Sem vložte zadání Vaší práce.

České vysoké učení technické v Praze
Fakulta informačních technologií
Katedra softwarového inženýrství

Diplomová práce

Datový sklad pro anketu ČVUT

Bc. Jiří Grill

Vedoucí práce: Ing. Stanislav Kuznetsov

5. května 2015

Poděkování

Chtěl bych poděkovat Ing. Stanislavu Kuznetsovovi za vedení práce a Ing. Michalu Valentovi, Ph.D. za ochotu a vytrvalost odpovídat na mé dotazy.

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. V souladu s ust. § 46 odst. 6 tohoto zákona tímto uděluji nevýhradní oprávnění (licenci) k užití této mojí práce, a to včetně všech počítačových programů, jež jsou její součástí či přílohou, a veškeré jejich dokumentace (dále souhrnně jen „Dílo“), a to všem osobám, které si přejí Dílo užít. Tyto osoby jsou oprávněny Dílo užít jakýmkoli způsobem, který nesnižuje hodnotu Díla, a za jakýmkoli účelem (včetně užití k výdělečným účelům). Toto oprávnění je časově, teritoriálně i množstevně neomezené. Každá osoba, která využije výše uvedenou licenci, se však zavazuje udělit ke každému dílu, které vznikne (byť jen zčásti) na základě Díla, úpravou Díla, spojením Díla s jiným dílem, zařazením Díla do díla souborného či zpracováním Díla (včetně překladu), licenci alespoň ve výše uvedeném rozsahu a zároveň zpřístupnit zdrojový kód takového díla alespoň srovnatelným způsobem a ve srovnatelném rozsahu, jako je zpřístupněn zdrojový kód Díla.

V Praze dne 5. května 2015 .....

Odkaz na tuto práci

Grill, Jiří. *Datový sklad pro anketu ČVUT.* Diplomová práce. Praha: České vysoké učení technické v Praze, Fakulta informačních technologií, 2015.

Abstrakt

Tato diplomová práce se zabývá reimplementací stávajícího datového skladu, jež shromažďuje záznamy získané ze školní aplikace Anketa. Pomocí této aplikace jsou na konci každého semestru ukládány záznamy z jednotlivých anket fakult Českého vysokého učení technického v Praze.

Cílem práce je vytvořit řešení, které bude možné dále integrovat do nově vznikajícího interního projektu „Datová čistota“ Fakulty informačních technologií.

Současně tato práce popisuje problematiku datových skladů, metadat a reportovacích nástrojů pro tvorbu reportů tvořených ze záznamů uložených v datových skladech.

Výsledným řešením je implementace umožňující podrobnou analýzu dat získaných z databáze aplikace Anketa za použití ETL skriptů.

Klíčová slova Datový sklad, Centrální datový sklad, Datové tržiště, Metadata, Reporting, Reportovací nástroje, Business Intelligence, OLAP, Mondrian, Pentaho

Abstract

This master thesis focuses on reimplementation of already existing data warehouse, which gathers and stores all the entries from the university-wide application Anketa. Main purpose of this data warehouse is to store the all data from particular surveys and according faculty at Czech Technical University in Prague at the end of each semester.

The main goal of this paper is to create a solution that would allow further integration into the newly established internal project at the Faculty of Information Technology.

Furthermore, this paper introduces the general issues of data warehouses and metadata.One standalone chapter is dedicated to reporting tools that are used to generate reports from all the entries stored in the data warehouse.

Resulting solution of this thesis is an implementation of the data warehouse that allows in-depth analysis of all the entries from the Anketa application database using the ETL scripts.

Keywords Data Warehouse, Stage, Enerprise Data Warehouse, Data Marts, Metadata, Reporting, Eeporting Tools, Business Intelligence, OLAP, Mondrian, Pentaho

Obsah

Seznam obrázků

Seznam tabulek

Úvod

Tato diplomová práce vznikla jako součást projektu interního projektu „Datová čistota“, který si klade za cíl vybudovat softwarovou infrastrukturu pro podporu informovaného rozhodování nejen pracovníku ČVUT v Praze.

Jelikož současné řešení datového skladu nad aplikací školní ankety nesplňuje parametry potřebné proto, aby mohlo být součástí nově vznikajícího projektu, bylo potřeba vytvořit řešení, které bude možné v budoucnu integrovat do již zmiňovaného projektu.

V rámci školní ankety mohou studenti na konci každého semestru hodnotit jak předměty, které měli daný semestr zapsané, tak vyučující, kteří tyto předměty vyučovali. Tím se výsledky anket stávají velmi cenným zdrojem informací a důležitou zpětnou vazbu nejen pro samotné učitele, ale i pro vedení školy.

Výsledkem práce bude vytvoření vhodné struktury přístupové vrstvy, pomocí které bude moci uživatel analyzovat data uložená v datovém skladu. A následně z těchto analýz vytvářet reporty.

Tuto práci, lze pomyslně rozdělit na dvě části. V první polovině se diplomová práce zaměřuje na teoretický základ datových skladů, na kterou v druhé polovině navazuje popis implementace nového řešení.

V první části diplomové práce bude také popsána problematika metadat a jejich uložení v datových skladech, na kterou bude navazovat kapitola, v rámci které budou srovnána vybraná řešení reportovacích nástrojů na tvorbu reportů pro koncové uživatele (učitelé, vedení školy apod). Na základě vzniklého srovnání, bude doporučen vhodný nástroj, který by měl být v budoucnu použit k tvorbě reportů.

Druhá, implementační část diplomové práce začíná vytyčením cílů, které by měly být dosaženy při implementaci datového skladu. Poté následuje analýza stávajícího řešení, na kterou navazuje návrh a realizace nové implementace, která přinese možnost integrace do interního projektu. V závěru druhé části bude nově vzniklé řešení otestováno a provedeno hodnocení dosažených cílů.

Kapitola 1

Datové sklady

Datový sklad je architektura databází navržených tak, aby bylo možné v nich ukládat data extrahovaná z transakčních systémů, úložišť operačních dat a externích zdrojů. Datový sklad kombinuje agregaci a sumarizaci dat do takové formy, která umožňuje dynamické analýzy záznamů, z kterých je dále možno vytvářet reporty pro předem definované byznys potřeby.

„A data warehouse is a subject-oriented, integrated, time-variant and non-volatile collection of data in support of management's decision making process.“[1]

Tato definice, kterou zavedl Bill Inmon je širokou veřejností přijímána jako jedna ze základních definic datového skladu. Jelikož toto řešení staví na základech této, je vhodné si ji přiblížit.

„Subject oriented“ znamená, že datový sklad může být použít k analýze konkrétní části byznysu, např prodej, nákup, zisk apod.

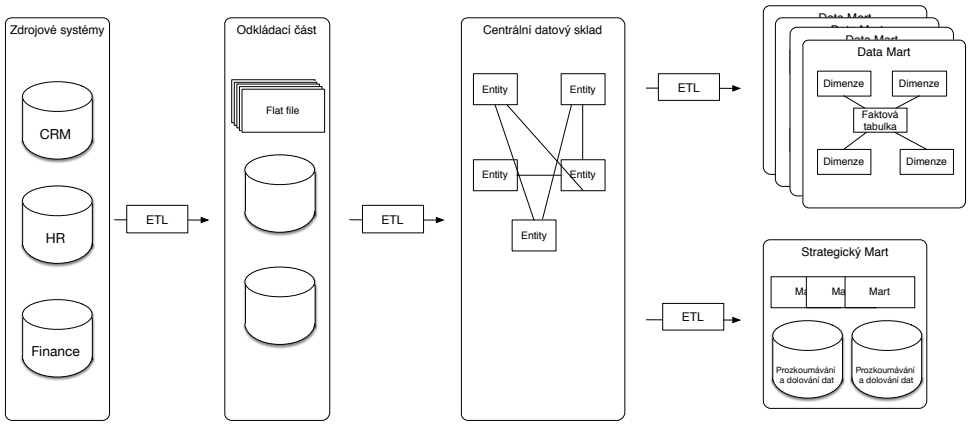
„Integrated“ značí schopnost datových skladů ukládat data z různých zdrojových systému, kde jedna entita v daném zdrojovém systému může znamenat něco jiného než ta samá entita v druhém zdrojovém systému a datový sklad má za úkol v rámci integrace dat sjednotit význam jednotlivých entit.

„Time - variant“ značí možnost ukládání historických dat v rámci datového skladu v řádu několika měsíců až let. Jelikož data obvykle přicházejí z transakčních systému, kde jsou uložena pouze ta aktuální, můžou se v průběhu času měnit (např adresa zákazníka). Datový sklad naopak podporuje uložení všech historických verzí těchto dat.

„Non-Volatile“ znamená, že jakmile jsou záznamy do datového skladu nahrány, už by se neměly měnit.

1.1 Obecná architektura datového skladu

Podle definice Billa Inmona byla vytvořena obecná architektura datového skladu. Jak tato architektura vypadá, je možné vidět na obrázku 1.1.



Obrázek 1.1: Diagram architektury datového skladu podle Inmona

Architekturu lze rozdělit na čtyři části. První částí nazvaná *zdrojové systémy* je tvořená zdrojovými systémy. Tyto systémy produkují transakční data, k jejichž dlouhodobému uchování slouží datový sklad.

Z těchto systémů jsou pomocí ETL skriptů přenesena/extrahována data do *odkládací části*, která se v angličtině nazývá „Stage“. Tato část může být tvořena databází nebo soubory typu csv, txt, xml apod.

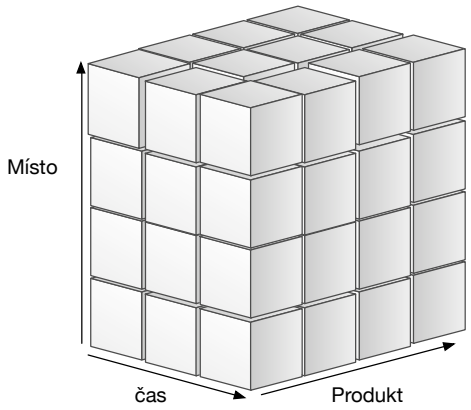
Třetí částí je *centrální datový sklad*, který se v angličtině nazývá „Enterprise Data Warehouse“, v rámci kterého jsou vytvořeny entity tak, aby odpovídaly zdrojovým systémům a zároveň poskytly ucelený pohled na ně samotné.

Tyto entity jsou v centrálním datovém skladu ukládány v normalizované podobě. Nejčastější forma normalizace, je třetí normální forma. Tento stupeň normalizace poskytuje eliminaci duplicit, nedělitelnost jednotlivých informací uložených ve sloupcích tabulek, správnou závislost hodnot záznamů v tabulce na primárním klíči tabulky a celkově přispívá ke správě databáze.[2].

Z centrálního datového skladu jsou poté data nahrána pomocí ETL skriptů do tzv. *datových tržišť*, jejichž anglický ekvivalent je „data marts“. V těchto tržištích, která představují určitý výsek z centrálního datového skladu, jsou poté uložena data v denormalizované podobě (opak normalizace). Denormalizace přispívá k lepší optimalizaci dat pro následnou analýzu. Jednotlivá datová tržiště poté slouží jako přístupová vrstva pro aplikace, pomocí kterých se provádí analýzy nad záznamy získaných ze zdrojových systémů .

Poslední prvek, který je postavený nad datovými tržišti se nazývá *datová kostka* neboli OLAP kostka. Tato kostka, jež může být uložena i ve více než třech dimenzích, slouží k analýze dat. Příklad OLAP kostky lze vidět na obrázku 1.2. Díky tomu, že lze takto vzniklou kostku různě natáčet, zanořovat se apod. má uživatel možnost získat údaje z různých úhlů pohledu podle toho, co s danou kostkou provede. Každá kostka se skládá ze dvou základních prvků.

První z nich jsou fakta tvořená faktovými tabulkami. Tyto tabulky nesou



Obrázek 1.2: Příklad OLAP kostky

numerické hodnoty (míry), na které se poté používají agregační funkce jako je funkce SUM nebo AVG. Pomocí těchto funkcí poté můžeme získat např. počet prodaných kusů určitého výrobku za daný čas apod.

Druhým typem jsou dimenze. Dimenze představují hierarchicky či logicky uspořádané údaje (např dimenze času, která může mít uspořádání rok-měsíc-týden-den nebo produkt, který může být uspořádán výrobce-název-kód apod.) do stromových struktur.

Faktové tabulky kromě měr nesou ještě odkazy na jednotlivé dimenze. Pokud ovšem dimenze obsahuje pouze primární klíč, může být tato dimenze uložena přímo ve faktové tabulce v podobě jednoho z jejích atributu jako „degenerovaná dimenze“.

1.2 Naplnění datového skladu daty

Aby mohly být datové sklady využívány, je potřeba je naplnit daty ze zdrojových systémů. To se děje za pomoci již zmíněných „ETL skriptů“, které se dají rozdělit do třech fází:

„Extract“ - V tomto kroku se data extrahují ze zdrojového systému. V rámci tohoto kroku může datový sklad extrahovat i z několik nehomogenních prostředí najednou.

„Transform“ - V tomto kroku dojde k ověření, očištění a integraci extrahovaných dat. V tomto kroku dochází ke sjednocení, dopočítání, či jinému dodání chybějících informací tak, aby výsledné informace měly požadovanou podobu pro uložení do datového skladu.

„Load“ - V tomto kroku dojde k nahráni dat do databáze datového skladu.

Kapitola 2

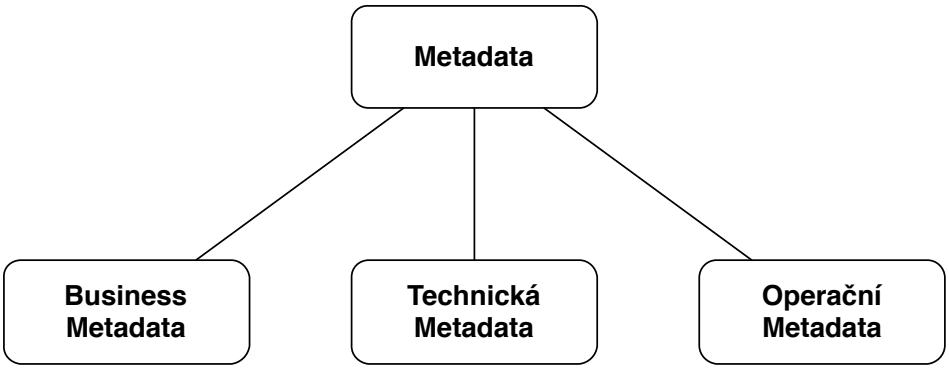
Metadata v datových skladech

Metadat jsou nedílnou součástí datových skladů stejně tak jako data samotná, protože bez metadat by data uložená v datovém skladu nedávala žádný smysl.

2.1 Definice a rozdělení metadat

Metadata lze chápat jako informace popisující různé aspekty informačních aktiv, které pomáhají zlepšit jejich využití během jejich životního cyklu. Jsou to právě metadata, která přemění informaci v majetek, který má pro společnost nějakou hodnotu. Obecně se dá říci, že čím hodnotnější jsou tyto informace o majetku, tím je zásadnější, aby metadata byla správně spravována, protože jsou to právě definice metadat, které pomáhají pochopit význam jednotlivých hodnot.

Metadata lze rozdělit do třech základních skupin tak, jak je zobrazeno na obrázku 2.1.



Obrázek 2.1: Rozdělení metadat

2.1.1 Business Metadata

Business metadata pomáhají netechnickým uživatelům systému, jako jsou například manažeři, testeři, obchodníci atd. Poskytují uživateli informace, kde můžou najít jednotlivé informace v datovém skladu, odkud tyto informace pocházejí a jak se tam dostaly. Dále popisují jejich kvalitu a jak lze tyto informace interpretovat.

2.1.2 Technická Metadata

Technická metadata naopak pomáhají technickým uživatelům při vytváření samotného systému. Do této kategorie spadají metadata týkající se názvů databázových systémů a jejich prvků jako jsou tabulky, sloupce, velikosti, datové typy obsažené v těchto tabulkách a povolené hodnoty. Do technických metadat lze dále zahrnout i informace týkající se datových modelů datového skladu apod.

2.1.3 Operační Metadata

Operační metadata podporují denní operace systému a správu celkového BI prostředí. Do těchto metadat se řadí logy serveru, ukázky vytíženosti procesorů a pamětí atd.

2.2 Důvody uložení metadat

Metadata se ukládají v takzvaném "Metadata repositáři", což si lze představit jako sdílenou databázi metadat napříč celým podnikem. Díky tomu mají jeho uživatelé přístup ke konzistentním a spolehlivým informacím. Důvody proč ukládat metadata v datovém skladu jsou následující:

2.2.1 Mapování informací

Mapování informací je velmi důležité, protože při nahrávání dat do datového skladu dochází k přenosu dat z operačního prostředí (zdrojové systémy) do datového skladu. Při tomto procesu dochází k mapování jednoho atributu na druhý, konverzi, změně v jmenných názvech, filtrování dat atd. Metadat nesoucí informace o transformacích, které byly prováděny s daty, poté umožní uživateli pochopit, co se s původními daty stalo a v jaké podobě se uložily do datového skladu.

2.2.2 Spravování dat v průběhu času

Toto je další velmi důležitý důvod, proč metadat ukládat. Jelikož doba, po kterou jsou data uložená v operačním prostředí je v porovnání s dobou, po kterou jsou data uložená v datovém skladu, velmi krátká. Díky tomu může

dojít ke změně datové struktury (klíče, hodnoty atd). Pro zpřehlednění změn dat v rámci delšího časového úseku jsou do repositáře ukládány metadata, která pomáhají uživateli najít správná data z různých časových úseků.

2.2.3 Verzování dat

Verzování dat souvisí s předchozím bodem tak, že při každé změně dat uložených v datovém skladu, by měl v datovém skladu vzniknout nový záznam (nová verze), aby pozdější uživatel neměl problém nalézt správný záznam. Jedním ze způsobů jak zajistit verzování dat uložených v datovém skladu je vytvoření proměnné, která značí časové rozmezí ("od - do") kdy je daný záznam platný. Druhým způsobem je, že se do proměnně uloží pouze začátek časové platnosti záznamu a konec se platnosti se vždy vypočítá ze začátku následující verze.

2.3 Obsah metadat popisující prvky datového skladu

2.3.1 Základní komponenty

V základních komponentech metadat jsou uloženy informace o tabulkách datového skladu, klíče k těmto tabulkám a jejich atributy.

2.3.2 Mapování

Obsahem metadat popisující mapování tabulek by měly být následující prvky:

- Identifikátor zdrojového pole
- Jednoduchá ukázka mapování
- Konverze atributů
- Konverze referenčních tabulek
- Změna názvu
- Defaultní hodnoty
- Logiku, která je použita při výběru z různých zdrojů
- algoritmus, jakým dochází ke změně

I tady by v případě změny metadat mělo docházet k verzování metadat.

2.3.3 Historie extrahování

Tento záznam v metadatech by měl umožnit uživatel zjistit, kdy byl konkrétní záznam vyextrahován z operačního prostředí a uložen do datového skladu.

2.3.4 Ostatní

Do této kategorie komponent metadat patří *Alias*, což znamená, že v rámci datového skladu se může typ záznamu jmenovat jinak, než je tomu v operačním prostředí. Díky tomu je možné snáze pochopit, co takový záznam představuje.

Dalším atributem, který patří do této skupiny je *Status*, který může sloužit jako indikátor toho, že daný záznam už není aktivní nebo je chybný.

Další atributy jsou tzv. měřitelné atributy o datech v datových skladech, které typicky obsahují:

- počet řádků, které se momentálně nachází v tabulce
- jak rychle daná tabulka roste
- charakteristika použití tabulky
- indexaci tabulky a její strukturu

2.3.5 Sumarizační algoritmy

Tato komponenta pomáhá pochopit uživateli, jak došlo k sumarizaci nebo jiné kalkulaci dat napříč jednotlivými vrstvami datového skladu.

2.3.6 Historie vztahů

Datové artefakty jsou výsledkem vztahů vzniklých mezi daty, které se vyskytují v datových skladech. Tyto artefakty jsou velmi důležité v následné interpretaci dat z datových skladů. Typická historie vztahů obsahuje následující informace:

- tabulky, které jsou obsažené ve vztahu
- datum vytvoření vztahu
- jaké parametry byly použity k vytvoření vztahu
- kardinality
- slovní popis vztahu
- kdo spravuje vzniklý vztah mezi tabulkami

2.3.7 Vlastník/Správce

Jelikož data do datových skladů přicházejí z operačního prostředí, je jasné že tím, kdo za ně odpovídá, je vlastník operačního prostředí.

Naopak ten, kdo se o data stará v rámci datových skladů a je tak za ně zodpovědný, je správce.

2.3.8 Přístup k datům

Tento atribut popisuje, kteří uživatelé/systémy mají přístup k datům uloženým v datových skladech.

2.3.9 Referenční tabulky

Referenční data/tabulky jsou taková data, která jsou uložena v externích tabulkách a obsahují často používané hodnoty. Obvykle jsou data v referenčních tabulkách uloženy v nestandardní podobě. Jak se data mění v průběhu času, je důležité měnit obsah referenčních tabulek.

2.3.10 Datový model

Tento atribut metadat představuje datový model datových skladů. Díky tomu může uživatel snáze pochopit závislosti mezi fyzickým a logickým modelem datového skladu.

V rámci této diplomové práce má čtenář možnost vidět ukázku některých typů metadat jako je model řešení, popis atributů tabulek, mapování tabulek datového skladu na tabulky zdrojového systému apod.

Kapitola 3

Reportovací nástroje v oblasti Business Intelligence

Vytváření reportů z analýz nad daty uložených v datových skladech je jedna z nejdůležitějších činností samotného BI oddělení firmy. Proto by měly reportovací nástroje splňovat určitý standard co se funkcionality týče, ale také naplňovat základní principy reportingu. V této kapitole budou nejprve popsány hlavní důvody, proč vůbec vytvářet reporty, následované popisem základních oblastní reportingu. V další části budou popsány základní principy, které by měly jednotlivé reporty dodržovat, za kterými bude následovat popis základních funkcí a komponent reportovacích nástrojů. Předtím než budou zhodnoceny jednotlivá vybraná řešení, která jsou v současné chvíli nabízena, budou ještě popsány jednotlivé přístupové vrstvy reportovacích nástrojů a také možnosti vizualizace dat.

3.1 Hlavní důvody reportingu

Následující odstavce by měly čtenáři objasnit, proč se reporting tvoří, resp. na jaké otázky se snažíme reportingem odpovědět.

3.1.1 Co se stalo?

Report se nám snaží odpovědět, co se vlastně stalo, jaký jsme měli příjem apod. Tento druh reportu je dnes běžný a měl by být bezpodmínečně splnitelný pomocí daného reportovacího nástroje.

3.1.2 Co se děje právě teď?

Jedná se o real-time analýza, která nám říká, jak si právě teď stojíme a díky tomu můžeme pružně reagovat na trendy a události, které se dějí právě teď.

Tato analýza se v poslední době dere do popředí a stává se velmi důležitou součástí reportovacích nástrojů.

3.1.3 Proč se to stalo?

Tato otázka, na kterou se snažíme najít odpověď pomocí reportovacího nástroje, vznikla určitou evolucí dvou předešlých otázek. Report by nám měl pomoci nalézt odpovědi na to proč se to stalo, co jsme udělali špatně nebo co danou akci pohání.

3.1.4 Co se stane potom?

Odpovědí na tuto otázku se snažíme najít něco o budoucnosti, která ale souvisí s tím, co se událo dříve. A díky tomu se můžeme rozhodnout, co je třeba udělat, aby se stalo to, co chceme.

3.2 Základní oblasti reportingu

Základní oblasti nebo typy reportingu, do kterých lze reporting rozdělit jsou:

3.2.1 Operační

Operační reporting se zabývá tím, co se událo dneska nebo včera či před nějakou dobou a vyhodnocuje to z pohledu managementu (light management, ředitelé), který má potřebu sledovat daný problém či jeho předchozí vývoj.

3.2.2 Strategický

Slouží top managementu nebo výkonným ředitelům firem, kteří určují strategii v období pěti nebo deseti let, tak aby bylo možno firmu řídit, co se týče vize či strategie firmy.

3.2.3 Taktický

Slouží senior managementu, který zkoumá to, jak se požadovaná oblast vyvíjela nebo jak se měla vyvíjet. Zároveň se srovnává s výhledem na další období (např příští rok).

3.3 Základní principy reportingu

Následující podkapitola popisuje principy, které by měl správný report dodržovat.

3.3.1 Míra transparentnosti

Data, samotný report, ale i navazující akce, by měly mít jasně definované procesy, které musejí být popsány, tzn musejí být průhledné. Konzument reportu by měl být schopný zjistit, jakým procesem se dané číslo v reportu objevilo.

3.3.2 Možnost auditovatelnosti

Výstup reportu by měl mít možnost uložení pro pozdější audit a zároveň by měla být umožněna správa vytvořených reportů.

3.3.3 Úplnost

Report by měl pokrývat co nejširší škálu výstupů, aby poskytl celistvý údaj o dané problematice a ne pouze o části.

3.3.4 Relevantnost

Dodržování tohoto principu představuje zobrazení konzumentovi reportu pouze relevantní data (např. konzument vidí pouze data, ke kterým má bezpečnostní prověrku).

3.3.5 Přesnost

Jedná se o jeden ze základních principů reportingu. Podle tohoto principu by měl report poskytovat přesná data o tom, co se stalo nebo co se stane. S tímto principem souvisí i datová kvalita.

3.3.6 Neutralita

Informace z reportů by měly být doručovány bez tendenčního zabarveni či pozdější úpravy dat.

3.3.7 Možnost srovnání

Konzument reportů by měl mít možnost porovnávat jednotlivé reporty a na základě toho vyvodit důsledky.

3.3.8 Časové zařazení

Report by měl mít možnost časového zařazení. Souvisí to také s včasným dodáním reportů pokud se jedná o opakující se činnost.

3.3.9 Přístupové vrstvy reportovacích nástrojů

V této podkapitole budou popsány možné přístupové vrsty reportovacích nástrojů.

3.3.10 Databáze

V tomto případě se přístupová vrstva nachází přímo nad daty, z kterých daný report vychází - databází. Velmi často už je v tomto případě vyřešeno zabezpečení, či data governance na úrovni datového skladu. Tato vrstva má následující vlastnosti:

- View vs materializace - zpřístupnění podkladových dat ve formě tabulek nebo normalizace denormalizovaných objektů.
- Denormalizace objektů - přidání nového termínu, pojmu či jeho atributů tak, aby měl větší přínos pro byznys.
- Speciální vztahové objekty - přidávají tabulkám vztahy, které se v databázi vyskytují a zároveň je těžké se k nim dostati pomocí BI nástroje.
- Možnost časových řezů
- Manuální vstupy uživatele
  - Speciální transformace
  - Zabezpečení - jedná se o skrývání jednotlivých tabulek či jejich atributů.

3.3.11 BI nástroje

V tomto případě se přístupová vrstva tvoří v samotném BI nástroji. To znamená, že daný nástroj má svůj vlastní přístup, jak data modelovat. Tím může vzniknout problém, protože každý reportovací nástroj, má svůj vlastní způsob, jak takový model vytvářet. Většinou se jedná o objektově orientovaný nástroj, kde jako objekty nefigurují pouze samotné tabulky databáze, ale patří tam i objekty, které celý daný model drží pohromadě. Jedná se tak např o bezpečnostní nebo objekty s metadaty. Některé nástroje umožňují automatické generování na základě vrstvy, která leží pod ním (databáze). BI nástroje zastoupené v této přístupové vrstvě dále obsahují následující vlastnosti a prvky:

- Tool dependant - přístupová vrstva závisí na tom, jaký BI nástroj si vybereme (každý má svá specifika)
- Vliv na tvorbu přístupové vrstvy v databázi - jedná se např. o podporu star, či snowflake schéma daným nástrojem.
- Atributy a dimenze - Objekty, které jsme schopni popisovat
- Metriky - Objekty, které jsme schopni měřit a počítat
- Speciální objekty - jedná se např o transformační objekty či objekty s metadaty

- Bezpečnost - BI nástroje nabízejí bezpečnost formou přístupu k samotnému nástroji, licenční politikou nebo definování zabezpečení samotné aplikace.

3.3.12 Společné vlastnosti pro obě přístupové vrstvy

- Jednotný přístup - přístupová vrstva se snaží unifikovat přístup k jednotlivým oblastem a objektům pro dané byznys uživatele

- Jednodušší analýza - měly by zjednodušit přístup uživatelů k daným informacím tak, aby se daly tyto informace snadno konzumovat a následně s nimi operovat

- Byznys termíny - operují v byznys pojmech, která jsou danému byznys odběrateli známá

- Flexibilita - měly by dodávat jistou míru flexibility pro konzumenta tak, aby mohl nadále tuto vrstvu rozvíjet a přizpůsobovat aktuální potřebě.

- Maximální využití přístupové vrstvy - přístupová vrstva by měla být maximálně přizpůsobitelná na vrstvu, která se nachází pod ní (databáze) a měla by maximálně využívat možnosti dané platformy, na které vrstva běží.

3.3.13 Strategie pro tvorbu přístupové vrstvy

Vychází z toho, jakým způsobem daný nástroj pracuje s daty a jaké techniky nabízí. Podle toho lze strategie rozdělit do následujících typů:

- OLAP - Jedná se o standardní technologie, kterou daná přístupová vrstva využívá

- MOLAP - Data jsou uložená v kostkách, s kterými daný nástroj umí pracovat a dále prezentovat. Díky tomu velmi vzrůstá výkon, jelikož je možné si napočítat komplexní výpočty, které se skládají ze spousty funkcí, či statistických výpočtů. Ale na druhou stranu do kostek je možné ukládat pouze limitované množství dat a cena takového řešení vzrůstá s počtem kostek.

- ROLAP - Jedná se o možnost analýzy na úrovni databáze a uživatel je tak schopen dostávat data přímo z databáze, tudíž můžeme pracovat s takovou velikostí dat, kterou dovoluje databáze. Tímto řešením jsme schopni vyrovnat vlastnosti a výkon BI nástroje k dané databázi. Naopak se snižuje odezva reportů a celkově je funkcionalita omezena.

3.4 Vizualizace dat

Vizualizace dat se dá chápat jako vytváření a studium reprezentace podoby dat, které byly extrahovány ze zdroje a uživatel má potřebu do nich zařadit nějaké své atributy a proměnné tak, aby tomu porozuměl. Vizualizace dat je tvořena např. pomocí různě silných čar, velikostí textury nebo formou barev. Díky tomu všemu lze datům dát jasnější význam.

3.4.1 Dashboard

Jedná se o vizualizaci, která udává aktuální stav důležitých metrik, ale tato vizualizace neuvádí nic o tom, za jak dlouho dosáhneme cíle procesu. Tento druh reportu díky tomu, že poskytují agregovanou informaci, slouží pro top management firmy. Naopak, čím se dashboard stává konkrétnějším (až do detailu transakčních dat), tím je určen pro nižší management firmy. Ukázku této vizualizace lze vidět na obrázku 3.1.

3.4.2 Scorecard (Balanced Scorecard)

Tento druh vizualizace nabízí oproti dashboardu zobrazení průběhu celého procesu. Jedná se o výkonnostní metriku používanou ve strategickém řízení firmy k identifikaci a zlepšení množství vnitřních funkcí a jejich externí výsledky. Tato metrika by měla poskytnou organizaci zpětnou vazbu při realizaci strategie a cílů. Ukázku této vizualizace, lze vidět na obrázku 3.2.

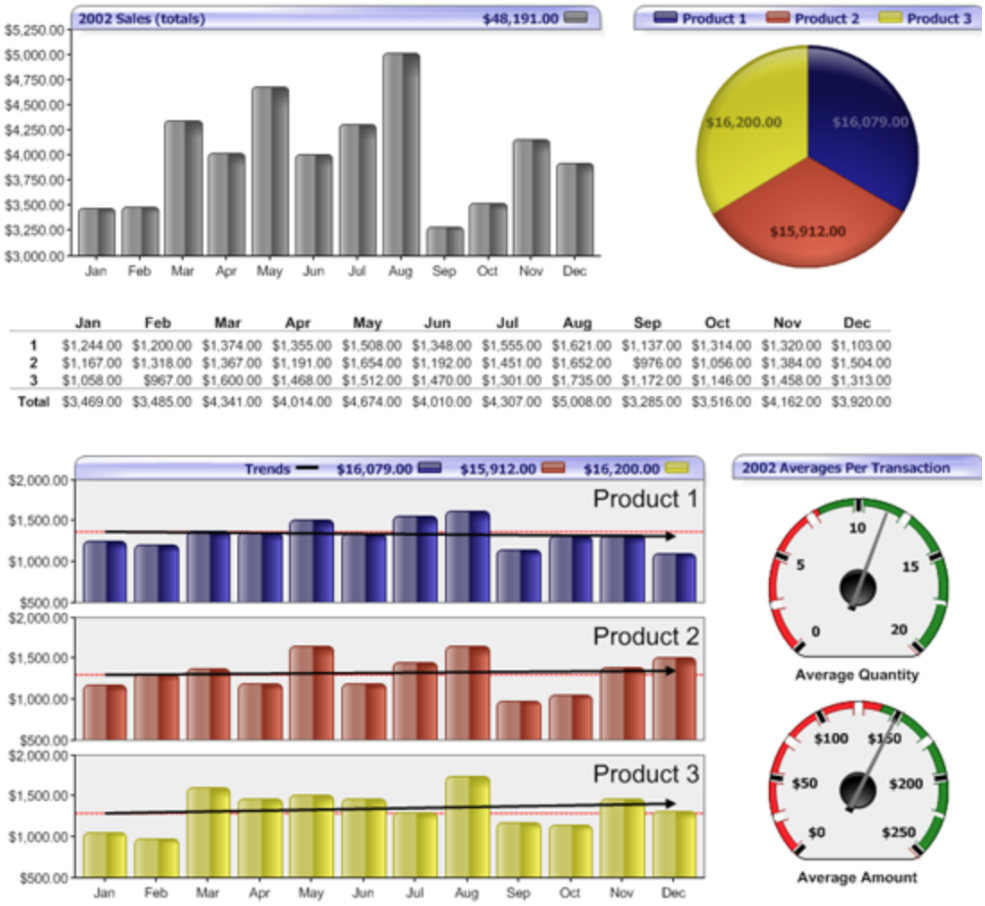
3.5 Zhodnocení vybraných řešení s možností BI reportingu

V následujících odstavcích jsou popsány vybrané komerční, ale i open source BI reportovací nástroje z pohledu, jak jsou schopné naplnit cíle reportingu (dodání informace, analýzy dat, integrace do podnikové i metadatové infastruktury).

Popisované nástroje byly vybrány na základě studie “Gartner Magic Quadrant”, jejíž výsledek je zobrazen na obrázku 3.3, tak aby v open source a komerční řešení bylo na obrázku co možná nejblíže.

3.5.1 MicroStrategy

Jako zástupce komerční verze BI řešení byl vybrán produkt od firmy MicroStrategy. Jedná se o standardní komerční BI platformu, která nabízí všechny možnosti, jak s daty pracovat a následně je i prezentovat, auditovat a doručovat uživateli. Tudíž všechny potřebné komponenty, které by měl správný BI reportovací nástroj obsahovat. Architektura MicroStrategy BI platformy je zachycena na obrázku 3.4.



	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
1	$1,244.00	$1,200.00	$1,374.00	$1,355.00	$1,508.00	$1,348.00	$1,555.00	$1,621.00	$1,137.00	$1,314.00	$1,320.00	$1,103.00
2	$1,167.00	$1,318.00	$1,367.00	$1,191.00	$1,654.00	$1,192.00	$1,451.00	$1,652.00	$976.00	$1,056.00	$1,384.00	$1,504.00
3	$1,058.00	$967.00	$1,600.00	$1,468.00	$1,512.00	$1,470.00	$1,301.00	$1,735.00	$1,172.00	$1,146.00	$1,458.00	$1,313.00
Total	$3,469.00	$3,485.00	$4,341.00	$4,014.00	$4,674.00	$4,010.00	$4,307.00	$5,008.00	$3,285.00	$3,516.00	$4,162.00	$3,920.00

Obrázek 3.1: Ukázka typu vizualizace dashboard (Zdroj: Google)

Firma MicroStrategy nabízí tři možnosti přístupu:

- MicroStrategy Desktop - Desktopová aplikace (tlustý klient)
- MicroStrategy Web - tenký klient využívající webové rozhraní
- Microstrategy Mobile - tenký klient využívající webové rozhraní

MicroStrategy představuje jednoho z top leaderů v tomto odvětví. Operují s největším počtem dat (průměrně více než 17TB dat) a v průměru instalaci řešení obhospodařuje více než 2500 uživatelů. Na druhou stranu je toto řešení v porovnání s ostatními poměrně drahé a jejich licenční politika je velmi komplexní a proto není jednoduché se v ní orientovat.

Business Priority Area	KPI	KPI Measure	Monthly Evaluation			Comment
			1	2	3	
Financial	1	Working Capital				Yes, Peckham remains very near to London.
	2	Debtor Days				All cash in hand, no income tax, no VAT.
	3	Creditor Days				Usually Thursday and Friday but Del avoids the pub on those days.
Customer	4	Distribution Coverage (%)				Dominant in Peckham, South London
	5	Quality Demerit Index				Dry-clean only raincoats and dodgy watches have affected performance – amongst others.
Process	6	Sales forecast accuracy (%)				Targets 100% success in supplying what nobody needs or wants.
Logistics	7	Warehouse Storage Cost (per piece)				Included in flat/garage rent at Nelson Mandela House.
	8	Transport Cost (per km)				Company policy to use 3 wheel Reliant Regal van for tyre economy.
	9	Stock Security				Constant screen monitoring by Grandad or Uncle Albert.
Learning	10	Training Days				Rodney aka Dave already has 2 GCSEs but studies Computers at night school.

Obrázek 3.2: Ukázka typu vizualizace scorecard (Zdroj: Google)

3.5.2 Jaspersoft Business Intelligence Suite (Community Edition)

Produkt od firmy Jaspersoft představuje původní open source produkt společnosti TIBCO. Tento produkt nabízí reporty založené na webovém přístupu, dasboardy a jejich analýzy a to vše za odpory cloud computingu, mobilního BI a Big Data. Kromě bezplatné verze tato společnost nabízí placené verze v několika rozděleních: Express, Professional a Enterprise. Tyto jednotlivé verze se rozlišují jednak různým způsobem licencování, podporou, ale také různým rozsahem poskytovaných funkcí.

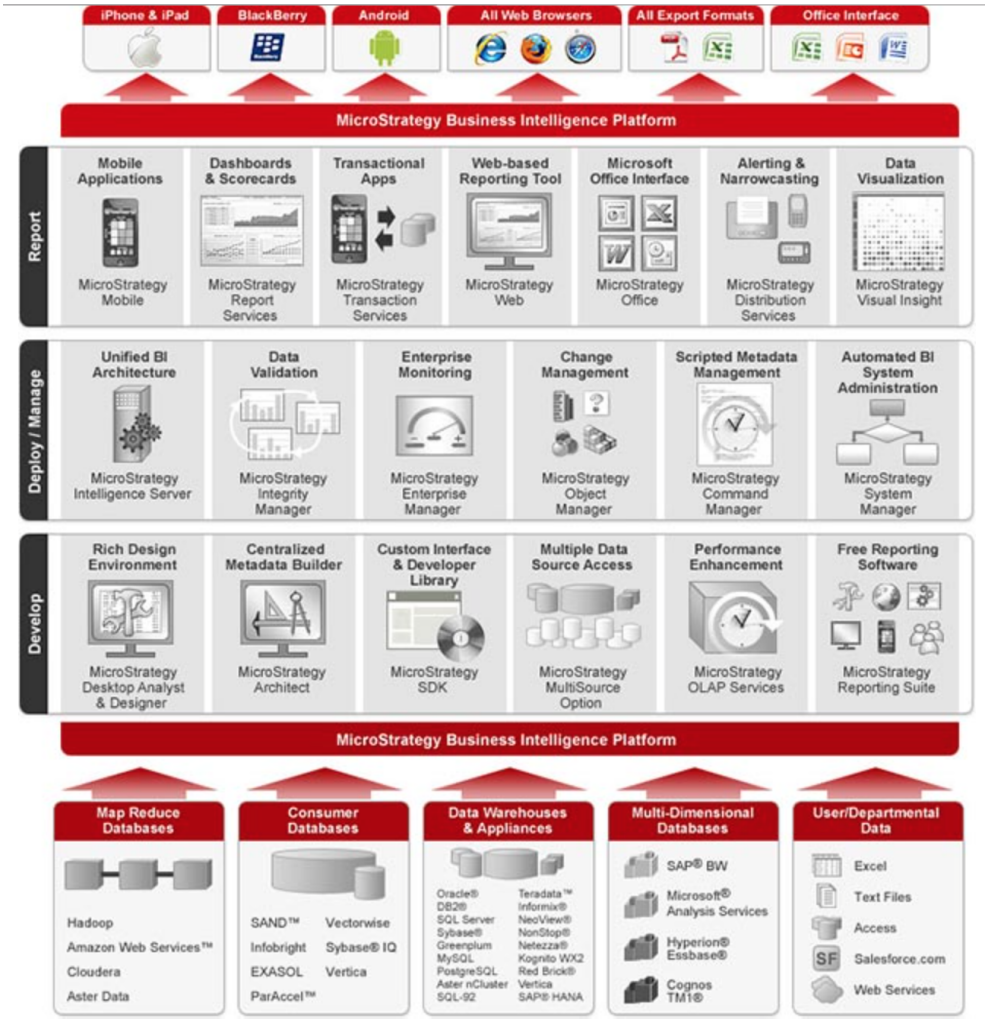
Architekturu tohoto řešení lze nalézt na obrázku 3.5. JasperReports Server je samostatný reportovací server, který poskytuje nepostradatelné informace jak v reálném čase tak i s pravidelnou četností. Výstup toho serveru je možné publikovat na webu, tisknout nebo exportovat do různých datových formátů. K vytváření sofistikovaných layoutů výstupů z JasperReports Serveru slouží Jaspersoft Studio. Jedná se o open source reportovací nástroj založený na programu Eclipse. Díky tomu mohou vytvářet uživatelé tabulky, obrázky, subreporty a mnoho dalšího. Uživatel s tímto nástrojem může přistupovat ke zdrojům pomocí JDBC, Hibernate, CSV souboru, JavaBeans, XML atd. Následně může poté publikovat výsledek v široké škále formátů jako jsou např. PDF, RTF, XML, XHTML, DOCX nebo OpenOffice. Druhým reportovacím nástrojem je Jaspersoft iReport, který poskytuje podobné možnosti jako Stu-

Obrázek 3.3: Gartner Magic Quadrant za rok 2014 zobrazující lídry BI (Zdroj: Google)

dio, ale je založen na programu NetBeans.

3.5.3 Pentaho Business Analytics Platform (Community Edition)

Reportovací nástroj Pentaho Reporting nabízí společnost Pentaho v rámci svého open source balíku Pentaho Business Analytics Platform. Díky tomu, že tento nástroj podporuje mnoho formátů, uživatel má tak možnost vytvářet reporty nad daty, které pocházejí např. z RDBMS, XML, ale také mnoho dalších formátů. Následně si uživatel může vybrat, v jakém formátu bude jeho report uložen (HTML dokument, MS Office Excel, PDF soubor, či prostý text). Samotný návrh výsledné podoby reportu má možnost uživatel vytvořit



Obrázek 3.4: Architektura BI platformy od společnosti MicroStrategy (Zdroj: Google)

Obrázek 3.5: Architektura BI platformy od společnosti Jaspersoft (Zdroj: Google)

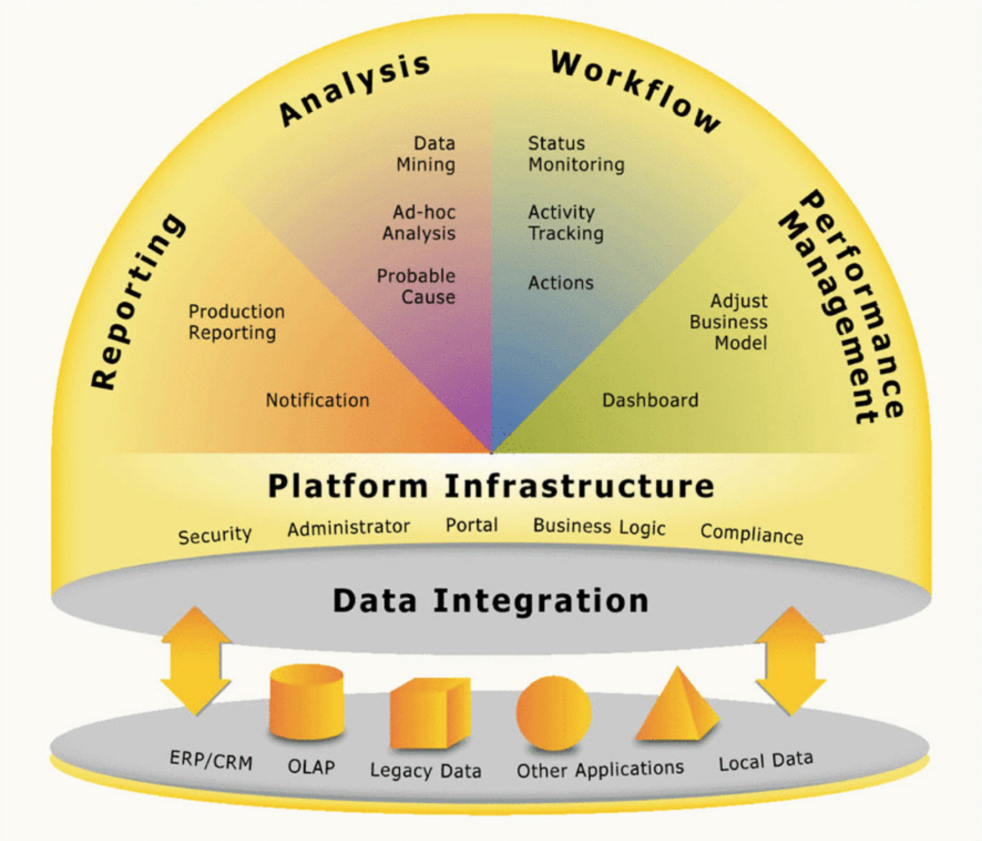
pomocí Pentaho Report Designer. Architekturu řešení společnosti Pentaho je možní vidět na obrázku 3.6

V následujících tabulkách jsou zhodnoceny vybraná řešení podle metrik určených v této kapitole. V tabulkách se vyskytují tři značky (x znamená, že danou vlastnost podporuje, - znamená, že danou vlastnost nepodporuje, ? znamená, že nebylo zjištěno, jestli danou vlastnost podporuje či nikoliv).

BI řešení	Pixel-perfect report	Ad-hoc dotazování	Integrace s MS-Office	Integrace s mobilní BI
MicroStrategy	x	x	x	x
Jaspersoft	x	x	x	x
Pentaho	x	x	x	x

Tabulka 3.1: Srovnání řešení v oblasti doručení informace (Information delivery)

Jak je patrné z jednotlivých srovnání řešení nabízené firmou MicroStrategy podporuje všechny zkoumané metriky. V rámci školního řešení připadají v úvahu pouze nekomerční řešení a v těch má navrch řešení od firmy Pentaho.



Obrázek 3.6: Architektura BI platformy od společnosti Pentaho (Zdroj: Google)

BI řešení	Historizace	Search-based data discovery	Pokročilá analýza	Olap analýza
MicroStrategy	x	x	x	x
Jaspersoft	x	?	x	x
Pentaho	x	x	x	x

Tabulka 3.2: Srovnání řešení v oblasti schopností a možnosti analýzy

Dalším důvodem, proč je doporučována zrovna tato platforma, je snadná instalace prostředí (instalace Jaspersoft řešení nebyla ani po opakovaných pokusech úspěšně dokončena na notebooku Macbook Pro (early 2013))

BI řešení	Dashboard	Scorecard
MicroStrategy	x	x
Jaspersoft	x	-
Pentaho	x	x

Tabulka 3.3: Srování řešení v oblasti vizualizace dat

BI řešení	Jeden přístupový bod do administrace	Vývojářské nástroje	Spolupráce	Podpora Big Data
MicroStrategy	x	x	x	x
Jaspersoft	x	?	x	x
Pentaho	x	-	-	x

Tabulka 3.4: Srovnání řešení v oblasti integrace

Kapitola 4

Cíl práce a specifikace požadavků implementační části

V následující kapitole specifikovány jednotlivé cíle této diplomové práce a požadavky na implementované řešení.

4.1 Specifikace cíle

Cílem diplomové práce je navrhnutí a implementování datového skladu nad systémem školní ankety. V první fázi dojde k analýze stávajícího řešení. Poté bude následovat návrh komponent datového skladu (ETL, EDW, DM) a jejich následná implementace za použití databáze PostgreSQL a softwaru Pentaho Data Integration.

4.2 Požadavky na výsledné řešení diplomové práce

Po konzultaci s Ing. Michalem Valentou, Ph.D byl sestaven následující seznam požadavků na výsledné řešení.

4.2.1 Vytvoření centrálního datového skladu

V rámci splnění tohoto požadavku by mělo dojít k vytvoření centrálního datového skladu (Enterprise Data Warehouse), pomocí kterého dojde ke sjednocení významu jednotlivých entit. Dalším důvodem pro použití tohoto řešení je snazší začlenění do celkového řešení v rámci projektu IP „Datová čistota“.

Zároveň tento datový sklad bude sloužit jako zdrojová databáze pro datová tržiště.

4.2.2 Vytvoření datového tržiště hodnocení učitelů

Toto datové tržiště bude sloužit k zobrazení výsledků hodnocení učitelů na základě odevzdaných anketních lístků na konci každého semestru. Uživatel, který bude mít možnost s tímto tržištěm pracovat by měl být schopný vytvořit report sestávající se z hodnocení vybraných učitelů a jejich hodnocení na základě anketních otázek položených v rámci jednotlivých anket.

4.2.3 Vytvoření datového tržiště hodnocení předmětů

Toto datové tržiště podobně jako hodnocení učitelů by mělo uživateli poskytnout možnost sestavit report sestávající se z otázek a příslušných odpovědí. Zároveň bude možnost vytvářet agregovaná hodnocení v rámci celých kateder či fakult.

4.2.4 Vytvoření datového tržiště statistik předmětů

Toto datové tržiště bude sloužit k analýze statistik jednotlivých předmětů jako je např statistika úspěšnosti dokončení předmětu nebo procentuální úspěšnost dokončení předmětu. Díky možnosti agregace předmětů do kateder a fakult bude možné určit procentuální úspěšnost studentů agregovaných přes katedry či fakulty.

4.2.5 Vytvoření možnosti analýzy entit datového skladu v průběhu času

Tento požadavek je jednou ze základních vlastností datových skladů. Umožní uživateli analyzovat hodnocení učitelů a předmětů v průběhu času tzn. uživatel bude moci sestavit report zobrazující hodnocení v průběhu několika semestrů.

4.2.6 Vytvoření řešení, které umožní jednoduchou správu i pro budoucí uživatele datového skladu

Tento požadavek vznikl ze strany autora práce, který povede v budoucnu ke snazší správě řešení vzniklého v rámci této práce. V tomto požadavku je zahrnuta podrobná a smysluplná dokumentace, která zaručí snadné zorientování se ve výsledném řešení. K naplnění tohoto požadavku by měl také přispět výběr vhodného reportovacího nástroje, který umožní pracovat s datovým skladem i lidem neseznámeným s touto problematikou.

Kapitola 5

Analýza současného řešení

V rámci této kapitoly bude provedena analýza současného řešení datového skladu podporující aplikaci Anketa ČVUT dostupnou na adrese `https://anketa.cvut.cz`

Současné řešení celé aplikace Ankety ČVUT je zpracováno v následujících bakalářských a diplomových pracích:

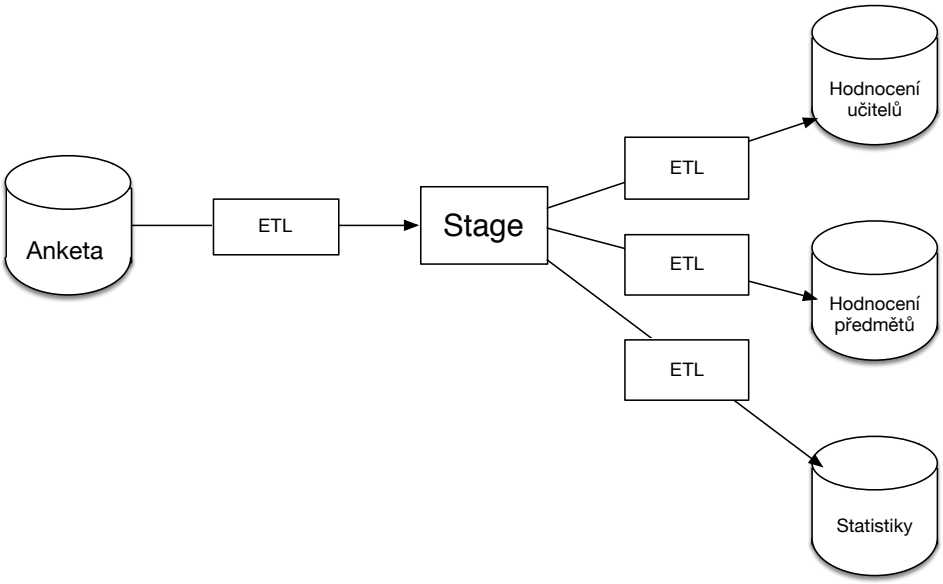
- Diplomová práce - Návrh a implementace OLAP prostředí nad archivy výsledků studentské ankety ČVUT, Bc. Jan Zámyslický[3]
- Bakalářská práce - Datový sklad nad výsledky ankety - nasazení, Pavlína Topinková[4]
- Bakalářská práce - Anketa ČVUT - refaktoring a rozšíření, Jan Stadler[5]
- Diplomová práce - Anketa ČVUT - přístup k aktuálním i historickým výsledkům anket, Bc. Pavlína Topinková[6]

Analýza popsána v této kapitole se především věnuje Diplomové práci Bc. Jana Zámyslického, Bakalářské a Diplomové práci Bc. Pavlíny Topinkové.

5.1 Architektura současného datového skladu

Na obrázku 5.1 je možné vidět architekturu současného řešení. Tato architektura se označuje *Independent Data Marts*[7], což se dá volně přeložit jako „nezávislá datová tržiště“ nebo také *Dimensional Data Store*[8].

Toto řešení spočívá v tom, že z odkládací části, která je v tomto případě tvořena exporty ze zdrojové databáze ankety, se vytvoří za pomoci ETL skriptů jednotlivá datová tržiště, která se dále používají k analýze získaných dat. Použití této architektury je vhodné, pokud řešení obsahuje pouze několik datových tržišť, čímž tímto řešení vskutku je. Nevýhodou této architektury je, že data v každém datovém tržišti se musí spravovat odděleně. Jelikož toto řešení je součástí projektu, jehož úkolem je vytvořit centralizovaný datový sklad

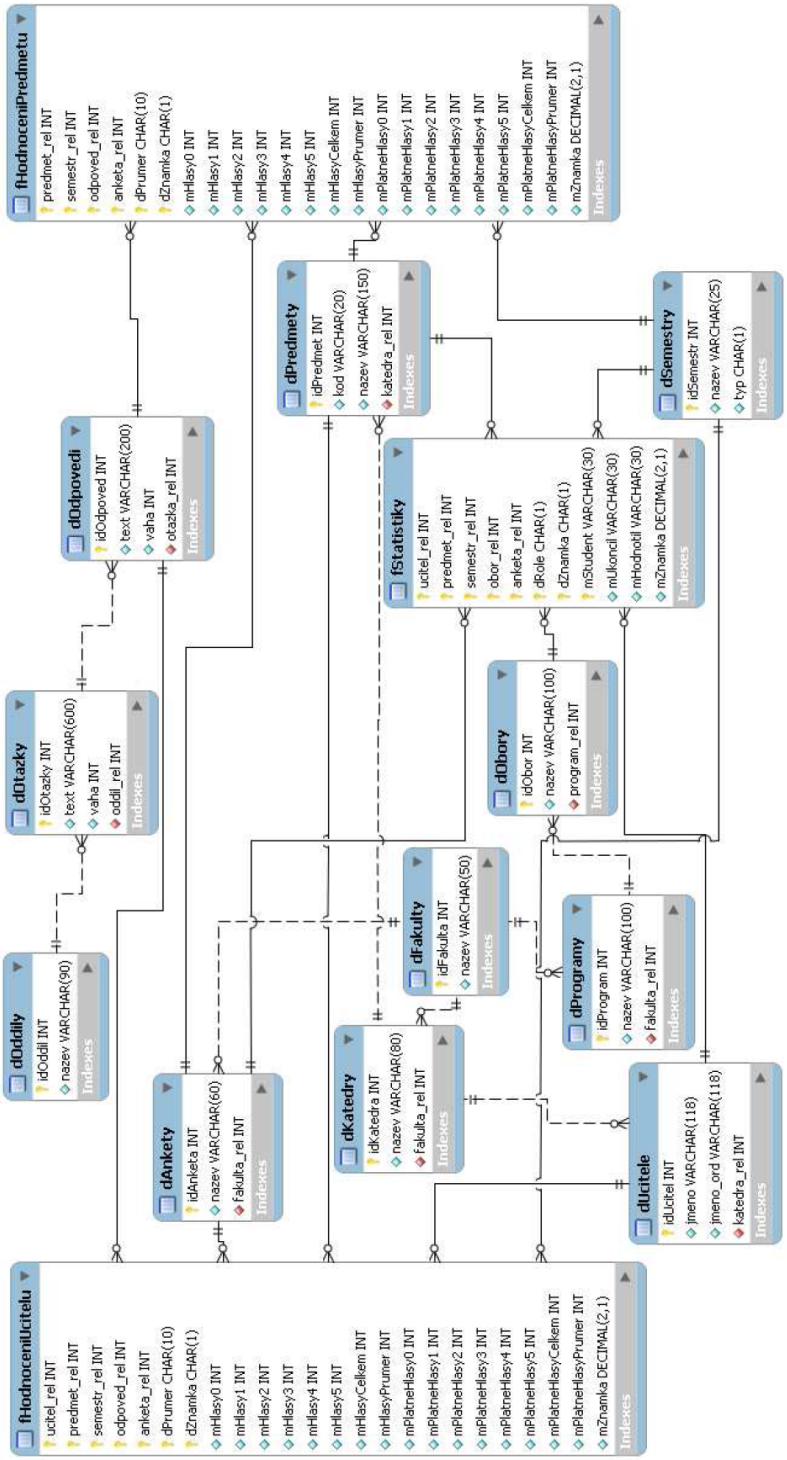


Obrázek 5.1: Diagram architektury současného řešení

nad informačními systémy ČVUT bude vybrána architektura *Dependent Data Marts*[7], což se dá volně přeložit jako „závislá datová tržiště“. Podrobnější popis této architektury se nachází v kapitole 6.

5.2 Datové schéma současného datového skladu

Na obrázku 5.2 je znázorněn entitní model datového schéma současného řešení. Entity začínající „f“ značí faktové tabulky a entity začínající „d“ značí dimenzionální tabulky. Jak je patrné z diagramu, výsledné řešení je navrhnuto v normalizované formě, která má za následek, delší časovou prodlevu výsledků na dotazy nad jednotlivými datovými tržišti. Řešení tržišť v této práci je v denormalizované formě, což by mělo přispět k rychlejším výsledkům dotazů, viz obrázky 6.15 nebo 6.17



Obrázek 5.2: Datový model současného řešení

Kapitola 6

Návrh a realizace nového řešení

V této kapitole bude podrobně popsána zdrojová databáze aplikace Anketa a následně zdokumentovány návrhy a implementace datového skladu.

6.1 Použité technologie

V rámci této podkapitoly budou postupně představeny všechny použité technologie při implementaci a následné práci s datovým skladem.

6.1.1 Zdrojový systém aplikace Anketa

Databáze podporující aplikaci Anketa využívá databázi Oracle.

6.1.2 Odkládací část pro exportované soubory zdrojového systému

V tomto kroku je prozatím využíván standardní file systém prostředí MAC OS X, ale lze tento file system nahradit za jakýkoliv jiný, či jakoukoliv databázi, kterou podporuje aplikace vybrána k vytvoření ETL skriptů.

6.1.3 Nástroj na vytváření ETL skriptů

Nástroj pro vytváření ETL skriptů, který bude použit pro tvorbu jednotlivých částí datového skladu, je alfou a omegou celého řešení. Hlavním požadavkem byla schopnost připojení k Oracle a PostgreSQL databázi, na které poběží výsledný datový sklad.

Jak bylo zmíněno v kapitole 3 na samotnou analýzu a následný reporting bude použita platforma od společnosti Pentaho a tak se logicky nabízelo využití programu **Pentaho Data Integration** (Kettle), jež má velmi intuitivní grafické prostředí s širokou škálou prvků usnadňující vytváření ETL skriptů. Tento program je napsán v Javě a tudíž by neměl být problém pro pozdějšího správce tohoto datového skladu upravovat a vytvářet nové skripty. Navíc

okolo tohoto programu je vytvořená velká komunita lidí, kterou doplňuje velmi dobře popsaná dokumentace [9].

Základními prvky tohoto programu jsou takzvané joby, které sdružují transformace. V těchto transformacích se provádí vytváří samotné ETL skripty.

6.1.4 Databáze použitá pro výsledné řešení datového skladu

Jak již bylo popsáno v kapitole popisující cíl této práce, výsledné řešení by mělo využívat databázi PopstgreSQL, která bude použita pro celkové řešení velkého projektu, jehož je součástí tato diplomová práce.

6.1.5 Nástroj na analýzu dat a vytváření reportů

Nástroj, který bude použit na analýzu a následné vytváření reportů bude z nabídky BI platformy firmy Pentaho. Server, na kterém poběží OLAP řešení se názývá **Mondrian**[10] a jedná se o server napsaný v Javě, který se nachází v základním balíku služeb poskytovaných BI řešením. Tento server využívá jazyk MDX, který umožňuje uživateli dotazovat se na data uložená v datových tržištích.

Pro analýzu nad daty uloženými v datových tržištích je vhodné mít na BI serveru nainstalovaný open source plugin Saiku Analytics[11], který umožňuje uživateli analyzovat a vytvářet analýzy bez jakékoliv znalosti jazyka MDX. Vytvořené analýzy se dále mohou použít do dalšího pluginu BI serveru Community Dashboard Editor, který umožňuje vytvářet dynamické reporty s možností interakce s koncovým uživatelem vytvořených reportů.

6.2 Návrh architektury datového skladu

Jak již bylo zmíněno v kapitole 4 řešení této práce by mělo obsahovat centrální datový sklad. Architektura obsahující toto řešení se nazývá *Centralized Data Warehouse Architecture*, což se dá volně přeložit jako „architektura centralizovaného datového skladu“.

Další důvod, který přispěl k výběru tohoto typu architektury,je popsán ve srovnávací studii[12] nejpoužívanějších architektur datových skladů, kde se autoři dotazovali respondentů z více než 400 různě velkých firem na řešení datových skladů a jejich úspěchy. Respondenti se skládali z různých pracovnách pozic (dwh manažeři, IS manažeři, dwh vývojáři, konzultanti). Jak je patrné z tabulky 6.1 (vyšší číslo znamená lepší výsledek) nejhůře dopadla architektura nezávislých datových tržišť. Naopak architektury centralizovaného datového skladu a Hub and spoke[1] dopadli velmi dobře.

Jednotlivé architektury byly srovnávány ve čtyřech vlastnostech:

[1]Hub and Spoke - architektura, kde jsou všechna data nejdříve shromážděna a poté přetransformována do přístupové vrstvy

Metrika	Nezávislá datová tržiště	Hub and spoke	Centralizovaný datový sklad
Informační kvalita	4.42	5.35	5.23
Systémová kvalita	4.59	5.56	5.41
Dopad na uživatele	5.08	5.62	5.64
Dopad na organizaci	4.66	5.24	5.30

Tabulka 6.1: Srovnání jednotlivých architektur datového skladu

- Informační kvalita - značí přesnost informací, jejich kompletnost a konzistentnost
- Systémová kvalita - zahrnuje flexibilitu systému, škálovatelnost a integraci
- Dopad na uživatele - tato vlastnost značí, jak rychle a snadno mohou uživatelé přistupovat k datům
- Dopad na organizaci - z hlediska byznys požadavků, využití business inteligence[2], podpora dosáhnutí byznys strategií, vylepšení komunikace a spolupráce napříč organizací

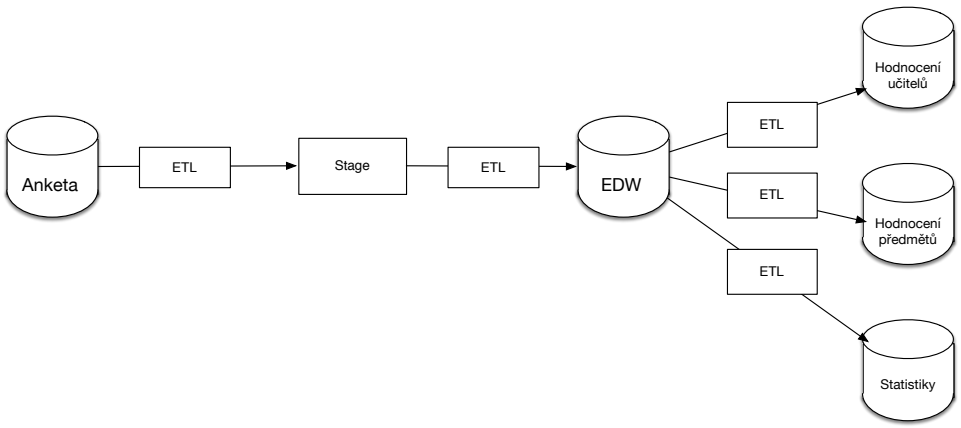
Na obrázku 6.1 je znázorněna architektura řešení. Oproti stávajícímu řešení se liší přidáním centrálního datového skladu z něhož se poté pomocí ETL skriptů vytváří jednotlivá datová tržiště, z kterých jsou dále vytvořeny datové kostky. Pomocí těchto kostek poté koncový uživatel provádí analýzy dat.

V první fázi dojde pomocí ETL skriptů k vyexportování jednotlivých tabulek z databáze aplikace Anketa do souborů typu csv. V další fázi dojde pomocí ETL skriptů k transformaci a nahrání záznamů do centrálního datového skladu. V poslední fázi dojde pomocí ETL skriptů k nahrání dat do jednotlivých datových tržišť. Během ETL skriptů jsou původní data přetransformována do takové podoby, aby vyhovovala zásadám jednotlivých části datového skladu.

6.3 Popis zdrojové databáze Anketa

V této sekci vycházím především z bakalářské práce Lukáše Frélicha[14], který se ve své práci zabýval modelem databáze školní ankety, a také z diplomové

[2]Business intelligence[13] (BI) jsou dovednosti, znalosti, technologie, aplikace, kvalita, rizika, bezpečnostní otázky a postupy používané v podnikání pro získání lepšího pochopení chování na trhu a obchodních souvislostech



Obrázek 6.1: Diagram navrhnuté architektury řešení datového skladu

práce Jana Zámyslického, který ve své práci využíval skoro tytéž tabulky databáze. Na obrázku 6.2 je zobrazeno datové schéma části ankety, která poté sloužila jako zdrojový soubor pro celý datový sklad. Tento diagram je pouze výřez z celkové databáze aplikace Anketa, která je mnohem složitější. Popis atributů jednotlivých tabulek zdrojové databáze lze nalézt v příloze B.

V diagramu se nachází několik typů tabulek:

- prefix „t“ značí tabulku databáze
- prefix „s“ značí materializovaný pohled typu snapshot
- prefix „v“ značí pohled

6.3.1 Materializovaný pohled *sKatedra*

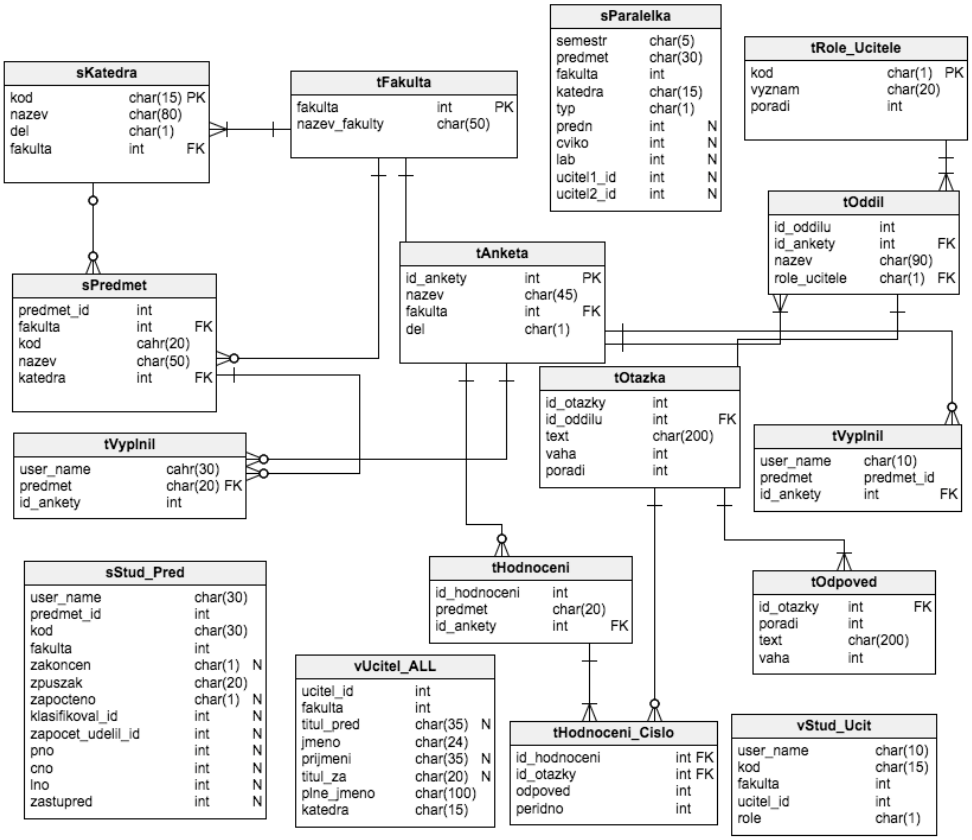
Tento pohled představuje seznam kateder, které náleží pod jednotlivé fakulty. Popis a význam jednotlivých sloupců tohoto materializovaného pohledu se nachází v tabulce B.1.

6.3.2 Materializovaný pohled *sParalelka*

Tento pohled představuje seznam částí předmětů (cvičení, laboratoř, přednáška) a vyučující, kteří danou část předmětu vyučují. Popis jednotlivých atributů tohoto pohledu, lze nalézt v tabulce B.2.

6.3.3 Materializovaný pohled *sPredmet*

Jentolivé předměty jsou vyučovány katedrami ČVUT, které patří pod fakulty. Informace o jednotlivých předmětech vyučovaných na ČVUT jsou uloženy



Obrázek 6.2: Datové schéma aplikace Anketa

v materializovaném pohledu popsaném v tabulce B.3. Tento materializovaný pohled vznikl sjednocením číselníků KOSu, FEL a FSI. Primárním klíčem tabulky *sPredmet* je dvojice (predmet_id, fakulta), což připouští možnost že na dvou fakultách se vyučuje předmět se stejným parametrem predmet_id. V rámci této tabulky by mělo zůstat zachováno pravidlo, že v rámci jedné katedry jsou názvy a kódy předmětu unikátní.

6.3.4 Materializovaný pohled *sStud_Pred*

V materializovaném pohledu *sStud_Pred* popsaném v tabulce B.4 je uložena informace o tom, jaký student si zapsal jaký předmět a jestli ho úspěšně dokončil či nikoliv. Primárním klíčem této tabulky je trojice (user_name, kod, fakulta).

6.3.5 Tabulka *tAnketa*

Tabulka *tAnketa* představuje seznam anket, které probíhají v rámci jednoho semestru na jednotlivých fakultách. Popis tabulky *tAnketa* lze nalézt v tabulce B.5.

6.3.6 Tabulka *tFakulta*

Tato tabulka představuje seznam jednotlivých fakult, jejíž parametry jsou popsané v tabulce B.6.

6.3.7 Tabulka *tHodnoceni*

V rámci jednotlivých anket mají studenti možnost vyplnit anketní lístky. Každý takovýto anketní lístek odpovídá hodnocení daného předmětu, které se skládá ze sady otázek, z nichž některé se vztahují k hodnocení učitele a některé k hodnocení samotného předmětu. Tyto anketní lístky jsou uloženy v tabulce *tHodnoceni* jejíž popis lze nalézt v tabulce B.7.

6.3.8 Tabulka *tHodnoceni_Cislo*

V rámci anketních lístku mají studenti možnost odpovídat na předem definovanou sadu otázek. Tyto odpovědi se zaznamenávají do tabulky *tHodnoceni_Cislo*. Detail této tabulky lze najít v tabulce B.8.

6.3.9 Tabulka *tOddil*

V rámci každé ankety, jsou jednotlivé anketní lístky rozděleny do oddílů. V rámci tohoto oddílu může potom student hodnotit učitele nebo předmět v roli, v které daný učitel vyučoval (zkoušející, cvičící apod.) Seznam oddílu je uložen v tabulce *tOddil*, jejíž detail je popsán v tabulce B.9.

6.3.10 Tabulka *tOdpoved*

Většina otázek v rámci aplikace Anketa má předdefinovanou sadu odpovědí, z kterých si může student vybrat. Tyto odpovědi se zaznamenávají v tabulce *tOdpoved*, jejíž detaily lze najíz níže v tabulce B.10.

Důležitým parametrem jednotlivých odpovědí jsou jejich váhy, jelikož určují o jak kladnou/zápornou odpověď se jedná. Váhy můžou nabývat hodnoty od 0 do 6, kdy váhy 0 a 6 jsou neutrální a váha 1 je nejkladnější a naopak váha 5 nejzápornější. Díky číselnému vyjádření odpovědí lze poté provádět analýzy nad odevzdanými hlasy.

6.3.11 Tabulka *tOtazka*

V rámci jednotlivých oddílu ankety jsou na anketním lístku položena sada otázek, na které může student odpovědět a tím ohodnotit daný předmět či daného učitele. Tyto otázky jsou zapsány v tabulce *tOtazka* jejíž popis udává tabulka B.11. V jednotlivých anketách se vyskytují dva druhy otázek. Jedny otázky se berou jako oficiální hodnocení předmětu/učitele a ostatní jsou otázky doplňovací.

6.3.12 Tabulka *tRole_Ucitele*

V rámci oddílu může student vyplnit sadu otázek příslušející k určité roli učitele či předmětu (cvičící, zkoušející, přednášející apod.) Popis jednotlivých atributů tabulky je popsán v tabulce B.12.

6.3.13 Tabulka *tVyplnil*

V databázi aplikace Anketa se nachází tabulka *tVyplnil*, do které se ukládají záznamy o studentech, kteří vyplnili anketní lístek daného předmětu v rámci konkrétní ankety. K identifikaci záznamu je použita trojice (user_name, predmet, id_ankety). Popis atributů této tabulky se nachází v tabulce B.13.

6.3.14 Pohled *vUcitel_All*

Pohled *vUcitel_All* vznikl sjednocením materializovaného pohledu *sUcitel*, který obsahuje informace o učitelích přebírané z KOSu a tabulky *tUcitel_local*, která obsahuje informace o učitelích, kteří se nenacházejí v KOSu. Tento pohled je pospán v tabulce B.14.

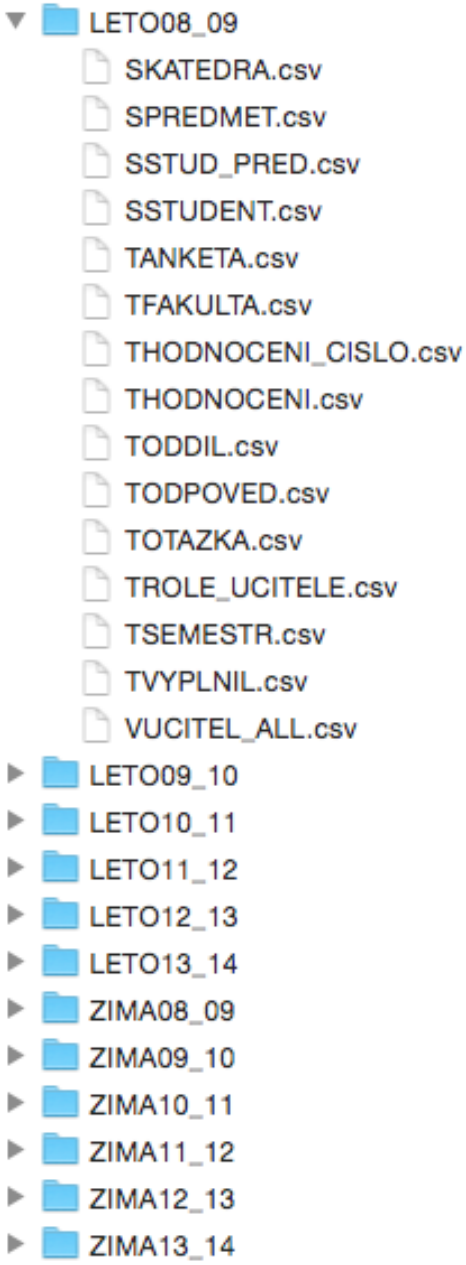
6.3.15 Pohled *vStud_Ucit*

Pohled *vStud_Ucit* představuje seznam, kde lze nalézt vztah student-učitel. Tento vztah se váže k předmětu a roli, v které daný učitel v předmětu figuroval. Popis atributů tohoto pohledu lze nalézt v tabulce B.15 .

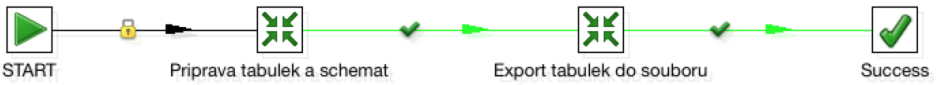
6.4 Návrh a implementace odkládací části

6.4.1 Návrh file systému použitého v odkládací části

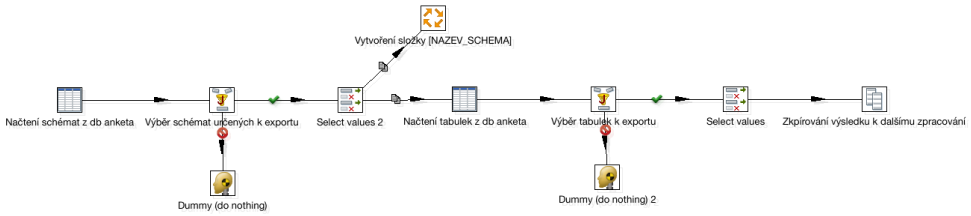
Stage neboli *odkládací část* má podobu souborů typu csv. Tento file system zachovává strukturu, která je ve zdrojové databázi. Strukturu tohoto file systemu je možné vidět na obrázku 6.3. Názvy jednotlivých složek, ve kterých jsou extrahované soubory odpovídají profilům zdrojové databáze. Tyto profily kopírují názvy semestrů, v rámci kterých jsou vytvářené tabulky z podkapitoly 6.3 (ZIMA08_09, LETO08_09 atd.).



Obrázek 6.3: Struktura odkládací části



Obrázek 6.4: ETL skript vytvářející odkládací část datového skladu



Obrázek 6.5: Přípravná fáze ETL skriptu k vyexportování tabulek ze zdrojové databáze do csv souborů

Obrázek 6.6: Finální fáze ETL skript, který provede samotný export tabulek do příslušných csv souborů

6.4.2 ETL skripty na vytvoření odkládací části ze zdrojové databáze

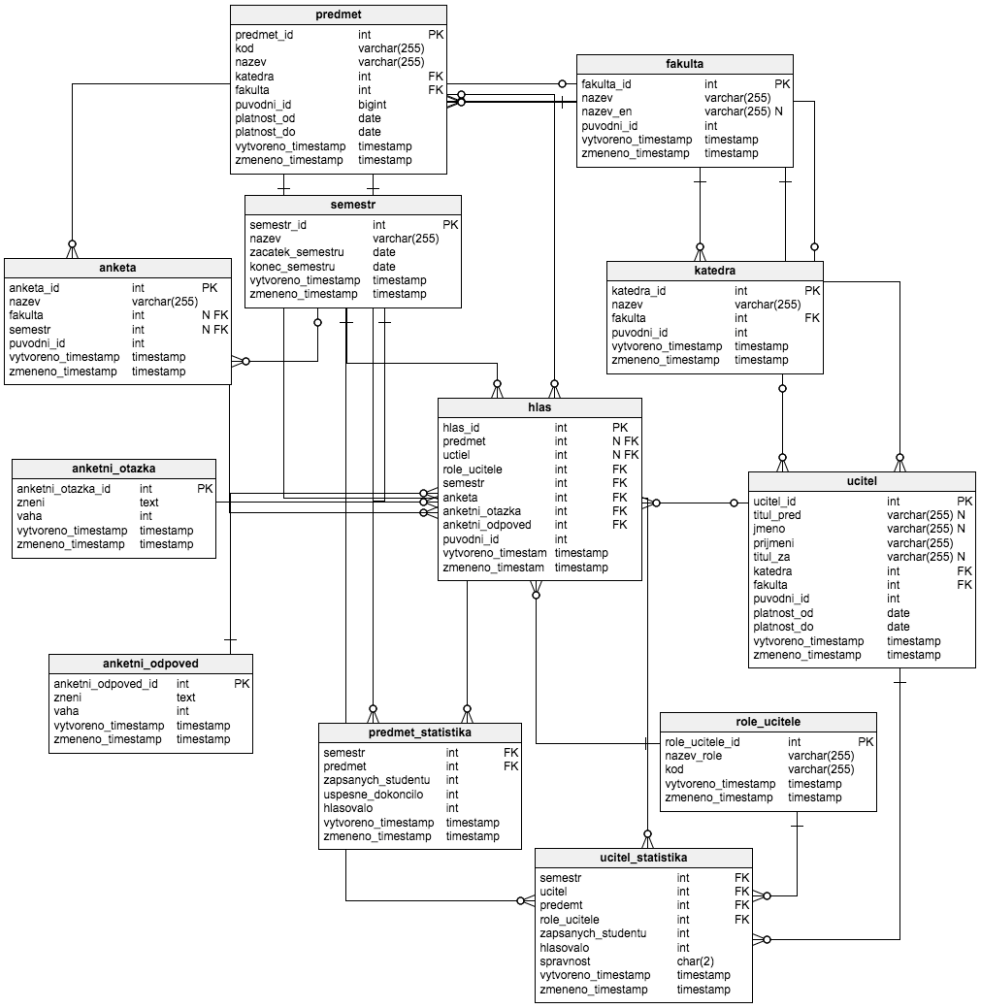
ETL skript, který vytvoří csv soubor z příslušné tabulky, je možné vidět na obrázku 6.4.

V první fázi dojde k vybrání požadovaného schéma, které ETL skript dostane zadaný jako parametr a poté v k vyexportování předem určených tabulek. V této fázi je důležité aby tabulky zdrojové databáze zachovávaly jmennou konvenci. První fázi celého skriptu je možné vidět na obrázku 6.5.

Ve druhé fázi poté dojde k samotnému exportu tabulek do csv souborů a pojmenování vzniklých souborů tak, aby je bylo možné dále použít v ETL skriptech k naplnění centrálního datového skladu daty. Tuto fázi je možné vidět na obrázku 6.6.

6.5 Návrh a realizace centrálního datového skladu

V rámci této podkapitoly bude ukázán návrh centrálního datového skladu (EDW) a také modely několika ETL skriptů, které jsou prováděny k exportování, transformaci a nahrání dat z odkládací části do centrálního datového



Obrázek 6.7: Datové model centrálního datového skladu

skladu. Všechny ETL skripty je možné nalézt v příloze D.

6.5.1 Návrh datového schéma centrálního skladu

Jak již bylo zmíněno v kapitole 1 centrální datový sklad by měl být navrhnut ve třetí normální formě. Jeho návrh je možné vidět na obrázku 6.7.

Většina tabulek kopíruje svými parametry tabulky zdrojové databáze a proto zde budou popsány pouze ty, které se výrazně odlišují či úplně chybí ve zdrojové databázi. (úplnou specifikaci všech tabulek lze nalézt v příloze C).

6.5.2 Atributy *vytvoreno_timestamp, zmeneno_timestamp*

Ve všech tabulkách centrálního datového skladu se nacházejí dva zcela nové atributy:

- vytvoreno_timestamp
- zmeneno_timestamp

Tyto dva atributy představují časový otisk, kdy byly do skladu nahrány. Navíc atribut *zmeneno_timestamp* má nastavený trigger, který při změně záznamu v tabulce přepíše časový otisk na čas, kdy byl tento záznam změněn. Výchozí hodnota parametru *zmeneno_timestamp* je nastavena na časový otisk vytvoření záznamu v tabulce.

6.5.3 Atribut *puvodni_id*

V některých tabulkách se vyskytuje nový atribut *puvodni_id*, který odkazuje na primární klíč záznamu ve zdrojové databázi. Tímto opatřením se zabrání tomu, aby se při změně atributů záznamu v původní databázi nenačetl do datového skladu nový záznam, ale došlo pouze k aktualizaci všech parametrů spojených s tímto primárním klíčem.

6.5.4 Atributy *platnost_od, platnost_do*

V tabulkách *predmet* a *ucitel* byly vytvořeny dva zcela nové atribty:

- platnost_od
- platnost_do

Tyto atributy byly vytvořeny proto, aby byla splněna jedna ze základních vlastností datových skladů - historizace. Jelikož může dojít k jevu, kdy učitel/ka si změní příjmení nebo získá nový titul během semestrů, a poté je potřeba tuto skutečnost promítnout do tabulky *ucitel*, kde vznikne pro učitele se stejnými identifikačními atributy nový záznam. K tomu slouží výše zmíněné atributy, které říkají od jakého a do jakého data je daný záznam platný a díky tomu je v rámci hodnocení přiřazen správný učitel. Je důležité zmínit že zimní i letní semestr mají pevně nastavené měsíce v rámci jednoho akademického roku - zimní semestr (září-březen) a letní semestr (duben-srpen).

U předmětu je toto opatřeni proto, aby došlo ke změně názvu předmětu při jeho změně ve zdrojové databázi.

6.5.5 Tabulka *hlas*

Záznam v tabulce představuje jeden odevzdaný hlas v rámci fakultní ankety, která probíhá na konci každého semestru. Atributy této tabulky lze vidět v tabulce 6.2.

Název sloupce	Popis
hlas_id	Unikátní identifikátor záznamu v tabulce hlas.
predmet	Cizí klíč do tabulky *predmet*, který určuje předmět, v kterém bylo hlasováno.
ucitel	Cizí klíč do tabulky *ucitel*, který určuje učitele, který byl daným hlasem hodnocen. Tato hodnota je NULL pokud se jedná o hodnocení předmětu.
role_ucitele	Cizí klíč do tabulky *role_ucitele*, který určuje v jaké roli byl daný učitel/předmět hodnocen.
anketa	Cizí klíč do tabulky *anketa*, který určuje v rámci které ankety, byl daný hlas odevzdán.
anketni_otazka	Cizí klíč do tabulky *anketni_otazka*, který určuje na jakou otázku bylo hlasem odpovídáno.
anketni_odpoved	Cizí klíč do tabulky *ankatni_odpoved*, který určuje jaká odpověd byla vybrána.
puvodni_id	Odkaz na konkrétní hlas ze zdrojové databáze v tabulce *thodnoceni_cislo*.
vytvoreno_timestamp	Časový otisk, kdy byl daný záznam vytvořen.
zmeneno_timestamp	Časový otisk, kdy byl daný záznam změněn.

Tabulka 6.2: Popis tabulky *hlas*

6.5.6 Tabulka *statistika_predmet*

Tato tabulka představuje statistiku o předmětech hodnocených v anketách. V této tabulce jsou uloženy záznamy o tom, kolik studentů se zapsalo na daný předmět a kolik jich úspěšně dokončilo daný předmět. Druhou statistikou je počet hlasujících studentů o daném předmětu. Z těchto údajů lze zjistit jakou průchodnost má daný předmět a také kolik procent ze zapsaných studentů hlasovalo v anketě o daném předmětu. Popis atributů této tabulky je možné vidět na obrázku 6.4.

6.5.7 Tabulka *predmet_statistika*

Tato tabulka představuje statistiku o předmětech hodnocených v anketách. V této tabulce jsou uloženy záznamy o tom, kolik studentů se zapsalo na daný předmět a kolik jich úspěšně dokončilo daný předmět. Druhou statistikou je počet hlasujících studentů o daném předmětu. Z těchto údajů lze zjistit jakou

Název sloupce	Popis
predmet	Cizí klíč ukazující do tabulky *predmet*, který identifikuje o jaký předmět se jedná.
semestr	Semestr, ve kterém byla daná statistika předmětu počítána.
zapsanych_studentu	Počet studentů zapsaných na předmět.
uspesne_dokoncilo	Počet studentů, kteří hlasovali o předmětu.
vytvoreno_timestamp	Časový otisk, kdy byl daný záznam vytvořen.
zmeneno_timestamp	Časový otisk, kdy byl daný záznam změněn.

Tabulka 6.3: Popis pohledu *statistika_predmet*

průchodnost má daný předmět a také kolik procent ze zapsaných studentů hlasovalo v anketě o daném předmětu. Popis atributů této tabulky je možné vidět na obrázku 6.4.

Název sloupce	Popis
predmet	Cizí klíč ukazující do tabulky *predmet*, který identifikuje o jaký předmět se jedná.
semestr	Semestr, ve kterém byla daná statistika předmětu počítána.
zapsanych_studentu	Počet studentů zapsaných na předmět.
uspesne_dokoncilo	Počet studentů, kteří hlasovali o předmětu.
vytvoreno_timestamp	Časový otisk, kdy byl daný záznam vytvořen.
zmeneno_timestamp	Časový otisk, kdy byl daný záznam změněn.

Tabulka 6.4: Popis pohledu *predmet_statistika*

6.5.8 Tabulka *ucitel_statistika*

Tato tabulka představuje statistiku o učitelích hodnocených v anketách. Jeden záznam v této tabulce představuje určitého vyučujícího předmětu v rámci konkrétní role. K této roli učitele v rámci předmětu se váže údaj o počtu zapsaných studentů a počtu studentů, kteří hodnotili daného učitele v dané roli. Jelikož nebylo snadné získat údaje o tom, kolik studentů přišlo do kontaktu s vyučujícím v konkrétní roli, byl zaveden příznak, který určuje o jak přesnou statistiku se jedná (sloupec *spravnost*).

Pokud byl údaj o zapsaných studentech zřejmý, má tento příznak hodnotu „A“. Pokud nebyl ve zdrojové databázi dohledán údaj o počtu zapsaných studentů ke konkrétní roli učitele v rámci předmětu, byl brán počet hodnotících studentů jako počet zapsaných. V tom případě má příznak hodnotu „N“. Po-

kud se anketní lístek sestával pouze z nepovinných odpovědí, nebylo možné určit, kolik přesně studentů odpovědělo k hodnocení daného učitele v dané roli. V tomto případě byl počet hodnotících brán jako maximum z odpovědí na jednotlivé otázky v anketním lístku (tento počet představuje minimální počet studentů, kteří hodnotili učitele) a příznak *spravnost* má hodnotu „NN“. Popis všech atributů tabulky *ucitel_statistika*, lze nalézt v tabulce 6.5. Primárním klíčem této tabulky je čtveřice tvořená sloupci (*semestr, ucitel, predmet, role_ucitele*).

Název sloupce	**Popis**
semestr	Semestr, ve kterém byla daná statistika pro učitele počítána.
predmet	Cizí klíč ukazující do tabulky *predmet*, který identifikuje předmět, který hodnotící učitel vyučoval.
ucitel	Cizí klíč ukazující do tabulky *ucitel*, který identifikuje konkrétního učitele, který byl hodnocen a pro kterého byla vytvořena statistika.
role_ucitele	Cizí klíč do tabulky *role_ucitele*, která určuje, ke které roli učitele se daná statistika vztahuje
zapsanych_studentu	Počet studentů zapsaných k dané roli učitele v rámci konkrétního předmětu.
hlasovalo	Počet studentů, kteří hodnotili daného učitele.
spravnost	Příznak, který určuje o jak přesnou statistiku se jedná.
vytvoreno_timestamp	Časový otisk, kdy byl daný záznam vytvořen.
zmeneno_timestamp	Časový otisk, kdy byl daný záznam změněn.

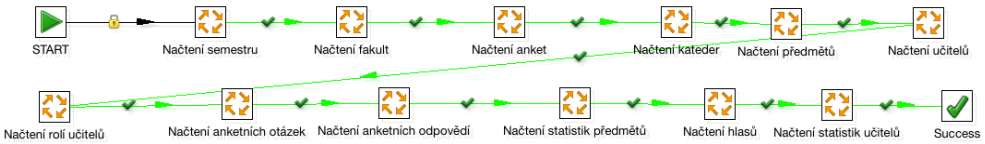
Tabulka 6.5: Popis pohledu *ucitel_statistika*

6.5.9 ETL skripty použité k naplnění datového skladu z odkládací části

ETL skript plnící centrální datový sklad z odkládací části probíhá dvoufázově, viz obrázek 6.8.

V první fázi dojde pouze k načtení textového souboru se seznamem složek, které se mají načíst do datového skladu. Tyto názvy složek odpovídají jednotlivým uživatelům ve zdrojové databázi (letní semestr 2008/2009 - LETO08_09, zimní semestr 2011/2012 - ZIMA11_12 atd.).

Obrázek 6.8: Dvoufázový ETL skript plnící centrální datový sklad z odkládací části



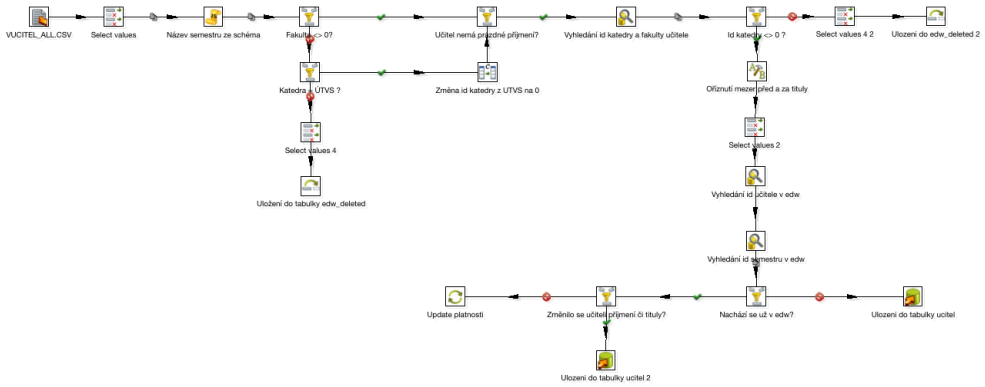
Obrázek 6.9: Skript, který nahrává tabulky z odkládací části do centrálního datového skladu v rámci jednoho semestru

Ve druhé fázi dojde ke zpracování samotného schéma. Schéma jsou zpracovávaná postupně, tzn pokud konfigurační soubor obsahuje deset semestrů, job *Zpracování schéma* proběhne desetkrát. Struktura této fáze je zobrazena na obrázku 6.9. Ve zdrojové databázi se nachází předměty, učitelé a fakulty, které ale patří pod katedru nebo fakultu, která se v původní databázi nenachází. Prozatímní řešení tohoto problému je takový, že dané záznamy jsou odkládány do speciální databáze, pro pozdější zpracování a úspěšnému nahrání do datového skladu. V důsledku toho dojde k nenahrání některých hlasů, které jsou ve zdrojové databázi.

V rámci jednotlivých podksriptů dochází k transformacím a nahrání dat z tabulek zdrojové databáze aplikace Anketa do tabulek centrálního datového skladu. Jednotlivé modely skriptů lze nalézt v příloze D. V této podkapitole budou pouze ukázány a popsány transformace nahrání učitelů ze souborů v odkládací části do tabulky *ucitel* v centrálním datovém skladu, viz obrázek 6.10.

Popis jednotlivých kroků transformace načtení učitelů do tabulky *ucitel*:

1. Načtení souboru *VUCITEL_ALL.csv*, který obsahuje seznam učitelů pro daný semestr.
2. Vytvoření názvu semestru z názvu schématu.
3. Rozhodnutí o tom, jestli fakulta není rovna 0.
   a) Pokud se fakulta rovná 0 a zároveň katedra je rovna „ÚTVS“ nebo „ÚTV“ je hodnota této katedry změněna na 0, která značí, že se jedná o Ústav tělesné výchovy a sportu. Tato „katedra“ je brána jako celoškolská instituce.
   b) Pokud se fakulta rovná 0 a katedra nerovná „ÚTVS“ ani „ÚTV“ je považována za neznámou a uložena do tabulky *ucitel* speciální



Obrázek 6.10: Transformace, při které dochází k nahrání učitelů do tabulky *ucitel*

databáze *edw_deleted*, která slouží pro pozdější opravy těchto záznamů.

4. Rozhodnutí o tom, jestli učitel má prázdné příjmení
   a) Pokud má prázdné příjmení, tak je pro nás v tuto chvíli nezajímavý a je uložen do tabulky *ucitel* databáze *edw_deleted*
5. Vyhledání id katedry a fakulty učitele v tabulce *katedra* centrálního datového skladu.
6. Rozhodnutí o tom, jestli tam katedra existuje
   a) Pokud neexistuje, je učitel uložen do tabulky již zmíněné výše.
7. Oříznutí mezer před a za tituly učitele.
8. Vyhledání id učitele a jeho atributů (jméno, příjmení, tituly,...) na základě jeho původního id v tabulce *ucitel* v centrálním datovém skladu.
9. Vyhledání id semestru v centrálním datovém skladu na základě jeho názvu.
10. Rozhodnutí o tom, jestli se daný učitel už v tabulce *ucitel* nachází.
    a) Pokud se v tabulce nenachází (nalezené id učitele je rovno NULL) dojde k jeho uložení do tabulky *ucitel*.
    b) Pokud se v tabulce již nachází dojde k rozhodnutí, jestli se všechny atributy nalezeného učitele shodují.
       i. Pokud se všechny atributy shodují, dojde k prodloužení platnosti tohoto záznamu o konečné datum aktuálního semestru.
       ii. Pokud se atributy neshodují, dojde k vytvoření nového záznamu s platností od začátku aktuálního semestru.

Obrázek 6.11: ETL skript nahrávající data do faktových a dimenzionálních tabulek datových tržišť

6.6 Návrh a realizace jednotlivých datových tržišť

V rámci této podkapitoly bude nejdříve popsán ETL skript (obrázek 6.11), kterým dojde ke spuštění jednotlivých ETL podskriptů na tvorbu dimenzionálních a faktových tabulek datových tržišť. Dále zde budou popsána datová schémata těchto tržišť.

Skript, který vytváří jednotlivé tabulky datových tržišť se spouští automaticky po nahrání centrálního datového skladu. Pro úspěšné dokončení tohoto skriptu musejí všechny dílčí skripty proběhnou bez chyby. Pokud se tak vyskytne chyba, uživatel je o tom dostatečně informován.

6.6.1 Datové tržiště sloužící k analýze statistik předmětů

V rámci tohoto datového tržiště může uživatel analyzovat statistiky předmětů (informace o vyplněnosti anketních lístků a průchodnosti daného předmětu). Granularitou faktové tabulky je záznam o předmětu tvořený počtem zapsaných studentů, počtem studentů, kteří úspěšně dokončili předmět a počet studentů, kteří hodnotili daný předmět v rámci fakultní ankety.

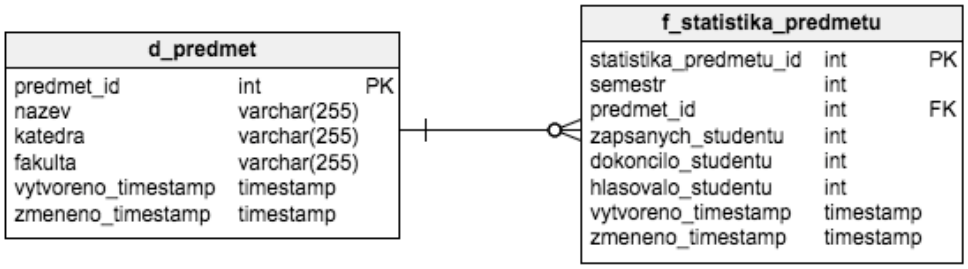
6.6.1.1 Datové schéma datového tržiště statistika předmětů

Datové schéma tohoto tržiště je velmi jednoduché, viz obrázek 6.12. Je tvořeno jednou dimenzionální tabulkou *d_predmet* a jednou faktovou tabulkou *f_statistika_predmetu* s jednou degenerovanou dimenzí - semestr. V rámci faktové tabulky musí platit, že dvojce (*semestr,predmet*) musí být unikátní. Popis atributů faktové tabulky, lze nalézt v tabulce 6.6.

Dimenzionální tabulka *d_predmet* představuje předmět, ke kterému se váže daná statistika o vyplněnosti ankety a jeho průchodnosti. Primárním klíčem tabulky je *predmet_id*, který je přebírán z centrálního datového skladu. Dále jako parametry této tabulky je název katedry a fakulty, kde se daný předmět vyučuje a název předmětu, který je tvořen kódem a názvem samotným.

Metriky faktové tabulky:

- zapsanych_studentu
- dokoncilo_studentu
- hlasovalo_studentu



Obrázek 6.12: Datový model datového tržiště statistika předmětů

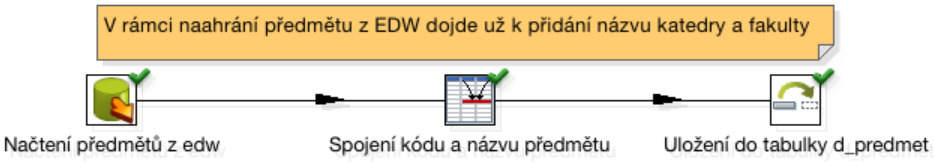
Název sloupce	Popis
statistika_predmetu_id	Unikátní identifikátor záznamu v tabulce.
semestr	Název semestru, ke kterému se statistika předmětu vztahuje.
predmet_id	Cizí klíč do tabulky *d_predmet*, který určuje předmět, kterému statistika přísluší.
zapsanych_studentu	Počet zapsaných studentů do předmětu.
dokoncilo_studentu	Počet studentů, kteří úspěšně dokončili předmět.
hlasovalo_studentu	Počet studentů, kteří hodnotili daný předmět.
vytvoreno_timestamp	Časový otisk, kdy byl daný záznam vytvořen.
zmeneno_timestamp	Časový otisk, kdy byl daný záznam změněn.

Tabulka 6.6: Popis tabulky *f_statistika_predmetu*

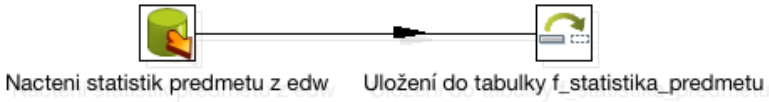
6.6.1.2 ETL skripty naplnění datového tržiště statistika předmětů daty

Pro úspěšné vytvoření datových tržišť je nutné, aby nejprve proběhl skript, v rámci kterého dojde k nahrání dat do dimenzní tabulky *d_predmet*. Tento skript je možné vidět na obrázku 6.13. Při této transformaci se nejdříve provede export záznamů z tabulky *predmet* v centrálním datovém skladu a následnému spojení kódu s názvem předmětu. V posledním kroku dojde do uložení záznamů do tabulky *d_predmet*.

Skript zobrazený na obrázku 6.14 nahrává data do faktové tabulky *f_statistiky_predmetu*. V rámci tohoto skriptu dojde pouze k extrahování záznamů z tabulky *predmet_statistika* v centrálním datovém skladu a uložení do fakotvé tabulky.



Obrázek 6.13: Skript, při kterém dojde k naplnění dimenzionální tabulky daty *d_predmet*



Obrázek 6.14: ETL skript, při kterém dojde k nahrání dat do faktové tabulky *f_statistika_predmetu*

6.6.2 Datové tržiště sloužící k analýze hodnocení předmětů

V rámci tohoto datového tržiště může uživatel provádět analýzy nad hodnoceními jednotlivých předmětů. Granularitou faktové tabulky je jeden hlas odevzdaný ve fakultní anketě vztahující se k hodnocení předmětů.

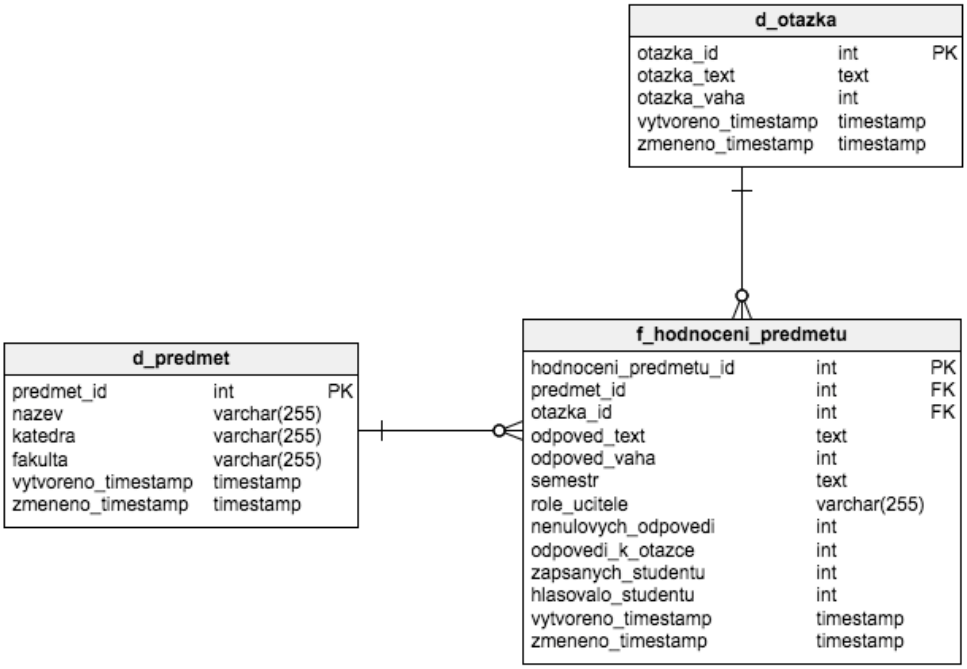
6.6.2.1 Datové schéma datového tržiště *hodnocení předmětů*

Jak je možné vidět na obrázku 6.15 toto datové tržiště se skládá z jedné faktové tabulky *f_hodnoceni_predmetu* a dvou dimenzionálních tabulek. První z nich je *d_predmet*, jejíž popis, lze nalézt v tabulce 6.7, a tabulkou *d_otazka*. Toto schéma dvou již zmíněných dimenzí ještě tři degenerované dimenze - dimenze odpovědi, semestru a role učitele.

Faktová tabulka *f_hodnoceni_predmetu* tvořící střed „star schema“ vzniká exportem záznamů z tabulky *hlas* z centrálního datového skladu. Popis atributů této tabulky lze nalézt v tabulce 6.8. Důležité je si povšimnout atributů *zapsanych_studentu* a *hodnotilo_studentu*. Tyto údaje se nemusí zdát na první pohled logické, ale díky nim je možné spočítat, kolik procent ze zapsaných studentů hodnotilo daný předmět. Následně je možné zjistit procento vyplněnosti anket v rámci kateder a fakult.

Metriky faktové tabulky:

- odpoved_vaha
- nenulovych_odpovedi
- celkovych_odpovedi
- zapsanych_studentu



Obrázek 6.15: Datový model datového tržiště *hodnocení předmětů*

Název sloupce	**Popis**
predmet_id	Unikátní identifikátor záznamu v tabulce. Tento identifikátor je přebírán z tabulky *predmet* v centrálním datovém skladu.
nazev	Spojený kód a název předmětu
katedra	Katedra, v rámci které je daný předmět vyučován.
fakulta	Fakulta, v rámci které je daný předmět vyučován.
vytvoreno_timestamp	Časový otisk, kdy byl daný záznam vytvořen.
zmeneno_timestamp	Časový otisk, kdy byl daný záznam změněn.

Tabulka 6.7: Popis dimenzionální tabulky *d_predmet*

- hlasovalo_studentu

V tomto datovém tržišti se objevuje nová dimenzionální tabulka *d_otazka*, která je sdílenou dimenzí jak pro datové tržiště *hodnocení předmětů*, tak pro datové tržiště *hodnocení učitelů*. Popis atributů lze nalézt v tabulce 6.9. Tato tabulka představuje seznam anketních otázek položených v rámci aplikace Anketa.

Název sloupce	Popis
hodnoceni_predmetu_id	Unikátní identifikátor záznamu v tabulce.
predmet_id	Cizí klíč do tabulky *d_predmet*, který určuje předmět, v kterém bylo hlasováno.
otazka_id	Cizí klíč do tabulky *d_otazka*, který určuje otázku, na kterou bylo odpovídáno.
odpoved_text	Textový popis odpovědi, která byla zvolena na zadanou odpověď. Jedná se o degenerovanou dimenzi.
odpoved_vaha	Číselná interpretace zvolené odpovědi.
semestr	Název semestru v rámci kterého bylo hlasováno. Jedná se o degenerovanou dimenzi.
role_ucitele	Název role učitele (v tomto případě se jedná o role typu cvičení, přednáška, hodnocení předmětu jako celku). Jedná se o degenerovanou dimenzi.
nenulova_odpoved	Příznak, který značí, jestli se jedná o nenulovou odpověď (1-ano, 2-ne).
odpovedi_k_otazce	Počet všech odpovědí odevzdaných v rámci jedné otázky v hodnocení předmětu. Tento údaj je zde proto, aby bylo možné získat informaci o tom, kolik procent ze všech hlasů k dané otázce hodnocení předmětu náleží konkrétní odpovědi.
zapsanych_studentu	Počet studentů zapsaných do hodnoceného předmětu.
hodnotilo_studentu	Počet studentů, kteří hodnotili daný předmět.
vytvoreno_timestamp	Časový otisk, kdy byl daný záznam vytvořen.
zmeneno_timestamp	Časový otisk, kdy byl daný záznam změněn.

Tabulka 6.8: Popis tabulky *f_hodnonceni_predmetu*

6.6.2.2 ETL skripty na naplnění datového tržiště hodnocení předmětů daty

Celé datové tržiště *hodnocení předmětů* je načteno v rámci velkého skriptu, který vytvoří všechny tři datová tržiště. Proces tohoto jobu je možné vidět na obrázku 6.11. Pro úspěšné dokončení ETL procesu, který vytvoří datové tržiště pro hodnocení předmětů je třeba, aby došlo k úspěšnému dokončení ETL skriptu nahrávající data do dimenzionálních tabulek *d_predmet* a *d_otazka* a úspěšnému vytvoření datového tržiště sloužícího k analýze statistik předmětů.

Název sloupce	Popis
otazka_id	Unikátní identifikátor záznamu v tabulce. Tento identifikátor je přebírán z tabulky *anketni_otazka* v centrálním datovém skladu.
otazka_text	Znění otázky.
otazka_vaha	Příznak značící, jestli se jedná o otázku počítanou do hodnocení předmětu či nikoliv (1-ano, 2-ne).
vytvoreno_timestamp	Časový otisk, kdy byl daný záznam vytvořen.
zmeneno_timestamp	Časový otisk, kdy byl daný záznam změněn.

Tabulka 6.9: Popis dimenzionální tabulky *d_otazka*

Načtení dimenzionální tabulky *d_predmet* probíhá v jobu *Dimenze předmět*, který provede transformaci *Načtení dimenze předmětu do Data Martu*. Model této transformace je zobrazen na obrázku 6.13.

Načtení dimenzionální tabulky *d_otazka* probíhá v transformaci s názvem *Načtení dimenze otázky do Data Martu*, která je součástí jobu *Dimenze otázka* v hlavním jobu na vytváření datových tržišť. V rámci této transformace je prováděno pouze extrahování záznamů z tabulky *anketni_otazka* z centrálního datového skladu a jejich uložení do dimenzionální tabulky.

Načtení faktové tabulky *f_hodnoceni_predmetu* probíhá v transformaci s názvem *Načtení faktové tabulky hodnocení předmětů do Data Martu*, která je součástí jobu *Faktová tabulka hodnocení předmětů*. Model této transformace je možné vidět na obrázku 6.16. Pro úspěšné dokončení této transformace je důležité mít úspěšně vytvořeno datové tržiště se statistikami o předmětech.

Popis kroků transformace načtení faktové tabulky*f_hodnoceni_predmetu*:

1. Načtení záznamů z tabulky *hlas* z centrálního datového skladu.

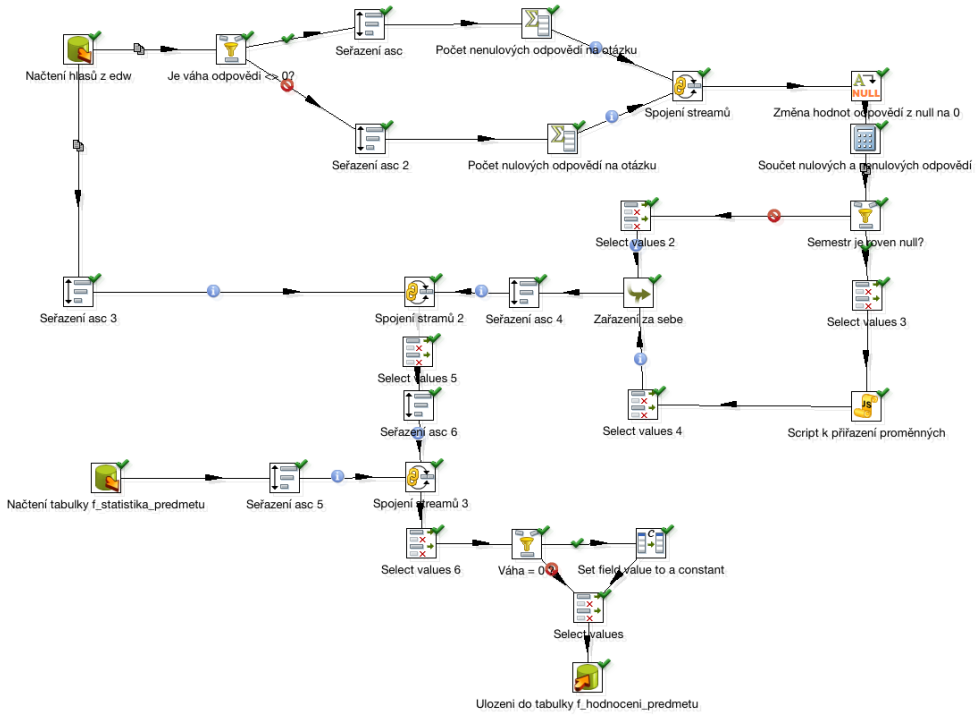
2. Výpočet počtu nulových a nenulových odpovědí v rámci jednotlivých otázek, které byly odevzdány v hodnocení daného předmětu za příslušný semestr.

3. Přepsání hodnot na 0 pokud se hodnoty nulových nebo nenulových odpovědí rovnají NULL.

4. Přiřazení vypočtených hodnot k původnímu streamu hodnoceních.

5. Načtení statistik o zapsání a počtu hodnotících studentů z faktové tabulky *f_statistika_predmetu*.

6. Spojení streamů.



Obrázek 6.16: ETL skript na naplnění tabulky *f_hodnoceni_predmetu*

7. Uložení vzniklých záznamů to tabulky *f_hodnoceni_predmetu* v datovém tržišti.

6.6.3 Datové tržiště sloužící k analýze hodnocení učitelů

V rámci tohoto datového tržiště může uživatel provádět analýzy nad hodnoceními jednotlivých učitelů.

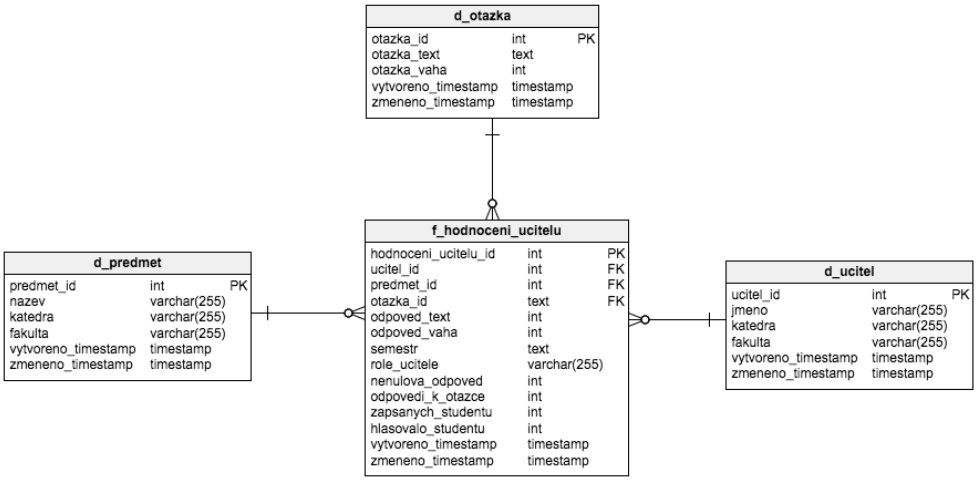
6.6.3.1 Datový model datového tržiště hodnocení učitelů

Toto datové tržiště je složeno z jedné faktové tabulky a tří klasických dimenzionálních tabulek *d_ucitel*, *d_predmet* (sdílená dimenze, která již byla popsána v tabulce 6.7). Tabulka *d_ucitel*, jejíž atributy jsou popsány v tabulce 6.10 představuje vyexportovaný seznam učitelů z centrálního datového skladu. Během ETL procesu, který přenesl záznamy z jedné databáze do druhé, došlo ke spojení titulů (před a za) se jménem a příjmením učitele a vyhledání názvu kateder a fakult, v rámci kterých daný učitel učil během semestru.

Jak je možné vidět z modelu na obrázku 6.17, faktová tabulka tohoto datového tržiště je velmi podobná faktové tabulce datového tržiště hodnocení předmětů. Granularitou této tabulky je jeden odevzdaný hlas v rámci fakultních anket ČVUT hodnotící konkrétního učitele v roli, ve které vyučoval daný

Název sloupce	Popis
ucitel_id	Unikátní identifikátor záznamu v tabulce. Tento identifikátor je přebírán z tabulky *ucitel* z centrálního datového skladu.
jmeno	Spojené tituly se jménem a příjmením vyučujícího.
prijmeni	Tento atribut je v tabulce uložen proto, aby bylo možné seřadit učitelé abecedně.
katedra	Katedra, v rámci které je daný učitel vyučuje.
fakulta	Fakulta, v rámci které je daný učitel vyučuje.
vytvoreno_timestamp	Časový otisk, kdy byl daný záznam vytvořen.
zmeneno_timestamp	Časový otisk, kdy byl daný záznam změněn.

Tabulka 6.10: Popis dimenzionální tabulky *d_ucitel*

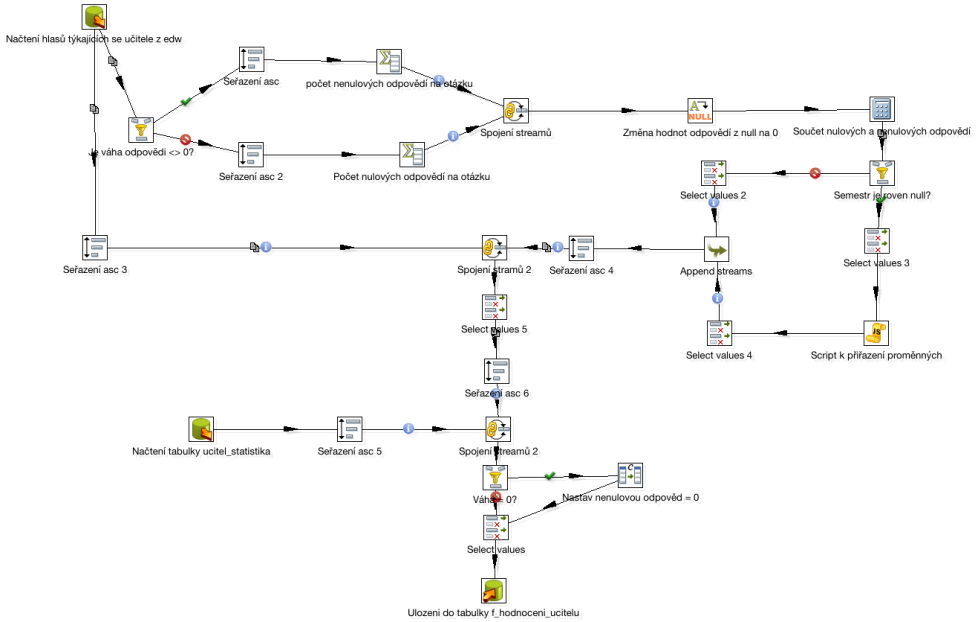


Obrázek 6.17: Datový model datového tržiště *hodnocení učitelů*

předmět. Podobně jako již zmíněná faktová tabulka také obsahuje údaje o počtu zapsaných a počtu hodnotících studentů. V tomto případě se jedná o počty patřící ke konkrétní roli učitele v rámci předmětu.

6.6.3.2 ETL skript na naplnění datového tržiště hodnocení učitelů daty

ETL skript na naplnění faktové tabulky znázorněné na modelu 6.16 je velmi podobný jako ETL proces, který vytváří faktovou tabulky datového tržiště hodnocení předmětů, proto zde nebudou popisovány jednotlivé kroky tohoto procesu.



Obrázek 6.18: Skript na naplnění tabulky *f_hodnoceni_ucitelu*

Obrázek 6.19: Skript, při kterém dojde k naplění dimenzionální tabulky *d_ucitel*

Proto, aby došlo k úspěšnému naplnění datového tržiště daty je nutné, aby byly naplněny všechny dimenzionální tabulky. Modely ETL skriptů sloužící k naplnění dimenzionálních tabulek *d_predmet*, *d_otazka* byly popsány výše. Skript, který naplní tabulku *d_ucitel* je možné vidět na modelu 6.19

6.7 Návrh OLAP kostek

Poté, co byla naplněna datová tržiště daty, mohou být vytvořeny OLAP kostky, pomocí nichž může uživatel provádět analýzy dat uložených v příslušných datových tržištích. Kostky byly vytvořeny za pomoci programu Mondrian Schema Workbench.[3] Za pomoci tohoto programu byly vytvořeny tři kostky,

[3]Mondrian Schema Workbench - open source program od společnosti Pentaho na vytváření a testování OLAP kostek, které po nahrání na Mondrian server umožňují analyzovat data.

jejich atributy byly posléze uloženy do xml souborů, které tyto kostky popisují, a nahrány na Mondrian server, kde lze pomocí jazyka MDX analyzovat data uložená v datových tržištích.

6.7.1 Kostka statistika předmětů

Jedná se o kostku postavenou kolem faktové tabulky *f_statistika_predmetu*, která slouží k zaznamenání statistik o jednotlivých předmětech vyučovaných na ČVUT (počet zapsaných studentů, počet studentů, kteří úspěšně dokončili předmět a počet studentů, kteří vyplnili anketní lístek k hodnocení předmětu). Tato kostka importuje jednu sdílenou dimenzi předmět a jednu degenerovanou dimenzi - semestr s hierarchií s úrovní Název semestru. Sdílená dimenze předmět má tři stupně hierarchie - Fakulta - Katedra - Název předmětu.

Míry této kostky byly zmíněny výše - počet zapsaných studentů, počet studentů, kteří úspěšně dokončili předmět a počet studentů, kteří hodnotili předmět v anketě. Z těchto měr se poté počítají tzv „Calculated Member“, ktré představují hodnoty určené matematickým výrazem, který obsahuje míry kostky.

Pomocí matematických výrazů je spočítána [Průchodnost předmětu], kterou vyjadřuje poměr studentů, kteří předmět dokončili, k počtu studentů, které si daný předmět zapsali.

Druhým výrazem, který se v rámci této kostky počítá je [Vyplněnost anketních lístků k hodnocení předmětů]. Tento výraz je určen jako poměr počtu hlasujících studentů k počtu studentů, kteří si předmět zapsali.

Na všechny míry v této kostce byly použity agregační funkce SUM, takže při agregování na vyšší úroveň, např z předmětů na katedry, nedochází ke zkreslení dat.

6.7.2 Kostka hodnocení předmětů

Tato kostka je postavená kolem faktové tabulky *f_hodnoceni_predmetu*, která slouží k zachycení hodnocení předmětů v rámci anket probíhajících na ČVUT. Tato kostka importuje tři sdílené dimenze (předmět,otázka a odpověď) a má dvě degenerované dimenze - role učitele a semestr. Dimenze semestr je stejná jako v předcházející kostce a role učitele je dimenze s hierarchií prvního stupně Role učitele. Sdílená dimenze předmět je stejná jako v předcházející kostce a dimenze otázka má hierarchii s dvěma stupni - Oficiální hodnocení předmětu - Otázka. První stupeň dimenze otázky značí, jestli otázky patřící pod tento stupeň, jsou brány jako oficiální hodnocení předmětu či se jedná o doplňující otázky. Dimenze odpověď má pouze jednostupňovou hierarchii *Odpověď*.

Míry této kostky jsou počet hlasů, počet nenulových odpovědí, počet odpovědí k otázce, součet vah odpovědí a míry zapsaných, hlasujících studentů. Míra počet hlasů představuje počet všech odevzdaných hlasů k hodnocení části předmětu (cvičení, přednáška, laboratoř, získání klasifikovaného zápočtu,

zkouška a hodnocení předmětu jako celku). Na tuto míru je použita agregační míra DISTINCT-COUNT, takže agregace do vyššího stupně dává očekávané výsledky.

Míra představující počet nenulových odpovědí je suma všech odpovědí, které mají nenulovou váhu. Tato suma se počítá pomocí agregační funkce SUM přes sloupec *nenulova_odpoved* v tabulce *f_hodnoceni_uucitelu*. Tento sloupec je pouze reprezentován 1 a 0, kde 1 značí nenulovou odpověď a naopak 0 značí odpověď s váhou nula.

Míra „počet odpovědí k otázce“ představuje poměrnou část odpovědi v procentech ze všech odpovědí k otázce. Počítá se pomocí agregační funkce AVG na sloupec *odpovedi_k_otazce* faktové tabulky. V tomto sloupci je uložen počet všech odpovědí k dané otázce v hodnocení předmětu.

Jak již název napovídá míra „součet vah“ slouží k výpočtu sumy vah odpovědí odevzdaných v rámci hodnocení předmětu. Míry vyjadřující počet zapsaných studentů a počet studentů, kteří hodnotili daný předmět jsou stejné jako u statistik předmětů.

Kostka obsahuje tři výrazy, které se počítají z měr popsaných výše. Prvním výrazem je výraz představující, kolik studentů ze zapsaných hodnotilo daný předmět. Počítá se stejně jako u statisti předmětů.

Dalším výrazem, který se počítá jako poměr sumy vah k počtu nenulových odpovědí, je průměrné hodnocení. Jelikož odpovědi uložené v Datových tržištích nabývají hodnot 0-5, kde nula představuje nerozhodné odpovědi typu „nevím, neumím se vyjádřit“, byly odpovědi s váhou vyloučeny z výpočtu průměrného hodnocení.

Posledním výrazem, který je počítán z měr je [Ze všech odpovědí otázky], který vyjadřuje poměr dané odpovědi vůči ostatním odpovědím na příslušnou otázku.

6.7.3 Kostka hodnocení učitelů

Tato kostka je tvořena ze stejných dimenzí, měr a vypočtených, výrazů jako kostka hodnocení předmětů. Navíc má dimenzi učitele, která představuje učitele, který byl daným hlasem hodnocen. Tento učitel má hierarchii o třech úrovních: Fakulta - Katedra - Jméno učitele.

Kapitola 7

Testování výsledného řešení

Aby se mohl datový sklad používat/testovat je potřeba nejdříve datový sklad vytvořit.

7.1 Vytvoření datového skladu a nahráni dat

Struktura datového skladu se vytváří pomocí tří sql skriptů. Centrální datový sklad je vytvořen pomocí skriptu *edw_create.sql*, dimenzionální a faktové tabulky datových tržišť se vytvářejí hromadně v rámci skriptu *dms_create.sql* a na vytvoření speciální databáze pro ukládání „poškozených“ záznamů zdrojové databáze se používá skript *edw_deleted.sql*. Vytvoření těchto tří databázi je nezbytnou podmínkou pro úspěšné vytvoření architektury datového skladu.

Pro vyexportování dat ze zdrojové databáze a následné naplnění těmito daty datový sklad se využívají skripty, které lze spustit z příkazové řádky nebo pomocí programu Pentaho Data Integration. Celý proces je zautomatizovaný a jediným uživatelským vstupem je zadání názvů profilů zdrojové databáze, které korespondují se semestry, jak bylo popsáno v kapitole 6.3. Zadání požadovaných profilů se provádí zapsáním do konfiguračního *schemas.txt*. Jedinou podmínkou pro tento konfigurační soubor je, že vybrané semestry musejí zachovávat pořadí jak ve smyslu akademickém tak i ve smyslu zachování pořadí dle jednotlivých let, tzn zimní semestr musí být před letním a nižší rok před vyšším.

7.2 Vytvoření OLAP kostek

K tomu, aby bylo možné analyzovat data uložená v datovém skladu, je nezbytné vytvoření a zaregistrování datových kostek na serveru Mondrian. Jak již bylo zmíněno dříve, tento server je součástí balíku Business Analytics Platform od společnosti Pentaho. Nejprve je nutné server spustit za pomoci pří-

slušného sh/bat skriptu. V současné chvíli běží tento server na lokální adrese *http://localhost:8080*.

Poté co je server spuštěn, mohou být nahrány příslušné xml soubory na již zmíněný server. Každá kostka má svůj xml soubpr, který obsahuje popis dimenzí, faktových tabulek a jejich napojení na tabulky datových tržišť.

7.3 Testování

Ve chvíli, kdy je datový sklad naplněn daty a konfigurační soubory kostek nahrané na serveru, mohou být prováděny analýzy dat. Jak bylo zmíněno v kapitole 6 nástrojem, který bude použit k analýze je Saiku Analytics.

7.3.1 Ověření správnosti dat uložených v datovém skladu

Testování správnosti výsledné implementace probíhalo srovnávání dat získaných z datového skladu s daty dostupnými na adrese `https://fit.cvut.cz/student/studijni/anketa`. Legendu ke všem prvkům tabulek lze nalézt na adrese `https://it.fit.cvut.cz/anketa/2010-2011-L/help/index.html`.

Analýzy tvořené pomocí nástroje Saiku Analytics jsou dynamické, proto legenda k jednotlivým prvkům tabulek zde nebude uvedena, ale význam prvků by měl být zřejmý z popisu sloupců.

V rámci tetování správnosti byly vytvořeny následující případy užití:

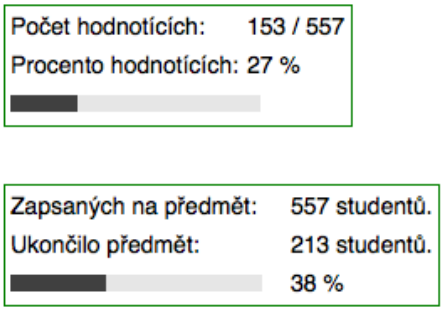
1. Statistiky předmětu BI-LIN v letním semestru 2013/2014.
2. Hodnocení předmětu BI-OMO v zimním semestru 2012/2013.
3. Hodnocení učitele Plicka Martin Ing. v letním semestru 2010/2011.

7.3.1.1 Statistiky předmětu BI-LIN v letním semestru 2013/2014

Statistiky dostupné na serveru FIT jsou zobrazeny na obrázku 7.1. Na obrázku 7.2 je statistika tohoto předmětu vytvořená pomocí Saiku Analytics. Statistiky se od sebe liší o několik studentů, což může být způsobeno odlišným zpracováním dat zdrojové databáze. Data uložená v tomto řešení se shodují s daty ve zdrojové databázi.

7.3.1.2 Hodnocení předmětu BI-OMO v zimní semestru 2012/2013

Hodnocení předmětu BI-OMO na serveru FIT je možné vidět na obrázku 7.3. V rámci analýzy dat uložených v datovém skladu, nelze v současné chvíli vytvořit takto komplexní hodnocení. Každopádně nástroj Saiku Analytics umožňuje vytvořit stejné hodnocení ve více krocích. Na obrázku 7.4 lze vidět hodnocení předmětu v rámci jednotlivých otázek. V tomto případě čísla u otázek



Obrázek 7.1: Statistiky předmětu BI-LIN v letním semestru 2013/2014 dostupné na webových stránkách FIT

Predmet	MeasuresLevel	letní semestr 2013/2014
BI-LIN - Lineární algebra	Zapsanych studentu	556
	Dokoncilo studentu	215
	Hlasovalo studentu	153
	Uspesnost [%]	38.669
	Hlasujicich studentu [%]	27.518

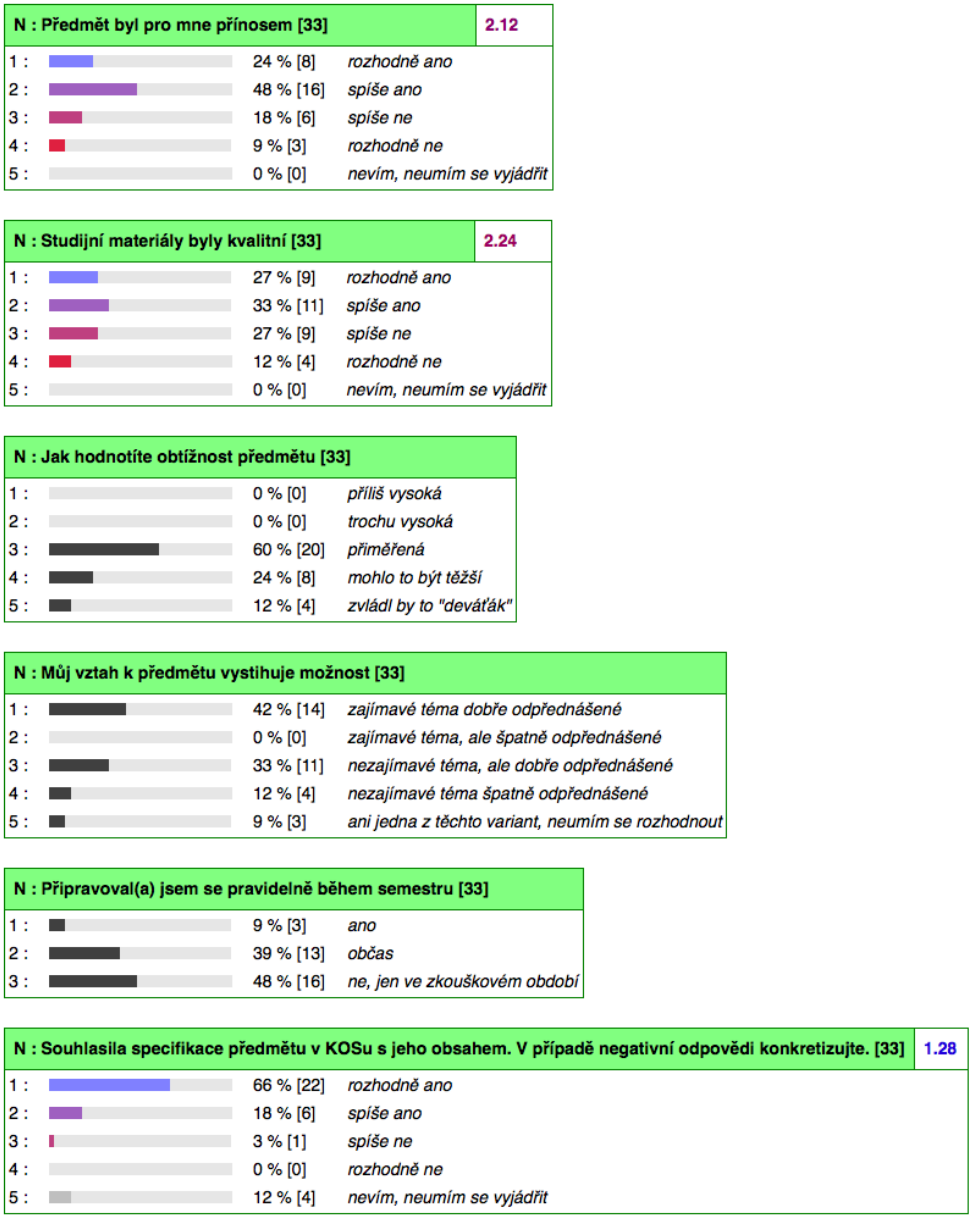
Obrázek 7.2: Statistiky předmětu BI-LIN v letním semestru 2013/2014 vytvořené pomocí Saiku Analytics

nepatřících do oficiálního hodnocení předmětu (úroveň otázky je rovna 0) jsou čísla nadbytečná.

Jednotlivé odpovědi na otázky, jejich počet a procentuální zastoupení, lze vidět na obrázku 7.5. Jak si čtenář může povšimnout, v současné chvíli není možné, aby byly možné odpovědi na otázky s nulovým počtem. Proto, aby bylo možné zobrazit i nezodpovězené odpovědi, by bylo nutné zasáhnout do současných modelů centrálního datového skladu a datových tržišť.

7.3.1.3 Hodnocení učitele Plicka Martin Ing. v letním semestru 2010/2011

Hodnocení učitele, které je dostupné na serveru FIT je možné vidět na obrázcích 7.6, 7.9. Jako v případě hodnocení předmětů nelze v současné chvíli sestavit takto komplexní reporty, ale lze podobného výsledku dosáhnout v několika krocích, jak je ukázáno na obrázcích 7.7, 7.8, 7.10, 7.11 . Při analýze celkového hodnocení učitele pomocí Saiku Analytics není v tuto chvíli možné zobrazit procentuální zastoupení jednotlivých vah odpovědí.



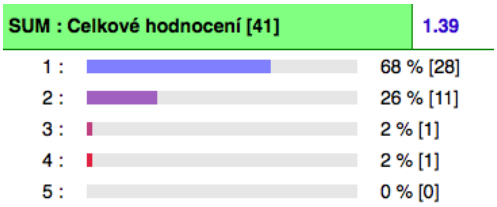
Obrázek 7.3: Hodnocení předmětu BI-OMO v zimním semestru 2012/2013 dostupné na webových stránkách FIT

		Prumerne hodnoceni
		BI-OMO - Objektové modelování
Oficialni hodnoceni	Otazka	zimní semestr 2012/2013
0	jak hodnotíte obtížnost předmětu	3.5
	můj vztah k předmětu vystihuje možnost	2.438
	připravoval(a) jsem se pravidelně během semestru	2.406
1	předmět byl pro mne přínosem	2.121
	souhlasila specifikace předmětu v kosu s jeho obsahem. v případě negativní odpovědi konkretizujte.	1.276
	studijní materiály byly kvalitní	2.242

Obrázek 7.4: Průměrné váhy odpovědí na otázky v rámci hodnocení předmětu BI-OMO v zimním semestru 2012/2013 vytvořené pomocí Saiku Analytics

		BI-OMO - Objektové modelování	
		zimní semestr 2012/2013	
Otazka	Odpoved	Ze vsech odpovedi otazky [%]	Pocet hlasu
jak hodnotíte obtížnost předmětu	mohlo to být těžší	25.	8
	přiměřená	62.5	20
	zvládl by to "deváťák"	12.5	4
		100.	**32**
můj vztah k předmětu vystihuje možnost	ani jedna z těchto variant, neumím se rozhodnout	9.38	3
	nezajímavé téma špatně odprednášené	12.5	4
	nezajímavé téma, ale dobře odprednášené	34.38	11
	zajímavé téma dobře odprednášené	43.75	14
		100.	**32**
připravoval(a) jsem se pravidelně během semestru	ano	9.38	3
	ne, jen ve zkouškovém období	50.	16
	občas	40.63	13
		100.	**32**
předmět byl pro mne přínosem	rozhodně ano	24.24	8
	rozhodně ne	9.09	3
	spíše ano	48.48	16
	spíše ne	18.18	6
		100.	**33**
souhlasila specifikace předmětu v kosu s jeho obsahem. v případě negativní odpovědi konkretizujte.	nevím, neumím se vyjádřit	12.12	4
	rozhodně ano	66.67	22
	spíše ano	18.18	6
	spíše ne	3.03	1
		100.	**33**
studijní materiály byly kvalitní	rozhodně ano	27.27	9
	rozhodně ne	12.12	4
	spíše ano	33.33	11
	spíše ne	27.27	9
		100.	**33**

Obrázek 7.5: Výpis odpovědí k jednotlivým otázkám v rámco hodnocení předmětu BI-OMO v zimní semestru 2012/2013 vytvořené pomocí Saiku Analytics



Obrázek 7.6: Hodnocení učitele Plicka Martin Ing. v rámci všech vyučovaných předmětů v letním semestru 2010/2011 dostupné na webových stránkách FIT

	letní semestr 2010/2011	
Jmeno ucitele	Prumerne hodnoceni	Pocet hlasu
Ing. Martin Plicka	1.39	42

Obrázek 7.7: Průměrné hodnocení učitele Plicka Martin Ing. v rámci všech vyučovaných předmětů v letním semestru 2010/2011 vytvořené pomocí Saiku Analytics

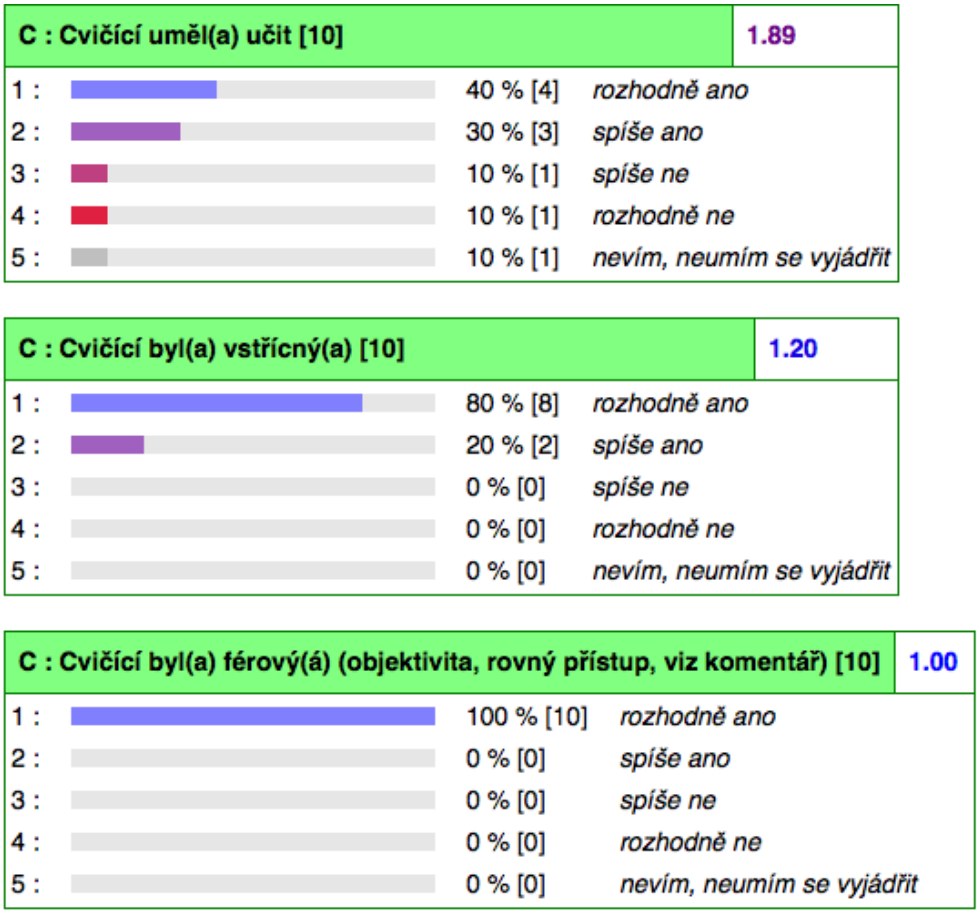
	letní semestr 2010/2011	
	Ing. Martin Plicka	
Odpoved	Prumerne hodnoceni	Pocet hlasu
rozhodně ano	1	28
spíše ano	2	11
spíše ne	3	1
rozhodně ne	4	1
nevím, neumím se vyjádřit	5	1

Obrázek 7.8: Zobrazení počtu jednotlivých odpovědí a vah (sloupec prumerne hodnoceni) učitele Plicka Martin Ing. v rámci všech vyučovaných předmětů v letním semestru 2010/2011 vytvořené pomocí Saiku Analytics

7.3.2 Zhodnocení naměřených hodnot v rámci testování

Jak je patrné při srovnání výsledků analýz nad daty nově vzniklého datového skladu, všechny naměřené hodnoty se shodují, respektive v některých případech jsou dokonce přesnější.

V rámci demonstrace možností analýzy datového skladu byla vytvořen sloupcový graf srovnávající počty zapsaných studentů s počty studentů, kteří úspěšně dokončili předmět MI-PAR. Statistika se týká posledních třech zimních semestrů. Výsledný graf je možné vidět na obrázku 7.12, kde červený sloupec představuje počet zapsaných studentů a okrový sloupec představuje počet zapsaných studentů.



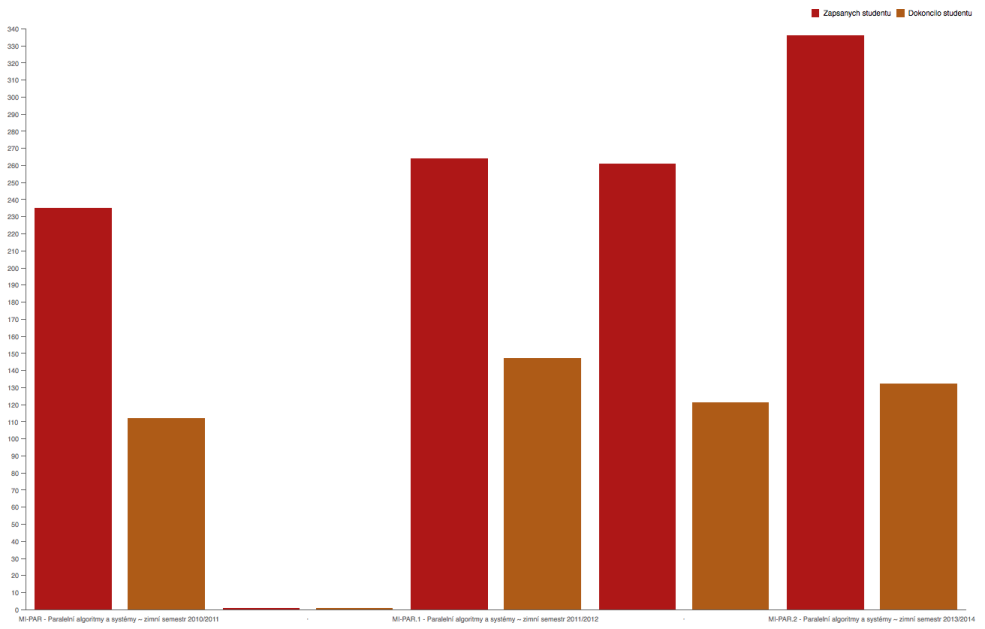
Obrázek 7.9: Hodnocení učitele Plicka Martin Ing. v rámci jednotlivých otázek v předmětu MI-GEN v letním semestru 2010/2011 vytvořené pomocí Saiku Analytics

		Ing. Martin Plicka	
		MI-GEN - Generování kódu	
Oficialni hodnoceni	Otazka	Pocet hlasu	Prumerne hodnoceni
1	cvičící byl(a) férový(á) (objektivita, rovný přístup, viz komentář)	10	1
	cvičící byl(a) vstřícný(a)	10	1.2
	cvičící uměl(a) učit	10	1.889

Obrázek 7.10: Průměrné hodnocení učitele Plicka Martin Ing. v rámci jednotlivých otázek v předmětu MI-GEN v letním semestru 2010/2011 vytvořené pomocí Saiku Analytics

		Ing. Martin Plicka	
		MI-GEN - Generování kódu	
Otazka	Odpoved	Pocet hlasu	Ze vsech odpovedi otazky [%]
cvičící byl(a) férový(á) (objektivita, rovný přístup, viz komentář)	rozhodně ano	10	100
		10	**100**
cvičící byl(a) vstřícný(a)	rozhodně ano	8	80
	spíše ano	2	20
		10	**100**
cvičící uměl(a) učit	rozhodně ano	4	40
	spíše ano	3	30
	spíše ne	1	10
	rozhodně ne	1	10
	nevím, neumím se vyjádřit	1	10
		10	**100**

Obrázek 7.11: Hodnocení učitele Plicka Martin Ing. v rámci jednotlivých otázek s odpovědmi v předmětu MI-GEN v letním semestru 2010/2011 vytvořené pomocí Saiku Analytics



Obrázek 7.12: Statistika porovnání počtů zapsaných studentů s počty studentů, kteří dokončili předmět MI-PAR v posledních třech letech

Kapitola 8

Možnosti rozšíření implementovaného datového skladu

Tak, jak je tento datový sklad implementován, tak jeho rozšíření je velmi jednoduché. Jedním z návrhů, jak rozšířit tento datový sklad je přidáním měr týkajících se známek studentů a jejich propojení s hodnocením předmětů a učitelů.

Další velkou oblastí, která zde byla pouze nastíněna je vytváření reportů z analýz dat uložených v datovém skladu, což se díky nedostatku času prozatím nepovedlo. Tyto reporty se týkají i současných statických stránek, které jsou dostupné na adrese `https://fit.cvut.cz/student/studijni/anketa`. Pokud by došlo k vytvoření šablon, které by poskytovaly dynamické řešení a umožňovaly např interakci z koncovým uživatelem, celá tato oblast by dostala nový rozměr.

Další možností, jak zdokonalit současné analýzy hodnocení předmětů je upravením stávajících ETL skriptů, které v současné době nezohledňují hlasování v rámci jednotlivých částí předmětů, jako je tomu u hodnocení učitelů.

Velkým přínosem, jak rozšířit možnosti analýz by bylo přidáním virtuálních OLAP kostek, pomocí kterých by bylo možné propojit hodnocení předmětu a učitelů s průchodností semestrů.

V současné chvíli nejsou v řešení implementovány studijní obory ani programy. Zařazením těchto oblastí do datového skladu, by se možnosti analýzy rozšířily a koncoví uživatelé by dostávali o to podrobnější informace.

Dalším prvkem, o který by mohlo být současné řešení rozšířeno, je úprava stávajících ETL skriptů tak, aby bylo možné v analýze dat vidět i odpovědi otázek, které nebyly odevzdány v rámci anketních lístků.

Závěr

Cílem této práce byla reimplementace současného řešení datového skladu tak, aby bylo možné výsledné řešení integrovat v rámci nově vznikajícího interního projektu. Tento hlavní cíl byl dle mého názoru naplněn.

Při implementaci jsem narazil na několik problémů týkající se neúplnosti dat a jejich „dopočítání“, které zkresluje výsledná data. Týká se to především anket, která nemají definovanou ani jednu povinnou odpověď. V tomto ohledu by možná stálo za zvážení zavést v budoucnu alespoň jednu povinnou odpověď, protože by to mělo velký dopad na přesnost analýz.

Pokud bych měl srovnat mnou implementované řešení s tím stávajícím, tak bych rád podotknul, že v rámci hodnocení učitelů se mi podařilo zpřesnit výstupy analýz, jelikož jsem vytvořil ETL skript, který správně (v rámci možností) počítá se studenty zapsaných k učiteli v rámci zkoušky. Z toho vyplývá, že procenta vyplněnosti role učitele „zkoušejíc“í by měla být přesnější než doposud.

Literatura

[1] Inmon, B.: Data Warehouse Definition [online]. [Citováno 18-04-2015]. Dostupné z: http://www.1keydata.com/datawarehousing/data-warehouse-definition.html

[2] 1keydata.com: Third Normal Form (3NF) [online]. [Citováno 18-04-2015]. Dostupné z: http://www.1keydata.com/database-normalization/third-normal-form-3nf.php

[3] Zámyslický, J.: Návrh a implementace OLAP prostředí nad archivy výsledků studentské ankety ČVUT [online]. [Citováno 18-04-2015]. Dostupné z: https://dip.felk.cvut.cz/browse/pdfcache/zamysj1_2009dipl.pdf

[4] Topinková, P.: Datový sklad nad výsledky ankety - nasazení [online]. [Citováno 18-04-2015]. Dostupné z: https://dip.felk.cvut.cz/browse/pdfcache/topinpa1_2010bach.pdf

[5] Stadler, J.: Anketa ČVUT - refaktoring a rozšíření [online]. [Citováno 18-04-2015]. Dostupné z: https://dip.felk.cvut.cz/browse/pdfcache/stadlja1_2012bach.pdf

[6] Topinková, P.: Anketa ČVUT - přístup k aktuálním i historickým výsledkům anket [online]. [Citováno 18-04-2015]. Dostupné z: https://dip.felk.cvut.cz/browse/pdfcache/topinpa1_2012dipl.pdf

[7] Corporation, T.: Independent Data Marts [online]. [Citováno 18-04-2015]. Dostupné z: http://goo.gl/DO3EnF

[8] Rainardi, V.: *Building a Data Warehouse With Examples in SQL Server*. Apress, ISBN 978-1-59059-931-0.

[9] Pentaho: Pentaho dokumentace [online]. [Citováno 18-04-2015]. Dostupné z: http://wiki.pentaho.com/x/KwU

[10] Pentaho: Mondrian dokumentace [online]. [Citováno 18-04-2015]. Dostupné z: http://mondrian.pentaho.com/documentation/

[11] Saiku: Saiku dokumentace [online]. [Citováno 18-04-2015]. Dostupné z: http://wiki.meteorite.bi/display/SAIK/Saiku

[12] Thilini Ariyachandra, H. J. W.: Which Data Warehouse Architecture Is Most Successful? [online]. [Citováno 18-04-2015]. Dostupné z: http://goo.gl/ZDddjD

[13] Wikipedia: Business Intelligence [online]. [Citováno 18-04-2015]. Dostupné z: http://cs.wikipedia.org/wiki/Business_Intelligence

[14] Frélich, L.: Databázový model aplikace Anketa ČVUT [online]. [Citováno 18-04-2015]. Dostupné z: https://dip.felk.cvut.cz/browse/pdfcache/frelil1_2009bach.pdf

Příloha **A**

Seznam použitých zkratek

BI Business Intelligence

CSV Comma Separated Values

DM Data Mart

DWH Data Warehouse

EDW Enteprise Data Warehouse

ETL Extract, Transform, Load

IS Informační systém

MOLAP Multidimensional Online Analytical Processing

MDX Multidimensional Expressions

OLAP Online Analytical Processing

ROLAP Relational Online Analytical Processing

XML Extensible Markup Language

Příloha B

Tabulky zdrojové databáze aplikace Anketa

Název sloupce	**Popis**
del	Příznak, který určuje jestli je daný záznam smazán či nikoliv.
fakulta	Cizí klíč do tabulky *tFakulta*, pod kterou daná katedra patří.
kod	Kód značící danou katedru.
nazev	Název dané katedry.

Tabulka B.1: Popis materializovaného pohledu *sKatedra*

Název sloupce	Popis
semestr	Kód semestru v rámci kterého je daný předmět vyučován.
predmet_id	Identifikátor předmětu, v rámci kterého je daná paralelka vyučována.
predmet	Kód předmětu.
fakulta	Kód fakulty, v rámci které je daný předmět vyučován.
katedra	Kód katedry, v rámci které je daný předmět vyučován.
typ	Kód značící o jaký typ paralelky se jedná (C,L,P).
predn	Kód pro paralelku typu přednáška.
cviko	Kód pro paralelku typu cvičení.
lab	Kód pro paralelku ty laboratř.
ucitel1_id	Identifikátor učitele, který danou paralelku vyučuje.
ucitel2_id	Identifikátor učitele, který danou paralelku vyučuje pokud je vyučována dvěma učiteli současně.

Tabulka B.2: Popis materializovaného pohledu *sParalelka*

Název sloupce	Popis
fakulta	Cizí klíč do tabulky *tFakulta* jejíž katedra vyučuje daný předmět.
katedra	Cizí klíč do materializovaného pohledu *sKatedra* v rámci, které je daný předmět vyučován.
kod	Kód předmětu.
nazev	Název předmětu.
predmet_id	Identifikátor předmětu, který je přebírán z KOSu

Tabulka B.3: Popis materializovaného pohledu *sPredmet*

Název sloupce	Popis
username	Uživatelské jméno studenta, který si daný předmět zapsal.
predmet_id	Identifikátor předmětu, který si daný student zapsal.
kod	Kód předmětu, který si student zapsal.
fakulta	Kód fakulty, v rámci které je daný předmět vyučován.
zakoncen	Příznak který značí, jestli daný student předmět dokončil či nikoliv. (NULL značí neukončeno, písmeno „A“ značí úspěšně ukončený předmět)
zpuszak	Kód pro způsob zakončení předmětu (Z,ZK, KZ, Z,ZK)
zapocteno	Příznak který značí, jestli daný student předmět dostal zápočet či nikoliv. (NULL značí nezapočteno, písmeno „Z“ značí obdržený zápočet)
klasifikoval_id	Identifikátor učitele, který klasifikoval studenta v daném předmětu.
zapocet_udelil_id	Identifikátor učitele, který udělil studentovi zápočet, pokud bylo možné ho v daném předmětu získat.
pno	Číslo paralelky typu přčdnáška, do které byl daný student zapsán (pokud předmět měl přednášku).
cno	Číslo paralelky typu cvičení, do kterého byl daný student zapsán (pokud předmět měl cvičení).
lno	Číslo paralelky typu laboratoř, do které byl daný student zapsán (pokud předmět měl laboratoř).
zastupred	Identifikátor učitele, který zastupoval přednášejícího pokud se tak stalo.

Tabulka B.4: Popis materializovaného pohledu *sStud_Pred*

Název sloupce	Popis
id_ankety	Unikátní identifikátor ankety.
nazev	Název ankety.
fakulta	Cizí klíč do tabulky *tFakulta*, v rámci které probíhá daná anketa.

Tabulka B.5: Popis tabulky *tAnketa*

Název sloupce	Popis
fakulta	Unikátní identifikátor fakulty.
nazev_fakulty	Název fakulty.

Tabulka B.6: Popis tabulky *tFakulta*

Název sloupce	Popis
id_ankety	Cizí klíč do tabulky *tAnketa*, v rámci které byl odevzdán anketní lístek.
id_hodnoceni	Unikátní identifikátor anketního lístku.
predmet	Kód předmětu, který určuje předmět, v rámci kterého byl anketní lístek odevzdán.

Tabulka B.7: Popis tabulky *tHodnoceni*

Název sloupce	Popis
id_hodnoceni	Cizí klíč do tabulky *tHodnoceni*, která představuje anketní lístek, v rámci které bylo dané hodnocení odevzdáno.
id_otazky	Cizí klíč do tabulky *tOtazka*, který odkazuje na otázku, která byla zodpovězena v rámci daného hlasu.
odpoved	Číšlo odpovědi určující, která odpověď byla zvolena. Společně s id otázky určuje konkrétní odpověď.
peridno	Id učitele, který byl hodnocen v rámci daného hlasu. Pokud se jedná o hodnocení předmětu jako celek je PERIDNO rovno 0 a pokud se jedná o odpověď na otázku týkající se studia na ČVUT je PERIDNO rovno -1.

Tabulka B.8: Popis tabulky *tHodnoceni_Cislo*

Název sloupce	Popis
id_ankety	Cizí klíč do tabulky *tAnketa*, který ukazuje na konkrétní anketu, ke které se daný oddíl vztahuje.
id_oddilu	Unikátní identifikátor oddílu.
nazev	Název oddílu.
role_ucitele	Cizí klíč do tabulky *tRole_Ucitele*, který ukazuje na konkrétní roli, která je hodnocena v rámci celého oddílu.

Tabulka B.9: Popis tabulky *tOddil*

Název sloupce	Popis
id_otazky	Cizí klíč do tabulky *tOtazka*, který ukazuje na konkrétní otázku, v rámci které bylo odpovězeno jednou z nabízených odpovědí.
poradi	Pořadí odpovědi v rámci konkrétní otázky. Společně s id otázky určuje konkrétní odpověď na danou otázku.
text	Textová forma odpovědi.
vaha	Váha odpovědi.

Tabulka B.10: Popis tabulky *tOdpoved*

Název sloupce	Popis
id_oddilu	Cizí klíč do tabulky *tOddil* určující konkrétní oddíl, v kterém se daná otázka nachází.
id_otazky	Unikátní identifikátor otázky.
text	Textová forma otázky.
vaha	Váha otázky, která může nabývat hodnot 0 a 1. Otázka s váhou 1 se započítává do celkového hodnocení učitele nebo předmětu a naopak otázka s váhou 0 nikoliv.

Tabulka B.11: Popis tabulky *tOtazka*

Název sloupce	Popis
kod	Kód představující roli učitele.
nazev	Název role učitele.

Tabulka B.12: Popis tabulky *tRole_Ucitele*

Název sloupce	Popis
use_rname	Uživatelské jméno studenta, ketrý vyplnil anketní lístek.
predmet	Kód předmětu, v rámci kterého daný student vyplnil anketní lístek.
id_ankety	Cizí klíč do tabulky *tAnketa*, v rámci které byl anketní lístek odevzdán.

Tabulka B.13: Popis tabulky *tVyplnil*

Název sloupce	Popis
fakulta	Identifikátor fakulty, v rámci které daný učitel vyučuje. Spolu s ucitel_id jednoznačně určuje daného učitele.
jmeno	Jméno učitele.
katedra	Identifikátor katedry, v rámci které daný učitel vyučuje.
plne_jmeno	Plné jméno učitele i s tituly.
prijmeni	Příjmení daného učitele.
titul_pred	Tituly učitele, které učitel užívá před svým jménem.
titul_za	Tituly učitele, který učitel užívá za svým jménem.
ucitel_id	Identifikátor učitele převzatý z KOSu.

Tabulka B.14: Popis pohledu *vUcitel_All*

Název sloupce	Popis
user_name	Uživatelské jméno studenta.
kod	Kód předmětu.
fakulta	Identifikátor fakulty. Společně s kódem předmětu jasně identifikuje předmět.
ucitel_id	Identifikátor učitele.
role	Kód pro roli, ve které daný učitel figuroval v předmětu.

Tabulka B.15: Popis pohledu *vStud_Ucit*

Příloha C

Tabulky centrálního datového skladu a jejich mapování na zdrojovou databázi

Sloupec	Popis	Zdroj	Transformace
nazev	Název ankety.	tAnketa.nazev	Nezměněno.
puvodni_id	Původní identifikátor ankety.	tAnketa.id_ankety	Nezměno.

Tabulka C.1: Mapování tabulky *anketa* na zdrojový systém

Sloupec	Popis	Zdroj	Transformace
zneni	Znění anketní odpovědi.	tOdpoved.text	Převedeno na malá písmena.
vaha	Váha dané anketní odpovědi.	tOdpoved.vaha	Pokud je vaha = null, dojde ke změně na vaha = 0.

Tabulka C.2: Mapování tabulky *anketni_odpoved* na zdrojový systém

Sloupec	Popis	Zdroj	Transformace
zneni	Znění anketní otázky.	tOtazka.text	Převedeno na malá písmena.
vaha	Váha dané anketní otázky.	tOtazka.vaha	Pokud je vaha = null, dojde ke změně na vaha = 0.

Tabulka C.3: Mapování tabulky *anketni_otazka* na zdrojový systém

Sloupec	Popis	Zdroj	Transformace
nazev	Název fakulty.	tFakulta.nazev_fakulty	Nezměněno.
puvodni_id	Původní id fakulty.	tFakulta.fakulta	Nezměněno.

Tabulka C.4: Mapování tabulky *fakulta* na zdrojový systém

Sloupec	Popis	Zdroj	Transformace
puvodni_id	Kód představujici označení hodnoceni ve zdrojovém systému.	tHodnoceni_Cislo.id_hodnoceni	Nezměněno.

Tabulka C.5: Mapování tabulky *hlas* na zdrojový systém

Sloupec	Popis	Zdroj	Transformace
nazev	Název katedry.	sKatedra.nazev	Nezměněno.
puvodni_id	Původní id katedry.	sKatedra.kod	Nezměněno.

Tabulka C.6: Mapování tabulky *katedra* na zdrojový systém

Sloupec	Popis	Zdroj	Transformace
kod	Kód předmětu.	sPredmet.kod	Nezměněno.
nazev	Název předmětu.	sPredmet.nazev_fakulty	Nezměněno.
puvodni_id	Původní id předmětu.	sPredmet.predmet_id	Nezměněno.

Tabulka C.7: Mapování tabulky *predmet* na zdrojový systém

Sloupec	Popis	Zdroj	Transformace
nazev_role	Název role.	tRole_Ucitele.vyznam	Nezměněno.
kod	Kód role.	tRole_Ucitele.kod	Nezměněno.

Tabulka C.8: Mapování tabulky *role_ucitele* na zdrojový systém

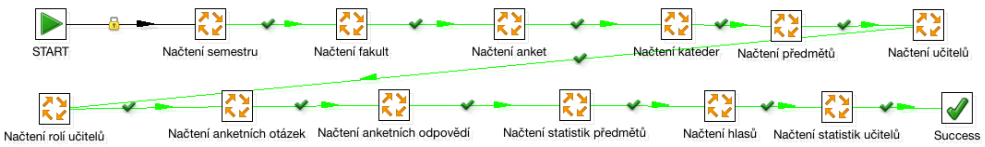
Sloupec	Popis	Zdroj	Transformace
titul_pred	Titul před jménem.	vUcitel_All.titul_pred	Nezměněno.
jmeno	Jméno učitele.	vUcitel_All.jmeno	Nezměněno.
prijmeni	Příjmení učitele.	vUcitel_All.prijmeni	Nezměněno.
titul_za	Titul za.	vUcitel_All.titul_za	Nezměněno.
puvodni_id	Původní id učitele.	vUcitel_All.ucitel_id	Nezměněno.

Tabulka C.9: Mapování tabulky *ucitel* na zdrojový systém

Příloha D

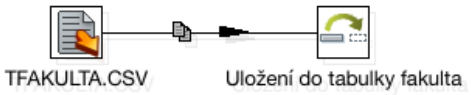
ETL procesy k vytvoření centrálního datového skladu

Obrázek D.1: Dvoufázový ETL skript na vytváření centrálního datového skladu z odkládací části



Obrázek D.2: ETL skript, který nahrává tabulky z odkládací části do centrálního datového skladu v rámci jednoho semestru

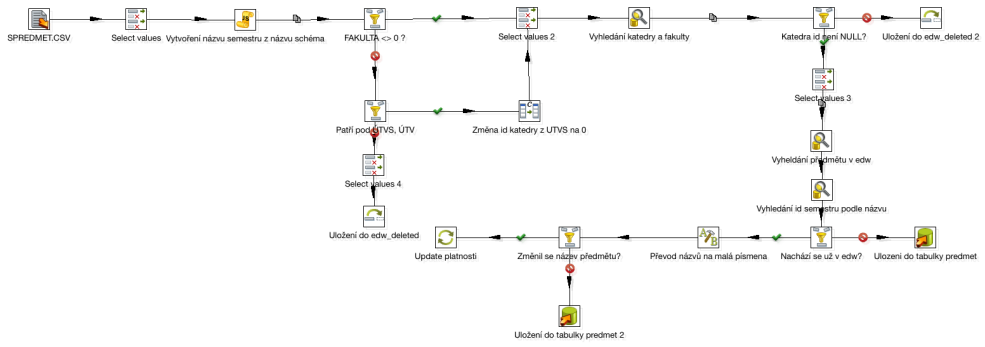
Obrázek D.3: ETL skript nahrávající data do tabulky *semestr*



Obrázek D.4: ETL skript nahrávající data do tabulky *fakulta*

Obrázek D.5: ETL skript nahrávající data do tabulky *anketni_otazka*

Obrázek D.6: ETL skript nahrávající data do abulky *katedra*

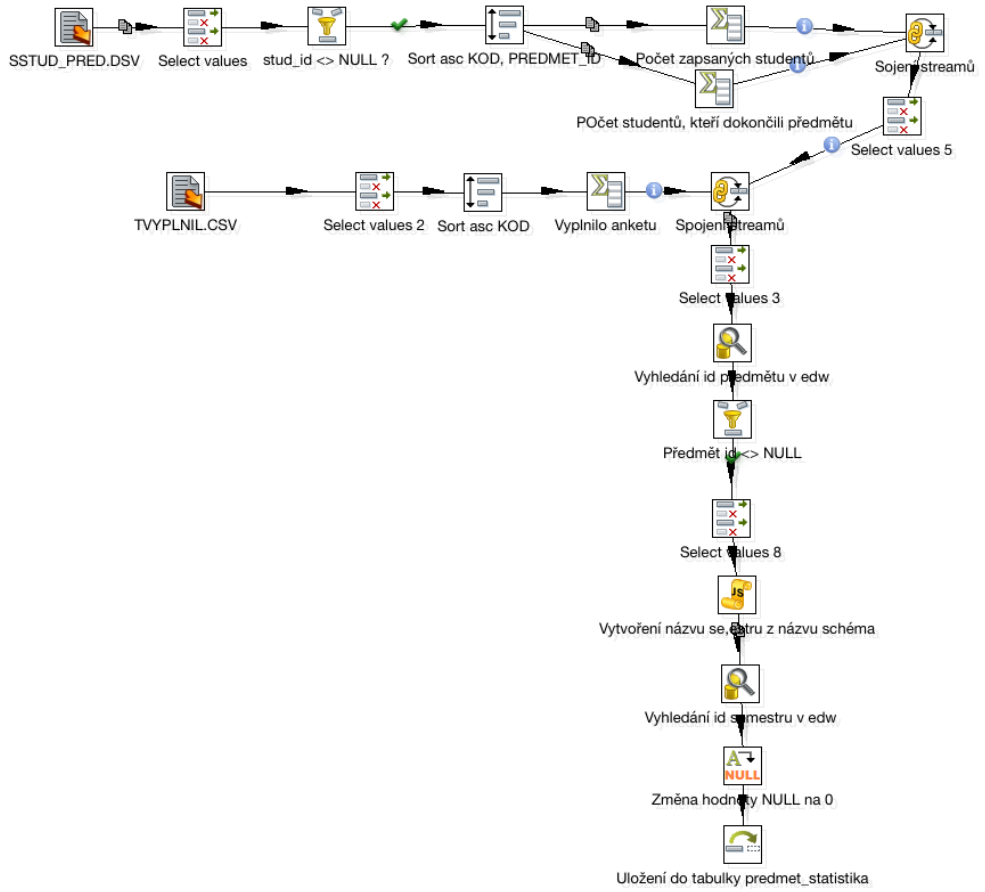


Obrázek D.7: ETL skript nahrávající data do tabulky *predmet*

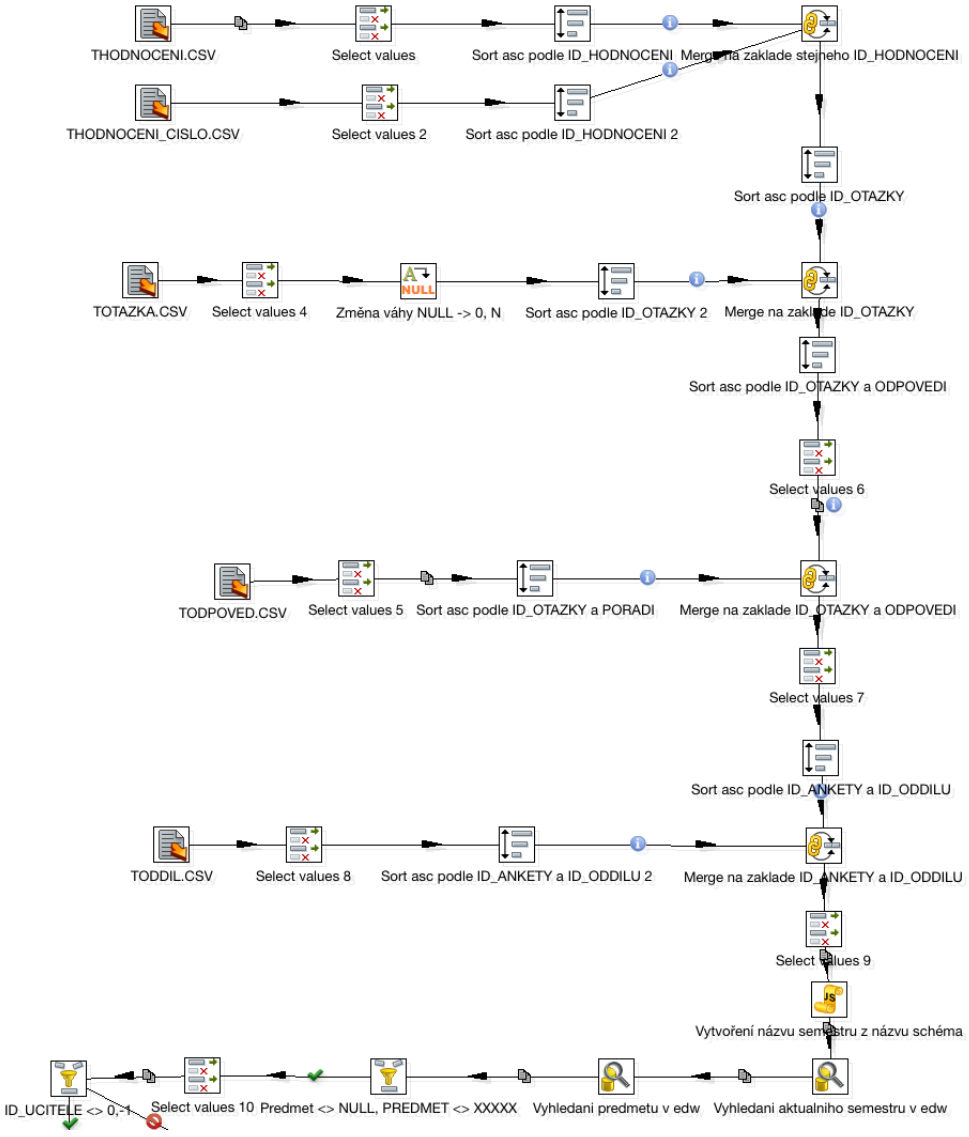
Obrázek D.8: ETL skript nahrávající data tabulky *role_ucitele*

Obrázek D.9: ETL skript nahrávající data do tabulky *anketni_otazka*

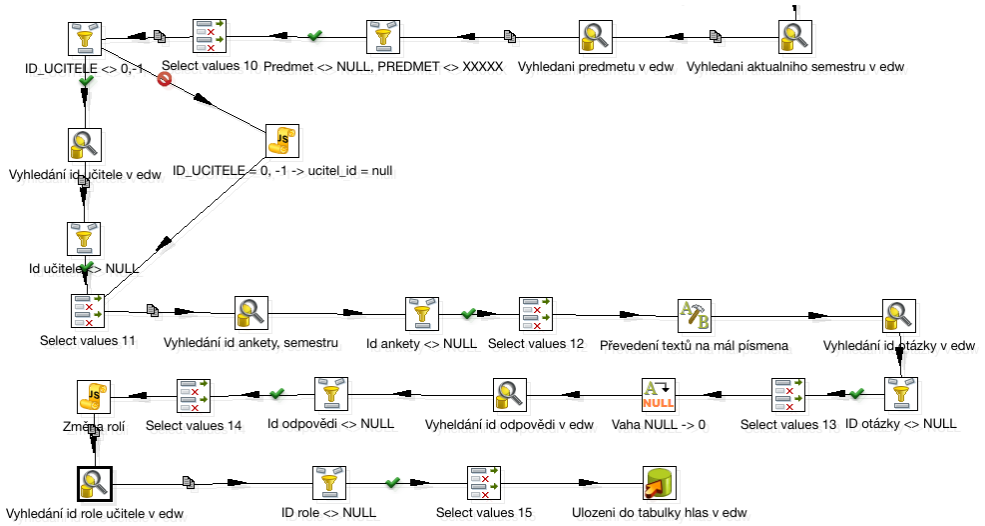
Obrázek D.10: ETL skript nahrávající data do tabulky *anektni_odpoved*



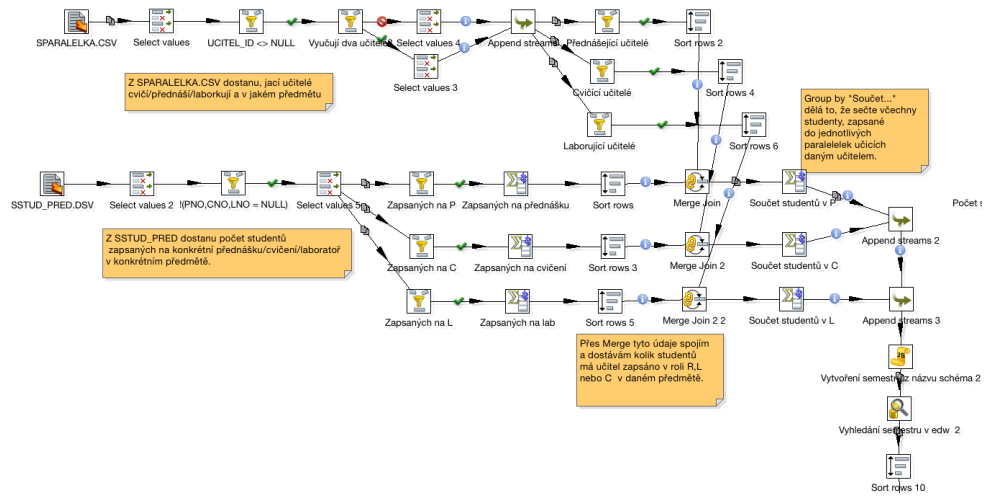
Obrázek D.11: ETL skript nahrávající data do tabulky *predmet_statistika*



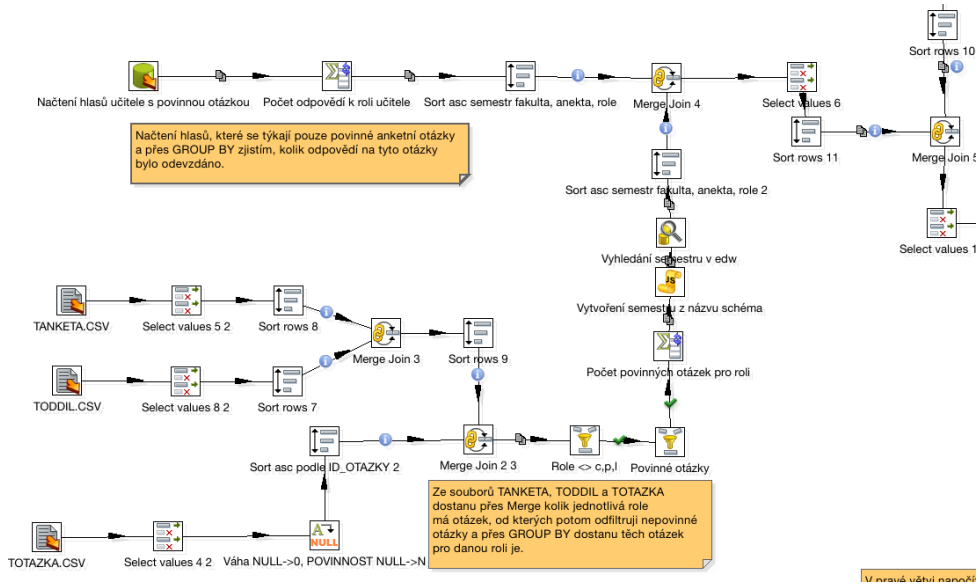
Obrázek D.12: ETL skript nahrávající data do tabulky *hlas* část první



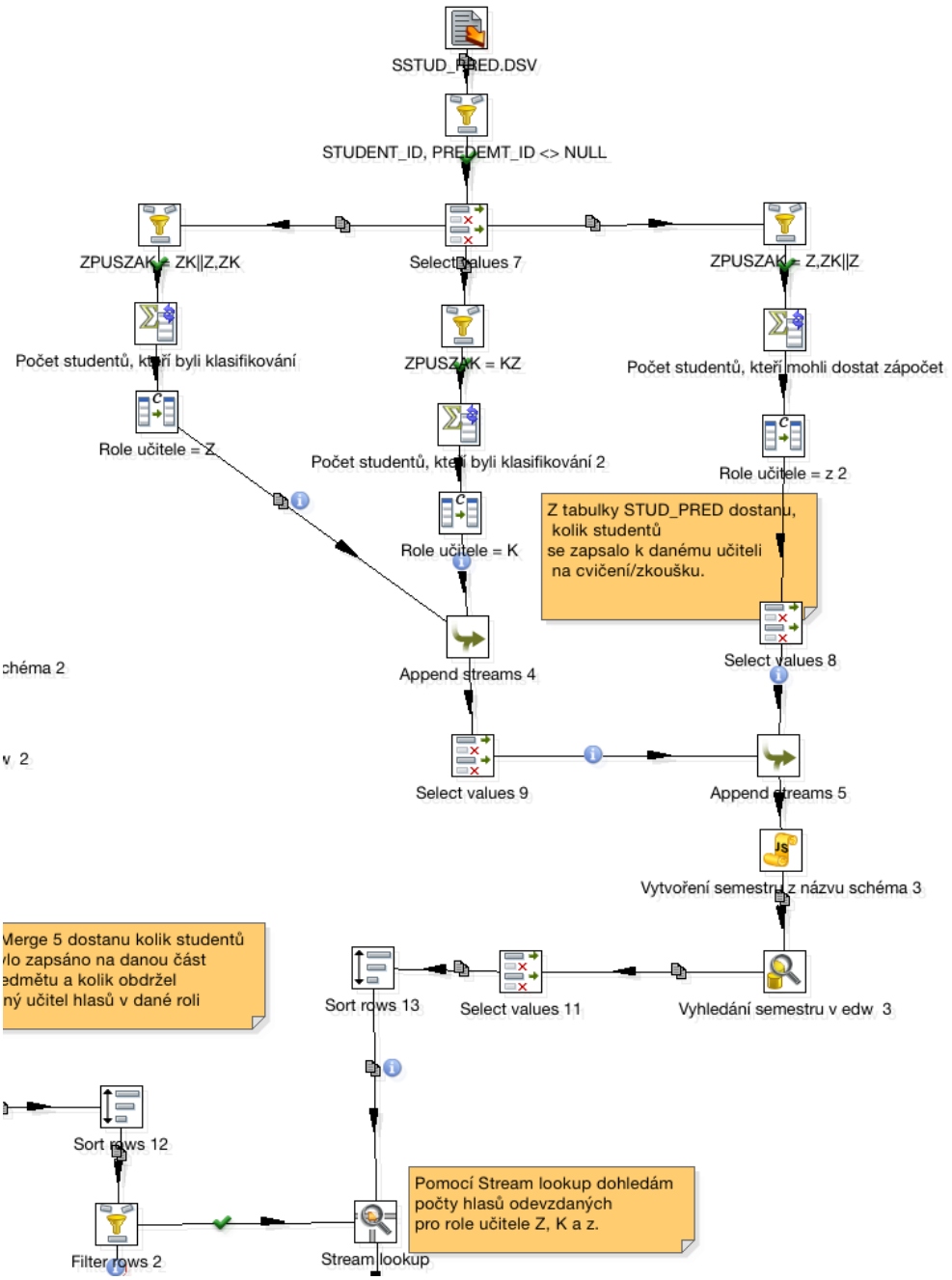
Obrázek D.13: ETL skript nahrávající data do tabulky *hlas* část druhá



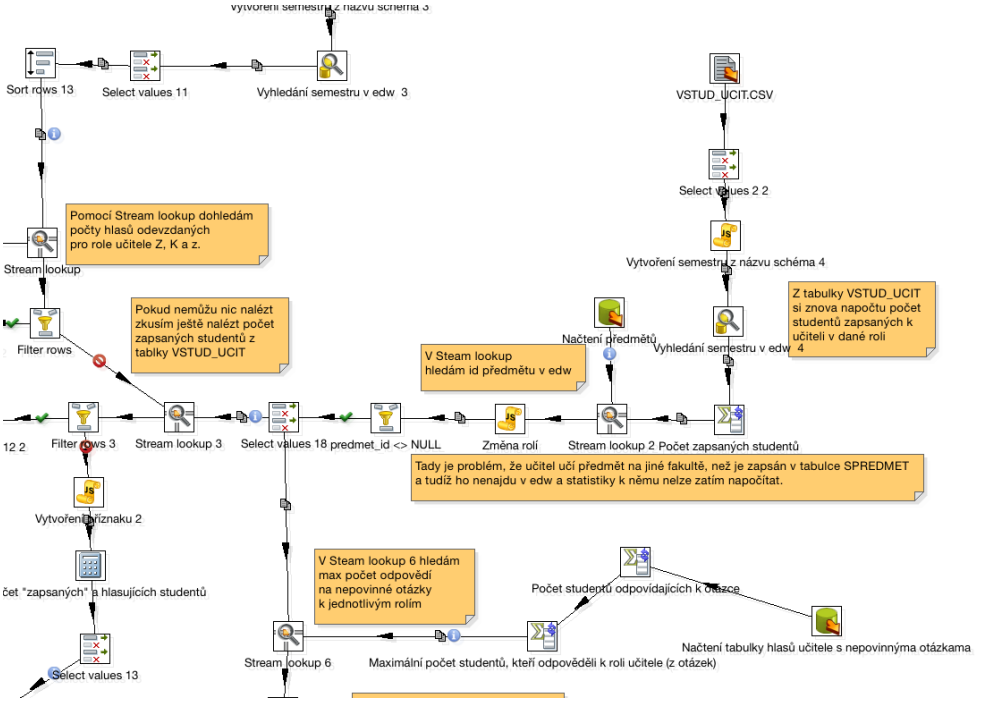
Obrázek D.14: ETL skript nahrávající data do tabulky *ucitel_statistika* část první



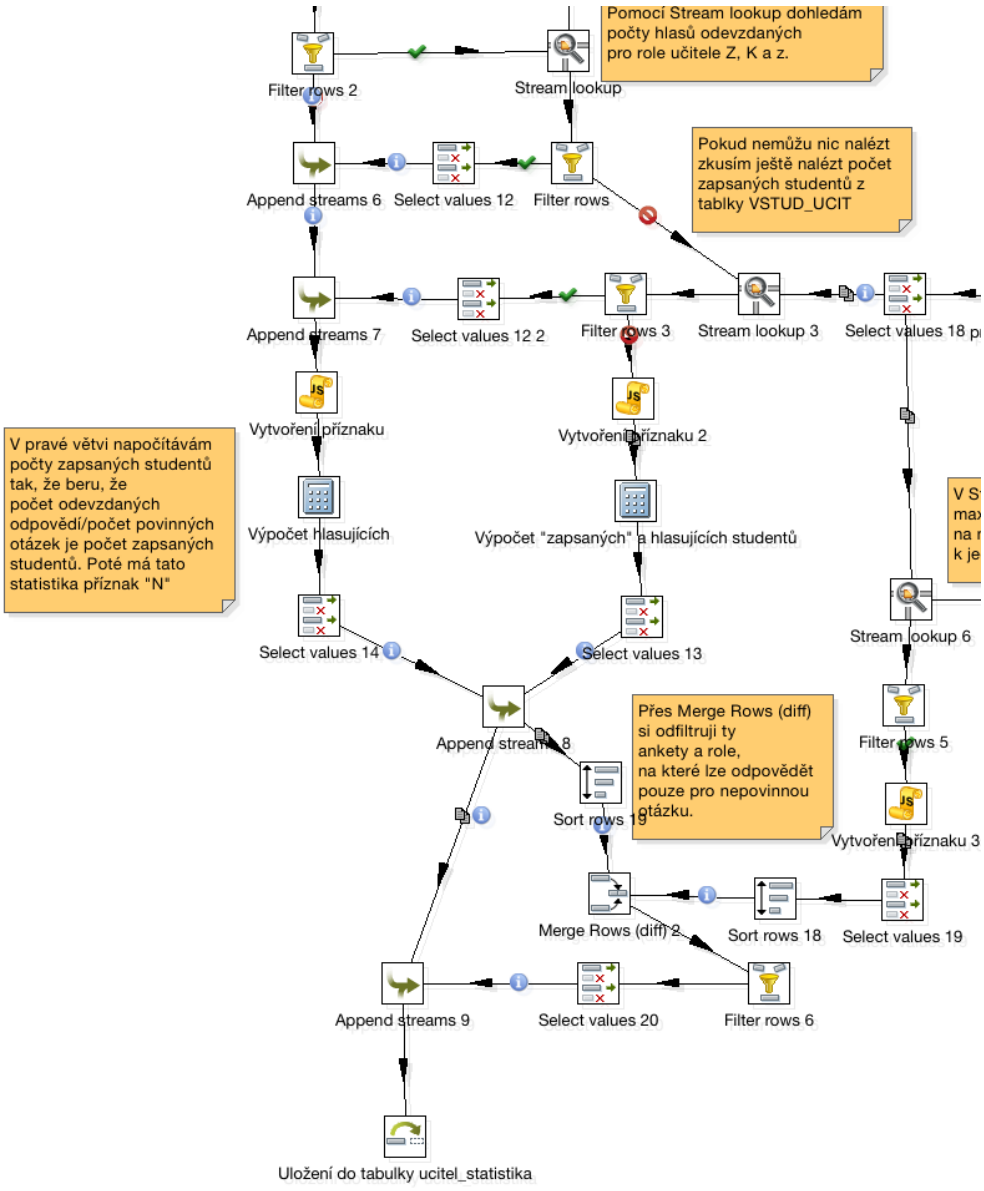
Obrázek D.15: ETL skript nahrávající data do tabulky *ucitel_statistika* část druhá



Obrázek D.16: ETL skript nahrávající data do tabulky *ucitel_statistika* část třetí



Obrázek D.17: ETL skript nahrávající data do tabulky *ucitel_statistika* část čtvrtá



Obrázek D.18: ETL skript nahrávající data do tabulky *ucitel_statistika* část pátá

Příloha E

Mapovaní tabulek datových tržišť na tabulky centrálního datového skladu

Sloupec	Popis	Zdroj	Transformace
otazka_id	Identifikátor otázky.	anketni_otazka.otazka_id	Nezměněno
otazka_text	Znění otázky.	anketni_otazka.zneni	Nezměněno
otazka_vaha	Váha otázky určující jestli se jedná o otázku z oficiální hodnocení či nikoliv	anketni_otazka.vaha	Nezměněno.

Tabulka E.1: Mapování tabulky *d_otazka* na centrální datový sklad

Sloupec	Popis	Zdroj	Transformace
predmet_id	Identifikátor předmětu.	predemt.predmet_id	Nezměněno
nazev	Kód a název předmětu.	predmet.nazev predmet.kod	Spojení kódu a názvu
katedra	Název katedry předmětu.	katedra.nazev	Nezměněno.
fakulta	Název fakulty předmětu.	fakulta.nazev	Nezměněno.

Tabulka E.2: Mapování tabulky *d_predmet* na centrální datový sklad

Sloupec	Popis	Zdroj	Transformace
ucitel_id	Identifikátor učitele.	ucitel.ucitel_id	Nezměněno
jmeno	Spojené jméno, příjmení a tituly	ucitel.titul_pred ucitel.jemno ucitel.prijmeni ucitel.titul_za	Spojení titulů, jména a příjmení.
katedra	Název katedry učitele.	katedra.nazev	Nezměněno.
fakulta	Název fakulty učitele.	fakulta.nazev	Nezměněno.

Tabulka E.3: Mapování tabulky *d_ucitel* na centrální datový sklad

Sloupec	Popis	Zdroj	Transformace
predmet_id	Identifikátor předmětu.	predmet.predmet_id	Nezměněno
otazka_od	identifikátor otázky.	otazka.otazka_id	Nezměněno.
odpoved_text	Text odpovědi.	anketni_odpoved.zneni	Nezměněno.
odpoved_vaha	Váha odpovědi.	anketni_odpoved.vaha	Nezměněno.
semestr	Název semestru.	semestr.nazev	Nezměněno.
role_ucitele	Název role predmetu.	role_ucitele.nazev_role	Nezměněno.
zapsanych_studentu	Počet studentů zapsaných do předmětu.	predmet_statistika.zapsanych_studentu	Nezměněno.
hlasovalo_studentu	Počet studentů, kteří hlasovali v předmětu.	predmet_statistika.hlasovalo	Nezměněno.

Tabulka E.4: Mapování tabulky *f_hodnoceni_predmetu* na centrální datový sklad

Sloupec	Popis	Zdroj	Transformace
ucitel_id	Identifikátor učitele.	ucitel.ucitel_id	Nezměněno.
predmet_id	Identifikátor předmětu.	predmet.predmet_id	Nezměněno
otazka_od	identifikátor otázky.	otazka.otazka_id	Nezměněno.
odpoved_text	Text odpovědi.	anketni_odpoved.zneni	Nezměněno.
odpoved_vaha	Váha odpovědi.	anketni_odpoved.vaha	Nezměněno.
semestr	Název semestru.	semestr.nazev	Nezměněno.
role_ucitele	Název role učitele.	role_ucitele.nazev_role	Nezměněno.
zapsanych_studentu	Počet studentů zapsaných k učiteli v dané roli a předmětu.	ucitel_statistika.zapsanych_studentu	Nezměněno.
hlasovalo_studentu	Počet studentů, kteří hlasovali v dané role a předmětu.	ucitel_statistika.hlasovalo	Nezměněno.

Tabulka E.5: Mapování tabulky *f_hodnoceni_ucitelu* na centrální datový sklad

Sloupec	**Popis**	**Zdroj**	**Transformace**
predmet_id	Identifikátor předmětu.	predmet.predmet_id	Nezměněno
semestr	Název semestru.	semestr.nazev	Nezměněno.
zapsanych_studentu	Počet studentů zapsaných na předmět v rámci semestru.	predmet_statistika.zapsanych_studentu	Nezměněno.
hlasovalo_studentu	Počet studentů, kteří hlasovali o předmětu v daném semestru.	predmet_statistika.hlasovalo	Nezměněno.
dokoncilo_studentu	Počet studentů, kteří dokončili předmět.	predmet_statistika.uspesne_dokoncilo	Nezměněno.

Tabulka E.6: Mapování tabulky *f_statistika_predmetu* na centrální datový sklad

Příloha F

Obsah přiloženého CD

- `README.txt` ..... stručný popis obsahu CD
- `Implementace` ..... zdrojové kódy implementace
  - `Dumpy jednotlivých databází` .. Dumpy jednotlivých částí datového skladu
  - `Ostatní` ..... Ostatní soubory implementace
    - `Soubory pro Schema Workbench` .. XML soubory na tvorbu OLAP kostek
    - `SQL skripty na tvorbu DW` ..... SQL skripty na tvorbu datového skladu
    - `Stage` ..... Csv soubory s exporty tabulek za jednotlivé semestry
  - `Soubory Pentaho Data Integration` ..... Spustitelné soubory s ETL skripty
- `text` ..... text práce
  - `DP_JG.pdf` ..... text práce ve formátu PDF
  - `DP_JG.ps` ..... text práce ve formátu PS
  - `LaTEx.zip` ..... zdrojová forma práce ve formátu LaTeX