Dynamicky generované tabuľky dimenzií a ich využitie v marketingu
Dynamically generated dimension tables and their use in marketing

Robert Halenár

Abstract: The paper describes the possibility of using data warehouses and their potential in the field of marketing. Describes basic methods transfer data from routine information systems and databases into the data warehouse structures. The databases containing customer data is present time a large number of diverse data that is in the process of ETL homogenized and clean. Often follows a table filled with an extremely high volume. Article describes how to reduce the time dimension table, the accuracy of time customer data remains unchanged.

Key words: Data warehouse, table, dimension, ETL

Abstrakt: Článok opisuje možnosť využitia dátových skladov a ich potenciálu v oblasti marketingu. Popisuje základné spôsoby prevodu údajov z bežných informačných systémov a ich databáz do štruktúr dátového skladu. V databázach obsahujúcich údaje o zákazníkoch sa vyskytuje veľké množstvo rôznorodých časových údajov, ktoré sa pri procese ETL homogenizujú a čistia. Nezriedka má takto naplnená tabuľka extrémne veľký objem. Článok opisuje spôsob zmenšenia tabuľky časovej dimenzie, pričom presnosť časových údajov o zákazníkovi zostáva nezmenená.

Kľúčové slová: Dátový sklad, tabuľka, dimenzia, ETL

1. Úvod

Dopyt po čerstvých údajov v dátových skladoch bola vždy silná požiadavka zo strany užívateľov. Tradične aktualizácia dátových skladov bola vykonávaná v off-line móde. V takom dátovom sklade, sú údaje získané zo zdrojov, transformované, vyčistené, a nakoniec načítané do skladu. Tento súbor aktivít sa odohráva počas pravidelných dávkových aktualizácií, zvyčajne v noci, aby sa zabránilo preťaženiu zdrojových produkčných systémov mimoriadnou pracovnou záťažou tohto pracovného toku. Zaujímavé je, že záťaž vynaložená na tento proces bola jedným zo základných dôvodov pre zriadenie dátových skladov. Okamžité šírenie zmien, ktoré sa konajú je technicky nemožné, či už kvôli súčasnej povahe

predmetných zdrojov, alebo jednoducho kvôli obmedzeniu prevádzkových zdrojových systémov. Vo väčšine prípadov je dátový sklad zvyčajne aktualizované každých 24 hodín.

Potreba budovania dátových skladov je tým väčšia, čím zložitejšie algoritmy dolovania dát sú použité. (1)

za zdroj. V takom prípade sa transakcia pri zdrojovej strane sa stáva pružnejšou, pretože dáta, ktoré sa zobrazujú v mieste zdroja webu sú k dispozícii okamžite. Obchodné potreby – napr. rastúca konkurencia, potreba väčšieho predaja, lepšie sledovanie zákazníka alebo cieľ, presné monitorovanie na akciovom trhu, a tak ďalej – generujú vysoký dopyt po presných správach a výsledkoch založených na aktuálnych údajoch, a nie na údajoch zo včera. Zvyčajne sa proces ETL vykonáva počas noci, pretože v dopoludňajších hodinách je situácia zložitejšia, navyše ak vezmeme do úvahy, že pobočky organizácie môžu byť rozdelené v miestach s úplne rôznymi časovými zónami. Na základe týchto skutočností, sa dátové sklady vyvinuli na "aktívne" alebo "živé" produkčné dátové systémy pre užívateľov. V prevádzke sa začínajú správať a reagovať ako samostatné operačné systémy. Funkcie, ktoré boli predtým nedostupné (napríklad on-demand žiadosti o informácie) sú teraz použiteľné pre koncových užívateľov. V súčasnosti je aktualizácia stanovuje v intervaloch rádovo minútového oneskorenia, a nie hodinového alebo dokonca celodňového. V dôsledku toho sa v tradičných ETL procesoch čoraz viac používa pojem "reálny čas" alebo "takmer v reálnom čase". Takže je trend, aby údaje pohybujúce sa od zdroja smerom do dátového skladu, boli menšieho objemu, častejšie, a rýchlejším tempom.(2)

2. Metódy ETL

ETL trh už zareagoval na tieto nové požiadavky. Hlavný predajcovia ETL už dodávajú riešenia schopné pracovať "v reálnom čase" s ich tradičnými platformami. V praxi, také riešenia zahŕňajú softvérové balíky, ktoré umožňujú použitie ľahkých transformácií „on-the-fly“, aby sa minimalizoval čas potrebný na vytvorenie určitých správ. K oneskoreniu medzi jednotlivými transakciami dochádza na strane operačného procesu a času oneskorenia, ktorý sa šíri do cieľového miesta je len pár minút, zvyčajne päť až pätnásť. Takáto odpoveď je charakterizovaná ako "takmer v reálnom čase“.

Tradičné ETL procesy sa používali na zavádzanie údajov do dátového skladu a to ako pre prvotné napĺňanie na začiatku fungovania skladu, až po celú dobu prevádzky skladu v off-line režime. Zdá sa, že dátové sklady sa stali obeťou svojho úspechu: užívatelia sú spokojní s

údajmi, ktoré sú jeden deň staré a stlačením tlačidla získajú čerstvé údaje - pokiaľ možno, s okamžitými správ. Tento druh žiadosti je technicky náročné z rôznych dôvodov. Po prvé, zdrojové systémy nesmú byť preťažované extra úlohami migrácie dát smerom do skladu. Po druhé, nie je zrejmé, ako môže byť vykonávané aktívne šírenie dát, a to najmä so staršími systémami. Problém sa stáva ešte ťažšie riešiteľným, pretože je často krát nutná re – konfigurácia softvéru zdrojov, aby bola schopná reagovať správne na novú úlohou, vzhľadom k potrebe:

(a) nižšieho času pre nasadenie a testovanie, a

(b) vyšších nákladov na správu, údržbu, a monitorovanie nového prostredia. Dlhodobá vízia pre systémy dátových skladov fungujúce „takmer v reálnom čase“ spočívajú v samo – ladiacich sa architektúrach, kde sú užívateľské požiadavky na čerstvosť splnené na najvyššiu možnú mieru bez toho, aby narušili požiadavky administrátorov na priepustnosť a dostupnosť ich systémov.

Viac pragmatický prístup predstavujú poloautomatické prostredia, kde sa užívateľ posudzuje a delí do kategórií podľa požiadaviek na čerstvosť a úplnosť požadovaných dát.

Všeobecné ETL architektúry pracujúce na takmer reálnom čase v sklade údajov sú založené na databázových zdrojoch obsahujúcich nástroje, ktorý tlačia získané dáta do dočasného skladu. Potom pripravia podklady pre proces transformácie do transformačnej funkcie – tzv. pripravený dátový formát. Transformácia prebieha v oblasti DPA- data processing area (oblasť spracovania dát), kde sú dáta prevedené a vyčistené a potom sú dáta exportované ďalej „loaderom“. Loader načíta dáta do dátového skladu priamo do tabuliek faktov a dimenzií.(1)

3. Tabuľky dimenzií

Pri navrhovaní dimenzionálneho modelu väčšinou dostávame len jednu, alebo len malé množstvo tabuliek faktov, ale veľké množstvo tabuliek dimenzií. Rozmery stola môžu byť uvedené ako referenčné údaje pre tabuľku faktov, kde sa nachádzajú popisy a ďalšie statické informácie o výrobku. Výrobok sa považuje za rozmer, pretože v takejto tabuľke, sa nachádza všetko o výrobku, ako napríklad celé meno produktu, dodávateľov a rozmer palety. V tabuľke

faktov je stĺpec s názvom "product_key," ktorého hodnoty sa používajú na načítanie všetkých informácií o výrobku z tabuľky dimenzií.

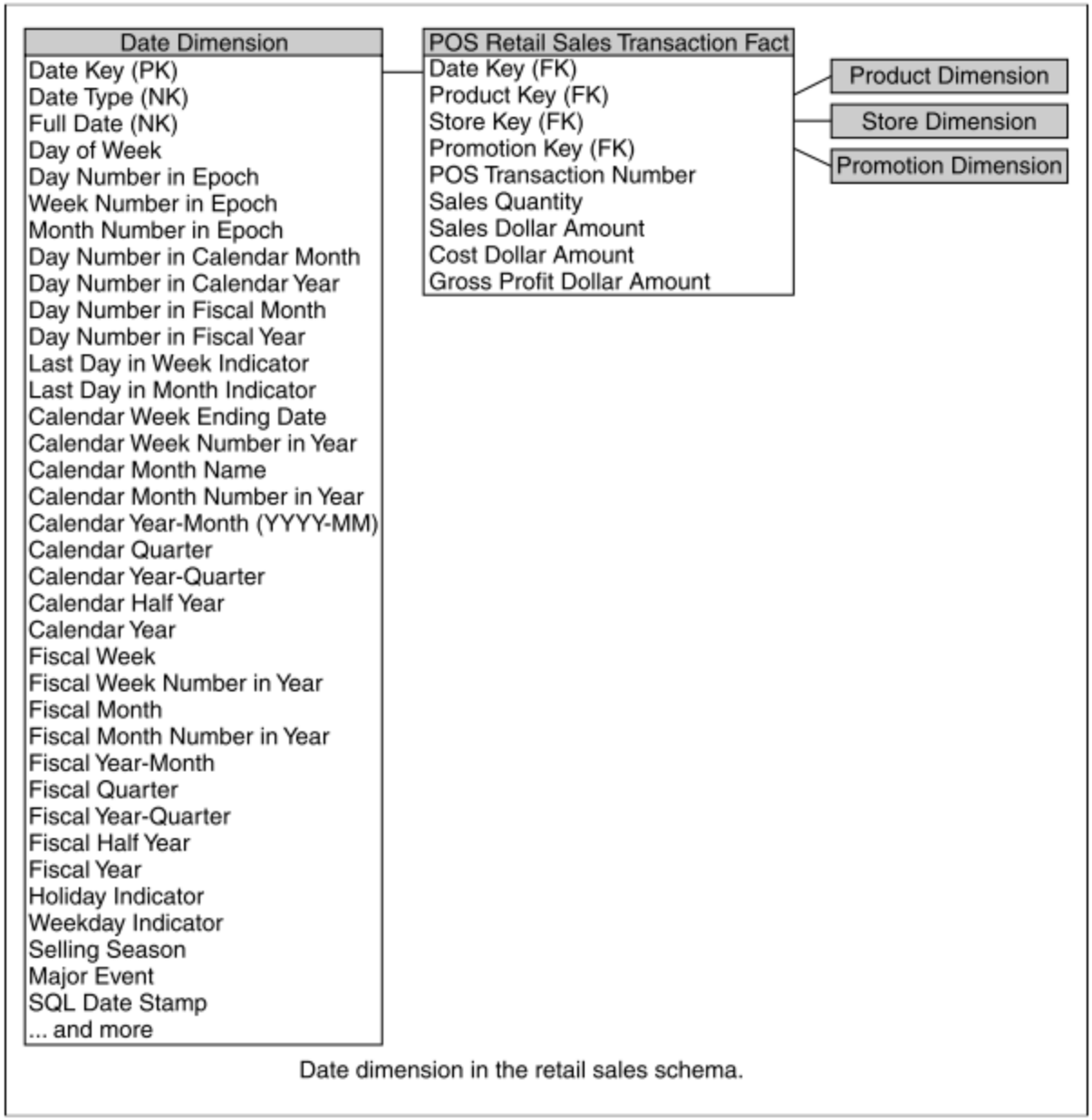
Ak si nie ste istí, či sa jedná o dáta faktov alebo dimenzií, položte si nasledujúce otázky: Sú dáta relatívne statické? Popisujú niečo? Tabuľky dimenzií majú tendenciu obsahovať viac textových polí, ktoré popisujú kótovaný objekt, zatiaľ čo tabuľky faktov majú tendenciu obsahovať viac číselných údajov. Napríklad, tabuľka faktov môže obsahovať milióny riadkov, zatiaľ čo tabuľka dimenzií môže mať iba niekoľko riadkov (napr. časová dimenzia by nemusela mať viac ako 52 riadkov, ak boli dáta uložené týždenne po dobu jedného roka). Alebo región môže obsahovať len 15 riadkov, v prípade, že krajina mala iba 15 regiónov.

Tabuľky dimenzií nemusia mať malé rozmery, pretože môžu uchovávať údaje o 50 000 výrobkoch, alebo údaje o 5 miliónoch zákazníkov. To všetko môžu byť dimenzie výrobku.(1)

4. Tabuľky časových dimenzií

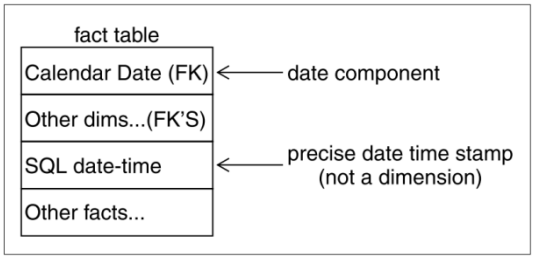
Prakticky každá tabuľka faktov poukazuje prostredníctvom referencií cudzích kľúčov na jednu alebo viac tabuliek časových dimenzií. Merania sú prevádzané v určitých intervaloch, pričom za jednotku času je možné urobiť viac meraní.

Najzákladnejšou tabuľkou časovej dimenzie sú dáta kalendára s delením na jeden deň. Prekvapujúco majú tieto tabuľky relatívne veľké množstvo atribútov, pozri Obr.1. Len niektoré z týchto atribútov môžu byť generované priamo v SQL vyjadrení. Prázdniny, dovolenkové obdobia, fiškálne periódy, pracovné dni, čísla týždňov, posledný deň v mesiaci musia byť súčasťou tabuľky časovej dimenzie. Dimenzia obsahujúca údaje kalendára má mnoho nezvyčajných atribútov. Je to jedna z mála tabuliek dimenzií, ktoré sú kompletne spracované už na začiatku projektu výstavby dátového skladu. Taktiež väčšinou tieto údaje väčšinou nepochádzajú z konvenčných zdrojov. Najvhodnejší spôsob ako vygenerovať ucelenú tabuľku časových dimenzií je urobiť ju ručne za jedno popoludnie. Tabuľka obsahujúca údaje o dňoch z desiatich rokov môže mať viac ako 4000 riadkov.(3)



Obrázok 1. Atribúty potrebné pre tabuľku dimenzií dátumov(4)

V niektorých tabuľkách faktov sú potrebné merania času ktoré súd pod rozlíšením jedného dňa, v minútach, alebo dokonca v sekundách. Človek nedokáže ručne takúto tabuľku spracovať, pretože jeden rok obsahuje viac ako 31 miliónov sekúnd. V praxi požadujeme údaje kalendára a súčasne môžeme chcieť presné odpovede na minúty a sekundy. Tiež môžeme požadovať výpočty presných časových intervalov porovnávaním dvoch presných časov v tabuľke faktov. Z týchto dôvodov odporúčame navrhovať tabuľku podľa Obr.2.



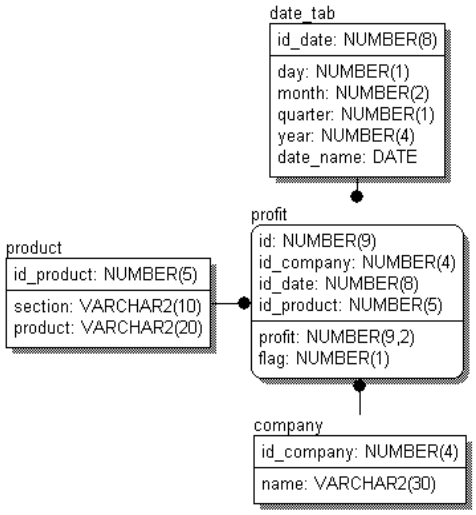
Obrázok 2. Návrh tabuľky faktov na uchovávanie presných časových meraní (4)

Tiež, na rozdiel od rozmeru kalendárnych dní, vo väčšine prostredí je len veľmi málo popisné atribúty pre konkrétne minútu alebo druhý v rámci dňa.

Ak podnik nemá už dobre definované atribúty pre čas rezy v rámci jedného dňa, ako posun názvy vysielaných reklamných časov, ďalší time-of-day rozmer môžu byť pridané do dizajnu, kde je tento rozmer definovaný ako počet minút (alebo dokonca sekúnd) po polnoci. Tým by sa táto doba-dňový rozmeru mať buď 1440 záznamy, ak obilia boli minút alebo 86.400 záznamy, ak obilia boli sekúnd. Prítomnosť tejto dobe denného dimenzie neodstráni potrebu SQL date-time stamp bolo popísané vyššie. (5)

5. Aplikácia v dátovom sklade

V našom výskume sme aplikovali architektúru podľa obrázku 3. Principiálne sa jedná o subjektovo orientovanú architektúru, takže dáta sú organizované podľa subjektu a nie podľa výrobcu, oddelenia, alebo fyzickej adresy. Takže napríklad dáta z výroby sú nezávislé na dátach z predaja, reklamy a pod. (6)



Obrázok 3. Architektúra dátového skladu

Náš dátový sklad je zameraný na dáta predaja, konkrétne zisk v závislosti na zákazníkovi a produkte. Pozostáva z troch tabuliek dimenzií a jednej tabuľky faktov. Tabuľka časovej dimenzie môže obsahovať aj viac ako státisíce až milióny záznamov. Príklad je v tabuľke č.1.

Tabuľka 1. Pôvodná tabuľka časových dimenzií

id_date	*day*	*month*	*quarter*	*year*	*date_name*		*id_date*	*day*	*month*	*quarter*	*year*	*date_name*
1	1	January	1	2012	01.01.2012		26	9	February	1	2012	09.02.2012
2	2	January	1	2012	02.01.2012		27	10	February	1	2012	10.02.2012
3	3	January	1	2012	03.01.2012		28	11	February	1	2012	11.02.2012
4	4	January	1	2012	04.01.2012		29	12	February	1	2012	12.02.2012
5	5	January	1	2012	05.01.2012		30	13	February	1	2012	13.02.2012
6	6	January	1	2012	06.01.2012		31	5	April	2	2012	05.04.2012
7	7	January	1	2012	07.01.2012		32	6	April	2	2012	06.04.2012
8	8	January	1	2012	08.01.2012		33	7	Apríl	2	2012	07.04.2012
9	9	January	1	2012	09.01.2012		34	8	Apríl	2	2012	08.04.2012
10	10	January	1	2012	10.01.2012		35	9	Apríl	2	2012	09.04.2012
11	11	January	1	2012	11.01.2012		36	10	Apríl	2	2012	10.04.2012
12	12	January	1	2012	12.01.2012		37	11	Apríl	2	2012	11.04.2012
13	13	January	1	2012	13.01.2012		38	12	Apríl	2	2012	12.04.2012
14	14	January	1	2012	14.01.2012		39	13	Apríl	2	2012	13.04.2012
15	15	January	1	2012	15.01.2012		40	14	Apríl	2	2012	14.04.2012
16	16	January	1	2012	16.01.2012		41	15	Apríl	2	2012	15.04.2012
17	17	January	1	2012	17.01.2012		42	16	Apríl	2	2012	16.04.2012
18	18	January	1	2012	18.01.2012		43	17	Apríl	2	2012	17.04.2012
19	2	February	1	2012	02.02.2012		44	18	Apríl	2	2012	18.04.2012
20	3	February	1	2012	03.02.2012		45	19	Apríl	2	2012	19.04.2012
21	4	February	1	2012	04.02.2012		46	20	Apríl	2	2012	20.04.2012
22	5	February	1	2012	05.02.2012		47	21	Apríl	2	2012	21.04.2012
23	6	February	1	2012	06.02.2012		48	22	Apríl	2	2012	22.04.2012
24	7	February	1	2012	07.02.2012		49	23	Apríl	2	2012	23.04.2012
25	8	February	1	2012	08.02.2012		50	24	Apríl	2	2012	24.04.2012

Kompresným algoritmom môžeme dosiahnuť oveľa menšiu tabuľku, ktorá obsahuje len zlomok pôvodnej tabuľky časovej dimenzie. Je nutné si uvedomiť, že pôvodná tabuľka časovej dimenzie nemusí byť zoradená a preto musíme najprv usporiadať dáta v tabuľke a až potom komprimovať.
redukovaná tabuľka časovej dimenzie je skoro 10 krát menšia ako pôvodná tabuľka, pozri Tab. 2.

Tabuľka 2. Redukovaná tabuľka časových dimenzií

id_date	*day*	*month*	*quarter*	*Year*	*date_name*
1	1	January	1	2012	01.01.2012
2	18	January	1	2012	18.01.2012
3	2	February	1	2012	02.02.2012
4	13	February	1	2012	13.02.2012
5	5	April	2	2012	05.04.2012
6	24	Apríl	2	2012	24.04.2012

Ak potrebujeme dáta z tabuľky kvôli výpočtom, musíme najskôr spustiť algoritmus a znovu naplniť tabuľku tak, aby obsahovala pôvodné dáta, ako je v tabuľke č.3. v tabuľkách časových dimenzií, kde nie je presne zadefinovaný najmenší časový interval, ho najskôr musíme definovať a následne spustiť proces znovu naplnenia. Celý process beží automaticky, zakiaľ s nevyskytne systémová chyba.

Napriek tomu je nutné dôkladne otestovať všetky alternatívy. (8)

Tabuľka 3. Redukovaná tabuľka časových dimenzií v procese znovu napĺňania

id_date	*day*	*month*	*quarter*	*Year*	*date_name*
1	1	January	1	2012	01.01.2012
...					
...	*space*	*for*	*data*	*population*	
...					
2	18	January	1	2012	18.01.2012
3	2	February	1	2012	02.02.2012
...					
...	*space*	*for*	*data*	*population*	
...					
4	13	February	1	2012	13.02.2012
5	5	April	2	2012	05.04.2012
...					
...	*space*	*for*	*data*	*population*	
...					
6	24	Apríl	2	2012	24.04.2012

Musíme dať zvlášť pozor na to, aby sa obnovili všetky závislosti s pôvodnými údajmi ostatných tabuliek faktov.

6. Záver

V názornom príklade je tabuľka časových dimenzií veľmi zjednodušená. V skutočnosti dátový sklad uchováva veľké množstvo záznamov a ich závislostí na čase. Preto je potrebné tabuľky najskôr homogenizovať, resp. zvážiť pre aký účel budú dimenzie slúžiť, aby bol celý proces znovu naplnenia automatizovaný. Vhodnou dátovou štruktúrou môžeme nie len redukovať uchovávanie veľkého množstva údajov, ale následne aj zrýchliť procesy a umožniť tak ich využitie v reálnom čase.

Kompresný algoritmus je možné použiť tiež v dátovom sklade, kde sa nachádzajú údaje potrebné k analýze spotrebiteľského nákupného koša.(7)

Článok bol prezentovaný na konferencii Nové trendy v marketingu 2012.

Zoznam bibliografických odkazov

(1) Hobbs, L., Hillson, S., Lawande, S., Smith, P.: *Oracle database 10g Warehousing*. Elsevier, Inc. USA Oxfort, 2005. Pp. 837. ISBN 1-55558-322-9.

(2) Kebisek, M.,Schreiber, P., Halenar, I.: *Knowledge Discovery in Databases and its application in manufacturing.* In International workshop Innovation Information Technologies – Theory and Practice; 2010 September 06-10, Dresden, Germany, pp. 204-207. ISBN 978-3-941405-10-3

(3) Kimball, R., Caserta, J.: *The data warehouse ETL toolkit*. Wiley Publishing, Inc. USA Indianapolis, 2004. Pp. 491. ISBN 0-764-57923-1.

(4) Kozielski S, Wrembel R, *New Trends in Data Warehousing and Data Analysis*. Springer; 2009. ISBN 978-0-387-87430-2

(5) Reeves, L.: *A Manager's Guide to Data Warehousing*, Published by Wiley Publishing, Inc.; 2009. ISBN: 978-0-470-17638-2

(6) Sivers, F.: *Building and Maintaining a Data Warehouse*, CRC Press; 2008. ISBN 978-1-4200-6462-9.

(7) Trnka A: *Market basket analysis with data mining methods*. In ICNIT 2010: International Conference on Networking and Information Technology; 11-12 June 2010, Manila, Philippines. ISBN 978-1-4244-7578-0

(8) Zeman, J., Tanuska, P., Kebisek, M.: *The Utilization of Metrics Usability To Evaluate The Software Quality*. In: ICCTD 2009 : International Conference on Computer Technology and Development. 13-15 November 2009, Kota Kinabalu, Malaysia. IEEE Computer Society, 2009. - ISBN 978-0-7695-3892-1

Adresa autora

Ing. Robert Halenár, PhD.
Univerzita sv. Cyrila a Metoda v Trnave
Fakulta masmediálnej komunikácie
Nám. J. Herdu 2
917 00 Trnava
robert.halenar@ucm.sk