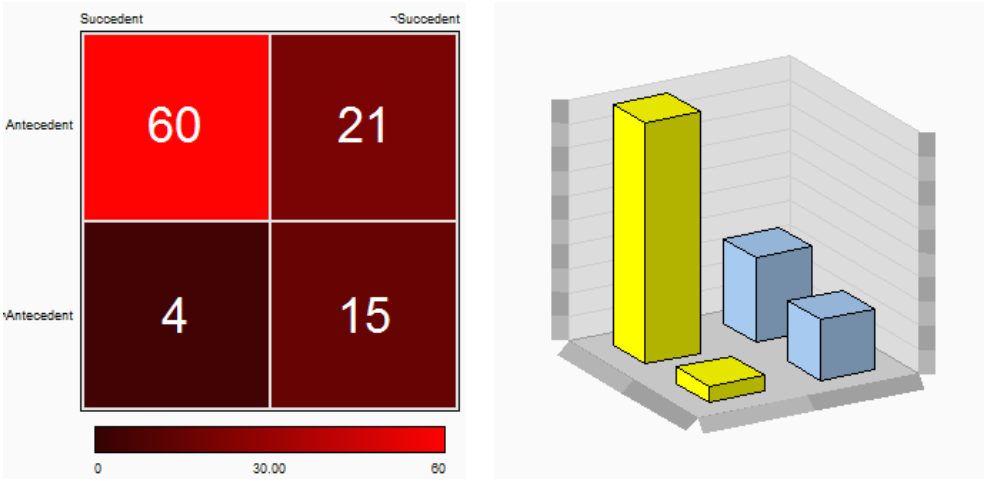
Pomocí asociačních pravidel budeme hledat nějaké závislosti v datech. Po nalezení či nenalezení konkrétních vztahů se na dané oblasti zaměříme podrobněji.

Jaký vliv mají kombinace atributů ze skupiny Osobní údaje a Střední škola na počet získaných kreditů za první semestr?

Kvantifikátor: Fundovaná implikace
Podpora: 1 %
Spolehlivost: 0,7

Původně jsme zkoumali pouze neúspěšné studenty, tedy ty, kteří mají méně než 15 kreditů, ale nebyla nalezena žádná silná hypotéza (pouze menší než 50 %), což bude ovlivněno tím, že mnoho studentů zanechá studií již na počátku a ti nemají žádné specifické charakteristiky. Proto bylo zadání upraveno a zaměřili jsme se na vliv lepšího průměru z maturity a percentilu ze Scia.

Výsledkem je hypotéza, která říká, že studenti, kteří jsou z gymnázií nebo mají průměr z maturity do 1,25 a jsou přijati na základě olympiád nebo percentilu ze Scia nad 95, mají pravděpodobnost 75 %, že získají z prvního semestru více než 30 kreditů.



Obrázek 3.1: Nejvýznamnější hypotéza – čtyřpolní tabulka

Tato hypotéza platí pro 60 případů, kdy antecedent splňuje celkem 81 případů (průměr z maturity do 1,25 nebo gymnázium a olympiády nebo Scio nad 96 percentil) a pouze sukcedent (více než 30 kreditů) splňuje 64 případů. Jedná se tedy o poměrně silnou hypotézu.

Má vliv kombinace atributů ze skupiny Střední škola a Přihláška na zvýšení relativní četnosti vysokého počtu kreditů a úspěšného zakončení těžkých předmětů?

Kvantifikátor: Above Average
Podpora: 0,5 %
Spolehlivost (rozdílu skupin): 0,1

Ano, skutečně se prokázalo, že to vliv má, konkrétně atributy Průměr z matematiky a percentil ze Scio testů. Nejvýznamější hypotéza říká, že studenti s průměrem 1,0 získají 2,5 krát častěji více než 25 kreditů a ukončí BI-ZMA za A vůči ostatním studentům. Další hypotézy potvrzují, že studenti s vynikajícím průměrem z matematiky (1,0 až 1,5) a percentilem ze Scio (nad 90 percentil) o více než 100 % častěji ukončí předměty ZMA, UOS nebo LIN za A nebo B a získají více než 25 kreditů za první semestr studia.

Je nutné si však uvědomit, že se jedná o porovnání relativních četností. Existuje stále mnoho studentů s vynikajícím prospěchem ze střední školy, kteří mají známky horší nebo studium zanechají hned v prvním semestru. Je jich však relativně méně než těch, kteří mají prospěch horší.

Jak se chovají frekvence rozložení známek jednotlivých předmětů v závislosti na atributech Přihlášky a Studia?

Procedura: CF Miner
Podpora: 20
Počet klesajících schodů: 4

S rostoucím počtem bodů ze Scia, výsledků přijímací zkoušky nebo přijetí na základě olympiád roste i počet studentů s lepšími známkami z PA1. Nejvíce studentů má A (42) a nejméně E (15). Podmnožina 159 studentů dokáže změnit tvar histogramu známek předmětu PA1, tedy z rovnoměrně rozloženého na klesající.

Při percentilu ze Scia nad 96, průměru z maturity 1, průměru z matematiky do 1,25 nebo více než 31 kreditů za první semestr se mění histogram také u ZMA z rostoucího na rovnoměrný pro 347 studentů.

Tato otázka je spíše potvrzení toho, co nám vyšlo u otázek předchozích. Bohužel nebyla nalezena žádná podmnožina, která by byla dostatečně relevantní a obracela histogram z rostoucího na klesající (např. u předmětu UOS).

Jaký má vliv pohlaví na výsledné známky z předmětů?

Procedura: KL Miner
Podpora: 20

Ženy vs. muži, kdo je na tom lépe? Obecně se předpokládá, že ženy jsou pečlivější, muži mají lepší analytické myšlení. Ve skutečnosti jsou obě skupiny poměrně vyvážené.

V předmětu Počítačové sítě se vede lépe studentům, kteří mají poměrně rovnoměrně rozdělené známky, studentky mají v 50 % případů C nebo D a A mají jen ve 4 %. O něco hůř jsou na tom také v programovacích předmětech PA1 a PA2, kdy mají méně často známky A a B.

Jednoznačně lépe se studentkám vede v předmětu Automaty a gramatiky, kde dosahují nejčastěji známky B, studenti pak známky C. Nejlepší známka A je pro obě skupiny stejně častá. V Právu a informatice více často dosahují na známku A, o pár procent lépe jsou na tom i v Základech matematické analýzy. Výrazně lépe jim jde Matematická logika, kde mají častější známky A, B i C.

Naprosto vyrovnaná je situace v předmětech Úvod do operačních systémů a Lineární algebra.

Existuje závislost mezi průměrem z matematiky a matematickými předměty?

Procedura: KL Miner
Podpora: 20

Ano, skutečně existuje, nejvýznamnější je závislost u předmětů BI-MLO a BI-ZMA. Zajímavé je, že u předmětu Matematická logika jsou nejlepšími studenty ti s nejlepším průměrem a naopak. Dokonce i studenti s průměrem kolem 2,0 mají nejčastěji známku C.

U předmětu Základy matematické analýzy to platí pouze v případě nejhorších studentů. Studenti s průměrem 1,0 jsou rovnoměrně rozděleni mezi všechny známky. Pokud však vezmeme v úvahu relativní počet vůči známkám, jednoznačně vítězí studenti s průměrem 1, protože téměř žádní jiní studenti A z předmětu nezískali.

Lineární algebra nemá tak významnou závislost jako předchozí předměty, nicméně opět platí, že student s lepším průměrem má lepší známku. Je to trochu zkresleno tím, že hodně studentů na první pokus neuspěje.

Závislosti takové, že studenti s lepším průměrem mají lepší známku, ale už to neplatí o špatných známkách, jsou kupodivu u předmětů AAG a PST. Je vidět, že tito studenti mají lepší předpoklady, ale již se smazává ta hranice, kdy studenti s horším průměrem měli horší známky.

3.3 Analýza oblastí dat

Na základě výše uvedených hypotéz si podrobně zanalyzujeme data pro konkrétní témata, tedy přijímací řízení a střední školy. Existují samozřejmě i další oblasti pro analýzu, avšak cílem této práce je zaměřit se na vliv výsledků ze střední školy a přijímacího řízení na studium na fakultě.

Použitý dataset je tentýž jako u asociačních pravidel, tedy první studia studentů, aby byl vliv předchozího studia co největší, a známky z jejich prvních zápisů, protože se zaměříme na postup z prvního semestru do druhého.

3.3.1 Střední škola

Pomocí asociačních pravidel jsme se snažili zjistit nějakou závislost mezi studijními výsledky podle typu střední školy, tedy jak moc se liší student gymnázia a střední průmyslové školy. Nebyla nalezena žádná relevantní hypotéza, proto se domnívám, že si tito studenti nevedou příliš odlišně, což si zkusíme nyní ověřit.

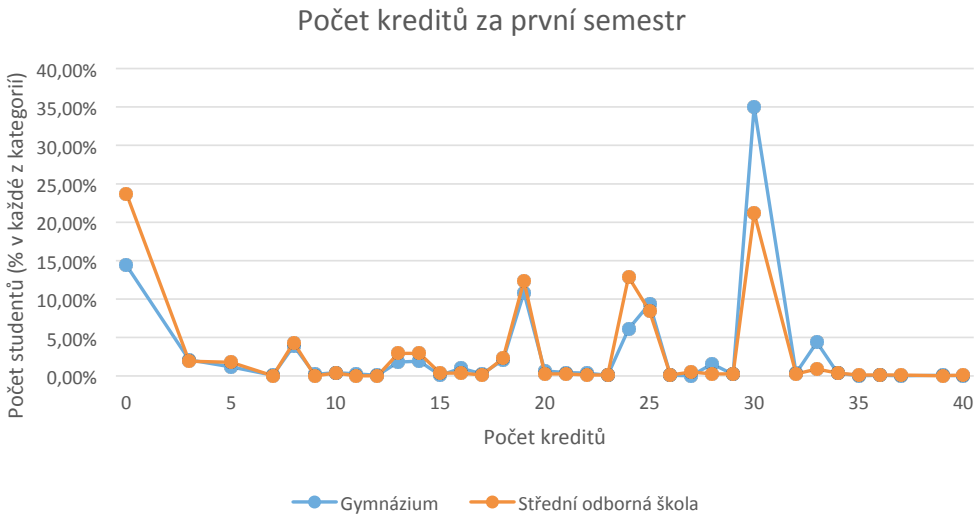
Střední škola může mít celkem 5 typů: gymnázium, střední odborná škola, střední odborné učiliště s maturitou, konzervatoř a ostatní. První dvě kategorie mají podobný počet studentů (784 a 1096), další se pohybují v řádu jednotek a jsou zanedbatelné, proto se zaměříme pouze na první dvě. U přibližně 30 % studentů tento údaj není znám.

Budeme brát v úvahu tři důležitá kritéria, kterými jsou: počet zapsaných studentů, počet studentů s více než 15 kredity po prvním semestru a počet absolventů. Pokud budeme uvádět procenta, jsou absolventi vztaženi k úspěšným studentům, nikoli k celkově zapsaným.

Typ školy	**Zapsaní**	**Úspěšní**	**Absolventi**
střední odborná škola	784	486	87
gymnázium	1096	806	192

Tabulka 3.3: Přehled podle typu střední školy

Obě kategorie jsou poměrně vyvážené, střední odborná škola má ovšem vyšší úmrtnost po prvním semestru (38 %) než gymnázia (26 %) a méně absolventů.



Obrázek 3.2: Vliv typu střední školy na počet kreditů za první semestr

Pokud se podíváme podrobněji na rozložení kreditů pro oba typy škol (viz obrázek 3.2), jsou poměrně vyvážené, rozdíl je zejména ve skupině s 0 kredity a 30 kredity. Studenti s 0 kredity nastoupili buď na dvě školy zároveň a pokračují na druhé, nastoupili pouze kvůli statusu studenta nebo v průběhu semestru

nastoupili či dali přednost práci. Studenti gymnázií nemají žádnou specifikaci a předpokládá se pokračování na vysoké škole, na kterou se opravdu zaměřili, protože jednoznačně převažují ve skupině s více než 30 kredity (až do 44). Studenti středních odborných škol mají větší pravděpodobnost, že si najdou práci hned po maturitě a tedy studia nedokončí (toto bohužel daty podloženo není, protože nemáme žádnou informaci o tom, proč studenti získají 0 kreditů).

Při porovnání úmrtnosti studentů, kteří získali aspoň 1 kredit, se rozdíl zmenšuje. Gymnázium má úmrtnost 14 %, střední odborná škola pak 18 %.

Rozdíly mezi studenty, pokud studovali gymnázium nebo střední odbornou školu nejsou tedy tak významné, aby určovaly zásadní rozdíl mezi tím, zda student uspěje nebo neuspěje. Gymnazisté jsou ovšem obecně lepší v počtu kreditů nad 30.

Jelikož nebyl nalezen významný rozdíl mezi typem škol, zaměříme se na konkrétní střední školy a úspěšnost jejích studentů v prvním semestru a na počet absolventů na FIT. Podrobnější informace o středních školách rozepsané podle jednotlivých ročníků jsou v dashboardech (viz kapitola 4.2.3).

Název školy	**Zapsaní**	**Úspěšní**		**Absolventi**	
Gymnázium, Kladno	17	16	94,12 %	6	37,50 %
Gymnázium, České Budějovice	16	15	93,75 %	1	6,67 %
Střední škola a VOŠ aplikované kybernetiky s.r.o.	28	25	89,29 %	6	24,00 %
První české gymnázium v Karlových Varech	17	14	82,35 %	3	21,43 %
Gymnázium a SOŠ dr. Václava Šmejkala, Ústí nad Labem	17	14	82,35 %	5	35,71 %
Gymnázium, Praha 4, Na Vítězné pláni	19	15	78,95 %	2	13,33 %
SPŠ sdělovací techniky, Praha 1, Panská 3	63	49	77,78 %	11	22,45 %
Gymnázium Dr. Josefa Pekaře, Mladá Boleslav	19	14	73,68 %	3	21,43 %
SPŠ strojní a elektrotechnická a VOŠ, Liberec	19	14	73,68 %	2	14,29 %
Gymnázium, Praha 6, Arabská	30	22	73,33 %	7	31,82 %

Tabulka 3.4: Deset nejlepších středních škol podle úspěšnosti studentů v prvním semestru studia

V tabulce 3.4 je vypsáno 10 škol s nejvyšší úspěšností studentů po prvním semestru. Jedná se o ty školy, které mají alespoň 15 studentů. Překvapivě mají studenti mimopražských škol velmi dobrou úspěšnost, relativně lepší než školy pražské. Například Střední průmyslová škola elektrotechnická (Praha 2, Ječná 30), která se umisťuje na prvních příčkách v počtu podaných přihlášek a zapsaných studentů, má úspěšnost v prvním semestru pouze 52 %.

3.3.2 Přijímací řízení BSP

Předchozí hypotézy (viz kapitola 3.2.3) prokázaly vztah studentů, kteří byli přijati na základě olympiád, Scio testů s vysokým percentilem a vysokým průměrem a výsledky v následujícím studiu. Jelikož je klíčový opět první semestr, zaměříme se zejména na to, jak si vedly jednotlivé skupiny v získávání kreditů za první semestr.

Zkoumané skupiny:

- Všichni
- Olympiády – studenti přijatí na základě olympiád
- Scio nad 90 – studenti, kteří měli percentil ze SCIA vyšší než 90
- Scop nad 95 – studenti, kteří měli percentil ze SCIA vyšší než 95
- Přijímací zkouška nad 90 bodů
- Přijímací zkouška nad 95 bodů
- Průměr z matematiky 1,0
- Průměr z matematiky do 1,5
- Průměr z maturity 1,0
- Průměr z maturity do 1,5

Ačkoli jsou některé skupiny podmnožiny, pro naše hypotézy to postačí, neboť se můžeme zaměřit na širší skupinu a podívat se, jak se chová pro zpřísnění podmínek. První skupina zahrnuje všechny studenty, tudíž se bere jako průměrná hodnota a ostatní kategorie jsou s ní porovnávány.

Z tabulky 3.5 můžeme vyčíst, že nejmenší úmrtnost mají studenti s vysokým percentilem ze Scio testů, pokud je nad 95, je úspěšnost dokonce téměř 93 %, nad 90 percentil je stále na druhém pořadí v tabulce. Dalším významným kritériem je průměr z matematiky (do 1,5 úspěšnost 83 %) a za ním jsou studenti přijatí na základě olympiád (78 %). Nejhůře ze zkoumaných skupin dopadl průměr z maturity. Paradoxně studenti s horším průměrem mají vyšší úspěšnost než studenti s průměrem 1,0. Každopádně toto kritérium není vhodné pro rozhodnutí o přijetí. Také studenti s nejvíce body z přijímací

zkoušky si nevedou tak dobře, jak by se dalo předpokládat (kolem 72 %, což je pouze o 2 % více než průměrný student).

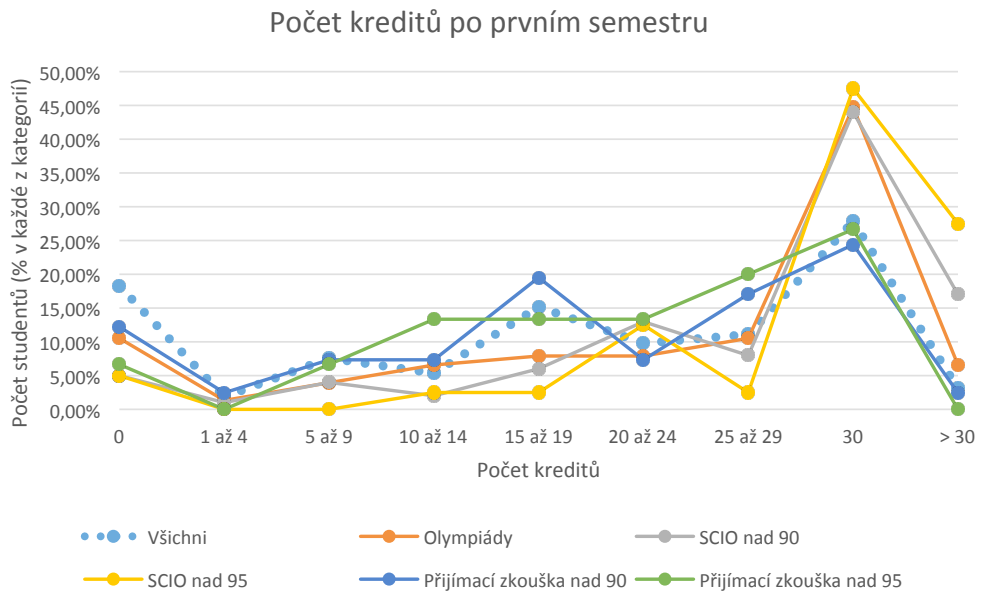
	Zapsáno	**Úspěšnost po 1. semestru**	
		Počet	**Procento**
Všichni	2976	1993	66,97 %
Olympiády	76	59	77,63 %
Scio nad 90	100	88	**88,00** %
Scio nad 95	40	37	**92,50** %
Přijímací zkouška nad 90	41	29	70,73 %
Přijímací zkouška nad 95	15	11	73,33 %
Průměr matematika 1,0	171	144	84,21 %
Průměr matematika do 1,5	350	292	83,43 %
Průměr maturita 1,0	80	49	61,25 %
Průměr maturita do 1,5	191	127	66,49 %

Tabulka 3.5: Vliv způsobu přijetí na úspěšnost 1. semestru

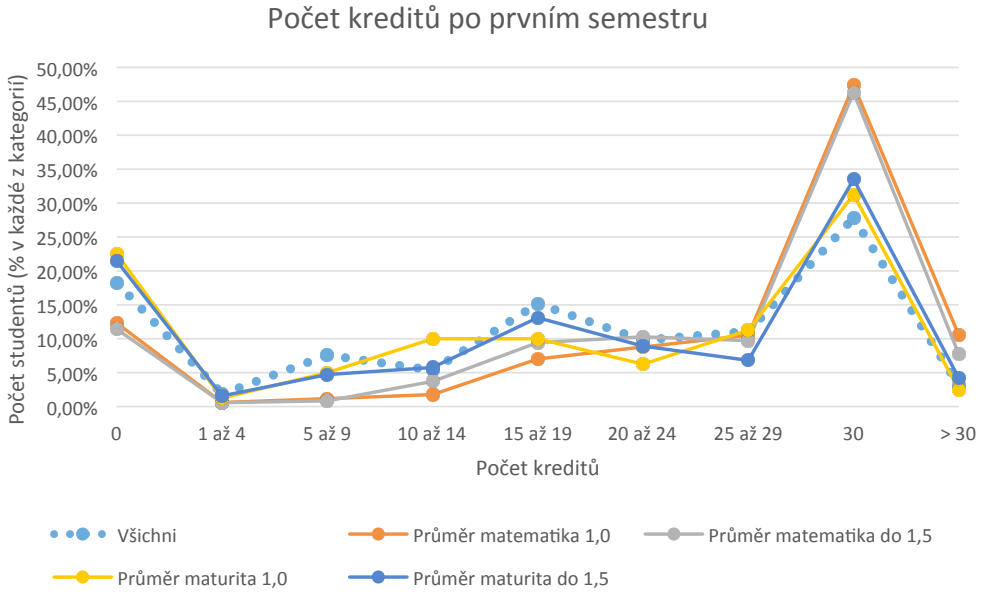
Podívejme se na konkrétní rozložení kreditů (skupiny po cca 5 kreditech pro větší přehlednost). Jelikož mají skupiny různý počet studentů, vzala jsem jejich relativní čísla, tedy procento v konkrétní skupině, tedy 45 % studentů, kteří mají 30 kreditů a přijetí na základě olympiád znamená 45 % ze všech, kterým byla prominuta zkouška kvůli olympiádám. Aby nedošlo k přeplnění grafu, rozdělila jsem ho na dva – první zobrazuje klasické podmínky pro přijetí, které se používají v těchto letech, druhý pak zobrazuje průměry z maturity a matematiky.

Oba grafy potvrzují to, co jsme částečně vyčetli již z tabulky. Všechny skupiny mají nižší počet studentů, kteří získali 0 kreditů. Nejlepší rozložení má Scio nad percentil 95 (pod 15 kreditů minimum a nad 30 kreditů mají nejvíce ze všech skupin), v těsném závěsu je percentil nad 90. Nejhůř dopadly přijímací zkoušky a průměr z maturity. V grafu je vidět, že relativně mají nevíce ze všech skupin v kategorii 10 až 14 kreditů, ale to je dáno malým počtem studentů v této kategorii.

Nejlepšími měřítky jsou tedy Scio testy nad percentil 90, překvapivě také průměr z matematiky do 1,5 a přijetí na základě olympiád.



Obrázek 3.3: Vliv způsobu přijetí na počet kreditů za první semestr, 1. část



Obrázek 3.4: Vliv způsobu přijetí na počet kreditů za první semestr, 2. část

3.3.3 Přijímací řízení MSP

Spolu s tématem přihlášek souvisí také přijímací řízení do magisterského studijního programu Informatika. Sice nemáme k dispozici tolik údajů jako u bakaláře, ale můžeme se aspoň zaměřit na rozdíly mezi našimi bakaláři a těmi z jiných škol.

Pokud se podaří doplnit do dat chybějící sloupec (viz kapitola 2.2.1.2) Předchozí vysoká škola, může v dalších let být analýza podrobnější o konkrétní školy, ze kterých naši magistři přišli.

První studenti magisterského programu nastoupili v zimním semestru 2010/2011, tedy ve druhém roce života fakulty. O dva roky později měla fakulta první absolventy jak bakalářského, tak magisterského programu. První studenti, kteří měli šanci vystudovat bakaláře na fakultě, nastoupili do magisterského programu v zimním semestru 2012/2013. Tyto analýzy proto proběhnou pouze pro dva ročníky, tedy 2012/2013 a 2013/2014.

	Absolvent FIT	**Jiný absolvent**	**Celkem**
2012/2013	100	97	197
2013/2014	179	59	238
Celkem	279	156	435

Tabulka 3.6: Přehled studentů MSP podle předchozího studia

První rok je situace velmi vyvážená, polovina studentů absolvovala FIT, druhá pochází ze školy jiné. Další rok již začínají převažovat naši absolventi, což se dá očekávat i v dalších letech kvůli nárůstu jejich počtu. V akademickém roce 2011/2012 absolvovalo na bakaláři 117 studentů, o rok později 221. Další rok se dá očekávat další nárůst, protože jen v zimním semestru tohoto roku (2013/2014) absolvovalo 20 studentů (o rok dříve pouze 3).

Opět se podíváme i na úspěšnost těchto studentů po prvním semestru. Úspěšnými studenty budeme myslet ty, kteří dosáhli na hranici 20 kreditů, ačkoli mnozí studenti s 18 nebo 19 kredity také postoupili do dalšího semestru na základě žádosti přes studijní oddělení, pokud jim chyběl předmět Paralelní algoritmy a systémy.

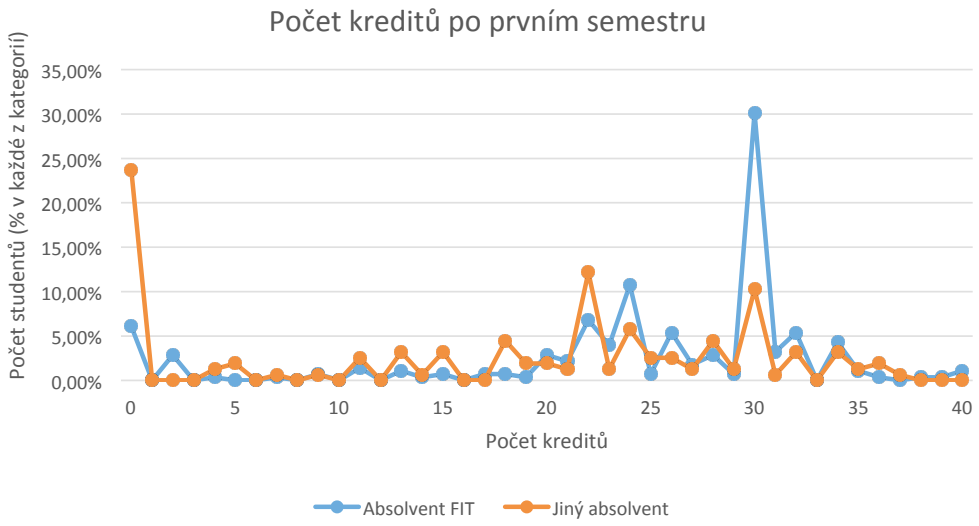
Jelikož žádný z ročníků ještě nemá absolventy, podíváme se místo nich na úspěšnost v prvním zápise předmětu MI-PAR. Pro první z ročníků je to konkrétně MI-PAR.1, pro druhý pak MI-PAR.2. Předmět MI-PPR.2 (Paralelní programování), který se vyčlenil z MI-PAR.2 nebudeme uvažovat, neboť stěžejní stále zůstává Paralelní algoritmy a systémy. Úspěšností tohoto předmětu pak budeme myslet poměr zapsáno/dokončeno, ale pouze u úspěšných stu-

dentů, kteří mají více než 20 kreditů, nebudeme tedy brát v úvahu všechny studenty.

		Úspěšnost po 1. semestru		Úspěšnost v MI-PAR Zapsáno	Ukončeno	
12/13	Absolvent FIT	93	93,00 %	84	63	75,00 %
	Jiný absolvent	61	62,89 %	56	29	51.79 %
	Celkem	154	78,17 %	140	92	65,71 %
13/14	Absolvent FIT	142	79,33 %	137	82	59,85 %
	Jiný absolvent	26	44,07 %	25	12	48,00 %
	Celkem	168	70,59 %	162	94	58,02 %
Celkem		**322**	**74,02 %**	**302**	**186**	**61,59 %**

Tabulka 3.7: Přehled studentů MSP podle předchozího studia

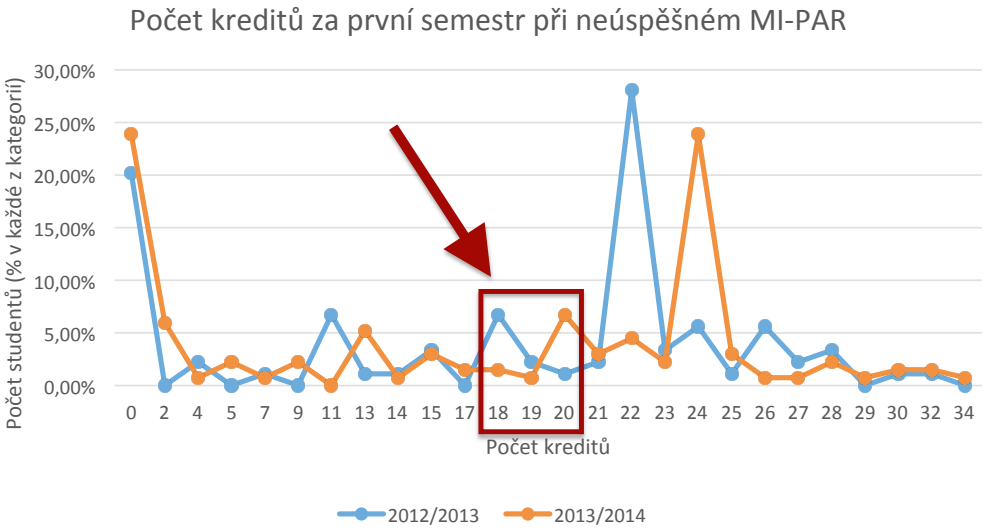
Z tabulky 3.7 vyplývá, že naši absolventi jsou lepší v celkové úspěšnosti i v předmětu MI-PAR. Opět se ale podíváme na rozložení pro jednotlivé kredity, které vidíme na obrázku 3.5.



Obrázek 3.5: Vliv předchozího studia na počet kreditů

Rozdíl je velmi patrný, absolventi bakalářského programu na fakultě převažují v části nad 23 kreditů, ostatní mají kreditů méně. Je také mnohem více absolventů jiných fakult, kteří získají 0 kreditů. Nejvýraznější rozdíl je u 30 kreditů, tedy splnění všech předmětů doporučeného průchodu programem. Celkem 30 % studentů, kteří absolvovali FIT, získá 30 kreditů, 17 % má dokonce kreditů více. Absolventi jiných fakult mají 30 kreditů pouze v 10 % případů, stejně tak i nad 30 kreditů. Z toho tedy jednoznačně vyplývá, že si absolventi bakalářského programu Informatika vedou v navazujícím programu lépe, než ti ostatní.

Ještě pro zajímavost se zaměříme na rozdělení předmětu MI-PAR na dvě části: dvoukreditový MI-PPR.2 a šestikreditový MI-PAR.2. Cílem bylo dát šanci studentům, kteří měli 18 kreditů a nyní mohou vypracovat semestrální práci za 2 kredity a dosáhnout tak na hranici 20 kreditů.



Obrázek 3.6: Rozdíl před a po rozdělení předmětu MI-PAR

Na obrázku 3.6 můžeme vidět, že záměru bylo dosaženo, celá křivka se posunula o 2 kredity vpravo. Celkem MI-PAR.2 nedokončilo 134 studentů, z toho 32 získalo 0 kreditů. Pokud vezmeme v úvahu ty, kteří měli více než 1 kredit, zůstane nám 102 studentů a z nich 75 ukončilo předmět MI-PPR.2, což je téměř 74 %.

3.4 Shrnutí

Na základě analýz dat byly zjištěny zajímavé informace o předchozích studiích na středních školách a jejich vztahu ke studiu na fakultě a o přijímacím řízení jak na bakalářské tak magisterské studium. Některé výsledky jsou zajímavé, jiné jsou jen potvrzení domněnek.

Shrnutí nejzajímavějších myšlenek:

- Podle histogramu rozložení známek je v magisterském studijním programu jediný těžký předmět MI-PAR, a to ještě nemá téměř žádná A ani E
- BI-PAI není lehký předmět
- Osobní údaje mají na studium minimální vliv
- Žádná podmnožina studentů není natolik specifická, aby dokázala obrátit histogram známek pro předmět BI-UOS (tedy nejvíce A, nejméně E)
- Studentky excelují v BI-MLO, horší jsou v programovacích předmětech a sítích
- Největší závislost mezi průměrem z matematiky a výsledky z předmětu je u BI-MLO
- Nejlepší studenti jsou ti, kteří mají vysoký percentil ze Scio testů a dobrý průměr z matematiky
- Studenti s excelentními výsledky z přijímací zkoušky jsou průměrní
- Mezi studenty z gymnázií a středních odborných škol je minimální rozdíl, nad 30 kreditů za první semestr převažují gymnazisté
- Mezi nejúspěšnější střední školy (tedy nejvíce procent studentů postoupilo do dalšího semestru) patří mimopražské
- Absolventi bakalářského studijního programu Informatika si vedou mnohem lépe v navazujícím studiu než ti, kteří přichází z jiných fakult
- Rozdělení předmětu MI-PAR na MI-PAR.2 a MI-PPR.2 pomohlo mnohým studentům dosáhnout na hranici 20 kreditů po prvním semestru

Samozřejmě je daleko více informací, které lze z dat vyčíst, některé zajímavé pro konkrétní skupiny či konkrétní roky. Proto byly vytvořeny dashboardy, které poskytují vhled do uložených dat pomocí jednoduchých grafů. Popíšeme si je v následující kapitole.

Kapitola 4

Dashboardy

Základní podoba dashboardů pochází od palubních desek aut, které zobrazují jednoduché a přehledné informace o stavu vozidla, což se ujalo také v oblasti informačních technologií, kdy se pod tímto pojmem myslí uživatelské rozhraní, které zobrazuje data tak, aby byla pro uživatele snadno čitelná. Nejčastěji se jedná o tabulky a grafy přizpůsobené potřebám dané společnosti, resp. koncového uživatele. Neexistuje jednotný návod na to, jak vytvořit dashboard (právě kvůli jeho nutnému přizpůsobení pro konkrétní případy), platí však základní pravidla. Dashboard by měl být:

- Jednoduchý
  Uživatel musí na první pohled vidět, co má dashboard sdělit, pokud to tak není, je špatně navržen. Neměl by obsahovat rozptylující prvky ani velké množství textu.
- Snadno čitelný
  Ideální je použití grafů, tabulek, nabídek a podobných komponent, které jsou shrnuté do logických celků. Vizuální informace jsou pro uživatele mnohem lépe čitelné než odstavec textu.
- Jednostránkový
  Tato charakteristika souvisí s výše uvedenými, jednoduchý a snadno čitelný dashboard by měl být pokud možno jednostránkový (měl by se vejít na obrazovku).
- V reálném čase
  Musí obsahovat aktuální data, tedy být např. napojený na databázi, narozdíl od reportů, které mají nejčastěji podobu pdf.

Na FIT jsou přehledná data ve formě dashboardů velmi důležitá pro průběžné sledování studentů, přihlášek, apod. Cílových skupin bude několik, jednak vedení fakulty, garanti předmětů a učitelé, tak střední školy či průmysloví partneři. Jelikož zde nebudou konkrétní data, není důvod, proč by k těmto dashboradům či některým z nich nemohli přistupovat studenti nebo veřejnost.

V rámci této práce bylo vytvořeno hned několik dashboardů, které byly rozděleny podle logických celků. Ty mohou sloužit také pro různé cílové skupiny (např. dashboard o středních školách může být dostupný jako pro fakultu, tak pro ředitele konkrétní střední školy).

Existuje velké množství nástrojů, které slouží pro vytváření dashboardů, od klasického Excelu až po specializované programy. Následující dashboardy byly vytvořeny pomocí dalšího z programů rodiny Pentaho s názvem Community Dasboard Editor (CDE)[13], kde se mi prozatím bohužel nepodařilo zprovoznit české znaky, proto prosím o shovívavost při prohlížení návrhů.

4.1 Struktura

Existují tři základní prvky, ze kterých se dashboardy skládají, a to: layout, komponenty a zdroje dat (datasources).

Layout

Jedná se o základní rozvržení dashboardu – kolik zde bude řádek, kolik sloupců, apod. Je možné vybírat z předdefinovaných layoutů, ale pomocí editoru není těžké ho vytvořit ručně. V této práci jsou převážně použity layouty 2x2. Každá buňka obsahuje nadpis a může se nadále dělit (například tabulka a graf), lze jí přiřadit základní parametry jako je barva pozadí, obrázek nebo html kód. Jelikož je layout vnitřně implementován jako html, je ho možné také upravovat pomocí CSS.

Příklad pro dashboard Studium:

- Resource (CSS soubor)
- Row Header
  - Column Header_title
  - Column Year_title
  - Column Year_selector
  - Column Programme_title
  - Column Programme_selector
  - Column Form_title
  - Column Form_selector
- Row First
  - Column Overview

[13] http://www.webdetails.pt/ctools/cde.html

        * Row Overview_header
        * Row Overview_tabular
        * Row Overview_chart
    - Column Credits
        * Row Credits_header
        * Row Credits_chart

- Row Second
    - Column Study_time
        * Row Study_time_header
        * Row Study_time_chart
    - Column Subjects
        * Row Subjects_header
        * Row Subjects_chart

Komponenty

Každý prvek v layoutu je možné pojmenovat (vytvoří se pak samostatný div), aby bylo možné do něj umístit komponentu. Tou může být graf, tabulka, selektor nebo další specializované komponenty (mapy, Google Analytics, apod.). Každá komponenta má hodně možností nastavení, další je pak možné určovat přímo pomocí JavaScriptového kódu.

Zdroje dat

Aby bylo možné používat nějaká data, musí se definovat tzv. datasource (zdroj dat). Existuje velké množství typů, nám však bohatě postačí SQL dotazy přímo do našeho datového skladu. Tyto zdroje se poté přiřadí konkrétní komponentě.

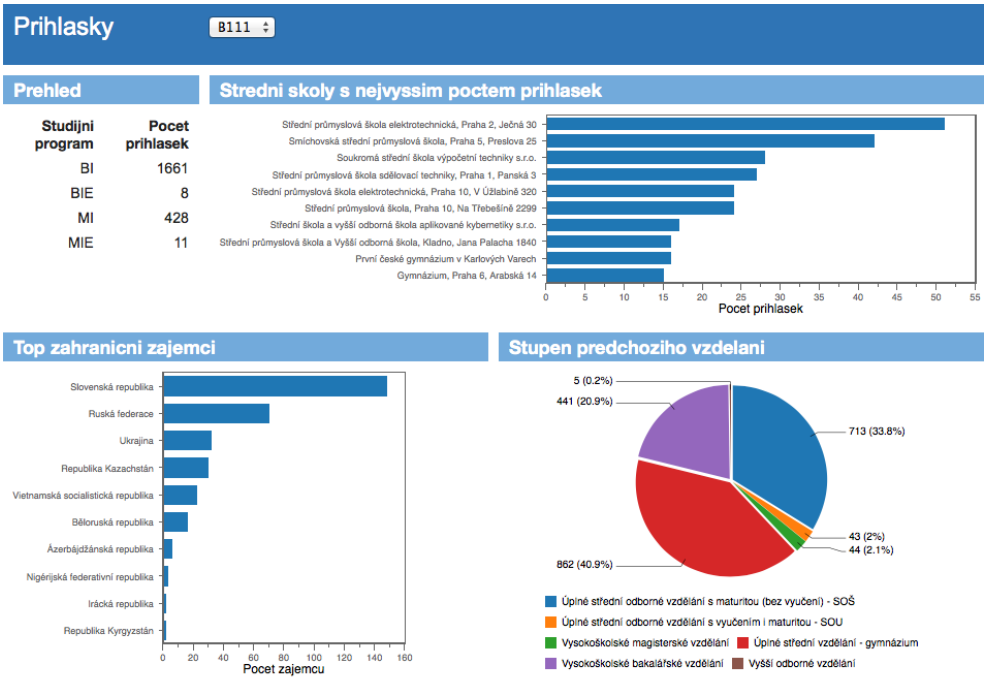
Některé SQL dotazy jsou velmi složité a mohou se v různých obměnách opakovat. Proto by bylo vhodné vytvořit z některých z nich pohledy do databáze, ze kterých by se pomocí parametrů získala požadovaná data. Parametry jsou mocné komponenty, které slouží jako proměnné. Je tak možné měnit data pro různé ročníky, studijní programy, apod.

4.2 Návrhy

Navrhované dashboardy se mohou místy překrývat, pokud to bude zvyšovat hodnotu celkové informace v nich obsažené, příp. může jít o zacílení na konkrétní cílovou skupinu, podmnožinu dat, apod.

4.2.1 Přihlášky

Jednoduchý dashboard, který má za úkol zobrazovat informace o zájemcích o studiu – odkud se hlásí a jaké mají dosavadní vzdělání.



Obrázek 4.1: Dashboard Přihlášky

Tento první dashboard si popíšeme podrobněji, protože obsahuje prvky, které se nadále opakují. Layout je podobný jako ukázkový layout popsaný výše, komponenty jsou poměrně dobře vidět z obrázku 4.1. Každá z komponent (selektor, tabulka, graf, apod.) vyžaduje datový zdroj, což jsou jednoduché i složité SQL dotazy jak si ukážeme níže.

Hlavička
První řádek obsahuje nadpis a nabídku akademického roku označeného podle semestru, do kterého se zájemci hlásí. Úplně první přihlášky jsou pod semestrem s kódovým označením B091, neboť v tomto semestru (zimní 2009/2010) nastoupili první studenti, kteří si podávali přihlášky na jaře roku 2009. Po výběru příslušného semestru se překreslí grafy s aktuálními daty. Tato nabídka je tvořena dvěma komponentami: Simple Parameter a Select Component. Vybraný semestr se použije jako proměnná *${year}* do SQL dotazu dalších prvků.

Přehled
Jedná se o jednoduchou tabulku, která zobrazuje, kolik přihlášek bylo podáno do kterých programů. Tabulky mohou mít stránkování a řazení podle jednotlivých sloupců. Kvůli jednoduchosti této (i dalších) tabulek byly tyto volby zakázány.

Střední školy s nejvyšším počtem přihlášek
Graf zobrazující střední školy, ze kterých se hlásilo daný rok nejvíce studentů. Je zde pouze 10 nejlepších škol, je ovšem možné dodat parametr určující tento počet. Dále je také možné vedle grafu zobrazit tabulku, která by zobrazovala na jedné straně 10 škol a umožnila by listování, tudíž by ukazovala všechny střední školy.

Top zahraniční zájemci
Velmi zajímavý údaj říká, jaké máme vlastně zahraniční zájemce (není myšleno studenty přes studijní programy). Opět je zde zobrazeno 10 nejčastějších zemí.

Stupeň předchozího vzdělání
Koláčový graf, který zobrazuje předchozí vzdělání zájemců. Je zde krásně vidět, že téměř každý rok se hlásí více studentů z gymnázií než ze středních odborných škol. Pro tuto komponentu si také uvedeme ukázkový SQL dotaz pro její datový zdroj.

```
SELECT a_prev_study_degree, count(id_application)
FROM d_application
JOIN d_time ON (d_application.id_time = d_time.id_time)
WHERE k_semester_code = ${year} AND a_prev_study_degree != ''
GROUP BY a_prev_study_degree
```

4.2.2 Přijímací řízení

Tematicky souvisí s přihláškami, ale kvůli přehlednosti nakonec vznikl vlastní dashboard s názem Přijímací řízení. Zobrazuje informace o samotném řízení, tedy kolik bylo přijato studentů a kolik se jich zapsalo do studia.

Hlavička
V hlavičce je opět obsažen nadpis a nabídka *Year_selector*, která umožňuje vybrat příslušný rok, resp. semestr, stejně jako u předchozího dashboardu.

Přehled
Jednoduchá tabulka, která zobrazuje počet přihlášek do jednotlivých programů (bakalářský a magisterský), počet přijatých a počet zapsaných stu-

dentů. Opět byly vypnuty veškeré prvky umožňující řazení nebo stránkování, neboť to u takto malé tabulky působí spíše rušivým dojmem.

Top střední školy
Střední školy s největším počtem přihlášek jsou doplněny o informaci, kolik studentů se zapsalo. Je zde zobrazeno 5 škol s nejvíce zapsanými studenty. Pořadí těchto škol je odlišné než u předchozího dashboardu, protože poměr přijatých ku zapsaných se může u jednotlivých škol lišit.

Varianta přijímací zkoušky
Velmi zajímavá komponenta je graf zobrazující výsledky studentů, kteří psali písemnou zkoušku z matematiky na fakultě. Pro jednotlivé varianty zkoušky je zde vidět poměr všech studentů a těch, kteří byli na základě zkoušky přijati. Podle tohoto grafu lze jednoduše poznat, jestli byly varianty vyvážené nebo naopak.

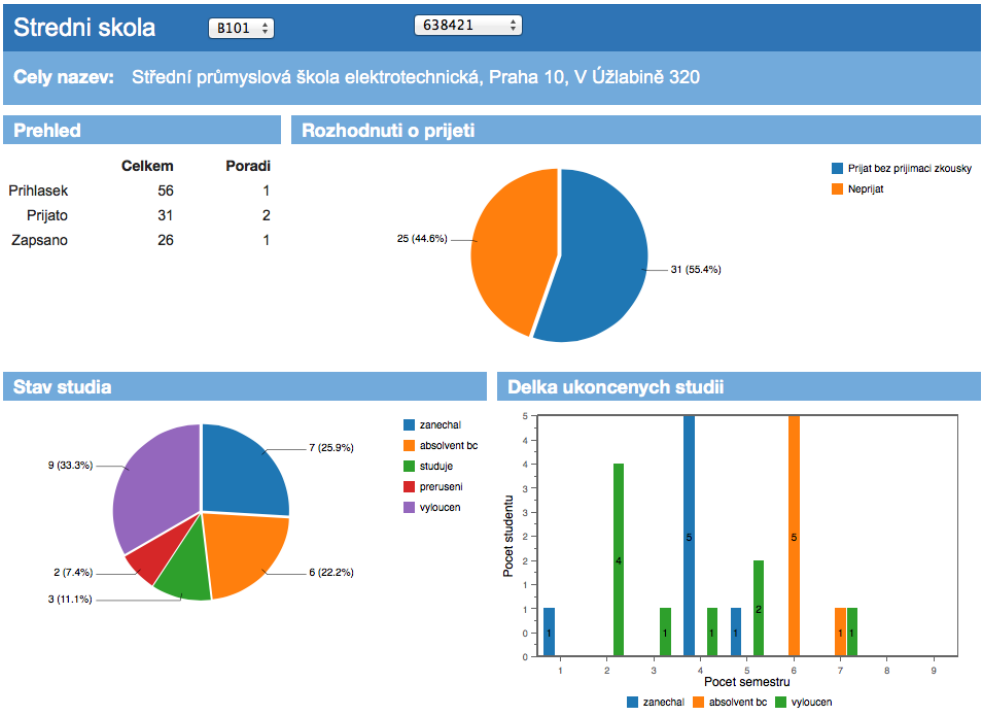
Rozhodnutí o přijetí
Tento atribut může nabývat několika hodnot, a to přijat na základě přijímací zkoušky, bez přijímací zkoušky (Scio, maturita, olympiády), mimo přijímací zkoušku (odvolání, přestupující) nebo nepřijat. Další zajímavý graf by byl, na základě čeho byl student přijat, avšak v tuto chvíli pro tuto komponentu nejsou dostatečná data.

4.2.3 Střední školy

Tento dashboard byl vytvořený tak, aby zobrazoval podstatné informace o konkrétní střední škole, kterou je možné vyhledat pomocí IZA. Primární účel byl zpřístupnění těchto informací ředitelům jednotlivých středních škol. Některé informace jsou velmi podobné již předchozím komponentám, rozdíl je v tom, že zde je zobrazena pouze daná podmnožina pro konkrétní střední školu.

Jsou zde zobrazeny informace pouze o bakalářských studentech, protože magistři nemusí vyplňovat informace o střední škole. Letos však budeme mít dostatečná data o prvních bakalářích, kteří jsou na fakultě již 5. rokem (tedy ve 2. ročníku MSP). Bude tedy možné ukázat průchod těch studentů, kteří po úspěšném ukončení bakalářského studia pokračují v navazujícím studiu.

Hlavička
Kromě výběru semestru je zde výběr IZA střední školy – vždy se zobrazují pouze ty školy, které měly více než jednu podanou přihlášku daný rok. Jelikož IZO je pouze číselný identifikátor přiřazený určité škole, je zde také podnadpis, který zobrazuje celý název školy.



Obrázek 4.2: Dashboard Střední škola

Tento dashboard je poslední, který používá semestr nástupu studentů, neboť je posledním, který se vztahuje k přihláškám.

Přehled

Podobná tabulka jako u předchozího dashboardu, která zobrazuje počet podaných přihlášek celkem, počet přijatých a zapsaných. Kromě těchto počtů je zde uvedeno pořadí, na kterém se škola umístila. Je to uvedeno pro každý řádek zvlášť, neboť škola s největším počtem přihlášek nemusí mít největší počet přijatých nebo zapsaných studentů, jak je ostatně vidět na příkladu pro Střední průmyslovou školu elektrotechnickou v Praze 10. Ta se umístila na prvním místě v počtu podaných přihlášek i zapsaných studentů pro rok 2010/2011, avšak v počtu přijatých ji předběhla Střední průmyslová škola sdělovací techniky, Praha 1, Panská 3.

Rozhodnutí o přijetí

Jednoduchý koláčový graf, který zobrazuje rozhodnutí o přijetí. Existuje více variant, avšak v roce 2010/2011 fakulta nevypisovala vlastní přijímací zkoušku, proto má na obrázku graf málo kategorií. Pokud by byl zvolen vyšší ročník,

byly by zase chudší následující grafy, protože by neobsahovaly žádné absolventy (první budou absolvovat tento semestr).

Stav studia
Každý student, který nastoupil ve zvoleném semestru, má uvedený stav studia, podle kterého lze určit, zdali úspěšně školu dokončil nebo nedokončil. Více než polovina studentů zapsaných v zimním semestru 2010/2011 školu nedokončila a studium zanechala nebo byla vyloučena (to se velmi těžko odlišuje, protože někteří studenti, kteří mají být vyloučeni sami ukončí studium, aby se jim zbytečně neodečítaly dny studií, které hradí stát). Necelá čtvrtina je nyní absolventy a přibližně osmina stále studuje.

Délka ukončených studií
Standardní doba studia je 6 semestrů, 8 semestrů bakalářského studia hradí stát. Jak dlouho ale trvalo studentům, než absolvovali? Nebo než byli vyloučeni či zanechali studií? To je právě zobrazeno v této komponentě, kde je vidět, že většina absolventů dělala státní zkoušku v 6. semestru, někteří v 7. Další budou jistě i v semestru 8., ale na tato data musíme ještě počkat. Některá čísla u ukončených studií bohužel nejsou úplně přesná, protože pokud student nezískal dostatek kreditů pro postup po prvním semestru, byl vyloučen až v semestru následujícím (proto je nejvíce studentů vyloučeno ve druhém semestru).

4.2.4 Studium

Nejsložitější navrhovaný dashboard zobrazuje informace právě o studiích. Jeho složitost je dána tím, že studia mohou mít různé programy i formy studia a zobrazovat všechno najednou by nemělo smysl. Průchod samotným studiem je důležitým ukazatelem pro fakultu, zejména úmrtnost studentů po prvním roce/semestru a krizové předměty.

Hlavička
Uživatel si může vybrat hned ze tří parametrů: semestr zápisu, studijní program a forma studia. Opět je možné vybrat pouze smysluplné kombinace, protože jednotlivé selektory využívají informace z předchozích jako parametry do svých dotazů. Není tedy možné vybrat například magistreský studijní program v kombinované formě. Výchozí je první semestr (zimní 2009/2010), bakalářský studijní program v prezenční podobě.

Přehled
Tato komponenta obsahuje kromě tabulky s předhledem také graf se stavy studia, který doplňuje informace v tabulce. Jelikož první rok byli přijati všichni studenti, není divu, že byla tak vysoká úmrtnost tohoto ročníku. Ostatní roky

Obrázek 4.3: Dashboard Studium

se poměr již zlepšuje. Mnohem lepší je situace na magisterském studiu, kde je více úspěšných než neúspěšných studií.

Počet kreditů za semestr

Kritérium, které určuje, zdali student postoupí nebo nepostoupí, je právě počet kreditů. Pro postup do druhého semestru jich musí student BSP nasbírat 15 (student MSP 20) a poté celkem 30 za akademický, resp. 40, pro postup do dalšího ročníku. Tento graf zobrazuje kolik studentů získalo kolik kreditů. Je vidět, že hned první rok studium zanechalo cca 100 studentů, neboť nezískali ani jeden kredit, 123 studentů ukončilo všechny předměty v doporučeném průchodu a 40 získalo 19 kreditů čili neukončilo dva předměty (zřejmě kombinace některých předmětů ZMA, UOS a PA1). Tento graf je možné zobrazovat po jednotlivých semestrech od nástupu do studia. S přibývajícími semestry přibývá studentů s vyšším počtem kreditů. Čím dál více jich dosáhne na 30 kreditů, mnozí mají i více. V 6. semestru se průměrný počet získaných kreditů pohybuje kolem 25, což může znamenat, že studenti si nechali na poslední semestr studia pouze bakalářskou práci.

Ukázka SQL dotazu pro vytvoření datového zdroje pro tuto komponentu. Jsou zde použity všechny parametry: *${year}* pro semestr zápisu, *${programme}* pro studijní program, *${form}* pro formu studia a *${semester}* pro určení semestru od nástupu do studia. Pro získání čísla semestru byla použita funkce *row_number()*, která očíslovala všechny semestry, ve kterých má nějaký student nastupivší daný rok nějaká studia. Pro výchozí nastavení parametrů vzniklo až 9 semestrů (tedy 5. rok studia BSP).

```
SELECT sum_ects, pocet
FROM
    (SELECT d_time.k_semester_code, sum_ects, count(id_student) AS
        pocet
    FROM d_study_time_ects
    JOIN d_study ON (d_study_time_ects.id_study = d_study.id_study)
    JOIN d_time ON (d_study_time_ects.id_time = d_time.id_time)
    WHERE k_education_start_semester_code = ${year} AND
        k_study_program = ${programme} AND k_study_form = ${form}
    GROUP BY k_semester_code, sum_ects
    ) sum
JOIN
    (SELECT row_number() OVER (ORDER BY d_time.k_semester_code ASC) AS
         semester, d_time.k_semester_code
    FROM d_study_time_ects
    JOIN d_study ON (d_study_time_ects.id_study = d_study.id_study)
    JOIN d_time ON (d_study_time_ects.id_time = d_time.id_time)
    WHERE k_education_start_semester_code = ${year} AND
        k_study_program = ${programme} AND k_study_form = ${form}
    GROUP BY d_time.k_semester_code) row_num
ON (sum.k_semester_code = row_num.k_semester_code)
WHERE semester = ${semester}
ORDER BY sum_ects
```

Délka ukončených studií

Tato komponenta je téměř identická s tou, která je použita v dashboardu Střední škola na straně 80, pouze jsou použité jiné parametry SQL dotazu pro datový zdroj. Z této komponenty je na první pohled vidět, že studenti bakalářského programu končí studia v 6. semestru nebo v 8. semestru (cca 30 %), kdežto studenti magisterského programu končí nejčastěji ve 4. a 5. semestru (50 %). Z mých vlastních zkušeností to přisuzuji tomu, že u bakalářského studia většina studentů opakuje nějaký předmět z letního semestru nebo chtějí stejně pokračovat na navazujícím magisterském studiu a nechtějí mít půl roku pauzu. U magisterského studia je to dáno náročností studia, kdy v posledním semestru, obzláště pokud student opakuje nějaký těžký předmět jako je například MI-PAR, nezbývá příliš mnoho času na diplomovou práci, proto studenti prodlouží o půl roku, aby ji stihli dopsat.

Předměty s nejnižší průchodností
Každý semestr se po uzavření Ankety ČVUT[14] sleduje, jaké měly předměty průchodnost (tedy poměr zapsaných studentů vůči těm, kteří předmět dokončili) k určení, zdali byly nastaveny správně podmínky pro úspěšné zakončení předmětu. Poté se předmět může a nemusí upravit. Je to velmi zajímavá statistika, které předměty byly kritické pro daný ročník. Rozdělení na ročníky je velmi důležité, protože podmínky pro získání zápočtů nebo i zkoušky se mohou každý rok měnit.

Jelikož mohou být zavádějící předměty, které mají velmi malý počet studentů, například pokud z 5 zapsaných dokončí předmět pouze 2, procento úspěšnosti bude velmi malé a zastíní se tak povinné předměty programu. Proto jsou vybrány pouze ty předměty, které má zapsáno více než 50 studentů. Pokud by se to ukázalo jako zajímavé měřítko, je možné z tohoto počtu udělat parametr a nechat na uživateli, jak velký počet studentů mají mít předměty, které chce zobrazit.

4.2.5 Předměty

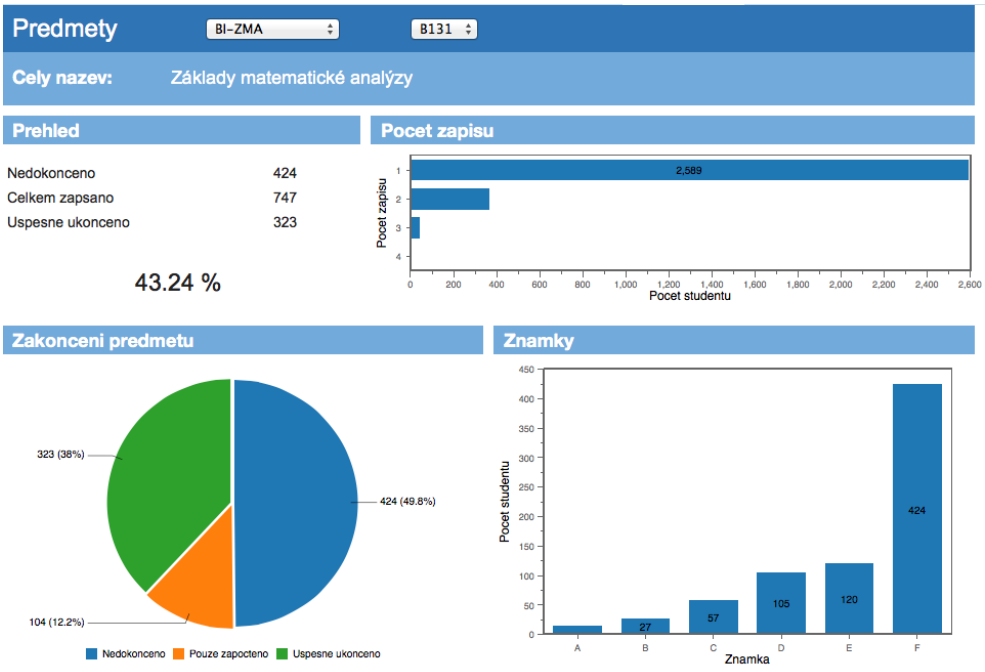
Pro garanty či vyučující určitých předmětů budou zajímavé statistiky z pohledu konkrétních předmětů v dané semestry. Jedná se o poměrně jednoduchý dashboard, který zobrazuje základní informace o průchodech předmětu.

Hlavička
Uživatel si může vybrat předmět podle jeho zkratky (parametr *${subject}*) a semestr, ve kterém byl předmět vyučován (parametr *${semester}*). Pod hlavním nadpisem je podnadpis, který zobrazuje celý název předmětu.

Přehled
Opět jednoduchá tabulka, ukazuje celkový počet zapsaných studentů a těch, kteří předmět dokončili a nedokončili. Pod ní je procento průchodu předmětem (poměr úspěšných a zapsaných).

Počet zápisů
Mnozí studenti mají problém s nějakým konkrétním předmětem a mnohdy tak velký, že kvůli němu studium musí ukončit. Naštěstí jsou pragmatičtí a včas si podají přihlášku, znovu nastoupí a předmět si zapíší. To mi přišlo jako velmi zajímavý ukazatel, proto jsem vytvořila tuto komponentu zobrazující, po kolikáté má student předmět zapsaný celkově (nikoli v rámci studia, kdy je maximální počet zápisů 2). Největší počet zápisů je 5, což má například předmět BI-UOS nebo BI-PA1.

[14] Anketa hodnocení studia ČVUT, http://anketa.cvut.cz



Obrázek 4.4: Dashboard Předměty

V datovém skladu bohužel nejsou zahrnuta studia v rámci Celoživotního vzdělávání[15], kde končí studenti v mezidobí, kdy ukončili studia a podáním nové přihlášky. Předměty si zde zapíší a poté si je nechají uznat, až znovu nastoupí. Pokud by se tato studia spárovala se studenty, získali bychom přesnější data.

Zakončení předmětu

Jednoduchý koláčový graf, který více méně graficky znázorňuje tabulku z přehledu. Navíc je zde ovšem uvedeno, kolik studentů získalo pouze zápočet, ale už nemají zkoušku. V dalším zápise si totiž mohou nechat zápočet uznat a jít rovnou ke zkoušce.

Známky

Opět jednoduchý graf, který zobrazuje rozložení známek pro předmět v daném semestru. Někdo trvdí, že by známky měly mít normální rozdělení, většina povinných předmětů se blíží ke krásné exponenciále. Velmi zajímavé jsou předměty, které v některých semestrech nevyužívají celou stupnici. Například

[15] Kdokoli se může přihlásit bez přijímací zkoušky do předmětu, za který zaplatí podle jeho kreditů. V rámci tohoto studia není možné získat titul, avšak ukončené předměty je možné si nechat uznat v dalším studiu.

BI-SKJ nebo MI-PAR úplně zavrhly známku A a jejich stupnice začíná až od známky B.

4.2.6 Absolvent

Pro tento dashboard prozatím nejsou v datovém skladu vhodná data, ale přesto si zde uvedeme návrh, protože věřím, že by byl velmi zajímavý.

Hlavička
V hlavičce by bylo vhodné použití těchto parametrů: semestr ukončení studia, studijní program a forma studia.

Přehled
Stejně jako u předchozích předhledů by zde byla jednoduchá tabulka, která by zobrazovala počet studentů přihlášených ke státní zkoušce, skutečně dostavených, úspěšných a neúspěšných, navíc by zde byl počet studentů, kteří absolvovali s vyznamenáním. Zajímavá by byla také statistika, kolik z absolventů bude nadále pokračovat na doktorském studiu.

Obory
Jednoduchý skládaný graf, který by zobrazoval počet studentů přihlášených ke státnicím a úspěšné absolventy v rámci jednotlivých oborů.

Známky
Státní závěrečná zkouška je hodnocená podobně jako zkouška z jakéhokoli jiného předmětu. Tato komponenta by zobrazovala známku celkovou, z ústní zkoušky a obhajoby.

Komise
Počty studentů, kteří uspěli nebo neuspěli u konkrétních komisí, což je zajímavé z podobného důvodu jako je graf ukazující varianty přijímací zkoušky, tedy jestli jsou komise vyvážené nebo naopak.

4.3 Možnosti rozšíření

Tyto základní návrhy umožňují velké množství nastavení a dalších rozšíření. Jejich používáním se také časem ukáže, které elementy chybí a bude je vhodné doplnit. Přesto si zde již nyní uvedeme možnosti rozšíření, které by bylo vhodné naimplementovat buď jako pokročilé funkce stávajících návrhů nebo jejich zlepšení docílit zajištěním určitých zdrojů dat.

Volba elementů

Téměř jakýkoli parametr, resp. atribut, použitý v databázi je možné vynést do dashboardu, ale pokud by se tak stalo, byly by velmi nepřehledné. Cíloví uživatelé mohou být různí a každého může zajímat něco trochu odlišného, proto by bylo vhodné ponechat uživateli volbu, které elementy ho zajímají (např. pomocí checkboxu), a ty by poté tvořily výsledný dashboard.

Interaktivita

Jelikož se jedná o pohledy do databáze, resp. datového skladu, je možné propojit různé dashboardy. Například u přehledu přihlášek je graf zobrazující střední školy s nejvyšším počtem přihlášek a po kliku na název školy by se zobrazil dashboard pro konkrétní školu.

Aktualita dat

Dasboardy jsou velmi mocné nástroje mimo jiné také v tom, že jsou přímo napojené na databázi s daty. Jelikož například data z přihlášek jsou nahrávaná z exportů jednou ročně, ztrácí zde dashboard své kouzlo. Bylo by proto vhodné zavést přímé napojení do zdrojových systémů a získávat průběžná data jako je to nyní možné např. přes KOSApi u studií studentů.

Doplnění chybějících dat

Již v popisu jednotlivých dashboardů zmiňuji, která data ve stávajícím řešení chybí a bylo by vhodné je doplnit. Jedná se zejména o informace ze státních závěrečných zkoušek, na základě čeho byl student přijat nebo informace o studiích v rámci celoživotního vzdělávání.

Závěr

Cílem této práce bylo nahrání dat z Přihlášky ČVUT do fakultního datového skladu, provedení analýz nad těmito daty a vytvoření dashboardů. Počátky práce provázela mnohá úskalí, protože prakticky jsem se s datovým skladem v této práci setkala poprvé a musela mnoho nastudovat. Dalším úskalím byla data samotná, protože každý rok se měnil formát dat, použité sloupce, a to vše mělo samozřejmě přímý vliv na jejich použitelnost. Získání a zpracování dat zabralo nejvíce času z celé práce. Nahrání dat (tedy vytvoření ETL procesů) bylo po rozlušení dat relativně jednoduché.

Po integraci dat v datovém skladu byly připraveny vhodné dotazy na spojení všech tabulek a bylo nutné se zamyslet, podle jakého subjektu data zobrazit – podle studenta, studií nebo zápisu předmětů. Také bylo nezbytné doplnit nějaká chybějící data pro analýzy – například počet kreditů ze jednotlivé semestry pro predikci postupu do druhého semestru nebo délku studia. Po získání exportu byly provedeny analýzy, které potvrdily domněnky nebo vnesly nový pohled na věc. Velmi překvapivé pro mě bylo, že studijní průměr z matematiky je poměrně vhodným prediktorem pro úspěšnost studentů nebo že předmět Právo a informatika se řadí mezi středně těžké předměty (nejvíce studentů má známku C). Naopak málo překvapivý výsledek je, že nejlepší studenti jsou ti, kteří mají hodně bodů ze Scio testů nebo fakt, že absolventi bakalářkého studijního programu jsou úspěšnější na magisterském programu než studenti jiných fakult.

Poslední část se zabývá dashboardy, jejichž vytvoření bylo také poměrně časově náročné. Každá komponenta je totiž navázána přímo do datového skladu a bylo tak nutné vytvořit stejný počet SQL dotazů kolik je komponent, přesně v takovém formátu, v jakém jsou vyžadovány. To znamenalo hodně transformací a spojení. Mnohdy se tak poměrně jednoduchý SQL stal složitým jen proto, že bylo nutné změnit formát výstupu.

V každé kapitole jsou také uvedeny nápady na vylepšení stávajícího řešení, které buď přesahovaly rámec této práce nebo je bylo možné navrhnout až po zkušenostech získaných při realizaci. Nejdůležitější je zaměřit se na původní data a doplnit vhodné sloupce, které se poté promítnou přes datový sklad i do analýz a dashboardů. I bez implementace těchto návrhů pevně věřím, že bude řešení této práce pro fakultu přínosem.

Literatura

[1] BERSON, A.; SMITH, S. J.: *Data Warehousing, Data Mining & OLAP*. New York : McGraw-Hill Companies, 1997, ISBN 0-07-006-272-2, 613 s.

[2] CARNEGIE, D. A.; WATTERSON, C.; ANDREAE, P.; aj.: *Prediction of Success in Engineering Study* [online]. Publikováno 20.4.2012 [cit. 2013-12-05]. Dostupné z: `http://80.ieeexplore.ieee.org.dialog.cvut.cz/stamp/stamp.jsp?tp=&arnumber=6201020`

[3] CHODNICKI, S.: *Creating Dashboards with CDE* [online]. Publikováno 28.6.2011 [cit. 2014-04-22]. Dostupné z: `http://type-exit.org/adventures-with-open-source-bi/2011/06/creating-dashboards-with-cde`

[4] DEKKER, G. W.: *Predicting students drop out: a case study* [online]. Publikováno 9.4.2010 [cit. 2013-12-02]. Dostupné z: `http://www.win.tue.nl/~mpechen/projects/edm/internshipreport_090409.pdf`

[5] HALL, M.; FRANK, E.; HOLMES, G.; aj.: *The WEKA Data Mining Software* [online]. ©2014, [cit. 2014-01-12]. Dostupné z: `http://www.cs.waikato.ac.nz/ml/weka/`

[6] HAYES, J. H.; DEKHTYAR, A.; HOLBROOK, A.; aj.: *Will Johnny-/Joanie Make a Good Software Engineer? Are Course Grades Showing the Whole Picture?* [online]. Publikováno 21.4.2006 [cit. 2013-12-03]. Dostupné z: `http://80.ieeexplore.ieee.org.dialog.cvut.cz/stamp/stamp.jsp?tp=&arnumber=1617344`

[7] HUMPRIES, M.; HAWKINS, M. W.; DY, M. C.: *Data Warehousing: Návrh a implementace*. Praha : Computer Press, 2001, ISBN 80-7226-560-1, 257 s.

[8] KIMBALL, R.; ROSS, M.: *The Data Warehouse Toolkit*. Indianapolis: Wiley, 2013, ISBN 978-1-118-53080-1, 564 s.

[9] KUZNETSOV, S.: *Datový sklad fakulty* [online]. Diplomová práce. Praha: České vysoké učení technické v Praze, Fakulta informačních technologií, 2013 [cit. 2014-02-05]. Dostupné z: https://dip.felk.cvut.cz/browse/pdfcache/kuznesta_2013dipl.pdf

[10] LABERGE, R.: *Datové sklady*. Brno: Computer Press, 2012, ISBN 978-80-251-3729-1, 350 s.

[11] MISHRA, T.; KUMAR, D.; GUPTA, D. S.: *Mining Students' Data for Performance Prediction* [online]. Publikováno 9.2.2014 [cit. 2014-03-15]. Dostupné z: http://80.ieeexplore.ieee.org.dialog.cvut.cz/stamp/stamp.jsp?tp=&arnumber=6783461&tag=1

[12] MÁRQUEZ-VERA, C.; CANO, A.; ROMERO, C.; aj.: *Predicting student failure at school using genetic programming and different data mining approaches with high dimensional and imbalanced data* [online]. Publikováno 26.8.2012 [cit. 2013-11-28]. Dostupné z: http://80.link.springer.com.dialog.cvut.cz/article/10.1007%2Fs10489-012-0374-8

[13] *Data Warehouse Architectures* [online]. ©2014, [cit. 2014-03-09]. Dostupné z: http://www.1keydata.com/datawarehousing/data-warehouse-definition.html

[14] *SQL Backend* [online]. ©2013, [cit. 2014-04-06]. Dostupné z: http://databrewery.org/cubes/doc/backends/sql.html

[15] POUR, J.; MARYŠKA, M.; NOVOTNÝ, O.: *Business Intelligence v podnikové praxi*. Praha: Professional Publishing, 2012, ISBN 978-80-7431-065-2, 276 s.

[16] TAURO, A.: *ETL vs. ELT: What's the Difference?* [online]. ©2013, publikováno 13.6.2013 [cit. 2014-04-14]. Dostupné z: http://blog.performancearchitects.com/wp/2013/06/13/etl-vs-elt-whats-the-difference/

[17] *Data Warehouse Definition* [online]. ©2014, [cit. 2014-03-23]. Dostupné z: http://www.1keydata.com/datawarehousing/data-warehouse-definition.html

[18] *Educational Data Mining* [online]. ©2014, [cit. 2013-12-03]. Dostupné z: http://www.educationaldatamining.org

[19] *Studijní a zkušební řád ČVUT* [online]. Publikováno 7.4.2009 [cit. 2014-04-20]. Dostupné z: http://intranet.cvut.cz/informace-pro-studenty/uredni-deska/resolveuid/c73936c3c7a51573550326543f105146

Příloha A

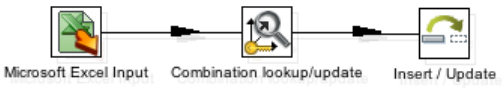
Obrázky a tabulky

A.1 ETL procesy

Následující kapitola obsahuje doplnění informací uvedených v kapitole 2.2.3, jedná se zejména o podrobný popis jednotlivých kroků transformací a datový model.

Úlohy App initialize tables a App additional data

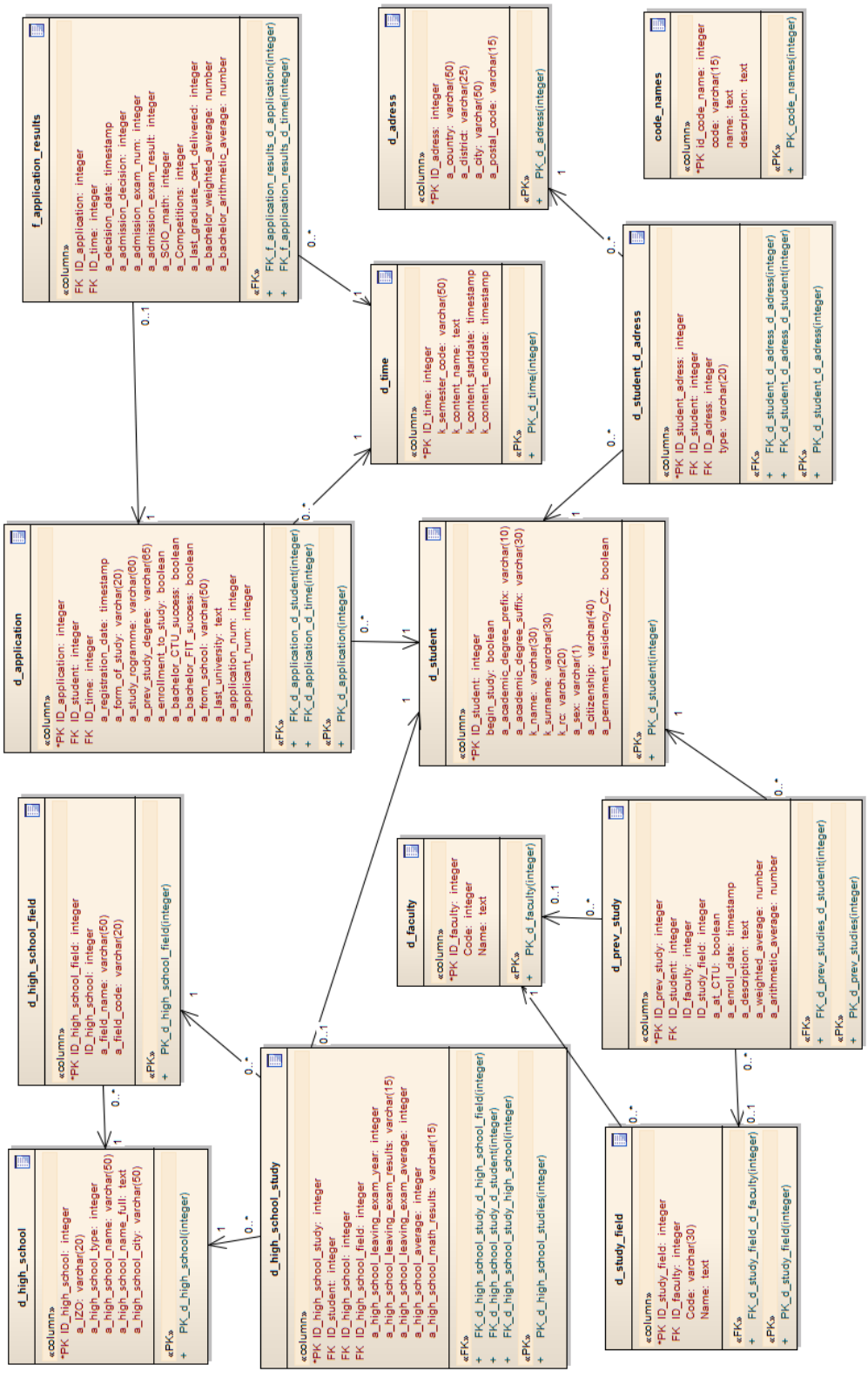
Úloha „App initialize tables“ obsahuje dvě transformace: „App d_study“ a „App create high schools“, úloha „App additional data“ obsahuje stejnojmennou tranformaci. Všechny tyto transformace mají stejný průběh, proto bude popsána pouze jedna z nich.



Obrázek A.2: Transformace App create high schools

Název	Popis
Microsoft Excel Input	Načti data
Combination lookup/ update	Vytvoř nový klíč, pokud neexistuje
Insert / Update	Nahraj data do určených tabulek

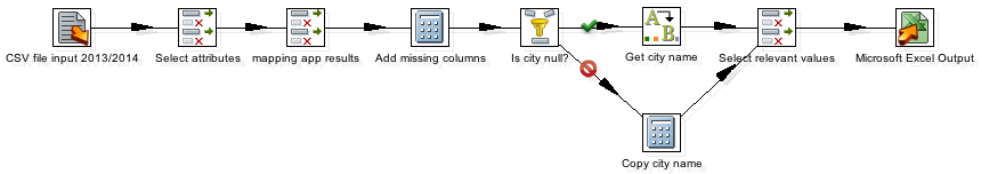
Tabulka A.1: Popis jednotlivých kroků transformace App create high schools



Obrázek A.1: Datový model

Úloha App load_file

Tato úloha obsahuje transformaci pro každý rok. Tyto transformace jsou velmi podobné, liší se pouze v mapování sloupců, proto si zde popíšeme pouze transformaci poslední, tedy přihlášky pro akademický rok 2013/2014.



Obrázek A.3: Transformace App load_file 2013_14

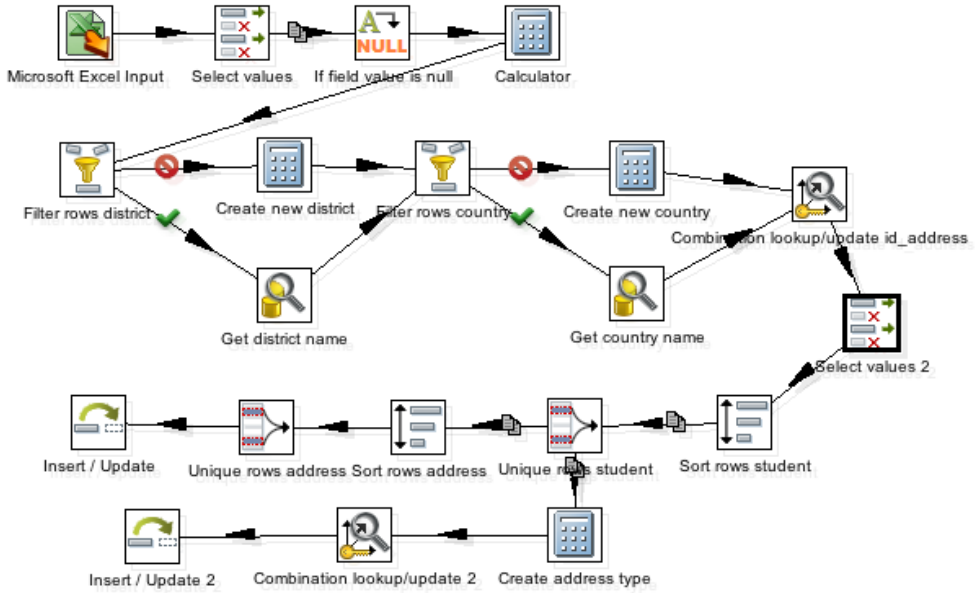
Název	Popis
CSV file input 2013/2014	Načtení souboru s daty (*.csv)
Select and rename attributes	Výběr a přejmenování atributů
Mapping app results	Mapování sloupců H1 až H10 k adekvátním atributům
Add missing columns	Přidání chybějících atributů (kvůli sjednocení formátu všech dat)
Is city null?	Rozdělení dat na ta, která mají vyplněný atribut „city“ a ta, která nemají
Get city name	Získání názvu města pomocí regulárního výrazů (některá obsahují i PSČ)
Copy city name	Získání názvu města z jiného sloupce
Select relevant values	Zahození pomocných sloupců
Microsoft Excel Output	Uložení předzpracovaných dat

Tabulka A.2: Popis jednotlivých kroků transformace d_load_file_2013_14

Úloha App d_student

Tato úloha (vč. jí odpovídající transformace) je kompletně popsána na straně 48.

Transformace d_address



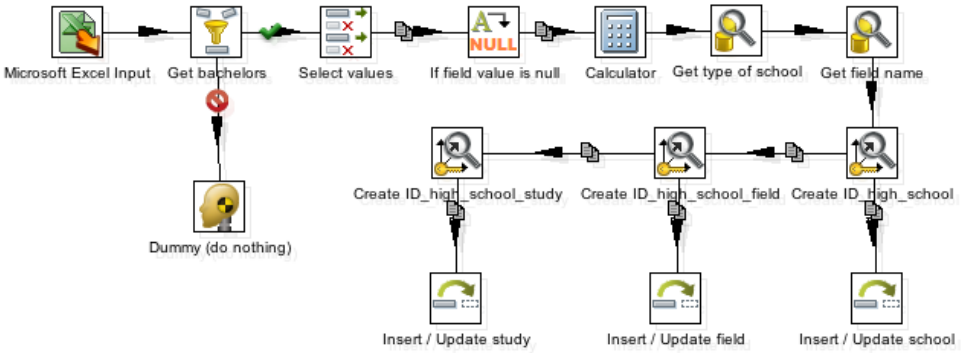
Obrázek A.4: Transformace App d_address

Název	Popis
Microsoft Excel Input	Načtení dat
Select values	Vybrání relevantních sloupců pro adresu
If field value is null	Nahrazení null hodnot odpovídajícími hodnotami (0, N/A)
Calculator	Vytvoření popisů „Země“ a „Okres“
Filter rows district	Rozdělení dat na ta, která mají vyplněný okres a která ne
Create new disctict	Nakopírování hodnot do atributu „a_district“
Get district name	Hodnoty nemáme, musíme je dohledat v „code_names“ pomocí kódu okresu
Filter rows country, Create new country, Get country name	Totéž pro zemi (atribut „a_country“)
Combination lookup/ update	Vytvoření nového klíče („ID_address“), pokud neexistuje
Select values 2	Zahození pomocných sloupců
Sort rows student	Seřazení dat podle „id_student“

Název	Popis
Unique rows	Získání unikátních studentů
Sort + unique rows addresses	Totéž pro adresy
Insert / Update	Nahrání příslušných dat do tabulky „d_address“
Combination lookup/ update	Vytvoření nového klíče („ID_student _address“), pokud neexistuje
Insert / Update	Nahrání příslušných dat do tabulky „d_student_d_address“

Tabulka A.3: Popis jednotlivých kroků transformace App d_address

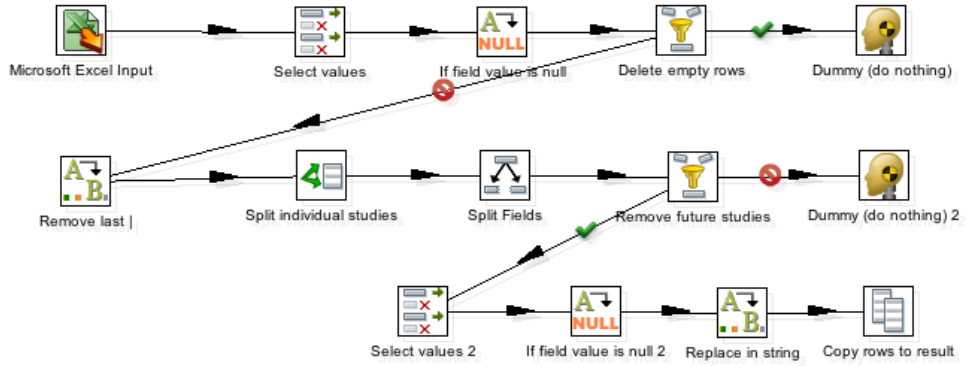
Transformace App d_high_school



Obrázek A.5: Transformace App d_high_school

Název	Popis
Microsoft Excel Input	Načtení dat
Get bachelors	Získání pouze přihlášek do bakaláře (magistři střední školu nevyplňují)
Dummy	Zahoď přihlášky do magisterského studia
Select values	Vyber relevantní hodnoty
If field value is null	Nahrazení null hodnot odpovídajícími hodnotami (0, N/A)
Calculator	Vytvoření popisů Typ střední školy“ a „Obor SŠ“

*[H]

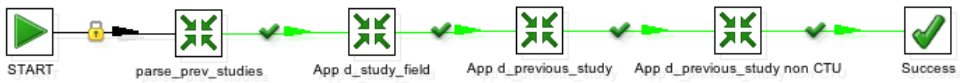


Obrázek A.7: Transformace parse_prev_studies

Název	**Popis**
Get type of school	Získání typu střední školy z tabulky „code_names“
Get field name	Získání názvu oboru z tabulky „code_names“
Create ID_high_school, ID_high_school_field, ID high_school_study	Vytvoření nových klíčů
Insert / Update	Nahrání dat do příslušných tabulek

Tabulka A.4: Popis jednotlivých kroků transformace App d_high_school

Úloha App prev_studies



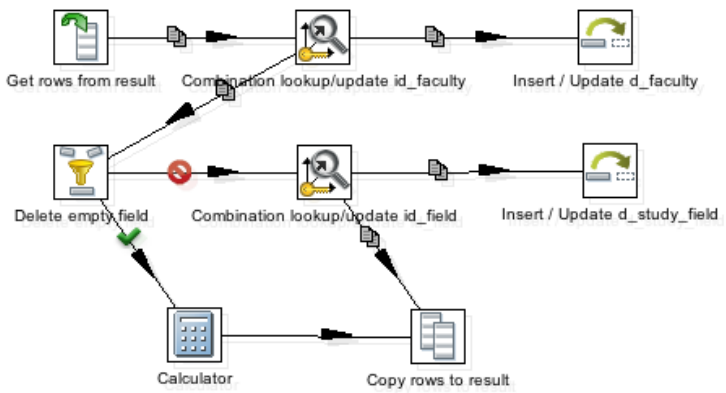
Obrázek A.6: Úloha App prev_studies

A.1.0.1 Transformace parse_prev_studies

Název	Popis
Microsoft Excel Input	Načtení dat
Select values	Vyber relevantní hodnoty
If field value is null	Nahrazení null hodnot odpovídajícími hodnotami (0, N/A)
Delete empty rows	Najdi prázdná předchozí studia
Dummy	Smaž je
Remove last |, Split individual studies, Split fields	Pomocí regulárních výrazů rozparsuj string z exportu na jednotlivá studia a následně jednotlivé položky v konkrétním studiu
Remove future studies	Pokud je datum registrace přihlášky menší než datum zápisu do studia, jedná se u budoucí studium (v době přihlášky neexistující)
Dummy	Smaž budoucí studia
Select values 2	Smaž dočasné sloupce
If field value is null	Nahrazení null hodnot odpovídajícími hodnotami (0, N/A)
Replace in string	Nahraď „—“ prázdnými znaky
Copy rows to result	Pošli data na výstup transformace

Tabulka A.5: Popis jednotlivých kroků transformace parse_prev_studies

Transformace App d_study_field



Obrázek A.8: Transformace App d_study_field

Název	Popis
Get rows from result	Načti data ze streamu (tzn. výstup předchozí tranformace)
Combination lookup/ update	Vytvoření nového klíče („ID_faculty“), pokud neexistuje
Insert / Update	Nahrání dat do tabulky „d_faculty“
Delete empty field	Smaž prázdný obor (N/A)
Combination lookup/ update	Vytvoření nového klíče („ID_field“), pokud neexistuje
Insert / Update	Nahrání dat do tabulky „d_study_field“
Calculator	Prázdnému oboru přiřaď ID 0
Copy rows to result	Pošli data na výstup transformace

Tabulka A.6: Popis jednotlivých kroků transformace App d_study_field

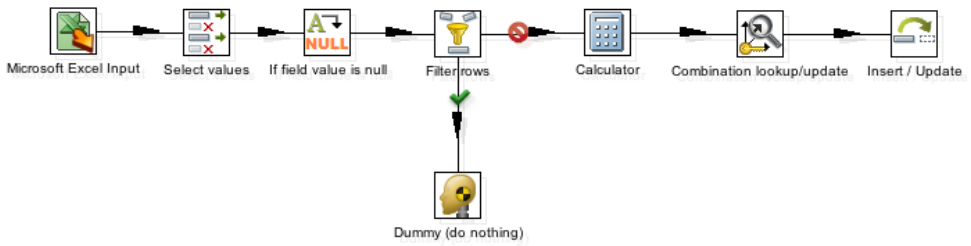
Transformace App d_previous_study

Obrázek A.9: Transformace App d_previous_study

Název	Popis
Get rows from result	Načti data ze streamu (tzn. výstup předchozí tranformace)
Calculator	Nastav atribut „a_at_CTU“ na true
If field value is null	Nahrazení null hodnot odpovídajícími hodnotami (0, N/A)
Combination lookup/ update	Vytvoření nového klíče („ID_prev_study“), pokud neexistuje
Insert / Update	Nahrání dat do tabulky „d_prev_study“

Tabulka A.7: Popis jednotlivých kroků transformace App d_previous_study

Transformace App d_previous_study non CTU

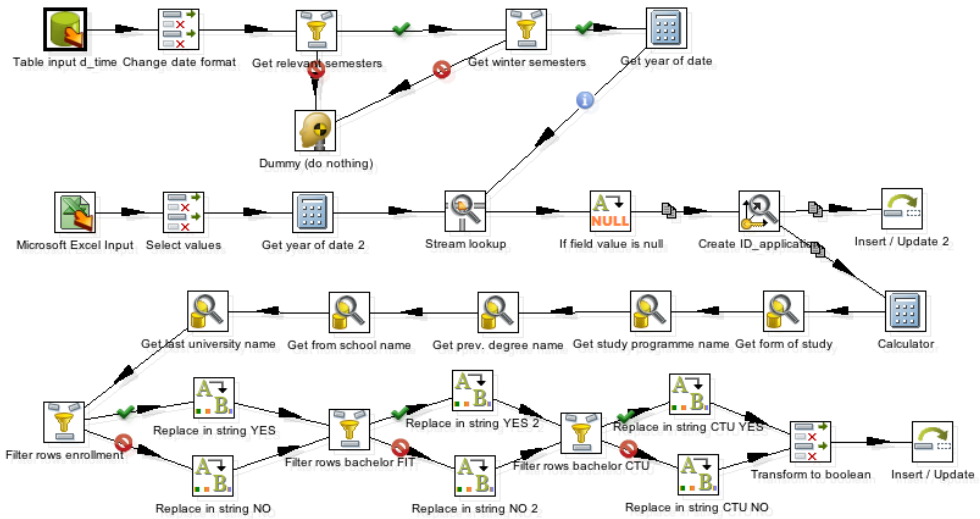


Obrázek A.10: Transformace App d_previous_study non CTU

Název	**Popis**
Get rows from result	Načti data ze streamu (tzn. výstup předchozí tranformace)
Select values	Vyber relevantní hodnoty
If field value is null	Nahrazení null hodnot odpovídajícími hodnotami (0, N/A)
Filter rows	Najdi prázdné popisy
Dummy	A smaž je
Calculator	Nastav atribut „a_at_CTU“ na false
Combination lookup/ update	Vytvoření nového klíče („ID_prev_study“), pokud neexistuje
Insert / Update	Nahrání dat do tabulky „d_prev_study“

Tabulka A.8: Popis jednotlivých kroků transformace App d_previous_study non CTU

Tranformace App d_application



Obrázek A.11: Transformace App d_application

Název	Popis
Table input d_time	Načti data z tabulky „d_time“
Change data format	Změň formát data na „dd.MM.yyyy“
Get relevant semesters	Najdi relevantní semestry pro FIT (tzn. ne starší než 2009/2010)
Get winter semesters	Najdi pouze zimní semestry
Dummy	Zahoď nepotřebné semestry
Get year of date	Získej rok z data začátku semestru
Microsoft Excel Input	Načti data
Select values	Vyber relevantní hodnoty
Get year of date 2	Získej rok z data registrace
Stream lookup	Porovnej roky a vrať „id_time“
If field value is null	Nahrazení null hodnot odpovídajícími hodnotami (0, N/A)
Create „ID_application	Vytvoření nového klíče („ID_application“), pokud neexistuje
Insert / Update	Nahrání dat do tabulky „f_application_results“
Calculator	Vytvoření popisů pro dohledání hodnot v tabulce „code_names“

Název	**Popis**
Get form of study, study programme...	Získání názvu atributu z jeho kódu
Filter rows enrollment, bachelor FIT + CTU	Rozdělení dat na ty, které mají vyplněnou hodnotu na významovou hodnotu true
Replace in string YES	Převod původní hodnoty na true
Replace in string NO	Převod původní hodnoty na false
Transform to boolean	Vytvoření boolenovské hodnoty ze stringu (příp. integeru)
Insert / Update	Nahrání dat do tabulky „d_application“

Tabulka A.9: Popis jednotlivých kroků transformace App d_application

Příloha B

Seznam použitých zkratek

FIT Fakulta informačních technologií

ČVUT České vysoké učení technické v Praze

ECTS European Credit Transfer System

PDI Pentaho Data Integration

BI Business Intelligence

DSA Data Staging Area

ETL Extract, Transform, Load

SCD Slowly Changing Dimension

CDE Community Dashboard Editor

BSP Bakalářský studijní program

MSP Magisterský studijní program

IZO Identifikační znak organizace

Příloha C

Obsah přiloženého CD

- `readme.txt` ..... stručný popis obsahu CD
- `src`
  - `analyzy` ..... podklady pro analýzy
  - `dashboardy` ..... podklady pro dashboardy
  - `datovy-sklad` ..... podklady pro datový sklad
- `text` ..... text práce
  - `DP_Hruba_Eliska_2014.pdf` ..... text práce ve formátu PDF
  - `DP_Hruba_Eliska_2014.tex` .. zdrojová forma práce ve formátu LaTeX