Vysoká škola ekonomická v Praze

Fakulta informatiky a statistiky

Katedra informačních technologií

Studijní program: Aplikovaná informatika

Obor: Informační systémy a technologie

Diplomant:	**Bc. Jakub Erlebach**
Vedoucí diplomové práce:	**doc. Ing. Ota Novotný, Ph.D.**
Oponent diplomové práce:	**Ing. Miloš Maryška**

Využití Business Intelligence při analýze návštěvnosti internetového obchodu

školní rok 2008/2009

Prohlášení

Prohlašuji, že jsem diplomovou práci zpracoval samostatně a že jsem uvedl všechny použité prameny a literaturu, ze kterých jsem čerpal.

V Praze dne 29.6 2009

…….
Podpis

Poděkování

Tímto bych chtěl poděkovat vedoucímu své diplomové práce, doc. Ing. Otovi Novotnému, Ph.D., za pomoc a veškerý čas, který mi věnoval.

Abstrakt

V současné době ekonomické krize je více než kdy jindy obtížné udržovat staré a nacházet nové odběratele vyráběných produktů. Stejně tak i s rozvojem prodeje zboží na internetu stále narůstá počet konkurentů, kteří také chtějí co největší podíl tržního koláče.

Cílem této práce je využít technologie a postupy Business Intelligence při analýze návštěvnosti internetového obchodu k získání takových informací a znalostí, které by následně měly napomoci k překonání těchto dvou zmiňovaných problémů, se kterými se dnes většina firem podnikajících na internetu potýká.

V první, teoretické části práce jsou nejprve popsány základní principy a postupy optimalizace internetových stránek jak z hlediska vyhledávačů, tak také uživatelů. Následující kapitola se věnuje teorii analýzy návštěvnosti, jsou zde stanovena jak obecná, tak i vlastní kritéria výběru nástroje této analýzy. Na základě těchto kritérií jsou následně nástroje popsány a vyhodnoceny. Poslední teoretická kapitola je věnována Business Intelligence, zejména pak teorii tvorby datového skladu a postupu tvorby BI projektu.

Úvod do praktické části zajišťuje popis současné situace internetového obchodu. V této části je zmíněna současná architektura řešení, postavení firmy na trhu a její vize do budoucna.

Na základě vybraného nástroje analýzy a teoretických postupů stanovených v předchozích částech je pak navrženo konkrétní BI řešení a popsán postup jeho implementace, některé výstupy a také náklady tohoto řešení.

Hlavním přínosem práce je získání nových a jinak těžko dostupných reportů o návštěvnosti, které by měly usnadnit a zlepšit optimalizaci stránek a napomoci tak firmě posílit svou pozici na trhu. Další přínos je možné spatřit v ukázce implementace BI řešení v měřítku malých podniků i s vyčíslením nákladů na ni.

Klíčová slova: SEO, Business Intelligence, Analýza návštěvnosti, Internetový obchod

Abstract

In the time of recent economic crisis is more than ever difficult to keep old and find new customers. Along with the growth of sales on the Internet there are also a growing number of competitors, which want their share of the market pie.

The aim of this thesis is to use technologies and processes of Business Intelligence with web analysis of an electronic store in order to get new information and knowledge that should later on help with overcoming the two mentioned problems most companies running an internet business are today dealing with.

In the first, theoretical part of this thesis there are described basic principles and processes of search engine optimization in term of both users and search engines. Next chapter follows the theory of web analysis. There are also described general as well as own selection criteria of the web analysis tool. Upon these criteria there are these tools described and evaluated. Last theoretical chapter deals with Business Intelligence, in particular with a data warehouse theory and with a process of a BI project.

Introduction to the practical part is formed by a description of a current e-shop situation. In this part there is also mentioned current architecture solution, standing of the firm and its future vision.

On the basis of selected web analysis tool and theoretical knowledge described in previous chapters there is than designed concrete solution and described process of its implementation, some of its outputs and also its costs.

The main contribution of this thesis I see in gaining new and otherwise hardly accessible reports about the visit rate, which should ease and improve the process of website optimization and therefore help the company to strengthen its market position. Another contribution would be in demonstration of the BI implementation on a small business scale also with its cost accounting.

Keywords: SEO, Business Intelligence, Web Analysis, Electronic shop

Obsah

1 ÚVOD 8

1.1 Úvod do problematiky 8
1.2 Vymezení tématu práce a důvod výběru tématu 8
1.3 Cíle práce 8
1.4 Cíloví čtenáři 9
1.5 Struktura práce 9
1.6 Vstupy 10
1.7 Výstupy 10
1.8 Předpokládané přínosy a omezení práce 11

2 SEO 12

2.1 Základní techniky a postupy optimalizace 13
2.1.1 Úprava obsahu stránek 13
2.1.2 Výběr klíčových slov 14
2.2 Externí faktory ovlivňující pozici stránek ve vyhledávačích 15
2.3 Placené kampaně 16

3 NÁSTROJE PRO ANALÝZU NÁVŠTĚVNOSTI 17

3.1 Kritéria výběru vhodného nástroje 18
3.1.1 Obecná kritéria 18
3.1.2 Vlastní kritéria 22
3.2 Popis nástrojů analýzy 23
3.2.1 Piwik 24
3.2.2 Google Analytics 26
3.2.3 TOPlist.cz 27
3.2.4 Další nástroje 29
3.3 Výběr nástroje analýzy 29

4 BUSINESS INTELLIGENCE 31

4.1 Datové sklady 32
4.1.1 Metody budování datového skladu 33
4.1.2 Transformační nástroje (ETL) 35
4.2 Postup projektu 35

5 SOUČASNÁ SITUACE 38

5.1 Architektura řešení 38
5.2 Situace v oblasti SEO 39
5.3 Postavení na trhu 40
5.4 Vize do budoucna 41

6 IMPLEMENTACE BI ŘEŠENÍ 42

6.1 Úprava nástroje analýzy 42
6.1.1 Úprava databáze 42
6.1.2 Úprava zdrojového kódu 42
6.2 Úvodní studie 43
6.2.1 Definování cílů a efektů BI 43
6.2.2 Katalog uživatelů 44
6.2.3 Analýza požadavků 44

6.2.4 Určení oblastí řešení a jejich priorit ... 45
6.2.5 Návrh BI architektury ... 46
6.2.6 Harmonogram řešení ... 46
6.3 Analýza BI řešení ... 47
6.3.1 Zdroje dat ... 47
6.3.2 Specifikace změn v produkčních databázích ... 48
6.4 Analýza datového skladu ... 48
6.4.1 Dimenze ... 49
6.4.2 Ukazatele a vztahy ukazatelů a dimenzí ... 54
6.5 Model datového skladu ... 55
6.6 ETL ... 56
6.7 Vytvoření OLAP kostek ... 59
6.8 Ukázka výstupů a jejich využití v praxi ... 60
6.8.1 Další možné sestavy ... 65
6.9 Náklady projektu ... 65
7 ZÁVĚR ... 67
7.1 Srovnání předpokládaných a dosažených cílů ... 67
7.2 Přínosy ... 68
7.3 Omezení ... 68
7.4 Návrhy na další řešení ... 68
LITERATURA ... 69
TERMINOLOGICKÝ SLOVNÍK ... 70
SEZNAM TABULEK ... 72
SEZNAM OBRÁZKŮ ... 73
PŘÍLOHY ... 74
Příloha 1 – úprava tabulky akcí nástroje Piwik ... 74
Příloha 2 – dekódování klíčových slov a zápis jejich pozic ... 75

1 Úvod

1.1 Úvod do problematiky

Business Intelligence (dále jen BI) je dnes již pojem velmi rozšířený nejen co se týče jeho zmiňování v odborných kruzích, ale také co se týče jeho implementace. S masivním rozmachem informačních systémů v devadesátých letech také rapidně narostl objem dat, která podniky generují a uchovávají. Dá se říci, že technologie BI jsou nyní již standardem pro každou větší a střední firmu využívající nějaký ten informační systém. Tak jednoznačné to už ale není u firem malých, které buď nemají data, nemají dostatek relevantních informací nebo nemají finance pro jejich implementaci.

1.2 Vymezení tématu práce a důvod výběru tématu

Tématem této práce bude využití technologií a postupů BI při analýze návštěvnosti internetového obchodu. Analýzou návštěvnosti rozumím chování návštěvníků na stránkách webu, důvody a zdroje jejich návštěv ve spojitosti s cílem webu, tj. prodejem výrobků. Důvodem výběru tohoto tématu je jednak moje dlouholetá činnost v oblasti webových technologií a jejich využití. V poslední době se také zaměřuji na optimalizaci webových stránek pro vyhledávače. Tuto oblast bych tímto rád také rozšířil o optimalizaci webu podle chování návštěvníků. Dalším důvodem mého výběru je snaha o rozvoj internetového obchodu, na kterém se podílím a který také bude mým zdrojem v této práci. V neposlední řadě je důvodem také můj zájem o oblast Business Intelligence a zvláště pak možnosti jejího využití v menších podnicích. Tyto poznatky bych pak rád aplikoval ve společnosti, kde pracuji a která se soustředí právě na malé a střední podniky.

1.3 Cíle práce

Práce si klade za cíl analyzovat návštěvnost internetového obchodu a připravit výstupy, které následně pomohou při lepší optimalizaci a funkčnosti webových stránek internetového obchodu.

Dílčí cíle

1) přiblížit a porovnat některé služby poskytující analýzu návštěvnosti
2) představit požadavky na datovou základnu
3) vyčíslit náklady na implementaci
4) ukázat možnosti využití BI i pro malé podniky

1.4 Cíloví čtenáři

Práce je určena především pro majitele nebo manažery internetových obchodů. Zároveň by měla být přínosná i pro SEO experty, kterým by měla poskytnout další rozměr k jejich práci. V neposlední řadě si své najdou i majitelé malých firem, kterým by měla sloužit jako ukázka využití technologií BI v měřítku malých firem.

1.5 Struktura práce

Práce bude rozdělena do šesti dílčích celků, přičemž první tři by měly poskytnout teoretický základ pro tuto práci. Další dva se budou věnovat přípravě implementace a závěrečná část již bude obsahovat konkrétní výstupy pro naplnění hlavního cíle práce.

1.5.1 SEO

V této kapitole přiblížím metodiku optimalizace pro vyhledávače. Zmíním, jaké postupy jsou nejúčinnější a jaké informace jsou potřebné pro správnou a účinnou implementaci.

1.5.2 Nástroje pro analýzu návštěvnosti

V této kapitole se zaměřím na dostupné nástroje pro analýzu návštěvnosti webových stránek. Tyto nástroje vyhodnotím, vypíšu jejich klady a zápory a následně vyberu jeden z nástrojů, který také využiji v kapitole 1.5.5. Vzhledem k tomu, že se zaměřuji spíše na menší firmy, budu analyzovat pouze nástroje, které jsou zdarma.

1.5.3 Teorie BI

Pro analýzu dat návštěvnosti bude nutné využít technologií BI. V této kapitole popíši možné přístupy k realizaci tohoto řešení a okomentuji zde také vybraný postup.

1.5.4 Současná situace

V této kapitole bude popsána současná situace internetového obchodu, základní nedostatky a problémy, s kterými se obchod setkává a k jejichž řešení pomohou použité postupy popsané v této práci.

1.5.5 Implementace BI řešení

Tato kapitola bude obsahovat postup implementace a bude se již konkrétně týkat analyzovaného internetového obchodu. Popíši zde, co všechno je nutné udělat k realizaci cílů. Popisované činnosti budou následující

- vypracování úvodní studie
- vytvoření datové základny
- návrh výstupů
- odhad pracnosti a nákladů implementace

Další část této kapitoly bude obsahovat praktické ukázky analýzy návštěvnosti. Budou zde ukázány výstupy a grafy, které následně budou sloužit odpovědným osobám při jejich rozhodování o strategii internetového obchodu.

1.5.6 Závěr

V závěru práce zhodnotím míru dosažení jak jednotlivých cílů, tak také cíle hlavního – tedy jak se podařilo analyzovat návštěvnost webu a zda tato analýza byla přínosná.

1.6 Vstupy

Primárními vstupy budou data získaná jednak z vlastní databáze internetového obchodu, dále pak data získaná z vybrané služby analyzující návštěvnost. Dalšími vstupy budou znalosti získané studiem relevantních předmětů na VŠE, znalosti získané z odborné literatury a znalosti z mé praxe v oblasti webových technologií.

1.7 Výstupy

Výstupem práce bude jednak implementace technologií BI pro daný internetový obchod a dále pak analýza získaných dat a jejich interpretace.

1.8 Předpokládané přínosy a omezení práce

Hlavním přínosem teoretické části by mělo být ucelení poznatků umožňující aplikaci analýzy návštěvnosti a to nejen pro mnou zkoumaný web, ale i pro weby podobné. Přínosem praktické části by pak mělo být získání nových poznatků a souvislostí, které následně umožní efektivnější optimalizaci stránek internetového obchodu, což by mělo vést jak ke zvýšení návštěvnosti, tak také ke zvýšení obratu a zisku firmy.

Omezením práce může být nedostatek dat pro vypracování některých analýz, případně obtížné propojení databáze internetového obchodu a online služeb monitorujících návštěvnost.

2 SEO

Tím, jak v posledních letech stále narůstá počet lidí nakupujících na internetu[3], tím roste důležitost být na internetu vidět. K tomuto účelu slouží SEO (Search Engine Optimization, česky optimalizace pro vyhledávače). SEO můžeme definovat jako:

> Proces zvyšování objemu a kvality návštěvnosti webových stránek z vyhledávačů skrze přirozené výsledky vyhledávání. Obvykle čím dříve je stránka umístěna ve výsledcích vyhledávání, nebo čím výše je hodnocena, tím více hledajících uživatelů tuto stránku navštíví [15].

> Nebo jako:

> Proces výběru cílených klíčových slov vztahujících se k dané webové stránce a zajištění, že se tato slova po zadání do vyhledávače dobře umístí [1].

Z výše uvedených definic je patrné, že klasické SEO je pouze technologický postup jak vyhovět algoritmům vyhledávačů a na oplátku v nich získat přední pozice. Já bych ale do tohoto pojmu ještě zahrnul pojem další a tím je usability, neboli použitelnost. Ta je definována jako

> Schopnost uživatele pochopit a umět používat webové stránky tak, aby mohl dosáhnout svých cílů (např. objednání zboží) [12].

Přidáním použitelnosti totiž přidáme k optimalizaci také lidský rozměr. Ten je totiž v konečném důsledku rozhodující. Pouze člověk rozhoduje o tom, zda provede akci, kterou bychom rádi, aby provedl. Samotná návštěva a následné opuštění webu nám žádné peníze nevydělá.

Ve výsledku již tedy nebudu v této práci chápat SEO pouze jako optimalizaci pro vyhledávače, ale také jako optimalizaci pro uživatele. Poměrně výstižně takovouto optimalizaci označil na svém blogu Dr. Peter J. Meyers, když ji nazval SUO (Search and User Optimization) [6].

Někdo by mohl předpokládat, že zaměření se na uživatele ještě neznamená dobré pozice ve vyhledávačích, ale spíše opak je pravdou. Dobré SEO je nejen dobré pro vyhledávače, ale i pro uživatele.

2.1 Základní techniky a postupy optimalizace

V této části představím nejdůležitější aspekty optimalizace pro vyhledávače, tj. úpravu obsahu a HTML (HyperText Markup Language, česky značkovací jazyk pro hypertext) kódu stránky a výběr klíčových slov.

2.1.1 Úprava obsahu stránek

Stránka z hlediska HTML členění má několik nejdůležitějších prvků. Tím hlavním je titulek. Dobrý titulek stránky je základem úspěchu. Je to totiž především titulek, který rozhoduje, zda návštěvník na odkaz klikne nebo ne. Takovýto titulek musí být relativně krátký (do 30 znaků) a měl by být výstižný a poutavý. V tomto případě, pokud bychom se soustředili pouze na vyhledávače, pak bychom do něj mohli napsat pouze změť několika klíčových slov a umísťovat se na slušné pozici. Ale kliknul by na tento odkaz někdo? Nedal by raději přednost před nějakou smysluplnou větou či souslovím?

Technika psaní titulků je v této oblasti velmi důležitá. Uživatelé většinou nemají příliš mnoho času a jejich rozhodování zda kliknout či ne probíhá v intervalu několika málo vteřin. Proto je důležité zaujmout a přesvědčit daného uživatele, že zrovna na vaší stránce najde to, co hledá.

S tímto rozhodováním také souvisí jak obsah, tak i popisek stránky. Vyhledávač se chová následovně. Při vyhledávání nějakého slova či slovního spojení vypíše titulek stránky a pod ním buď popis (v případě, že se klíčová slova v textu nevyskytují) – méně častý případ, případně vypíše část textu, ve kterém se hledaný výraz vyskytuje. Z tohoto plyne pak další důležitý aspekt optimalizace. A tím jsou kvalitní texty stránek.

Při psaní kódu stránek bychom se měli držet následujících pravidel

- kód by měl být dobře strukturovaný – měli bychom se držet struktury, jako bychom psali nějaký textový dokument, tzn. členit text na nadpisy a odstavce s tím, že hlavní nadpis stránky je pouze jeden
- kód by měl být čistý – v kódu by se nemělo vyskytovat příliš mnoho HTML značek, definic stylů a javascriptu
- vlastní text stránky by se měl v kódu vyskytovat co nejvýše – tzn., že všechen zbytečný kód by měl být odsunut na konec stránky

Pro psaní obsahu pak platí následující pravidla

- měli bychom vybrat maximálně 5 slov, která jednak vystihují danou stránku a která bychom použili my, kdybychom naši stránku hledali
- obsah by měl být dobře strukturovaný, srozumitelný a poutavý
- měli bychom také přesvědčit návštěvníka k nějaké akci, ať už je to koupě výrobku, registrace na webu či doporučení známým

Dalším důležitým aspektem optimalizace je výběr klíčových slov.

2.1.2 Výběr klíčových slov

Klíčové slovo je slovo nebo fráze popisující produkty nebo služby organizace (nebo nějaký jiný důležitý obsah stránek). Může také znamenat slovo nebo frázi, která je vložena jako dotaz do vyhledávače [4].

Existují tři hlavní faktory ovlivňující výběr klíčových slov.

1) Relevance

Pokud se rozhodneme optimalizovat na slova, která s naším webem nesouvisí, přitáhneme tak možná návštěvníky, kteří ale nenajdou, co hledali a ihned web opustí. Takováto snaha je pak tedy zbytečná. Na druhou stranu, pokud se nám podaří vybrat ta správná slova, pak je i mnohem větší šance, že návštěvník udělá to, co si od jeho návštěvy slibujeme.

2) Vyhledávanost

tedy to, kolikrát je dané klíčové slovo zadáno do vyhledávače. Tento faktor působí v částečné kontradikci s tím prvním. Většinou je to totiž tak, že ta nejvíce vyhledávaná slova jsou také nejvíce obecná a tudíž i méně relevantní. V rámci analýzy je tedy nutné alespoň částečně odhadnout, jak která slovní spojení jsou vyhledávaná. K tomu slouží několik nástrojů, např. Google Adwords Keyword Tool[1] či Wordtracker[2].

[1] https://adwords.google.com/select/KeywordToolExternal
[2] http://www.wordtracker.com/

3) Konkurence

Čím více stránek je optimalizováno na určité slovní spojení, tím větší úsilí je nutné investovat do jeho prosazení na přední místa. Je tedy nutné zvážit, zda se takové úsilí vyplatí vynaložit, nebo zda zvolit nějaké jiné, méně konkurenční slovo. Ke zjištění míry konkurence klíčového slova je možné například využít funkci vyhledávače Google, kdy při zadání allintitle:klíčové slovo, budou nalezeny pouze takové stránky, které mají toto slovo v titulku. Vzhledem k tomu, co již bylo řečeno o důležitosti titulku stránky v rámci optimalizace, je z těchto výsledků možné poměrně dobře odhadnout, jak je které klíčové slovo konkurenční.

Z výše uvedeného tedy vyplývá, že ideální klíčové slovo má být vysoce relevantní, vysoce vyhledávané a málo konkurenční. Bohužel v realitě najít takové slovní spojení je téměř nemožné a proto během analýzy klíčových slov je nutné určit priority a následně zvolit určitý kompromis mezi těmito třemi faktory.

Pokud se nám již podařilo dostat cíleného uživatele na náš web, přichází další fáze a to, přimět ho učinit to, co chceme, např. nakoupit zboží. To je ona zmiňovaná optimalizaci pro uživatele.

Pro správné pochopení chování je nutné toto chování zaznamenat a analyzovat. Musíme vědět, na kterou stránku návštěvník přišel, kam se vydal dále a kde skončil. Z těchto dat pak můžeme vycházet při dodatečných úpravách obsahu webových stránek.

2.2 Externí faktory ovlivňující pozici stránek ve vyhledávačích

Již byly zmíněny faktory, které závisí pouze na práci správce stránek. Existují ale také faktory, které jsou ovlivnitelné hůře. Mezi takové patří množství odkazů vedoucích z jiných internetových stránek a také jejich relevance. Tyto odkazy totiž většina vyhledávačů vyhodnocuje a utváří z nich jakýsi hodnotící algoritmus. Na jeho základě pak umisťuje stránky do pozic. Tento algoritmus je tajný, zejména z důvodu zneužívání, ale i tak je o něm něco známo. I zde je důležitá relevance. Pokud například budeme prodávat meče a odkaz povede ze stránek zaměřených na auta, pak tento odkaz nebude mít takovou váhu, jako by měl ze stránek podobného zaměření. Některé odkazy mohou dokonce znamenat penalizaci. Jsou to zejména ty z tzv. klikacích farem, což jsou stránky obsahující v podstatě pouze odkazy na

velké množství různých webů. Proto i zde je potřeba klást důraz zejména na relevanci a méně na množství.

2.3 Placené kampaně

Tato oblast již není přímou součástí SEO, ale spíše její přidruženou částí. Hlavní rozdíl oproti přirozené optimalizaci je ten, že zde nemusíme vynakládat úsilí pro dosažení pozic, ale stačí zde za tyto pozice zaplatit. Tento fakt ale také způsobuje mnohem větší potřebu monitoringu chování takto příchozích návštěvníků, jelikož co rozhodně nechceme, je platit za návštěvy, které nám nic nepřinesou. Co se týče tvorby PPC (Pay-Per-Click, česky platba za proklik) kampaně, i zde platí požadavky na klíčová slova, které jsem vysvětlil výše.

Princip Pay Per Click spočívá v tom, že inzerent za reklamu platí až ve chvíli, kdy zaúčinkovala. Inzerent tedy platí až ve chvíli, kdy na reklamu někdo klikl. Oproti klasickým bannerům nebo jiné plošné reklamě má tu výhodu, že zákazník má přesnou kontrolu nad svými výdaji, tedy že zaplatí jen tolik prokliků, potažmo návštěvníků, kolik chce. PPC reklama je velmi dobře měřitelná a plánovatelná [14].

Hlavní výhodou těchto kampaní je jejich okamžitý efekt. Při klasické optimalizace trvá alespoň tři měsíce, než se pozice klíčových slov zlepší a úsilí vložené do optimalizace se vrátí.

3 Nástroje pro analýzu návštěvnosti

Pojem analýza návštěvnosti je možné vykládat mnoha různými způsoby. Někdo si může pod tímto pojmem představit například jen sledování počtu návštěvníků webu, jiný zase pouze seznam a poměr zdrojů návštěvnosti, atd. Já vybral definici, která zřejmě nejvíce vystihuje analýzu návštěvnosti v tom smyslu, v jakém se jí budu zabývat já v této práci.

> Analýza návštěvnosti je monitorování a vykazování dat o použití internetové prezentace, díky nimž může firma lépe porozumět komplexním interakcím mezi jednáním návštěvníka a nabídkami zprostředkovanými internetovou prezentaci a tohoto porozumění využít pro zvýšení prodejů nebo loajality zákazníků [9].

Pokud se již pro takovouto analýzu rozhodneme, lze ji provést v podstatě třemi různými způsoby. Jedním z nich je využít některý z mnoha nástrojů, které jsou schopné tato data uschovávat a zpracovávat do sestav a přehledů. Druhým způsobem je si takový nástroj vytvořit svépomocí. A třetím je kombinace těchto dvou přístupů. Který přístup zvolit pak závisí hlavně na tom, co si od takovéto analýzy slibujeme. Vzhledem k faktu, že nedisponuji takovými časovými prostředky, abych si mohl takový nástroj vytvořit sám, využiji k mé analýze některý z již vytvořených nástrojů.

Tyto nástroje se liší jak z hlediska komplexnosti, tak i z hlediska jejich ceny. I tak ale mají společné prvky a tím jsou tyto základní přehledy

- počet návštěv a jejich rozklad z hlediska geografického, z hlediska prohlížeče, operačního systému, rozlišení obrazovky, atd.,
- členění návštěv podle zdrojů návštěvnosti,
- seznam klíčových slov, přes která návštěvníci na web přicházejí.

Takto nashromážděná data se pak následně dají využít k rozvoji webových stránek. Jednotlivé přehledy a sestavy jsou dobrým zdrojem k rozboru stránek a k jejich případné úpravě, zvláště s ohledem na optimalizaci jak pro vyhledávače, tak i pro uživatele.

Existují ale i požadavky, které jsou poměrně specifické a ne všechny nástroje jim vyhovují. Takové požadavky mohou být následující

- mít možnost spojit jednotlivé návštěvy s těmi, které si evidujeme na vlastním webu,
- exportovat přehledy a jednoduše je zapojit do vlastního systému,

- mít přehled o pohybu jednotlivých návštěvníků na webu,
- rychlost, aktuálnost a přehlednost jednotlivých sestav,
- množství přehledů.

Těchto požadavků může být celá řada, zmínil jsem tedy převážně ty, které jednotlivé nástroje odlišují.

3.1 Kritéria výběru vhodného nástroje

V této části představím nejprve obecná kritéria výběru nástroje webové analýzy a následně kritéria, která budu při výběru nástroje aplikovat pro potřeby této práce.

3.1.1 Obecná kritéria

Při výčtu obecných kritérií budu vycházet z faktorů a členění vypracovaných Zbyňkem Mužíkem [7].

Externí faktory

1) Účel investice do WA

Pří výběru WA aplikací je třeba začít od webových stránek, které budou pomocí ní analyzovány. V této fázi je vhodné si položit následující otázky:

- Jaké jsou strategické cíle analyzovaných webových stránek?
- Jak je úspěch těchto stránek měřen?
- Jaká je hodnota stránek pro uživatele?

Odpovědi na tyto otázky pomohou lépe pochopit faktický význam daných stránek, konkrétněji si uvědomit účel nasazení WA aplikací a zhodnotit celkový možný efekt tohoto nasazení.

2) Náklady WA řešení

Pro porovnání nákladů se doporučuje použít tříleté TCO jednotlivých řešení, zahrnující veškeré poplatky např. za licence, registraci, uložení dat, uživatelsky definované reporty, mimořádné projekty, přechodná zvýšení provozu na stránkách (relevantní u modelu ASP, kde cena koreluje s objemem provozu). Ceny se pohybují následovně.

- Jednoduchá WA řešení – začínají od open source řešení poskytovaných zdarma až po aplikace okolo 1000 USD nebo ASP řešení pod 1000 USD ročně. Poskytují však pouze základní funkcionalitu a nemusejí být vždy dostatečně škálovatelná.

- Pokročilá softwarová řešení – tyto produkty mají různé cenové politiky. Cena se může odvíjet od počtu serverů, domén nebo zobrazených stránek. Pro licenci za první rok zaplatí menší organizace okolo 10 000 USD a velké globální společnosti asi 50 000 USD. Tyto částky zahrnují jak poplatky za samotný produkt, tak cenu údržby, hardware, databáze, uživatelské podpory a vývoje vlastních specifických reportů.

- ASP. Nejčastější cenovou politikou je tzv. „cost per thousand“, tedy fixní náklady za 1000 zobrazených stránek. Náklady na zobrazení jedné stránky se typicky pohybují okolo 0,02 - 0,10 USD. Celkové náklady pro stránky, které mají méně než milion návštěvníků ročně, se pohybují okolo 15 000 USD ročně, ale cena rychle roste s rostoucím provozem na stránkách.

Je také vhodné určit předpokládané ROI WA projektu., které je poměrně jednoduché determinovat v případě komerčních stránek, ale pokud se daný web zabývá například uživatelskou podporou, nebo se jedná o interní web firmy, může být určení ROI velmi komplikované.

V případě hodnocení ROI je velmi důležité zohlednit časové hledisko, protože pokud nebude použití WA jen jednorázovým projektem, ale tyto nástroje mají být užívány trvale, je jasné že udržet počáteční ROI bude velmi těžké.

Interní faktory

1) Uživatelské požadavky na funkcionalitu

V organizaci jsou v zásadě tři skupiny uživatelů WA aplikací, kteří mají na tyto aplikace různé požadavky.

- Pracovníci zajišťující operativu IT – tato skupina využívá WA pro monitorování provozu, plánování technických kapacit, ladění výkonu a eventuelně zvažování některých bezpečnostních rizik. V rámci této skupiny jsou velmi populární různá open source řešení, která postačují jejím účelům, nemají však dostatečnou

funkcionalitu pro další skupiny uživatelů. Z funkcionality jde většinou pouze o (z technického pohledu) základní požadavky, jako např.:

- počet hitů za časovou jednotku,
- objem přenesených dat,
- síťové adresy uživatelů,
- nejpopulárnější stránky,
- atd.

- Vývojáři a designéři stránek – tato skupina patří k nejčastějším uživatelům WA, proto by měla být jejím požadavkům přikládána velká důležitost. Tvůrci stránek používají WA nástroje především k pochopení chování uživatelů na webu, tyto poznatky jim pomáhají při optimalizaci stránek. Vedle požadavků předchozí skupiny je tedy zajímá např.

  - kudy se uživatelé na stránky dostávají,
  - jaké jsou obvyklé sekvence stránek, které procházejí,
  - kde jejich návštěva končí,
  - atd.

- Business uživatelé – skupina uživatelů, zahrnující členy managementu, marketingového oddělení, atd., vidí web jako kanál, prostřednictvím něhož se snaží o naplňování cílů organizace. Z množiny funkcí WA nástrojů, používají pokročilé analýzy, jako např.:

  - profilování uživatelů,
  - analýzu klíčových slov zadaných do vyhledávače, které vedou uživatele na tyto stránky.
  - řízení marketingových kampaní,
  - atd.

Je třeba si uvědomit, že ačkoliv nejčastější uživatelé WA aplikací patří do druhé skupiny, většina komerčních nástrojů je zacílena právě na potřeby business uživatelů, protože pokročilé funkce, které používají, jsou na WA nástrojích nejdražší a právě tito lidé o investici do WA rozhodují. Požadavky třetí skupiny jsou klíčové pro rozhodnutí, zda organizace zvolí jednoduché nebo pokročilé řešení.

2) Požadavky na uživatelské rozhraní

Různé skupiny uživatelů preferují různá uživatelská rozhraní. Business uživatelé, na rozdíl od technicky orientovaných uživatelů často upřednostňují rozhraní, které je vizuálně přívětivé a intuitivní, což je však často na úkor použitelnosti. Je důležité nenechat se při rozhodování zlákat vzhledem uživatelského rozhraní, ale zaměřit se na to, zda uživatelům umožní zobrazit odpovědi na jejich otázky.

3) Dostupnost podpory, školení a dokumentace

Některé organizace disponují dostatkem kvalifikovaného personálu na to, aby si s provozem WA aplikace poradily samy, jiné však potřebují externí uživatelskou podporu k tomu, aby plně využili potenciál těchto nástrojů. Podpora je většinou lepší v oblasti ASP nástrojů, protože tyto společnosti jsou v kontaktu přímo s business uživateli. Existuje také několik konzultantských společností, které zákazníkům pomáhají definovat strategii v této oblasti a proces integrace WA do širšího kontextu CRM.

4) Životaschopnost řešení

Na trhu WA produktů se pohybuje mnoho výrobců a je velmi pravděpodobné, že velká část z nich bude silnými hráči odsunuta do okrajových částí trhu, některým z nich pohlcena, nebo úplně zanikne. Mnoho zákazníků, pro které je důležitá nižší cena řešení se může rozhodnout pro některou z těchto firem. Neustálé přeskakování od jednoho poskytovatele ke druhému se může ze začátku jevit jako velmi finančně výhodné, je ale třeba si uvědomit, že většinou dochází ke ztrátě historických dat a customizovaných reportů a musí neustále docházet k přeškolování uživatelů WA nástrojů.

Je tedy nutné při výběru poskytovatele přihlédnout k referencím, prověřit jeho finanční zdraví a to zda má zákazníky v našem oboru podnikání.

V případě open source řešení je situace v zásadě podobná. Je třeba dbát na to, aby se jednalo o živý a organizovaně vyvíjený produkt, v dostatečné fázi zralosti. V případě OS projektů jsou rizika ztráty datové kontinuity nižší, protože se dá předpokládat, že projekt jen tak nezanikne a i kdyby se tak stalo, dostupnost zdrojových kódů otevírá možnost řešit plynulý přechod na jiné řešení na aplikační úrovni.

5) Schopnost integrace s dalšími aplikacemi a DWH

I když nehodláme okamžitě integrovat WA s jinými aplikacemi, je vhodné brát zřetel na následující.

- Nejdůležitější je integrace s marketingovými nástroji, které řídí emailové kampaně, nakupování bannerových reklam a klíčových slov ve vyhledávačích.
- Některé nástroje umožňují export dat do MS Excelu, kde je následně možné provádět vlastní analýzy.
- Informace o produktech a zákaznících mohou být vloženy do WA nástroje, aby byly reporty srozumitelnější.
- Data z WA systému mohou být importována do CRM systému nebo datového skladu, aby mohla být použita pro vícekanálovou marketingovou analýzu.
- Schopnost práce s formáty na bázi XML umožňuje oboustrannou integraci s téměř jakýmkoliv systémem.

3.1.2 Vlastní kritéria

Vzhledem k tomu, že v rámci mé práce budu přistupovat k nástroji analýzy návštěvnosti spíše technologicky, tj. spíše jako k nástroji pro sběr dat, budou mě zajímat především následující kritéria.

1) Export dat z vybraného nástroje a možnost jejich propojení s relační databází internetového obchodu na bázi návštěvníků

Jedním z dílčích cílů mé práce je vytvoření datového skladu (teorií datového skladu se budu podrobněji zabývat v kapitole 4) pro potřeby následné analýzy. Pro naplnění tohoto skladu bude zapotřebí mít možnost exportovat data nasbíraná vybraným nástrojem.

Pro potřeby hlavního cíle práce, tedy analýzy internetového obchodu, budu potřebovat propojit data z vybraného nástroje z relační databáze obchodu. Tento požadavek je důležitý zejména proto, jelikož mi umožní spojit zdroje návštěvnosti a chování návštěvníka na webu s jeho nákupním košíkem a případnými objednávkami. Propojení dat bude realizováno pomocí atributu session ID, tj. dočasného unikátního klíče předávaného prohlížeči při připojení se na server (webové stránky). Je tedy nutné, aby vybraný nástroj tento atribut evidoval u každého návštěvníka.

2) Funkcionalita

Pro vytvoření výsledných reportů bude zapotřebí určitá funkcionalita vybraného nástroje. Jedná se zejména o

- statistiku vstupních stránek,
- statistiku výstupních stránek,
- evidenci pohybu návštěvníka na webu,
- statistiku klíčových slov, přes která uživatelé přicházejí,
- datum a čas jejich návštěvy a
- geografické rozložení návštěv.

3) Cena

Jak jsem se již zmínil v úvodu práce, realizace tohoto projektu bude probíhat v měřítku malých podniků. Takovéto firmy většinou nedisponují velkým množstvím kapitálu a proto je minimalizace nákladů jedním z jejich hlavním cílů.

Vzhledem k faktu, že na trhu nástrojů analýzy návštěvnosti panuje poměrně značná konkurence, je možné najít kvalitní nástroje, které jsou k dispozici zdarma. Proto do svého výběru zahrnu pouze tyto.

3.2 Popis nástrojů analýzy

Tato práce si neklade za cíl hodnocení všech dostupných nástrojů pro analýzu návštěvnosti, ale jejím cílem je najít takový nástroj, který bude vyhovovat kritériím stanovených v kapitole 3.1.2. Z toho důvodu budu testovat a popisovat pouze takové nástroje, které pokládám vzhledem k mé praxi v oblasti SEO analýzy za často používané a které představují jistý prototyp ve své kategorii. Těmito kategoriemi jsou webové služby poskytující analýzu návštěvnosti a open source nástroje. V tomto předvýběru také zohledním jedno z vlastních kritérií a tím je cena. Nástroji, které nejsou zdarma, se tedy vůbec zabývat nebudu.

V první kategorii webových služeb, jsem zvolil pro svou analýzu dva nástroje, které považuji za velmi rozšířené. Tím prvním jsou Google Analytics. Tento nástroj je velmi oblíben zvláště mezi SEO experty, dá se říci, že v podstatě v každé knize o SEO se s tímto nástrojem alespoň rámcově pracuje. Druhou službou je pak nástroj TOPlist.cz, oblíbený zejména v České

republice, který disponuje spíše méně sofistikovanou funkcionalitou, ale který je zároveň poměrně jednoduchý na ovládání.

V kategorii open source jsem zvolil nástroj Piwik. Podle jeho autorů má být tento nástroj jakousi open source alternativou ke Google Analytics. Tím je také v podstatě dána jak jeho funkcionalita, tak také vzhled sestav a možností práce s nimi. Tento nástroj vznikl jako pokračování jiného open source nástroje – phpMyVisites. V současné době (7. 3. 2009) je sice stále vyvíjen a jeho poslední dostupná verze je betaverze, ale z online dema na webových stránkách se dá předpokládat, že jeho funkcionalita je již alespoň v základní podobě dokončena.

Při popisu těchto nástrojů přiblížím postup jejich instalace, popíši uživatelské rozhraní a také zmíním jejich přednosti a nedostatky. Samozřejmě daný nástroj také vyhodnotím.

Vzhledem k cíli mé práce je pro výběr nástroje nutné, aby splňoval všechna mnou definovaná kritéria. Proto při jejich hodnocení budu volit pouze mezi hodnotou 0-nesplňuje a 1-splňuje, přičemž váha těchto kritérií bude stejná.

3.2.1 Piwik

Instalace

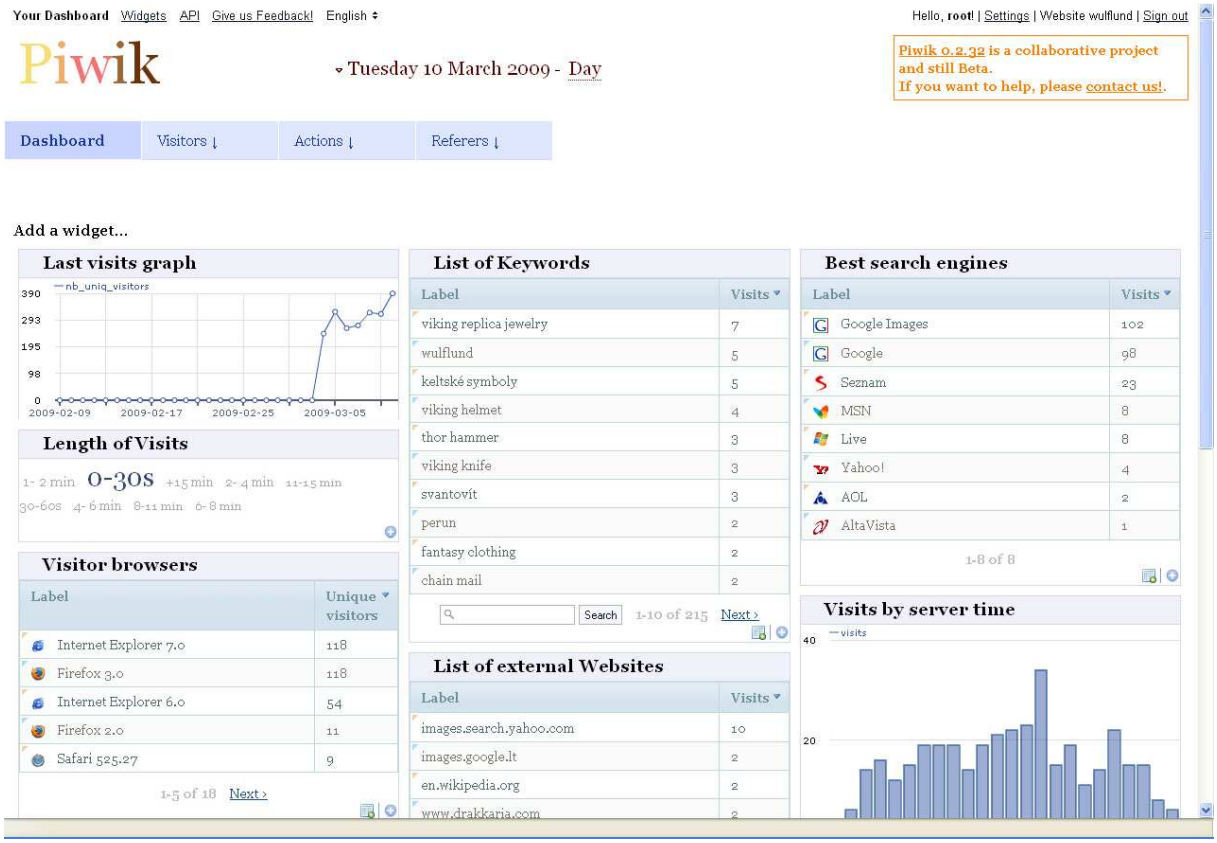
- stáhnout instalační soubor z internetu http://piwik.org
- rozbalit .zip soubor a nahrát na server do root adresáře
- sledovat kroky instalace (nutnost znát přístupové údaje do databáze - vytváří se tabulky v databázi)
- nakopírovat javascriptový kód do stránky

Pro úspěšnou instalaci je nutné mít na webovém serveru PHP verze alespoň 5.1 a mít nainstalovanou podporu MySQL. Na některých webových serverech, pokud je zapnut safe mód, může při instalaci docházet k problémům s právy se zápisem do adresářů. Doporučuji proto po rozbalení a nahrání aplikace na server ještě vytvořit adresáře tmp, tmp/cache a tmp/templates_c s nastavením práv adresářů na 777 a až poté přistoupit k instalaci.

Uživatelské rozhraní

Grafická stránka této služby je poměrně vyvedená. Menu i grafy jsou přehledné, orientace v prostředí je snadná. Oceňuji zde zvláště některé prvky, jako je třeba rychlý rozbor přístupů z hlediska jak vyhledávačů tak zároveň i klíčových slov. Zobrazení je možné buď pro jeden den, jeden týden, jeden měsíc nebo jeden rok. Další možností je export výstupů do PDF dokumentů, které mohou být užitečné zvláště v případě nutnosti reportovat průběh návštěvnosti např. vedení společnosti atd. Další nespornou výhodou je také možnost výběru a uspořádání sestav na hlavní stránce podle vlastních potřeb.

Obrázek č. 1 - Piwik



Výstupy

Množství výstupů je zde poměrně velmi rozmanité. Je možné zde nalézt jak klasické přehledy návštěvnosti za určité období, tak i mnohé pokročilé analýzy. Za zmínku stojí zejména rozklad vstupních stránek podle kategorií. Jako příklad mohu uvést např. kategorii jewelry, která se dá následně rozkliknout na podkategorie a zjistit tak zároveň i přehled vstupních stránek pro ně. A to vše v jednom přehledném souhrnu. Další zajímavostí je možnost konfigurace příchodů i z newsletterů a partnerských webů, což ne všechny služby poskytující

analýzu návštěvnosti nabízejí. Jako mínus považuji zejména nemožnost nastavení vlastního časového intervalu zobrazení přehledů a statistik. Další nevýhodou může pro někoho být neexistence českého překladu, ale u open source nástrojů je to vždy spíše otázkou času, kdy se česká lokalizace objeví.

Vyhodnocení na základě zadaných kritérií

- Možnost exportu dat a propojení s relační DB - 1

  U toho nástroje je možné propojit data z vlastní databáze z daty generovanými tímto nástrojem. To je umožněno zejména tím, že nástroj využívá stejnou databázi, jako analyzované webové stránky.
- Funkcionalita - 1

  Nástroj obsahuje požadovanou funkcionalitu.

3.2.2 Google Analytics

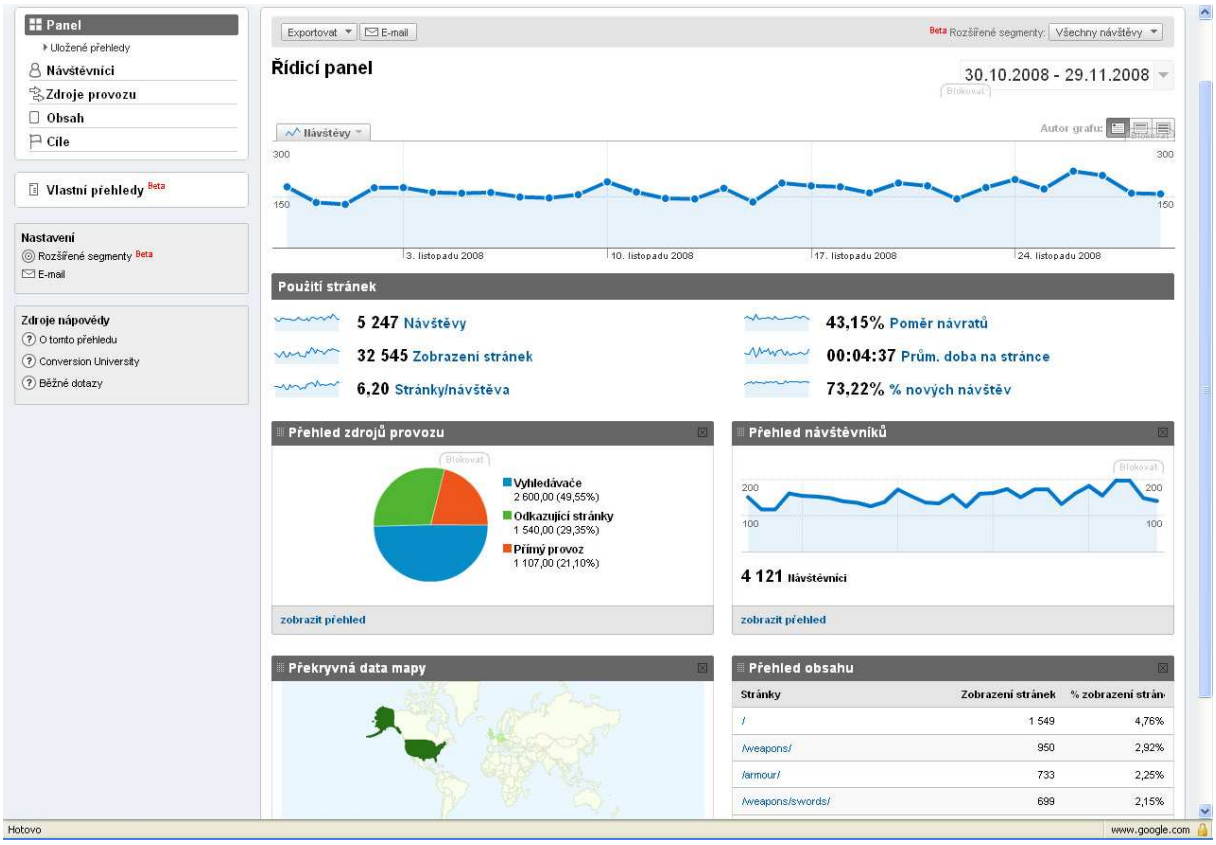
Instalace

- zaregistrovat se na http://www.google.com/analytics/
- vytvořit doménu pro svůj web
- nakopírovat kód vygenerovaný službou do kódu stránky

Uživatelské rozhraní

Z hlediska uživatelského prostředí je tato služba asi nejlepší a zpracováním konkuruje i mnohým komerčním nástrojům. Vše je zde konfigurovatelné, přehledné a poměrně rychlé. Jako velké plus považuji možnost přizpůsobit si obrazovky podle sebe, což umožňuje ušetřit poměrně značné množství času a mít tak vše pohromadě. Možností je také export jak do PDF, tak do XML. Zvláště export do XML může být pro někoho velmi užitečný, jelikož si pak tato data může libovolně přenášet i mezi ostatní software.

Obrázek č. 2 - Google Analytics



Výstupy

Výstupů zde můžeme nalézt opravdu velké množství. Od těch jednodušších až po ty sofistikovanější. Je pravda, že ty sofistikovanější výstupy vyžadují již určitou zkušenost, ale i tak je jejich využití velmi praktické.

Vyhodnocení na základě zadaných kritérií

- Možnost exportu dat a propojení s relační DB - 0

  U tohoto nástroje není možné propojit data vztahující se k jedinečnému návštěvníkovi s daty vlastními. Je ovšem možné pomocí XML přenášet souhrnné přehledy.

- Funkcionalita - 1

  Nástroj obsahuje požadovanou funkcionalitu.

3.2.3 TOPlist.cz

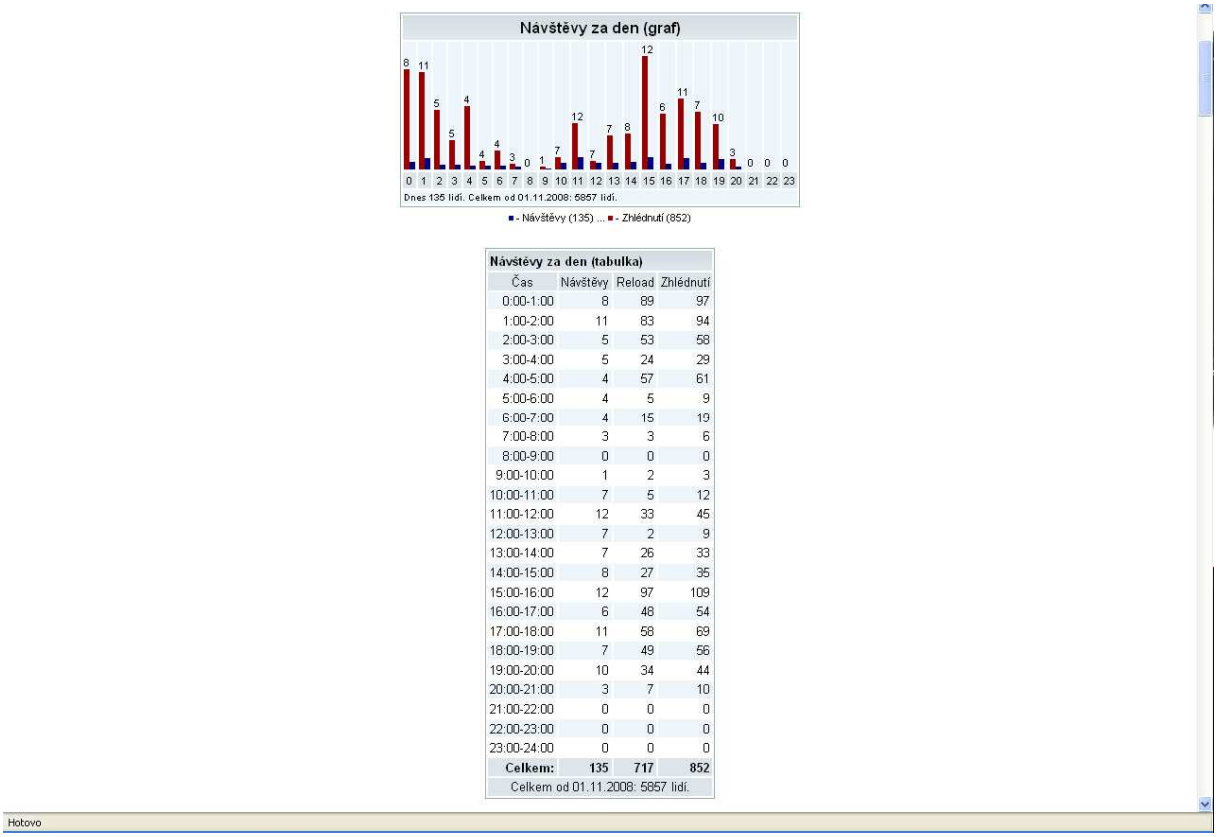
Instalace

- zaregistrovat se na http://toplist.cz (přes jednoduchý formulář)
- nakopírovat javascriptový kód do stránky

Uživatelské rozhraní

Uživatelské prostředí je zde velmi omezené. Skládá se v podstatě pouze z několika textových tabulek. Orientace v nich není příliš pohodlná nehledě na fakt, že žádný z těchto přehledů neposkytuje možnost tento přehled jakkoli si přizpůsobit. Na druhou stranu může být tento přehled pro někoho zajímavý právě pro svou jednoduchost a koncentraci údajů na jednu stránku. Uživatel pak nemusí složitě překlikávat a vyhledávat ty nejpodstatnější údaje, které každodenně využívá.

Obrázek č. 3 - TOPlist.cz



Výstupy

Množství výstupů z této služby je velmi omezené a zaměřené spíše na weby menší, kterým stačí pouze základní přehled. Z výstupů se můžeme dozvědět počet návštěvníků, vstupní i zobrazené stránky a následně některé údaje o návštěvnících jako je rozlišení obrazovky, prohlížeč, operační systém, IP adresa a země, ze které pochází.

Takto omezené výstupy mají samozřejmě i své nevýhody. Zmínil bych především nemožnost exportu dat a dále pak chybějící přehledy o pohybu návštěvníků na stránkách, nezobrazování klíčových slov, přes které lidé přišli a nemožnost filtrace jakýchkoli dat.

Ukládání denních přehledů pouze dva dny zpět pak tuto službu pro můj účel v podstatě diskvalifikuje. Nemožnost vrátit se zpět k některým detailním přehledům je opravdu velmi podstatná.

Vyhodnocení na základě zadaných kritérií

- Možnost exportu dat a propojení s relační DB - 0

  U toho nástroje není možné propojit data s daty vlastními, jelikož neumožňuje žádný export dat.
- Funkcionalita - 0

  Nástroj nemá požadovanou funkcionalitu.

3.2.4 Další nástroje

Navrcholu.cz

Tento nástroj je v České republice poměrně oblíbený. Původně jsem ho chtěl také otestovat, bohužel ale jeho plná verze je placená. Existuje sice zkušební měsíční verze, ale vzhledem k tomu, že bych tento nástroj následně stejně nevyužíval, tak jsem se rozhodl ho ani nepopisovat. Verze Lite, která je zdarma, obsahuje jen velmi málo přehledů a údajů. To byl také hlavní důvod, proč jsem se rozhodl popsat raději službu TOPlist.cz

AWstats

Tato služba patří do skupiny open source nástrojů. Je hojně využívána jako základní statistika společně s dodaným hostingem. Poskytuje poměrně slušný přehled o návštěvnosti, ale z hlediska analýzy, kterou chci provádět, mi opět přišel vhodnější nástroj Piwik, proto jsem popisoval ten.

3.3 Výběr nástroje analýzy

Tabulka č. 1 - Vyhodnocení nástrojů analýzy podle vlastních kritérií

	Piwik	Google Analytics	TOPlist
Funkcionalita	**1**	**1**	**0**
Možnost exportu dat a propojení s relační DB	**1**	**0**	**0**

Vzhledem k tomu, že stanovená kritéria splňuje pouze jeden nástroj, je finální výběr jednoduchý. Ke své analýze tedy budu využívat nástroj Piwik.

Pro provedení analýzy bude ale nutné učinit určité změny jak v databázi, tak ve zdrojovém kódu tohoto nástroje. Zejména proto, jelikož v základním nastavení neexistuje v databázi tohoto nástroje sloupec tabulky, označující unikátního návštěvníka webu, který by se dal propojit s tabulkou databáze internetového obchodu. Celý tento postup bude popsán v kapitole 6.1.

Takovéto propojení dat by pak mělo poskytnout podklady pro reporty, které umožní analyzovat chování uživatelů na webu a tím do optimalizace stránek přidat onen prvek použitelnosti zmiňovaný v kapitole 2.

Pro běžného uživatele či správce webu, bych ale doporučoval využít služby Google Analytics. Z dostupných zdarma nástrojů je totiž jednak uživatelsky nejpřívětivější a také pro běžné potřeby sledování návštěv plně dostačující.

4 Business Intelligence

Jak je ze zadání této práce patrné, k analýze návštěvnosti internetového obchodu budu využívat nástrojů a přístupů Business Intelligence. V této kapitole tedy představím základní teoretická východiska, přičemž se zaměřím primárně na teorii datových skladů a následně na postup projektu BI.

> Business Intelligence je sada postupů, procesů a technologií, jejímž cílem je účinně účelně podporovat rozhodovací procesy ve firmě. Představuje komplex aplikací (IS/ICT), které podporují analytické a plánovací činnosti podniků a organizací a jsou postaveny na specifických, tzv. OLAP (On-Line Analytical Processing) technologiích a jejich modifikacích [2].

Jak je patrné z výše popsané definice, Business Intelligence se skládá z mnoha různých částí. Jedná se zejména o [8]:

- Produkční, zdrojové systémy
- DSA (Data Staging Area) – dočasná datová úložiště
- ODS (Operational Data Store) – operativní datová úložiště
- ETL (Extraction Transformation Loading) – transformační nástroje
- EAI (Enterprise Application Integration) – integrační nástroje
- DWH (Data Warehouses) – datové sklady
- DMA (Data Marts) – datová tržiště
- OLAP (On-Line Analytical Processing)
- Reporting
- EIS (Executive Information Systems)
- Data Mining – dolování dat
- Nástroje pro zajištění datové kvality
- Nástroje pro správu metadat
- Ostatní

4.1 Datové sklady

Datový sklad je integrovaný, subjektově orientovaný, stálý a časově rozlišený souhrn dat, uspořádaný pro podporu potřeb managementu [8].

Z definice jsou patrné čtyři základní vlastnosti datového skladu [5]

- subjektová orientace

Data se do datového skladu zapisují spíše podle předmětu zájmu, než podle aplikace, ve které byla vytvořena. Při orientaci na subjekt jsou data v datovém skladu kategorizována podle subjektu, kterým může být např. zákazník, dodavatel, zaměstnanec, výrobek a podobně.

- integrace

Datový sklad musí být jednotný a integrovaný. To znamená, že data týkající se konkrétního předmětu se do datového skladu ukládají pouze jednou. Proto musíme zavést jednotnou terminologii, jednotné a konzistentní jednotky veličin. Ne vždy je tato úloha jednoduchá, protože data přicházejí do datového skladu z nekonzistentního a neintegrovaného operačního prostředí. Proto musí být data v etapě přípravy a zavedení upravena, vyčištěna a sjednocena. Pokud data nejsou konzistentní a důvěryhodná, datový sklad ztrácí svůj význam.

- časová rozlišenost

Data se ukládají do datového skladu jako série snímků, ze kterých každý representuje určitý časový úsek. Na rozdíl od operačního prostředí, kde jsou data platná v okamžiku přístupu, v datových skladech jsou platná pro určitý časový moment. Klíčové atributy v datovém skladu obsahují čas, který v operačních databázích být nemusí.

- neměnnost

V operačních transakčních databázích jsou data do databáze jednak vkládána, modifikována ale i mazána. Data v datovém skladu se obvykle nemění ani neodstraňují, jen se v pravidelných intervalech přidávají data nová. Proto je manipulace s těmito daty také mnohem jednodušší. V zásadě jsou povolené pouze dvě operace – zavedení dat do datového skladu a přístup k těmto datům.

4.1.1 Metody budování datového skladu

Celý projekt vytvoření datového skladu lze řešit dvojím způsobem [13]:

Velký třesk

Nejprve jsou provedeny detailní analýzy všech požadavků uživatelů a k nim patřičných zdrojů dat, je vypracován souhrnný návrh řešení a poté probíhá implementace. U velkých projektů může fáze realizace trvat i více než rok.

Výhody:

- od začátku je jasná datová struktura celého datového skladu

Nevýhody:

- dlouhá doba od zadání po předání do provozu – zadavatel může mít pocit, že se dlouho nic neděje a ztratí o dokončení datového skladu zájem
- možné změny zdrojů dat a požadavků ještě před ukončením implementace a předáním do ověřovacího provozu
- komplikované zohlednění změny priorit požadavků v průběhu implementace
- velké riziko neúspěšnosti projektu

Inkrementální metoda

V průběhu úvodní analýzy jsou určeny jednotlivé etapy a to podle priorit řešení požadavků, každá z etap by neměla trvat déle než tři až čtyři měsíce. Etapou se v tomto smyslu rozumí sled kroků detailní analýza, návrh řešení, implementace, testování, školení, ověřovací provoz, předání do provozu.

Obrázek č. 4 - Inkrementální metoda tvorby datového skladu

Výhody:

- zadavatel může v krátkém časovém úseku využívat výstupy pro další rozhodování – návratnost investic, možné i samofinancování dalších etap ze vzniklých úspor
- pořadí etap může být upraveno dle aktuálních potřeb společnosti
- snazší reakce na změny datových zdrojů v průběhu projektu

Nevýhody:

- možné komplikace při změně datové struktury – zásadní úpravy stávajících a již používaných datových struktur, doplnění historických dat

V současné době se téměř zcela ustoupilo od realizace datového skladu celého najednou a implementace se realizují inkrementálním způsobem.

I u inkrementálního způsobu existuje riziko přílišného nárůstu objemu prací v rámci etapy, když se nepodaří určit priority požadavků a „všichni chtějí všechno v prvním inkrementu“. Pak musí nastoupit řídící moc vedení projektu a striktně určit úlohy, které budou v etapě řešeny.

Vzhledem k tomu, že cílem této práce je pouze analýza dat návštěvnosti internetového obchodu, bude pro implementaci datového skladu zvolena inkrementální metoda, na jejímž základě by výsledné řešení mělo být mnohem jednodušší a rychlejší. Při návrhu datového skladu ale budou brány v potaz i požadavky na další inkrementy, aby bylo možné datový sklad následně rozšiřovat.

Po vybudování datového skladu je ještě nutné tento sklad naplnit daty. K tomuto účelu slouží transformační nástroje.

4.1.2 Transformační nástroje (ETL)

Tyto nástroje, také označované jako datové pumpy, by měly zajistit již zmiňovanou podmínku datového skladu a to integritu.

ETL označuje právě mechanizmus získávání dat z provozních systémů podniku (ekonomika, skladové hospodářství, výroba, odbyt atd.), jejich následné zpracování a poskytnutí aplikacím pro podporu rozhodování. Tato problematika je poměrně komplikovaná a představuje nezanedbatelný podíl (někdy až 70 %) v nákladech na budování systémů pro podporu rozhodování - nákladů ať finančních, tak i časových a lidských [11].

Během této transformace dochází též k úpravě dat, zejména co se týče jejich formátů, nahrazování hodnot, kontrole duplicit a dalších úprav.

4.2 Postup projektu

Jak již bylo výše zmíněno, řešení projektů v současné době je realizováno přírůstkovou metodou. Podle [8] existují následující postupy v rámci procesu tvorby BI řešení.

1) Úvodní studie BI

- definování cílů a efektů BI
- katalog uživatelů
- analýza připravenosti firmy na BI
- analýza požadavků
- určení oblastí řešení a jejich priorit
- analýza možnosti standardních BI aplikací
- návrh BI architektury

- návrh organizace řešení BI
- určení harmonogramu a ekonomických charakteristik projektu

2) Specifikace přírůstku

- vyhodnocení aktuálního stavu BI
- definování přírůstku
- určení harmonogramu a ekonomiky přírůstku

3) Analýza BI řešení

- interview, verifikace požadavků
- detailní analýza business požadavků oproti stávajícímu stavu
- detailní specifikace požadavků
- analýza kvality produkčních dat
- specifikace nároků na datové zdroje
- plánování kontroly a čištění datových zdrojů
- analýza potřeb a dostupnosti externích dat
- rekapitulace a hodnocení BI datových zdrojů
- specifikace změn v produkčních databázích

4) Modelování a návrh datového skladu a BI řešení

- dimenzionální analýza
- návrh vrstev řešení (DSA – dočasného úložiště, ODS – operativního úložiště, DWH – datového skladu, DMA – datových tržišť)
- modelování DWH/DMA
- návrh OLAP kostek a databáze
- návrh analytických a plánovacích aplikací
- návrh reportů
- návrh a realizace prototypů
- návrh analytických pravidel

5) Návrh technologické platformy přírůstku

- návrh technologické architektury

- fyzický návrh řešení DWH/DMA
- fyzický návrh uložení dat
- řešení objemu a nárůstu dat, sizing
- návrh podpory uživatelů v provozu BI
- specifikace přístupových práv
- návrh komplexního zajištění provozu BI aplikací

6) Návrh transformací – ETL

- specifikace transformací v různých vrstvách BI řešení – DSA, ODS,…
- specifikace detailních pravidel pro transformace a zajištění kvality dat
- řešení problémů chybějících dat a NULL hodnot
- řešení integritních problémů

7) Implementace BI řešení

- implementace databázových komponent aplikací BI
- implementace ETL
- implementace klientských aplikací nad OLAP kostkami
- customizace standardních aplikací
- vytvoření, popřípadě upgrade technologické infrastruktury
- nastavení definovaných přístupových práv
- prvotní migrace a čištění databází
- implementace nástrojů a aplikací pro Data Mining
- zpracování kompletní uživatelské a provozní dokumentace

Jelikož plánovaný projekt BI není příliš velkého rozsahu a tvůrce první části projektu pro analýzu návštěvnosti bude zároveň jak tvůrcem projektu, tak i jeho hlavním uživatelem, budou některé postupy vynechány. Jedná se zejména o postupy z části 3 jako je interview a také o postupy z části 7 jako definování přístupových práv a zpracování dokumentace.

5 Současná situace

Analyzovaný internetový obchod se skládá ze dvou částí. Jednou z nich je česká verze – drakkaria.cz, tou druhou je pak verze zahraniční – wulflund.com. Na českém trhu se obchod pohybuje již cca 5 let, na zahraničním pak cca 2 roky. Obchod je zaměřen na prodej historických předmětů a replik, jako jsou např. meče, šperky, amulety, oblečení, zbroj a jiné.

Obrázek č. 5- Náhled zahraniční verze internetového obchodu (www.wulflund.com)

Jeho hlavním cílem je udržet si ve svém oboru vedoucí postavení na českém trhu, v zahraničí se pak soustředit spíše na velké odběratele a vytvořit síť zastupitelských obchodů.

5.1 Architektura řešení

Vývoj internetového obchodu s menšími či většími úpravami probíhá již přibližně dva roky. Z původního záměru vybudovat pouze jednoduchý eshop se postupem času stala poměrně komplexní aplikace.

Technologicky je web řešen pomocí programovacího jazyka PHP, pro uložení dat je pak využívána databáze MySQL, která na web tohoto rozsahu pohodlně dostačuje a kterou nabízí naprostá většina poskytovatelů hostingových služeb.

Původní společný engine obou domén a databáze byly přepracovány v dubnu loňského roku. Důvodem tohoto kroku byla častá indexace stránek prohlížečem s českým obsahem, ale na doméně zahraniční. Z hlediska SEO bylo tedy nutné sjednat nápravu.

Tato úprava s sebou přinesla několik změn. Tou hlavní bylo rozdělení obou domén na různé hostingy. Z počátku databáze zůstala společná na jednom hostingu, z druhého hostingu se k ní připojovalo pomocí vzdáleného přístupu. Toto řešení, jak se posléze ukázalo, nebylo nejšťastnější, zejména vzhledem ke značnému zpomalení načítání stránek.

Náprava spočívala v duplikaci databáze na každý hosting. Ale i takovéto řešení s sebou přineslo mnohá úskalí.

Obsah stránek je z velké většiny tvořen pomocí CMS. Jedná se zejména o katalog výrobků, ale také o číselníky poštovného, obrázků, států, novinek a dalších prvků. Tato data, až na novinky využívají oba weby společně. Při rozdělení webů tedy muselo být zajištěno, aby tato data zůstala v obou databázích shodná. Této shody je zajištěno pomocí vzdáleného přístupu. Zpomalení v CMS je sice také patrné, ale již není tolik na závadu. Obrázky produktů jsou pak automaticky přenášeny pomocí FTP.

Existují také moduly, které společné nejsou. Typickým příkladem jsou objednávky. Slučování zde neproběhlo především z důvodu opětovného zpomalení webu pro uživatele. V tomto případě je tedy rozdílný obsah tabulky v databázi na drakkaria.cz a v databázi na wulflund.com.

Ovšem ani tento systém se neukazuje jako příliš pohodlný a efektivní. Dá se říci, že to je pouze dočasné řešení, vzhledem k selhání původního plánu, tedy jedné společné databáze a vzdáleného přístupu k ní.

5.2 Situace v oblasti SEO

Další oblast, kterou bych chtěl zmínit je SEO. Optimalizace webu byla, dá se říci, hlavní příčinou rozdělení stránek. Samotný postup optimalizace, který jsem zmínil v kapitole 2, však dosud uskutečněn nebyl. V současné době, jsou stránky připraveny pouze v rovině zdrojového kódu a jeho struktury, texty stránek, titulky a nadpisy optimalizovány zatím nebyly.

Pro analýzu návštěv momentálně využíváme nástroj Google Analytics, ale jeho využití je spíše pasivní a to hlavně v rovině počtu návštěvníků, klíčových slov přes která na web přicházejí a návštěvníků přicházejících z jiných webových stránek.

5.3 Postavení na trhu

Vzhledem k tomu, že prezentovaný obchod nemá kamennou prodejnu, pohybuje se pouze na internetovém trhu a přímou konkurencí mu jsou také pouze elektronické obchody.

Zde je ale situace obou částí obchodu, jak českého, tak zahraničního velmi odlišná. Zatímco česká verze je již dlouho zavedená a má v oblasti prodeje historických předmětů a replik poměrně dominantní postavení, tak verze zahraniční ještě zdaleka nedosahuje takového potenciálu a ani podílu na trhu.

Poměr návštěv je v současné době zhruba 1100:300 ve prospěch verze české. Přitom pokud se podíváme na obrázek č. 4, který zobrazuje poměr vyhledávání slov „swords“ (modrá křivka) a „meče“ (červená křivka) ve vyhledávači Google, můžeme vidět obrovský rozdíl v objemu vyhledávání těchto slov. Návštěvnost zahraniční verze by tedy, při dosažení podobného postavení jako má verze česká v rámci českého internetu, měla být přibližně stonásobná.

Obrázek č. 6 - Google Trends - porovnání objemu vyhledávání

Hlavním důvodem, proč tomu tak prozatím je, je prozatím nízká optimalizace zahraniční verze a také vysoká konkurence na anglická klíčová slova. Zatímco na výraz „meče“ najde Google 1,7 milionu výsledků, tak na výraz „swords“ jich už je 23,7 milionu. A umístit se mezi prvními třiceti již není bez velmi kvalitní optimalizace možné.

Podobná situace je i u ostatních klíčových slov.

5.4 Vize do budoucna

Jedním z hlavních úkolů bude v budoucnosti opět sloučit obě databáze. To jednak usnadní určité práce, které se již nebudou muset provádět dvakrát, ale především vše zpřehlední a sníží možnost chyb. V neposlední řadě též dojde ke zrychlení práce s CMS, což jistě jeho uživatelé ocení. K tomuto účelu by mělo posloužit vytvoření jednotné datové základny, které se budu konkrétněji věnovat v kapitole 6.

Další vývoj se bude týkat funkčnosti stránek. Hlavní oblastí zde bude skladové hospodářství. V současné situaci by bylo rozšíření o sklad poměrně složité a práce s ním velmi pomalá, jelikož i tento modul bude vyžadovat společnou databázi. Dalším omezením by byla jeho interakce s objednávkami a automatickým vyskladňováním. Jak jsem totiž zmínil výše, objednávky jsou v současné době oddělené.

Hlavním centrem zájmu bude ale optimalizace webu. V této oblasti bych pak rád využil analýzy a výstupů, které získám při zpracování této diplomové práce.

Počítám také s tím, že práce vynaložená na implementaci technologií BI bude možné využít i v jiných oblastech než jen SEO, zejména bych zmínil výpočty marží, nákladů a mnoho dalšího.

6 Implementace BI řešení

V této kapitole popíši vlastní realizaci řešení analýzy návštěvnosti internetového obchodu za využití nástrojů Business Intelligence. Při návrhu a implementaci budou využity doporučené postupy, principy a další teoretické poznatky blíže pospané v kapitole 4.

6.1 Úprava nástroje analýzy

Jak již bylo zmíněno v kapitole 3, pro propojení dat internetového obchodu s daty z nástroje analýzy je nutné provést určité změny jak v kódu, tak i v databázi vybraného nástroje Piwik.

K tomuto účelu jsem vybral atribut session_id, který je používán pro identifikaci návštěvníků v databázi internetového obchodu.

Je také možné zvolit opačný postup a přidat ID návštěvy (atribut idvisit) do vlastních tabulek. Výhodou tohoto postupu by bylo, že bychom nemuseli měnit zdrojové kódy a tabulky nástroje Piwik. Na druhou stranu, museli bychom provést změny v každé tabulce, která by měla být napojena na tabulku návštěv. Výsledkem takovéto úpravy by pak bylo nahrazení atributu session_id ve všech tabulkách eshopu atributem idvisit.

6.1.1 Úprava databáze

Úprava databáze je poměrně snadná. Stačí pouze vybrat databázi, kde jsou uložené tabulky vygenerované nástrojem Piwik a spustit následující dotaz.

```
ALTER TABLE piwik_log_visit ADD session_id VARCHAR( 100 ) NOT NULL
AFTER visitor_idcookie;
```

6.1.2 Úprava zdrojového kódu

Úprava kódu probíhá v souboru piwik/core/Tracker/Visit.php. Zde je nutné změnit pouze metodu handleNewVisit() jelikož u známé návštěvy již žádná změna v dané relaci nenastává.

Nejprve je nutné iniciovat relaci a zapsat její ID do proměnné.

```
Session_start();
$session_id = Session_id();
```

Následně je nutné přidat do pole visitorInfo řádek 'session_id' => $session_id. Výsledek pak bude vypadat následovně.

```
$this -> visitorInfo = array (
        'idsite'                => $this->idsite,
        'visitor_localtime'     => $localTime,
        'visitor_idcookie'      => $idcookie,
        'session_id'            => $session_id,
        'visitor_returning'     => $returningVisitor,
```

Pro potřeby této práce jsem prováděl ještě další úpravy kódu a databáze nástroje analýzy, které jsou ale již více specifické struktuře dat vlastního internetového obchodu a nedá se tedy předpokládat, že by byly v této podobě aplikovatelné i jinde. Z tohoto důvodu je uvádím pouze v příloze 1.

6.2 Úvodní studie

V úvodní studii budou popsány základní kroky přípravy implementace BI řešení. Jedná se zejména o definování cílů a efektů BI, definování uživatelů a jejich požadavků, návrh architektury a určení oblasti řešení a jejich harmonogramu.

Úvodní studie bude také dále sloužit jako východisko pro případné budoucí rozšiřování implementace BI v rámci internetového obchodu.

6.2.1 Definování cílů a efektů BI

Při definování cílů a efektů projektu budu vycházet z metody IRACIS [10]. Tato metoda identifikuje tři základní oblasti ukazatelů podle jejich vlivu na výkonnost podniku. Jedná se o IR (zvýšení příjmů), AC (snížení nákladů) a IS (zkvalitnění služeb).

1) Zvýšení příjmů

- Zvýšení počtu návštěvníků internetového obchodu
- Zvýšení konverzního poměru
- Identifikace a optimalizace slabých míst internetových stránek
- Lepší cílení marketingových kampaní
- Zvýšení efektivity řízení firmy

2) Snížení nákladů

- Efektivnější plánování skladových zásob
- Snížení času potřebného pro přípravu reportů
- Přesné vyhodnocování PPC kampaní

- Lepší segmentace zákazníků
- Optimalizace interních procesů

3) Zkvalitnění služeb

- Optimalizace webu internetového obchodu podle potřeb a chování návštěvníků
- Rozšíření nabídky produktů z nejvíce poptávaných kategorií
- Zkrácení dodacích lhůt

Vyhodnocení těchto očekávaných efektů implementace řešení BI proběhne po roce provozu navrženého řešení. Možnosti dosažení některých efektů budou vyhodnoceny v závěrečné části této práce.

6.2.2 Katalog uživatelů

V první fázi bude aplikace určena především pro potřeby SEO konzultanta, který na jejím základě bude provádět nutné změny k lepší optimalizaci stránek jak pro vyhledávače, tak i pro návštěvníky. Některé reporty ale budou sdíleny také s vedením společnosti.

Další fáze vývoje BI aplikace budou zaměřeny především na požadavky managementu obchodu, ale počítá se také s využitím výstupů pro potřeby zaměstnanců zejména v oblasti skladového hospodářství a komunikace s dodavateli.

6.2.3 Analýza požadavků

Hlavním požadavkem je získat nástroj umožňující analyzovat chování návštěvníků internetového obchodu. Navržené řešení by mělo poskytovat takové pohledy na data, které umožní SEO pracovníkovi reagovat na trendy a změny chování návštěvníků na webu, stejně tak jako pružně upravovat aktuální SEO kampaně.

Dalším požadavkem je sjednocení datové základny všech verzí internetového obchodu. Navržené řešení by mělo být schopno odhalit duplicity zejména v uživatelských účtech a objednávkách a poskytnout tak mnohem přesnější a ucelenější reporty. Na základě těchto dat by pak mělo být možné vytvářet podklady pro strategické a taktické rozhodování vedení společnosti.

Co se týče aktualizace dat v datovém skladu, tak u dat vztahujících se k analýze návštěvnosti je požadavek na aktualizaci jednou za 14 dní, u dat týkajících se provozu obchodu je tento požadavek jednou denně.

Navržené řešení by mělo také umožňovat pohledy na metriky z různých úhlů pohledu. Zejména pak z časového u metrik návštěvnosti a produktového u metrik objednávek a skladů.

Dalším požadavkem je určitá flexibilita navrhovaného řešení s tím, aby bylo v budoucnu možné toto řešení upravovat podle momentálních potřeb vedení firmy a dalších zainteresovaných osob.

6.2.4 Určení oblastí řešení a jejich priorit

Jak již vyplývá z analýzy požadavků, hlavní oblasti řešení budou dvě. Tou jednou je analýza návštěvnosti, tou druhou pak oblast provozu internetového obchodu. Navrhované řešení by mělo zejména umožňovat zjistit:

1) v oblasti analýzy návštěvnosti

- přes jaká klíčová slova a na jakou stránku návštěvníci přišli,
- na které pozici se ve vyhledávačích nachází odkaz na internetový obchod,
- jaké jsou souvislosti mezi pořadím v katalogu, počtem zobrazení stránky a počtem objednávek,
- jaké stránky nebyly vůbec navštíveny,
- na kterých stránkách uživatelé nejčastěji opustili web a na kterých ho opustili ihned poté, co na web přišli,
- jaký je konverzní poměr (že se z návštěvníka stane zákazník)
- a jak se vyvíjí návštěvnost a pozice klíčových slov v čase

2) v oblasti provozu obchodu

- jaký je aktuální stav zásob a kdy se dá předpokládat jejich vyčerpání,
- jaké jsou produktové marže,
- jak se vyvíjí obrat a zisk v čase,
- jaký je vliv slevových akcí na prodejnost výrobku
- a jaké jsou průměrné doby dodání jednotlivých produktů.

Prioritou je nejprve vytvořit řešení pro analýzu návštěvnosti.

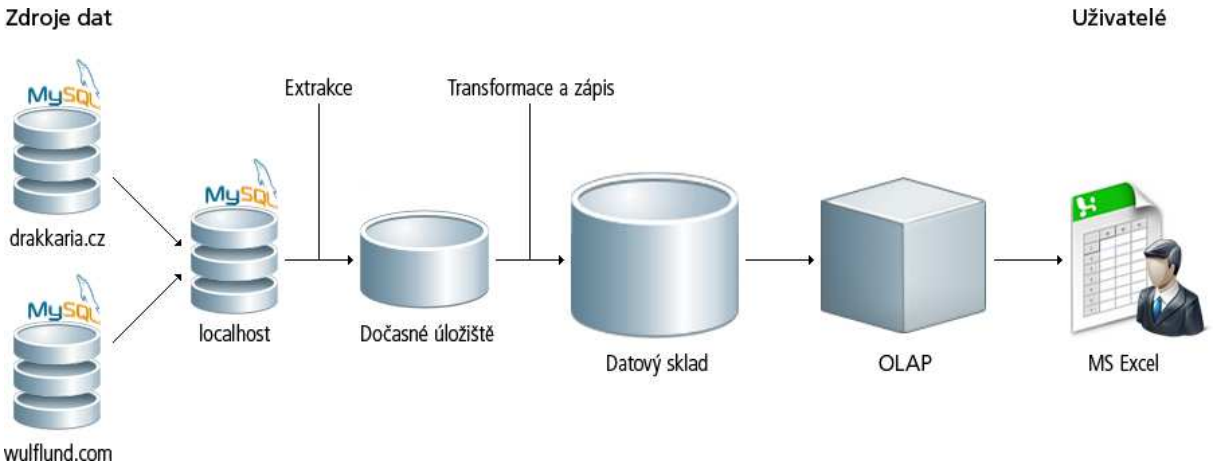
6.2.5 Návrh BI architektury

Česká i zahraniční verze internetového obchodu jsou provozovány na serverech umístěných u poskytovatelů hostingových služeb a data jsou ukládána v databázi MySQL. Data z těchto databází pak budou extrahována na vlastní počítač, kde je provozována jak databáze MySQL, tak i Microsoft SQL server ve verzi 2005. Lokální databáze MySQL je využita proto, jelikož je poměrně problematické extrahovat data z webové databáze přímo do lokálního MS SQL serveru (na lokálním počítači lze toto provést jednoduše přes ODBC) a také proto, jelikož v databázi je ještě před extrakcí do datového skladu nutné provést některé úpravy.

Licence MS SQL obsahuje i analytické nástroje pro implementaci BI aplikace, ale je omezena pouze pro využití k této práci. Pro další rozvoj BI řešení tedy bude nutné zakoupit novou licenci MS SQL. Další náklady na hardware či software by již vzniknout neměly.

Zamýšlení koncoví uživatelé aplikace jsou všichni vybaveni osobními počítači s nainstalovaným operačním systémem Windows XP nebo Windows Vista a kancelářským balíčkem Microsoft Office. Jako prezentační vrstvu tedy bude možné zvolit prostředí aplikace MS Excel.

Obrázek č. 7 - Návrh architektury řešení



6.2.6 Harmonogram řešení

Implementace řešení bude řešena přírůstkovou metodou a rozdělena na dvě etapy podle určených oblastí řešení.

V první etapě bude cílem implementovat takové řešení, které bude vyhovovat požadavkům pro analýzu návštěvnosti stanovených v kapitole 6.2.4 a které bude odpovídat cílům BI projektu. Tato část je také předmětem této práce.

Ve druhé etapě se bude implementovat řešení vyhovující požadavkům pro oblast provozu internetového obchodu.

6.3 Analýza BI řešení

Po vypracování úvodní studie je možné přikročit k analýze BI řešení. V této části se zaměřím především na popis zdrojů dat a také na změny, které bude nutné provést v produkčních databázích.

6.3.1 Zdroje dat

Základním zdrojem dat jsou dvě databáze internetového obchodu. Jedna obsahuje data pro českou, druhá pro zahraniční verzi.

Tato data jsou pro potřeby implementace plně dostačující. Jediným omezením je poměrně krátká doba sběru dat o návštěvnosti, k dispozici jsou data přibližně za dva a půl měsíce provozu.

V následující části zmíním zdrojová data, která budou využita v rámci implementace prvního přírůstku, tedy analýzy návštěvnosti.

Data o návštěvnosti

Data obsahují údaje jak o jednotlivých návštěvnících, tak i o jejich pohybu na webu internetového obchodu a jsou uložena na serveru v tabulkách piwik_site, piwik_log_visit, piwik_log_action a piwik_log_link_visit_action.

Data o objednávkách

Data obsahují údaje o objednávkách a jejich položkách. Uložena jsou na serveru v tabulkách objednavky a objednavky_polozky.

Data o uživatelích

Data jsou uložena na serveru v tabulce users.

Data o produktech katalogu

Data obsahují údaje o struktuře kategorií a s nimi spojenými produkty. Data jsou v tabulkách rozlišena podle jazykové verze. Uložena jsou na serveru v tabulkách katalog_kategorie, katalog_polozky, kategorie_lang a polozky_lang.

6.3.2 Specifikace změn v produkčních databázích

Aby bylo možné dostát požadavkům stanoveným v kapitole 6.2.4, zejména co se týče možnosti zjistit pozice klíčových slov a jejich vývoj v čase, je nutné upravit tabulku piwik_log_visit. Tato tabulka zaznamenává URL stránky, odkud uživatel přišel. V případě vyhledávače může takový záznam vypadat například takto - http://search.seznam.cz/?q=vy%F8ez%E1van%E9+so%9Aky&count=10&from=11. Toto URL v sobě skrývá jak klíčové slovo, tak i oblast pozic, kde se nacházel odkaz na internetový obchod. Zjištění této oblasti pak závisí na určitých parametrech obsažených v URL. PHP skript, který tato data z URL extrahuje, je možné nalézt v příloze 2. Součástí tohoto skriptu je i převod klíčového slova zakódovaného v URL. Výsledkem takového procesu by u výše zmíněného bylo klíčové slovo – vyřezávané sošky.

Ostatní změny ve struktuře tabulek budou provedeny až při transformaci dat v rámci datových pump.

6.4 Analýza datového skladu

Před samotnou implementací datového skladu je ještě na základě požadavků stanovených v předchozích kapitolách nutné stanovit:

- dimenze,
- ukazatele
- a vztahy ukazatelů a dimenzí.

Na základě této analýzy pak bude vytvořen model datového skladu.

6.4.1 Dimenze

Tabulka č. 2 - Dimenze

	Dimenze	Prvky
Web	Web	Web
Action	Akce	Akce
Time	Čas	Rok – Měsíc – Den
Category	Katalog Kategorie	Kategorie - Subkategorie
Product	Katalog Produkty	Kategorie - Produkt
VisitLocation	Návštěvy Oblasti	Kontinent – Země – Návštěva
VisitReferer	Návštěvy Zdroje	Typ zdroje – Zdroj – Návštěva
VisitUsers	Návštěvy Uživatelé	Uživatel - Návštěva
VisitEntryAction	Návštěvy Vstupní Akce	Akce - Návštěva
VisitExitAction	Návštěvy Výstupní Akce	Akce - Návštěva

Dimenze Web

Tabulka č. 3 - Dimenze web

	Struktura	Prvky	Počty	Identifikace
Struktura	1.	Celkem web	1	
	2.	Web	2	
Příklad		1. Celkem web 1.1. www.drakkaria.cz 1.2. www.wulflund.com		id_site
Zdroje		piwik_site		
Poznámka				

Dimenze Akce

Tabulka č. 4 - Dimenze akce

	Struktura	Prvky	Počty	Identifikace
Struktura	1.	Všechny akce	1	
	2.	Akce	18374	
Příklad		1. Všechny akce 1.1. weapons/index 1.2. weapons/swords/index		idaction
Zdroje		piwik_log_action		
Poznámka				

Dimenze Čas

Tabulka č. 5 - Dimenze čas

	Struktura	Prvky	Počty	Identifikace
Struktura	1.	Celkem čas	1	
	2.	Rok	3	
	3.	Měsíc	12	
	4.	Den	31	
Příklad		1. Celkem čas 1.1. Rok 2009 1.1.1. Leden 1.1.1.1. 1. 1. 2009 1.1.1.2. 2. 1. 2009 1.1.2. Únor 1.1.3. Březen		date visit_server_time
Zdroje		Nejsou. Vytvořeno v rámci tvorby časových dimenzí v analytických službách MS SQL serveru.		
Poznámka				

Dimenze Kategorie

Tabulka č. 6 - Dimenze kategorie

	Struktura	Prvky	Počty	Identifikace
Struktura	1.	Všechny kategorie	1	
	2.	Kategorie úroveň 1	24	
	3.	Kategorie úroveň 2	61	
	4.	Kategorie úroveň 3	45	
Příklad		1. Všechny kategorie 1.1. zbroj 1.1.1. přilby 1.1.1.1. starověké přilby 1.1.1.2. ostatní přilby 1.1.2. brnění 1.1.3. štíty		kat_id
Zdroje		katalog_kategorie, kategorie_lang		
Poznámka				

Dimenze Produkty

Tabulka č. 7 - Dimenze produkty

	Struktura	Prvky	Počty	Identifikace
Struktura	1.	Všechny produkty	1	
	2.	Kategorie	130	
	3.	Produkt	2903	
Příklad		1. Všechny kategorie 1.1. starověké přilby 1.1.1. ŘÍMSKÁ PŘILBICE 1.1.2. KELTSKÁ PŘILBA, replika 1.2. ostatní přilby 1.3. příslušenství		item_id
Zdroje		katalog_kategorie, kategorie_lang, katalog_polozky, polozky_lang		
Poznámka				

Dimenze Návštěvy Oblasti

Tabulka č. 8 - Dimenze návštěvy oblasti

	Struktura	Prvky	Počty	Identifikace
Struktura	1.	Všechny návštěvy	1	
	2.	Kontinent	6	
	3.	Země	97	
	4.	Návštěva	125795	
Příklad		1. Všechny návštěvy 1.1. eur 1.1.1. cz 1.1.1.1. 504545 1.1.1.2. 505145 1.1.2. de 1.2. afr 1.3. oce		idvisit
Zdroje		piwik_log_visit		
Poznámka				

Dimenze Návštěvy Zdroje

Tabulka č. 9 - Dimenze návštěvy zdroje

	Struktura	Prvky	Počty	Identifikace
Struktura	1.	Všechny návštěvy	1	
	2.	Typ zdroje	3	
	3.	Zdroj	775	
	4.	Návštěva	125795	
Příklad		1. Všechny návštěvy 1.1. Vyhledávače 1.1.1. Google 1.1.1.1. 14583 1.1.1.2. 25964 1.1.2. Seznam 1.2. Přímá návštěvnosti 1.3. Odkazující stránky		idvisit
Zdroje		piwik_log_visit		
Poznámka				

Dimenze Návštěvy Uživatelé

Tabulka č. 10 - Dimenze návštěvy uživatelé

	Struktura	Prvky	Počty	Identifikace
Struktura	1.	Všichni uživatelé	1	
	2.	Uživatel	5395	
	3.	Návštěva	12795	
Příklad		1. Všichni uživatelé 1.1. Jakub Erlebach 1.1.1. 5145 1.1.2. 29856 1.2. Milan Soukup		user_id
Zdroje		piwik_log_visit, users		
Poznámka				

Dimenze Návštěvy Vstupní Akce

Tabulka č. 11 - Dimenze Návštěvy vstupní akce

	Struktura	Prvky	Počty	Identifikace
Struktura	1.	Všechny akce	1	
	2.	Akce	18374	
	3.	Návštěvy	125795	
Příklad		1. Všechny akce 1.1. weapons/index 1.1.1. 58935 1.1.2. 48752 1.2. weapons/swords/index		visit_entry_action
Zdroje		piwik_log_action, piwik_log_visit		
Poznámka				

Dimenze Návštěvy Výstupní Akce

Tabulka č. 12 - Dimenze návštěvy výstupní akce

	Struktura	Prvky	Počty	Identifikace
Struktura	1.	Všechny akce	1	
	2.	Akce	18374	
	3.	Návštěvy	125795	
Příklad		1. Všechny akce 1.1. weapons/index 1.1.1. 3568 1.1.2. 95760 1.2. weapons/swords/index		visit_exit_action
Zdroje		piwik_log_action, piwik_log_visit		
Poznámka				

6.4.2 Ukazatele a vztahy ukazatelů a dimenzí

Tabulka č. 13 - Ukazatele a vztahy ukazatelů a dimenzí

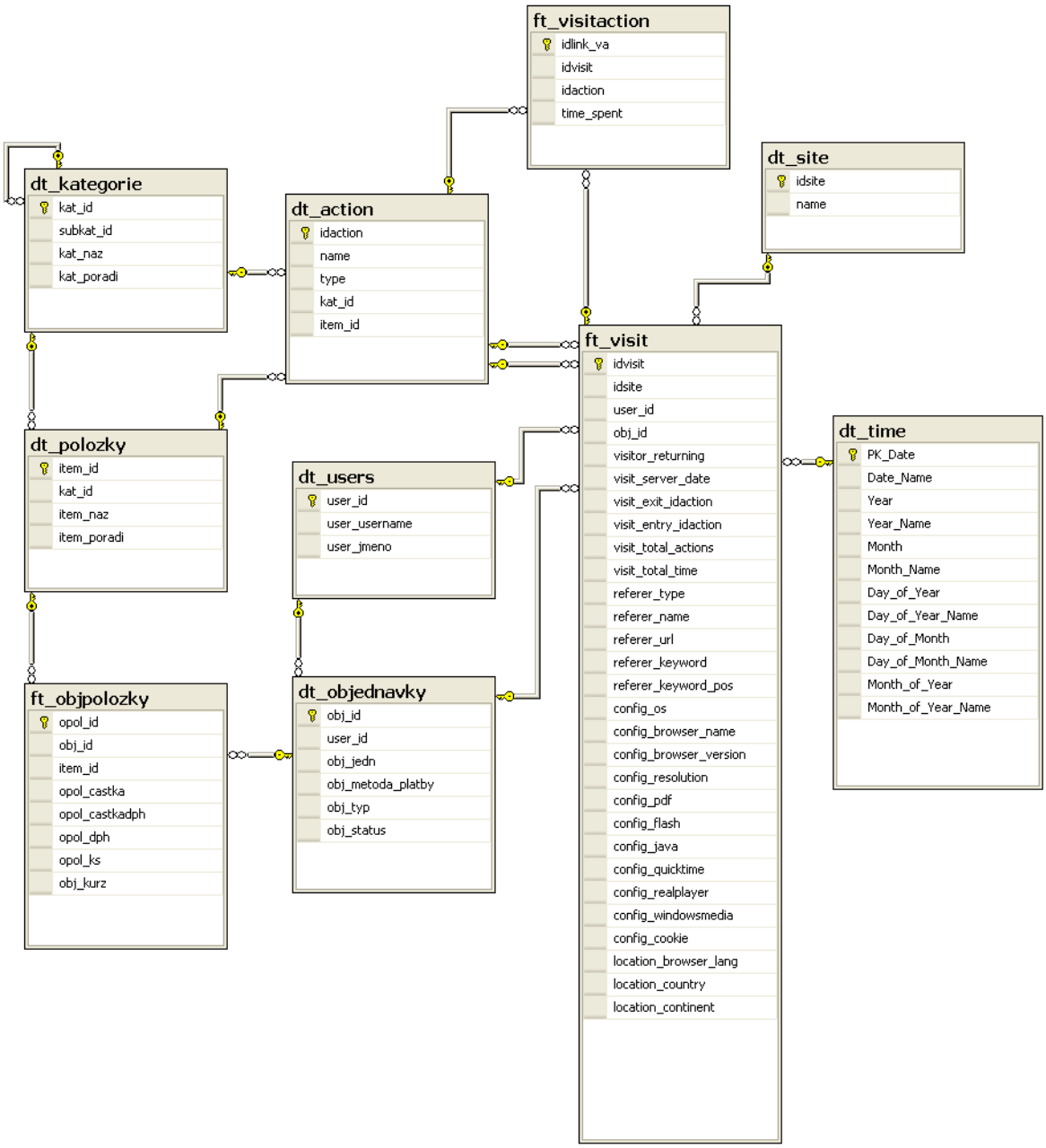
ID	Ukazatel	Jednotka	Zdroj	Výpočet	Pravidla	Web	Akce	Čas	Katalog Kategorie	Produkty	Návštěvy oblasti	Návštěvy zdroje	Návštěvy uživatelé	Návštěvy vstupní akce	Návštěvy výstupní akce
totalCZK	Částka v Kč za položku objednávky bez DPH	Kč	ft_objpolozky	Sum(opol_castka*opol_ks *obj_kurz)		X	X	X	X	X	X	X	X	X	X
totalCZK2	Částka v Kč za položku s DPH	Kč	ft_objpolozky	Sum(opol_castkadph*opol _ks*obj_kurz)		X	X	X	X	X	X	X	X	X	X
unit	Počet kusů	Ks	ft_objpolozky	Sum(opol_ks)		X	X	X	X	X	X	X	X	X	X
orderItemCount	Počet položek objednávky		ft_objpolozky	Count(opol_id)		X	X	X	X	X	X	X	X	X	X
visitCount	Počet návštěv		ft_visit	Count(idvisit)		X	X	X			X	X	X	X	X
visitActionCount	Počet navštívených stránek		ft_visitaction	Count(idlink_va)		X	X	X	X	X	X	X	X	X	X
timeSpent	Průměrná doba na stránce	sec	ft_visitaction	AVG(time_spent)		X	X	X	X	X	X	X	X	X	X

6.5 Model datového skladu

Po provedení analýzy je již možné přistoupit k tvorbě modelu datového skladu. Z analýzy dimenzí a ukazatelů je patrné, jaká má být jeho struktura a jaké mají být vazby mezi tabulkami dimenzí a tabulkami faktů.

Výsledný fyzický model datového skladu je znázorněn na obrázku č. 7.

Obrázek č. 8 - Fyzický model datového skladu



6.6 ETL

Po vytvoření fyzického modelu datového skladu je nutné naplnit tento sklad daty. K tomuto účelu sloučí ETL nástroje. Ty jsou obsaženy jako projekty Integration Services v rámci nástroje SQL Server Business Intelligence development studio.

Vytváření datových pump se skládá z posloupnosti datových toků seskupených do tzv. control flow. U každého datového toku musí být definovány jak datové zdroje pro čtení (fáze extraction), tak i datové zdroje pro zápis (fáze load). Mezi čtením dat a jejich zápisem do datového skladu pak ještě mohou probíhat různé transformace dat (fáze transformation) jako je čištění, odstranění duplicit, vytváření nových sloupců tabulek, atd.

Na obrázku č. 10 je zobrazen control flow vytvořený pro potřeby této práce. Z obrázku je patrné, že úvodním krokem je vytvoření dočasných tabulek potřebných při transformaci dat a dále následují datové toky sloužící k samotné transformaci a nahrání dat do datového skladu. Při vytváření datových pump bylo také nutné řešit několik problémů. Tím hlavním byly duplicity dat, které se ve zdrojové databázi vyskytovaly ve dvou podobách.

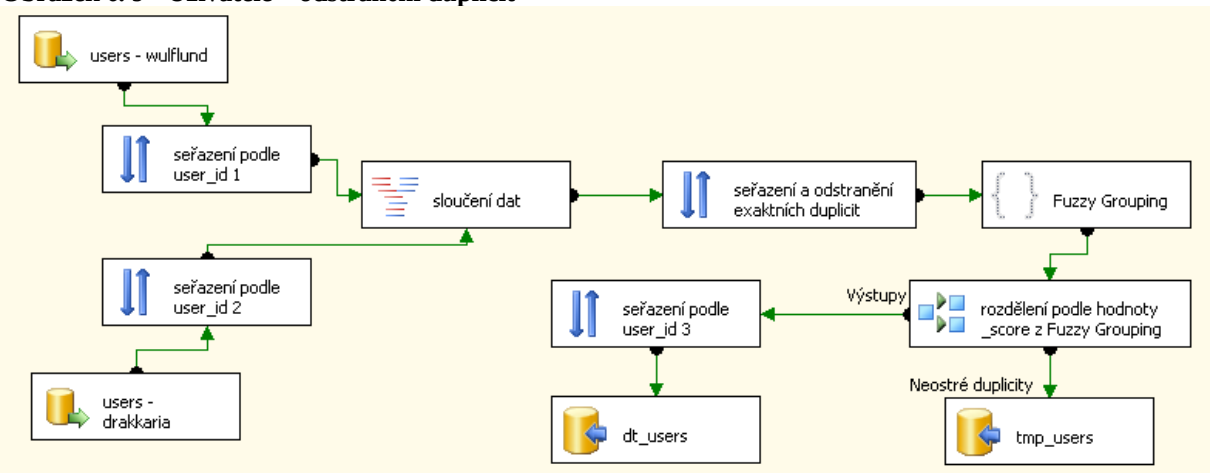
Tou první byly duplicitní záznamy o uživatelích. Tyto záznamy se vytvořily buď tím, že si některý uživatel postupem času založil více uživatelských účtů, anebo tím, že některý uživatel nakupoval jak bez předchozí registrace uživatelského účtu, tak následně již jako zaregistrovaný. K identifikaci těchto duplicit byla vybrána funkce Fuzzy grouping (viz obrázek č. 8). Tato funkce podle zadaných kritérií, které jsou nastaveny buď s parametrem Exact (přesná shoda), anebo s parametrem Fuzzy (fuzzy shoda), testuje, zda se dané záznamy podobají a následně je rozděluje na záznamy duplicitní a záznamy neduplicitní. Tuto funkci je poměrně obtížné nastavit tak, aby jednak nesloučila záznamy, které duplicitní nejsou a dále neponechala záznamy, které duplicitní jsou. Z hlediska přesnosti dat je ale dobré raději nastavit přísnější typ shody a některé duplicity tam zanechat, než sloučit odlišné záznamy.

Druhou podobou duplicit pak byly záznamy se shodným ID, které jednak mohly být duplicitní, ale také nemusely. Tyto duplicity vznikly tím, že obě databáze internetových obchodů využívají stejnou strukturu tabulek a každá má tedy svou vlastní řadu ID záznamů. Odstranění těchto duplicit bylo řešeno dvěma různými způsoby. První způsob se týkal tabulek obsahujících data o návštěvnosti. Tyto tabulky byly vytvořeny až po rozdělní obou verzí na odlišné hostingy. Neexistovaly zde tedy žádné společné záznamy z doby minulé. Z tohoto důvodu bylo možné vyřešit duplicity tak, že po extrakci dat ze zdrojových databází bylo u

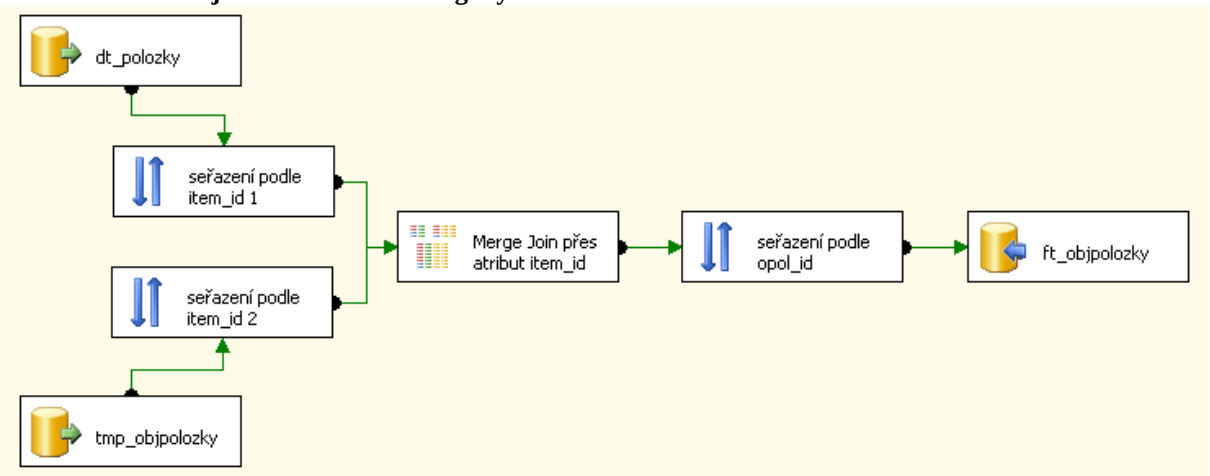
jedné části dat, konkrétně u dat ze zahraniční verze, zvýšeno ID tak, aby tato data po sloučení s daty z české verze již žádné záznamy se stejným ID mít nemohla. Druhý způsob řešení se týkal tabulek obsahující data o provozu databáze internetového obchodu. U těchto dat nebylo možné aplikovat první způsob, jelikož v tabulkách existovaly záznamy ještě z dob společné databáze a prostým zvýšením ID by tak vznikly nové duplicity. Bylo tedy nutné nejprve odstranit záznamy se stejným ID a ještě dalšími parametry, které zaručovaly shodu záznamů a následně vytvořit nová ID pro záznamy, které shodné nebyly.

Dalším problémem byla referenční integrita datového skladu. Ve zdrojových databázích se totiž vyskytovaly záznamy, které tuto integritu nesplňovaly a při jejich prostém přesunu do datového skladu by proces ETL skončil chybou. Tento problém byl vyřešen za pomoci již nahraných tabulek a také funkce Merge Join (viz obrázek č. 9). Pomocí této funkce bylo možné spojit extrahovaná data přes cizí klíč s daty, která se již v datovém skladu nacházela a vynechat tak záznamy, které by porušovaly referenční integritu tabulek.

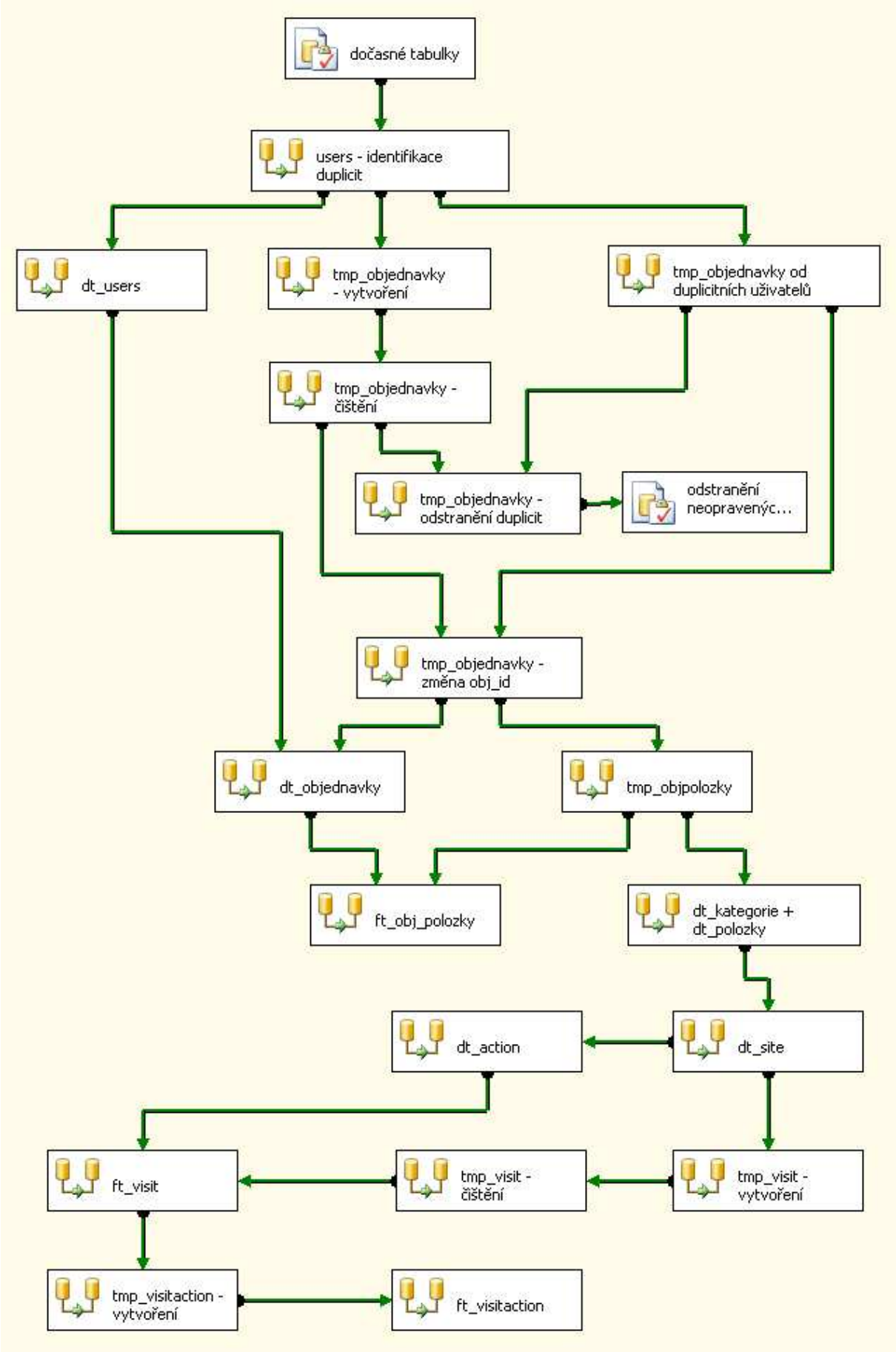
Obrázek č. 9 - Uživatelé – odstranění duplicit



Obrázek č. 10 - Zajištění referenční integrity



Obrázek č. 11 - ETL control flow



6.7 Vytvoření OLAP kostek

Tvorba OLAP kostek bude probíhat stejně jako tvorba datových pump v rámci nástroje SQL Server Business Intelligence development, jako projekt Analysis Services.

Vytvoření OLAP kostky se provádí ve čtyřech základních krocích.

Tím prvním je definování datového zdroje, ze kterého se budou čerpat data pro OLAP kostku. Definování datového zdroje lze učinit pomocí průvodce, kdy nejprve definujeme připojení databázovému serveru, následně vyplníme přihlašovací údaje a v posledním kroku nově vytvořený datový zdroj pojmenujeme.

Ve druhém kroku definujeme pohledy na datový zdroj. K jejich vytvoření opět použijeme průvodce. Výsledný pohled na datový zdroj nemusí obsahovat všechny tabulky datového skladu, stejně tak je možné měnit vztahy mezi těmito tabulkami a jejich atributy, podle potřeb zamýšlené OLAP kostky.

Třetím krokem je vytvoření dimenzí, kde opět využijeme integrovaného průvodce. Pokud si chceme ušetřit část práce, je dobré při tvorbě dimenzí zaškrtnout pole Auto build a ze seznamu vybrat položku Create attributes and hierarchies. Software pak následně vytvoří jak atributy dimenze, tak i případné její hierarchie. Tento výstup je samozřejmě možné následně upravit podle vlastních potřeb. Při tvorbě dimenzí pak vycházíme z analýzy, která byla provedena v kapitole 6.4.1.

V posledním kroku se již vytváří samotná OLAP kostka. Opět je možné využít průvodce a i zde platí doporučení pro zaškrtnutí pole Auto build a výběr položky Create attributes and hierarchies. Jedním z kroků při vytváření kostky je také určení tabulek, zda se jedná o tabulku faktů nebo tabulku dimenzí, po kterém následuje výběr ukazatelů, které byly definovány v kapitole 6.4.2. Z této kapitoly také vycházíme při kontrole vztahů mezi dimenzemi a ukazateli. Tyto úkony se provádějí po vytvoření kostky v záložce Dimension usage. Pokud jsou plánovány nějaké ukazatele, které se dopočítávají, lze tak učinit v záložce Calculated members.

Po vytvoření OLAP kostky je již možné přistoupit k vlastnímu výpočtu jejích dat a k vytváření sestav.

6.8 Ukázka výstupů a jejich využití v praxi

Pro vytváření sestav byla vybrána, jak již bylo zmíněno, aplikace Microsoft Excel.

V aplikaci Excel bylo nejprve nutné vytvořit připojení k datovému zdroji OLAP a následně nadefinovat základní kontingenční tabulky a grafy. Tyto tabulky a grafy je pak možné si různě přizpůsobovat, měnit si pohledy na data a granularitu dat tak, aby se dosáhlo požadovaných sestav.

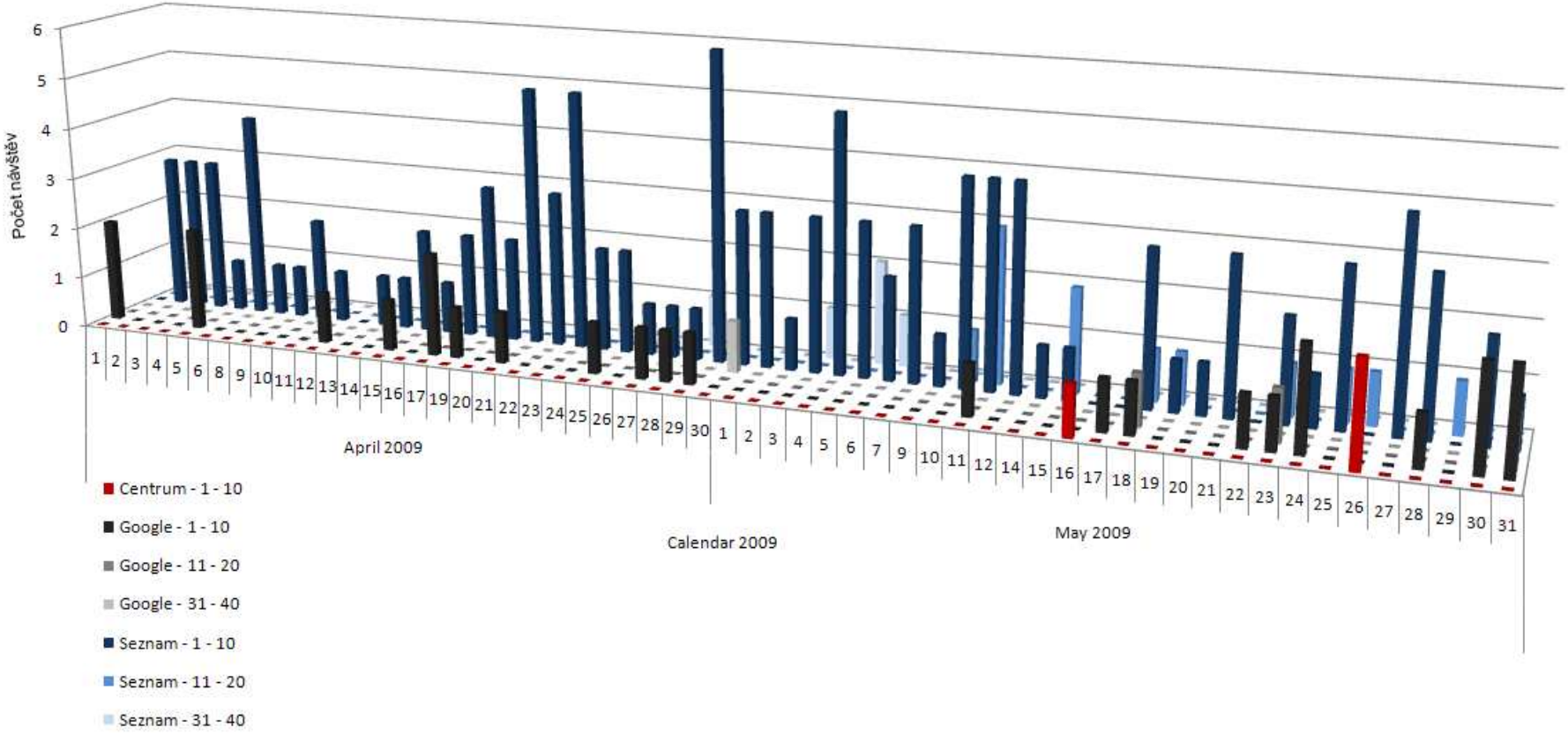
V následující části popíši ukázky tabulek a grafů, které byly vytvořeny pro potřeby této práce.

Obrázek č. 11 zobrazuje vývoj návštěvnosti na klíčové slovo *keltské symboly* v závislosti na čase, názvu vyhledávače a pozici klíčového slova ve vyhledávači. Z obrázku je patrné mírné snížení návštěv z první strany vyhledávače Seznam v měsíci květnu, které je velmi pravděpodobně dáno poklesem pozice klíčového slova v tomto vyhledávači. Tato sestava je velmi užitečná zejména při optimalizaci stránek pro zvolené klíčové slovo, jelikož poskytuje přehled o tom, zda se klíčové slovo nachází na předních stranách vyhledávače a zda přes něj na web někdo přichází. Lze z ní tedy vyčíst dva z hlavních ukazatelů úspěšnosti optimalizace.

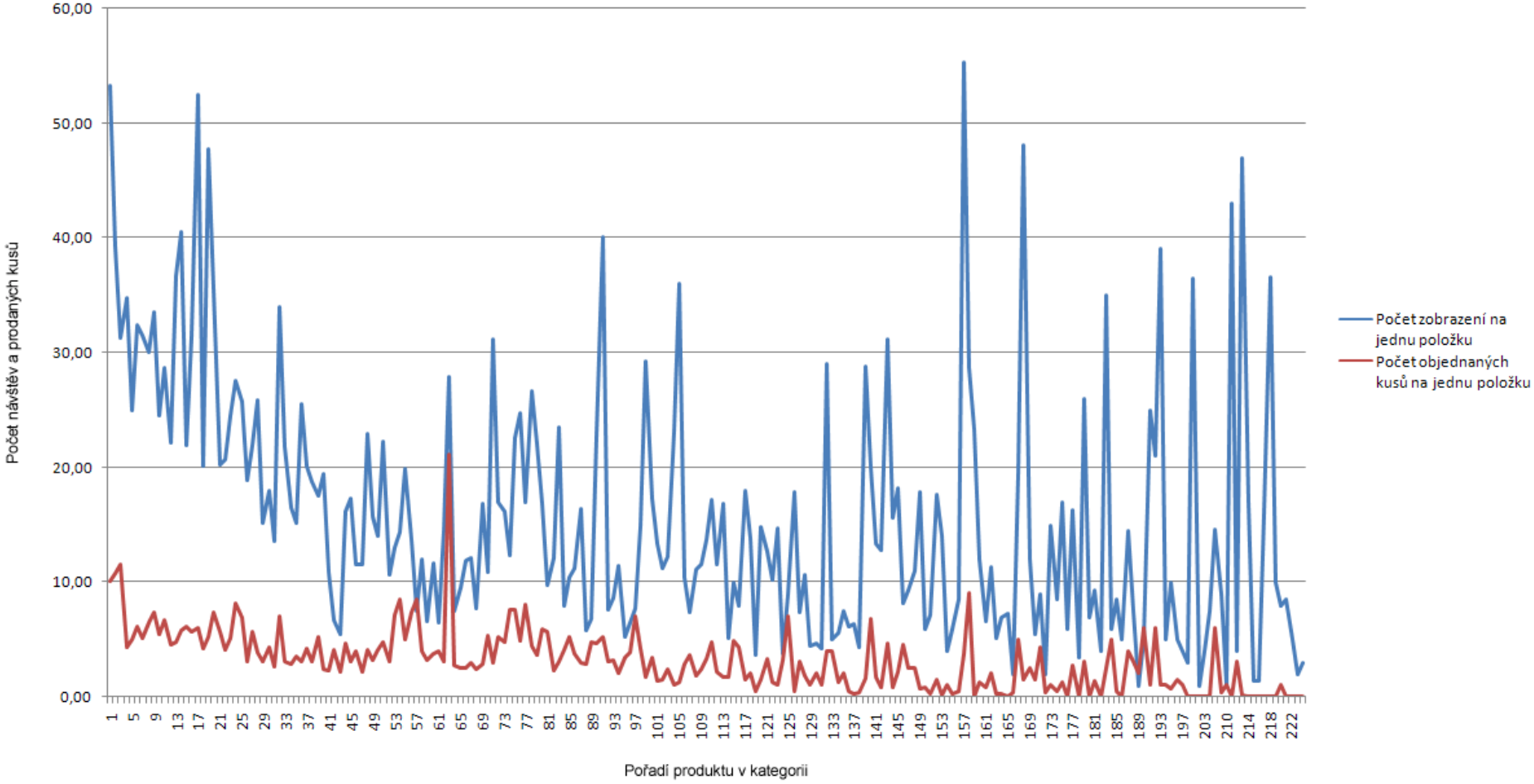
Obrázek č. 12 zobrazuje vliv pořadí výrobku v kategorii na počet zobrazení a počet prodaných kusů. Zatímco vliv pořadí na počet zobrazení je poměrně značný, tj. nejvíce je zobrazovaných výrobků na prvních dvou stranách (stránkování produktů na webu je 20 položek na stránku), tak vliv pořadí na počet prodaných kusů již tak značný není. Je tedy možné usuzovat, že návštěvníci, pokud mají o něco konkrétního zájem, jsou schopni nalézt produkty i ze zadních částí katalogu.

Obrázek č. 13 zobrazuje klíčová slova podle míry opuštění stránek, tj. jaké procento návštěvníků z celkového počtu návštěv stránky opustí web ihned po příchodu na něj. Po identifikaci těchto klíčových slov je ještě dobré podívat se na ně trochu detailněji. Tuto sestavu zobrazuje obrázek č. 14, kde je možné vidět, na jaké stránky návštěvníci přes dané klíčové slovo přišli. Zajímavá je situace zejména u klíčového slova *stojánek na náušnice*, přes které se návštěvníci dostanou na stránku produktu stojánku na náušnice. Dalo by se tedy předpokládat, že návštěvníci našli, co hledali, přesto 85% z nich ihned opouští web. To může být způsobeno buď špatným popisem, nekvalitním obrázkem nebo nepřiměřenou cenou. Ke snížení míry opuštění u této stránky tak bude třeba trochu experimentovat s těmito faktory s cílem návštěvníky na webu udržet.

Obrázek č. 12 - Vývoj návštěvnosti na klíčové slovo *keltské symboly* v závislosti na čase, názvu vyhledávače a pozici klíčového slova ve vyhledávači



Obrázek č. 13 - Vliv pořadí výrobku v kategorii na počet zobrazení a počet prodaných kusů



Obrázek č. 14 - Klíčová slova podle míry opuštění stránek

Klíčové slovo	Počet návštěv	Počet opuštění	Procento opuštění
products	70	70	100,00%
pendants	35	35	100,00%
price	15	15	100,00%
historical	11	11	100,00%
leather	27	25	92,59%
mayové aztékové inkové	19	17	89,47%
pompadůrka	21	17	80,95%
egyptian cross	30	24	80,00%
stojánek na naušnice	15	12	80,00%
bohyně plodnosti	18	14	77,78%
quetzalcoatl	35	27	77,14%
český znak	17	13	76,47%
kotva	18	13	72,22%
bubny	17	12	70,59%
sas	16	11	68,75%
drákulův hrad	35	24	68,57%
talisman online	43	28	65,12%
věstonická venuše	48	29	60,42%
mesicni znameni	25	15	60,00%
květinový stolek	33	19	57,58%
žralok kladivoun	66	37	56,06%
panna	27	15	55,56%
kolibrik	116	64	55,17%
kříž života	26	14	53,85%
ametystová geoda	55	29	52,73%
venuše z mélu	21	11	52,38%
karel iv.	25	13	52,00%
jelení shozy	26	13	50,00%
socha	51	25	49,02%
spartský král	35	17	48,57%
aztec mask	129	61	47,29%
aztec sculptures	32	15	46,88%
egyptští bohové	43	20	46,51%
metal	24	11	45,83%
aztec masks	122	55	45,08%
ametystové geody	29	13	44,83%
bohyně athéna	102	45	44,12%
tudor footwear	48	21	43,75%
fibula viking	39	17	43,59%
pavouci	53	22	41,51%
bohyně lásky	30	12	40,00%
vyrezavani ze dreva	31	12	38,71%
mosaz	29	11	37,93%
hamsa	55	20	36,36%

Obrázek č. 15 – Klíčová slova podle míry opuštění stránek (rozepsáno podle stránek webu)

Klíčové slovo - stránka	Počet návštěv	Počet opuštění	Procento opuštění
bohyně plodnosti	**18**	**14**	**77,78%**
drevene-vyrobky/sosky-a-reliefy/bohyne-ziva-bohyne-lasky-a-plodnosti.html	17	14	82,35%
images_items/bohyne-ziva-bohyne-lasky-a-plodnosti_2.jpg	1	0	0,00%
egyptian cross	**30**	**24**	**80,00%**
images_items/ankh-egyptian-cross-of-live_2.jpg	1	0	0,00%
images_items/ankh-s-insigniemi_2.jpg	1	0	0,00%
jewelry/amulets-and-talismans/magic-pendants/index	28	24	85,71%
leather	**27**	**25**	**92,59%**
armour/leather-armour/?page=2	1	1	100,00%
armour/leather-armour/carved-leather-bracers.html	1	1	100,00%
armour/leather-armour/fantasy-bracers-with-turquoise.html	1	1	100,00%
armour/leather-armour/leather-arm-bracers---whole-arm.html	1	1	100,00%
armour/leather-armour/leather-bracers---black.html	1	1	100,00%
armour/leather-armour/leather-bracers---special.html	1	1	100,00%
armour/leather-armour/leather-bracers-with-dragons.html	1	1	100,00%
armour/leather-armour/leather-cuirass.html	1	1	100,00%
armour/leather-armour/leather-vambraces-35-mm-thick.html	1	1	100,00%
armour/leather-armour/pair-of-brown--leather-bracers.html	1	1	100,00%
kontakt.php	1	0	0,00%
kuze---brasnarstvi/vlasove-spony/kozena-spona-do-vlasu---celtic-iii.html	1	0	0,00%
leather-products/bags/bag-for-nobleman-with-belt.html	1	1	100,00%
leather-products/bags/leather-bag---brown---black.html	1	1	100,00%
leather-products/bags/leather-bag-celtic-cross.html	1	1	100,00%
leather-products/bags/leather-bag-for-women---celtic-cross.html	1	1	100,00%
leather-products/bags/medieval-leather-pouch---big.html	1	1	100,00%
leather-products/hair-brooches/celtic-leaf-hair-clasp.html	1	1	100,00%
leather-products/hair-brooches/hand-carved-leather-hair-brooch-i.html	1	1	100,00%
leather-products/hair-brooches/leather-hair-brooch---celtic-design.html	1	1	100,00%
leather-products/wallets/black-leather-celtic-wallet.html	1	1	100,00%
leather-products/wallets/case-for-documents---thor-hammer.html	1	1	100,00%
leather-products/wallets/document-case---paleolithic-inspiration.html	1	1	100,00%
leather-products/wallets/index	1	1	100,00%
leather-products/wallets/leather-celtic-wallet.html	1	1	100,00%
leather-products/wallets/viking-style-leather-wallet.html	1	1	100,00%
sitemap.php	1	1	100,00%
mayové aztékové inkové	**19**	**17**	**89,47%**
knihy-a-cd/symboly-inku-mayu-a-azteku.html	14	13	92,86%
plastiky/amerika/index	5	4	80,00%
pompadůrka	**21**	**17**	**80,95%**
images_items/pompadurka---francouzska-lilie_2.jpg	2	0	0,00%
kuze---brasnarstvi/brasny-mesce/pompadurka---francouzska-lilie.html	19	17	89,47%
stojánek na náušnice	**15**	**12**	**80,00%**
sperky/doplnky---krabicky/index	1	0	0,00%
sperky/postribrene-sperky/sperky-de-luxe/stojanek-na-nausnice-a-prsteny.htr	14	12	85,71%

6.8.1 Další možné sestavy

Kromě výše zmíněných sestav je samozřejmě možné vytvářet i mnohá další. Jednou z možností by byla identifikace stránek s produkty, které ještě nebyly zobrazeny nebo které byly zobrazeny velmi málokrát. Tento seznam by pak bylo dobré projít a identifikovat produkty, které chceme prodávat a ty pak na stránkách obchodu nějakým způsobem propagovat.

V případě, že vedeme PPC kampaně, pak by rozhodně neměla chybět sestava objednávek a zobrazení stránek filtrovaná podle přístupů z těchto kampaní. Na jejím základě by pak bylo možné přesně vidět, přes která klíčová slova uživatelé přišli a zda si na webu něco objednali. Na základě těchto dat je pak velmi snadné optimalizovat PPC kampaň a soustředit se pouze na ta klíčová slova, která přinášejí největší zisk.

Jedním z dalších možných rozšíření by bylo přizpůsobovat obsah webu uživatelům na základě jejich předchozího chování. Pokud by uživatel navštěvoval zejména stránky s výrobky mečů, bylo by možné po přihlášení někam do layoutu webu umístit odkazy např. na nové produkty v této kategorii, apod.

6.9 Náklady projektu

Jedním z dílčích cílů této práce bylo také vyčíslení nákladů na implementaci. Tyto náklady jsou v podstatě dvojího druhu. Jednak náklady na software a dále náklady na práci při implementaci.

Do nákladů na software byly zahrnuty kromě nákladů na pořízení SQL serveru, který pravděpodobně většina firem nevlastní, také náklady na Microsoft Excel, který je sice poměrně rozšířen, ale mnoho menších firem může využívat jeho open source alternativu v podobě OpenOffice Calc. V případě, že někdo již ať jeden či druhý software vlastní, snižují se náklady na projekt o danou částku.

Tabulka č. 14 - Náklady projektu

Položka	**Částka s DPH**
Nástroj pro analýzu návštěvnosti - Piwik	0
MySQL / PostgreSQL databáze	0
Microsoft Office Basic 2007	4350[3]
Microsoft SQL Server Standard Edition	55700[4]
Implementace BI řešení (100 hodin * 500Kč/hod) - Instalace nástroje pro analýzu návštěvnosti 10 hodin - Příprava dat 10 hodin - Analýza datového skladu 25 hodin - Model datového skladu 5 hodin - ETL 25 hodin - Tvorba OLAP kostky 10 hodin - Příprava sestav 15 hodin	50000
Celkem	**110050 Kč**

Celkové náklady jsou tedy cca 110 tisíc korun. Tato částka může být pro některé menší podniky poměrně velkým výdajem. Řešením by mohlo být nasazení některého z open source OLAP nástrojů s výslednou úsporou jedné poloviny nákladů. Otázkou ale zůstává, zda jsou tyto nástroje dostatečnou alternativou nástrojům komerčním.

[3] http://alza.cz/kancelarsky-balik-oem-microsoft-office-basic-2007-cz-mlk-d71852.htm (14. 6. 2009)
[4] http://www.devshop.sk/microsoft/i61507-sql-server-2005 (14. 6. 2009)

7 Závěr

Cílem této práce bylo analyzovat návštěvnost internetového obchodu za využití technologií a postupů Business Intelligence.

V úvodu své práce byl nejprve definován pojem optimalizace pro vyhledávače, který byl rozšířen i o optimalizaci pro uživatele. Následně byly popsány základní postupy při optimalizaci, ovlivňující externí faktory a přiblížena problematika placených kampaní.

Další část se týkala výběru nástroje analýzy. Nejprve byla popsána obecná kritéria výběru a následně také kritéria výběru reflektující požadavky na nástroj analýzy z hlediska této práce.

V další kapitole byla nastíněna teoretická východiska pro BI. Zejména se jednalo o způsoby tvorby datového skladu a o proces tvorby BI projektu.

Následující část popisovala současnou situaci analyzovaného internetového obchodu, architekturu řešení, jeho postavení na trhu a také vize do budoucna. Obsah této kapitoly byl využit zejména při tvorbě úvodní studie projektu.

Na základě předchozích kapitol pak bylo přistoupeno již k samotnému řešení projektu analýzy návštěvnosti.

7.1 Srovnání předpokládaných a dosažených cílů

Hlavní cíl

- Využití BI při analýze internetového obchodu – kapitola 6

Dílčí cíle

- přiblížit a porovnat některé služby poskytující analýzu návštěvnosti – kapitola 3
- představit požadavky na datovou základnu – kapitoly 6.3 a 6.4
- vyčíslit náklady na implementaci – kapitola 6.9
- ukázat možnosti využití BI i pro malé podniky – kapitola 6

Lze tedy konstatovat, že jak hlavní, tak i dílčí cíle této práce byly úspěšně splněny.

7.2 Přínosy

Z hlediska zkoumaného internetového obchodu spatřuji přínos práce zejména v získání nových a jinak nedostupných analýz týkajících se návštěvnosti stránek s možností jejich využití v rámci následné optimalizace těchto stránek jak pro uživatele, tak i pro vyhledávače.

Z hlediska obecného spatřuji hlavní přínos v ukázání možností využití BI v rámci malých firem a ve vyčíslení nákladů na implementaci těchto řešení. Dále je také možné předpokládat, že při implementaci řešení analýzy návštěvnosti jiného internetového obchodu s využitím nástroje Piwik, bude možné aplikovat některé z postupů a výstupů této práce.

7.3 Omezení

Jedním z omezení práce byla chyba ve zdrojovém kódu internetového obchodu, která spočívala v odstraňování ID uživatele z tabulky identifikující návštěvníka při jeho odhlášení. Z tohoto důvodu pak nebylo možné využít data o uživatelích při sestavování reportů.

Dalším omezením pak byla poměrně krátká doba sběru dat, nedostatečná pro vyšší vypovídací hodnotu některých reportů.

7.4 Návrhy na další řešení

Další řešení v rámci internetového obchodu již bylo naznačeno v úvodní studii projektu v rámci kapitoly 6.2. Jedná se zejména o implementaci druhého inkrementu tvorby datového skladu, tj. dat o provozu podniku a s tím souvisejících reportů.

Poměrně zajímavé by také mohlo být rozpracovat a zhodnotit open source OLAP nástroje. V případě jejich dostatečné funkčnosti a použitelnosti by se tak naskytla velmi dobrá alternativa ke komerčním nástrojům zvláště pro malé firmy, které nechtějí či nemohou investovat tisíce do nákupu software.

Literatura

[1] Crucial Marketing: Search Engine Optimization. 2007. Dostupné z internetu <http://www.marketingterms.com/dictionary/search_engine_optimization/>

[2] ČSSI on line (Česká společnost pro systémovou integraci) [online]. 2006

[3] Factum Invenio: Nakupování na internetu. 2008. Dostupné z internetu <http://factum.cz/tz302.html>

[4] Grappone, J., Cousin, G.: SEO. Brno: Zoner Press, 2007. ISBN – 978-80-86815-85-5.

[5] Lacko, L.: Business Intelligence v SQL serveru 2005, Brno: Computer Press 2006, ISBN 80-251-1110-5

[6] Dr. Peter J. Meyers: SUO. 2009. Dostupné z internetu <http://www.usereffect.com/topic/suo-search-and-user-optimization>

[7] Mužík, Z.: Web Analytics. Praha: Vysoká škola ekonomická, 2006.

[8] Novotný, O., Pour, J., Slánský, D.: Business Inteligence. Praha: Grada 2005. 256s. ISBN 80–247–1094–3.

[9] PETERSON, E. T.: Web Site Measurement Hacks, O'Reilly, 2005, ISBN 0–596–00988–7

[10] Jesse Purdue: IRACIS - A Roadmap to Business Intelligence ROI. 2009. Dostupné z internetu <http://ezinearticles.com/?IRACIS---A-Roadmap-to-Business-Intelligence-ROI&id=673225>

[11] Martin Schiller: Co se skrývá pod zkratkou ETL? 2003. Dostupné z internetu <http://www.systemonline.cz/clanky/co-se-skryva-pod-zkratkou-etl.htm>

[12] U. S. Department of Health & Human Services: What is usability? 2009. Dostupné z internetu <http://www.usability.gov/basics/whatusa.html>

[13] Iva Valečková: Na co nezapomenout při budování datového skladu. 2009. Dostupné z internetu <http://www.sophias.cz/cs/co-delame/publikacni-cinnost/na-co-nezapomenout-v-projektu-budovani-datoveho-skladu.html>

[14] Wikipedia: PPC. 2009. Dostupné z internetu <http://cs.wikipedia.org/wiki/PPC>

[15] Wikipedia: Search Engine Optimization. 2009. Dostupné z internetu <http://en.wikipedia.org/wiki/Search_engine_optimization>

Terminologický slovník

Business Intelligence

Business Intelligence (BI) je sada postupů, procesů a technologií, jejímž cílem je účinně a účelně podporovat rozhodovací procesy ve firmě. Představuje komplex aplikací IS/ICT, které podporují analytické a plánovací činnosti podniků a organizací a jsou postaveny na specifických, tzv. OLAP (On-Line Analytical Processing) technologiích a jejich modifikacích.

Datový sklad

Datový sklad (Data Warehouse) je integrovaný, subjektově orientovaný, stálý a časově rozlišený souhrn dat, uspořádaný pro podporu potřeb managementu.

ETL

ETL jsou nástroje, které umožňují propojit různé databáze a přenášet data mezi těmito databázemi. Přitom se provádějí operace, jako jsou transformace datových typů a formátů, oprava dat, výpočet agregací, doplňování primárních klíčů, atd. Tyto nástroje umožňují vytvořit datové pumpy (data stages) tzn. definovat zdrojové databáze, cílové databáze a nastavit transformační proces.

OLAP

OLAP (On-Line Analytical Processing) je informační technologie založená především na koncepci multidimenzionálních databází. OLAP je základní technologie, na níž jsou založeny aplikace business intelligence. Jejím hlavním principem je několikadimenzionální tabulka umožňující rychle a pružně měnit jednotlivé dimenze a měnit tak pohledy uživatele na modelovanou ekonomickou realitu.

SEO

Proces výběru cílených klíčových slov vztahujících se k dané webové stránce a zajištění, že se tato slova po zadání do vyhledávače dobře umístí.

PPC

Princip Pay Per Click spočívá v tom, že inzerent za reklamu platí až ve chvíli, kdy zaúčinkovala. Inzerent tedy platí až ve chvíli, kdy na reklamu někdo klikl. Oproti klasickým

bannerům nebo jiné plošné reklamě má tu výhodu, že zákazník má přesnou kontrolu nad svými výdaji, tedy že zaplatí jen tolik prokliků, potažmo návštěvníků, kolik chce.

Analýza návštěvnosti

Analýza návštěvnosti je monitorování a vykazování dat o použití internetové prezentace, díky nimž může firma lépe porozumět komplexním interakcím mezi jednáním návštěvníka a nabídkami zprostředkovanými internetovou prezentaci a tohoto porozumění využít pro zvýšení prodejů nebo loajality zákazníků.

Seznam tabulek

Tabulka č. 1 - Vyhodnocení nástrojů analýzy podle vlastních kritérií 29
Tabulka č. 2 - Dimenze 49
Tabulka č. 3 - Dimenze web 49
Tabulka č. 4 - Dimenze akce 49
Tabulka č. 5 - Dimenze čas 50
Tabulka č. 6 - Dimenze kategorie 50
Tabulka č. 7 - Dimenze produkty 51
Tabulka č. 8 - Dimenze návštěvy oblasti 51
Tabulka č. 9 - Dimenze návštěvy zdroje 52
Tabulka č. 10 - Dimenze návštěvy uživatelé 52
Tabulka č. 11 - Dimenze Návštěvy vstupní akce 53
Tabulka č. 12 - Dimenze návštěvy výstupní akce 53
Tabulka č. 13 - Ukazatele a vztahy ukazatelů a dimenzí 54
Tabulka č. 14 - Náklady projektu 66

Seznam obrázků

Obrázek č. 1 - Piwik 25
Obrázek č. 2 - Google Analytics 27
Obrázek č. 3 - TOPlist.cz 28
Obrázek č. 4 - Inkrementální metoda tvorby datového skladu 34
Obrázek č. 5- Náhled zahraniční verze internetového obchodu (www.wulflund.com) 38
Obrázek č. 6 - Google Trends - porovnání objemu vyhledávání 40
Obrázek č. 7 - Návrh architektury řešení 46
Obrázek č. 8 - Fyzický model datového skladu 55
Obrázek č. 9 - Uživatelé – odstranění duplicit 57
Obrázek č. 10 - Zajištění referenční integrity 57
Obrázek č. 11 - ETL control flow 58
Obrázek č. 12 - Vývoj návštěvnosti na klíčové slovo *keltské symboly* v závislosti na čase, názvu vyhledávače a pozici klíčového slova ve vyhledávači 61
Obrázek č. 13 - Vliv pořadí výrobku v kategorii na počet zobrazení a počet prodaných kusů 62
Obrázek č. 14 - Klíčová slova podle míry opuštění stránek 63
Obrázek č. 15 – Klíčová slova podle míry opuštění stránek (rozepsáno podle stránek webu) 64

Přílohy

Příloha 1 – úprava tabulky akcí nástroje Piwik

Soubor piwik/core/Tracker/Action.php

```
class Piwik_Tracker_Action implements Piwik_Tracker_Action_Interface {
    // inicializace proměnných
    private $kat_id;
    private $item_id;
    …
    private function generateInfo() {
        // rozdělení názvu akce (předáno javascriptem – viz níže)
        $this->kat_id = StrTok($actionName,",");
        $this->item_id = StrTok(",");
        $actionName = ""; // vynulování názvu akce, aby se do něj mohlo uložit URL
    }
    …
    private function loadActionId() {
        $category = $this->kat_id;
        $item = $this->item_id;
        if($idAction === false) {
            Piwik_Tracker::getDatabase()->query(
                INSERT INTO ". Piwik_Common::prefixTable('log_action'). "(
                name,type, kat_id, item_id ) VALUES (?,?,?,?)",
                array($name,$type,$category,$item)
            );
            $idAction = Piwik_Tracker::getDatabase()->lastInsertId();
        }
    }
}
```

Úprava javascriptového kódu ve stránce internetového obchodu

```
piwik_action_name = '<?php echo $kat_id.",".$item_id; ?>';
```

Příloha 2 – dekódování klíčových slov a zápis jejich pozic

```
// piwik_log_visit - keywords_positions
$sql = "SELECT * FROM piwik_log_visit_back WHERE referer_url!='' AND
    referer_name!='' AND referer_name!='Google Images' AND referer_type=2";
@$vysledek = mysql_query($sql);
while ($zaznam = mysql_fetch_array($vysledek)) {
    $url = parse_url($zaznam[referer_url]);
    parse_str($url[query],$query);

    if ($query['amp;start']>0) {
        $min = $query['amp;start']+1;
        $max = $query['amp;start']+10;
    }
    elseif ($query['amp;from']>0) {
        $min = $query['amp;from'];
        $max = $query['amp;from']+$query['amp;count']-1;
    }
    elseif ($query['amp;b']>0) {
         $min = $query['amp;b'];
         $max = $query['amp;b']+10;
    }
    else {
         $min = 1;
         $max = 10;
    }

    $sql = "UPDATE piwik_log_visit SET referer_keyword_pos='".$min." - ".$max."'
        WHERE idvisit=$zaznam[idvisit]";
    @$vysledek2 = mysql_query($sql);
}

// piwik_log_visit - keywords
$sql = "SELECT * FROM piwik_log_visit";
@$vysledek = mysql_query($sql);
while ($zaznam = mysql_fetch_array($vysledek)) {
    $referer_keyword = iconv('utf-8', 'cp1250', urldecode($zaznam[referer_keyword]));
    $referer_keyword = iconv('cp1250', 'utf-8', $referer_keyword);

    $sql = "UPDATE piwik_log_visit SET
        referer_keyword='".addslashes($referer_keyword)."' WHERE
        idvisit=$zaznam[idvisit]";
    @$vysledek2 = mysql_query($sql);
}
```