Ukážka predspracovania v programe MS Excel

V tejto časti opisujeme ukážku predspracovania dát. Dáta pre projekt, môžeme mať z viacerých zdrojov alebo môžu byť vygenerované. Naše ukážkové dáta pochádzajú z transakčných informačných systémov, kde záznamy reprezentujú nákupy zákazníkov v sieti predajní. Sieť predajní je zameraná na predaj drobného kovového materiálu, náradia a stavebného materiálu.

Z transakčných systémov sme získali nasledujúce dáta.

ID	PRODUKT	CENA	DATUM	SKUPINA	NAZOV_PREDAJNE	OKRES	KRAJ
295	podlozka M12	7,5	1.1.2004	spotrebny material	MIBU zeleziarstvo	Banská Bystrica	Banskobystricky kraj
296	hmozdina M8	0,5	1.1.2004	spotrebny material	MIBU zeleziarstvo	Banská Bystrica	Banskobystricky kraj
297	hmozdina M10	15,9	1.1.2004	spotrebny material	MIBU zeleziarstvo	Banská Bystrica	Banskobystricky kraj
298	samorezna skrutka M10	4	1.1.2004	spotrebny material	MIBU zeleziarstvo	Banská Bystrica	Banskobystricky kraj
299	samorezna skrutka M12	5,1	1.1.2004	spotrebny material	Zeleziarstvo	Roznava	Kosicky kraj
300	lanko	32,8	1.1.2004	spotrebny material	Zeleziarstvo	Roznava	Kosicky kraj
301	brusny kotuc	118	1.1.2004	spotrebny material	Zeleziarstvo	Roznava	Kosicky kraj

Ešte pred spracovaním dát potrebujeme identifikovať základné informácie a to:

- merateľný fakt
- dimenzie

Fakt

Pod pojmom fakt, rozumieme numerickú merateľnú jednotku. V našom prípade to bude cena predaného produktu. V tomto prípade máme predaje v rámci SR, ale ak by sme mali predaje napr. v rámci Európy a Ázie, tak by mali viacero mien, reprezentujúcich predaje. V takom prípade je potrebné komunikovať s manažmentom a stanoviť vhodnú menu napr. Euro.

Dimenzia

Dimenzie, sú popisy obchodnej činnosti, obsahujú logicky alebo hierarchicky usporiadané údaje. Ak sa pozrieme na tabuľku vyššie, tak nájdeme dimenziu produktov, dátumov, miesta predaja a samotných značiek predajní

Máme identifikované potrebné informácie, ktoré teraz použijeme na rozdelenie transakcií na jednotlivé tabuľky dimenzií a faktu.

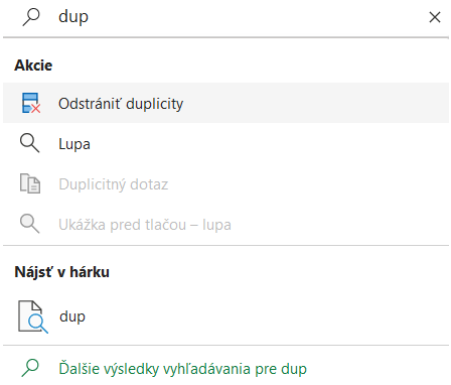
Dimenzia produktov

Hierarchia tejto dimenzie pozostáva zo samotného produktu a skupiny produktov. V programe Excel si vytvoríme nový hárok, pomenujeme ho **produkty**. Do bunky **A1** si vložíme popis ID_PRODUKTU. Do stĺpca B vložíme najjemnejšiu úroveň hierarchie. V prípade tejto dimenzie to je samotný produkt. Skopírujeme teda stĺpec PRODUKT z hárka transakcie a vložíme ho do stĺpca **B** hárku produkty. Rovnako vložíme aj stĺpec SKUPINA. Medzivýsledok vidíme nižšie.

	A	B	C
1	ID_PRODUKTU	PRODUKT	SKUPINA
2		samorezna skrutka M12	spotrebny material
3		lanko	spotrebny material
4		brusny kotuc	spotrebny material
5		rezny kotuc	spotrebny material
6		zavlacka 6mm	spotrebny material
7		poistny kruzok M10	spotrebny material
8		poistny kruzok M12	spotrebny material

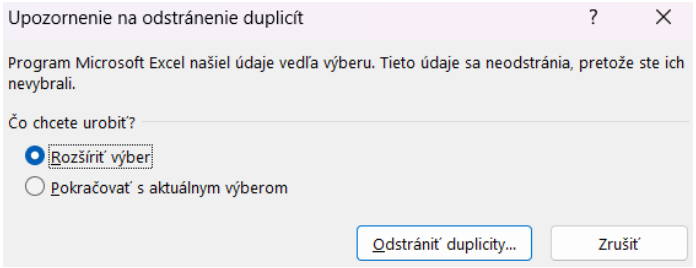
Obrázok 1 Začiatok dimenzie PRODUKTY

Označíme stĺpec B, v hornom menu prejdeme do karty Údaje, kde nájdeme možnosť **Odstrániť duplicity** alebo vyhľadáme pomocou vyhľadávania.



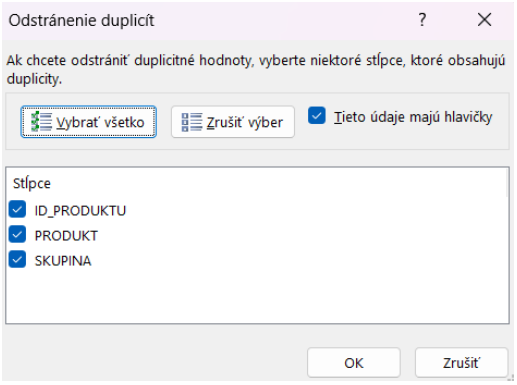
Obrázok 2 Odstránenie duplikátov – krok 1

V novom okne, necháme možnosť Rozšíriť výber a klikneme na **Odstrániť duplicity**.



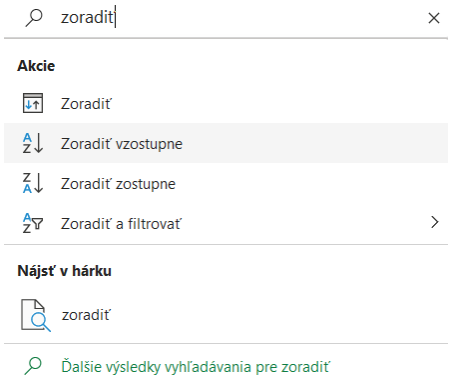
Obrázok 3 Odstránenie duplikátov – krok 2

V ďalšom kroku, môžeme zrušiť výber pre stĺpec ID_PRODUKTU a klikneme na **OK**. Zrušenie výberu je potrebné iba ak by boli v stĺpci hodnoty.



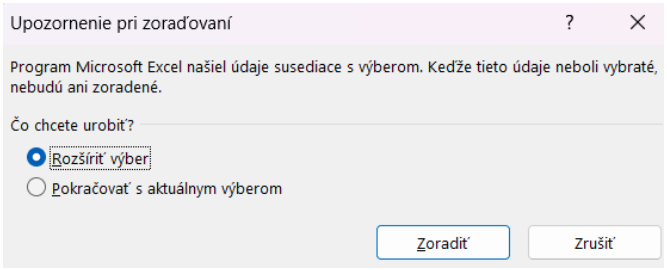
Obrázok 4 Odstránenie duplikátov – krok 3

Po odstránení duplikátov, potrebujeme zoradiť tento zoznam. Označíme stĺpec B a v karte Údaje nájdeme tlačidlo **Zoradiť od A po Z** alebo vo vyhľadávači nájdeme **Zoradiť vzostupne**.



Obrázok 5 Zoradenie – krok 1

Otvorí sa nové okno, v ktorom len klikneme na Zoradiť.



Obrázok 6 Zoradenie – krok 2

Máme zoradený zoznam produktov a skupín do ktorých patria. Nakoniec len vygenerujeme identifikátory. Do bunky **A2** vložíme číslo 1 a do bunky **A3** vložíme číslo 2. Označíme bunky A2 a A3, kurzorom prejdeme do spodného pravého rohu, kurzor sa zmení na symbol **+**. Pomocou dvojkliku sa vygeneruje postupnosť čísel až po koniec hárka.

	A	B	C
1	ID_PRODUKTU	PRODUKT	SKUPINA
2	1	brusny kotuc	spotrebny material
3	2	brusny papier	spotrebny material
4		cement 50kg	stavebny material
5		dlazdice 30x30 1 m2	stavebny material
6		dlazdice 50x50 1 m2	stavebny material
7		farba 1kg	stavebny material
8		hmozdina M10	spotrebny material
9		hmozdina M12	spotrebny material
10		hmozdina M8	spotrebny material
11		lanko	spotrebny material

Obrázok 7 Vygenerovanie identifikátorov

Výsledok sú identifikátory, až po koniec hárku.

	A	B	C
43	42	vrtak na zelezo	spotrebny material
44	43	zavitova tyc M10	spotrebny material
45	44	zavitova tyc M12	spotrebny material
46	45	zavitova tyc M8	spotrebny material
47	46	zavlacka 4mm	spotrebny material
48	47	zavlacka 5mm	spotrebny material
49	48	zavlacka 6mm	spotrebny material
50			
51			

Obrázok 8 Výsledná dimenzia

Takto upravená dimenzia je pripravená na vloženie do databázy. Tento postup môžeme opakovať aj pre dimenziu predajní a miest.

Dimenzia predajní

Výsledná dimenzia je zobrazená na obrázku nižšie. Postup je rovnaký ako pri dimenzií produktov.

	A	B
1	ID_PREDAJNE	NAZOV_PREDAJNE
2	1	Baumax
3	2	Hornbach
4	3	Kovomat
5	4	Lapal
6	5	Mega zeleziarstvo
7	6	MIBU zeleziarstvo
8	7	Proving
9	8	Tesco
10	9	Urob si sam
11	10	Zaleziarstvo merkur
12	11	Zeleziarstvo

Obrázok 9 Výsledok úpravy dimenzie predajní

Dimenzia lokalít

Znovu postupujeme ako pri dimenzií produktov. Najprv vkladáme najmenší celok. V tomto prípade je najmenší okres a potom nasleduje kraj. Výslednú dimenziu vidíme na obrázku nižšie.

	A	B	C
1	ID_MIESTA	OKRES	KRAJ
2	1	Banská Bystrica	Banskobystricky kraj
3	2	Bardejov	Presovsky kraj
4	3	Bratislava IV	Bratislavsky kraj
5	4	Bratislava V	Bratislavsky kraj
6	5	Brezno	Banskobystricky kraj
7	6	Cadca	Zilinsky kraj
8	7	Detva	Banskobystricky kraj
9	8	Dolny Kubin	Zilinsky kraj
10	9	Galanta	Trnavsky kraj
11	10	Kosice	Kosicky kraj
12	11	Levice	Nitriansky kraj
13	12	Liptovský Mikuláš	Zilinsky kraj
14	13	Martin	Zilinsky kraj
15	14	Michalovce	Kosicky kraj
16	15	Myjava	Tremciansky kraj
17	16	Nové Zámky	Nitriansky kraj
18	17	Pezinok	Bratislavsky kraj
19	18	Piešťany	Trnavsky kraj
20	19	Považská Bystrica	Tremciansky kraj
21	20	Presov	Presovsky kraj
22	21	Prievidza	Tremciansky kraj
23	22	Roznava	Kosicky kraj
24	23	Ružomberok	Zilinsky kraj
25	24	Snina	Presovsky kraj
26	25	Topoľčany	Nitriansky kraj
27	26	Trencin	Tremciansky kraj
28	27	Trnava	Trnavsky kraj
29	28	Vranov nad Topľo	Presovsky kraj
30	29	Zvolen	Banskobystricky kraj
31	30	Žilina	Zilinsky kraj

Obrázok 10 Dimenzia miest

Časová dimenzia

V tejto dimenzií máme dostupný len dátum. Ale samotný dátum obsahuje množstvo informácií, ktoré vieme extrahovať. Začneme vložením stĺpca dátum do nového hárka. Vymažeme duplikáty, prípadne zoradíme. Vytvoríme si hlavičku pre väčšie časové celky. To znamená že nám pribudne MESIAC, KVARTAL a ROK. Tieto hodnoty vieme ručne vložiť alebo môžeme použiť funkcie Excelu.

Mesiace

Na tvorbu popisov existuje viacero spôsobov. Jedným je že pomocou priradenia hodnoty bunke (=B2), vložíme dátum do bunky a potom pravým tlačidlom myši otvoríme zoznam a vyberieme položku Formátovať bunky. V zozname prejdeme do **Vlastné** a vložíme formát:

- [$-sk-SK]mmmm-yyyy;@

Tento formát reprezentuje slovenský názov mesiaca a rok. Výsledná hodnota pre 1.1.2004 bude január-2004. Tieto texty vieme následne upravovať napr. pomocou funkcií ako UPPER alebo PROPER.

Kvartály

Aj pre kvartály máme viacero možností ako ich získať. Stačí napríklad vložiť tento formát:

- ="Q" &INT((MONTH(B2)+2)/3) & "_" & YEAR(B2)

Získame tak texty ako Q1_2004. Bunka B2 reprezentuje prvý výskyt dátumu, z ktorého chceme extrahovať tieto údaje.

Roky

Pre roky môžeme použiť časť vyššie použitého výrazu YEAR(B2) alebo môžeme si zvoliť vlastný formát cez **Formátovanie buniek**.

	A	B	C	D	E
1	ID_DATUM	DATUM	MESIAC	KVARTAL	ROK
2		1.1.2004	január-2004	Q1_2004	=YEAR(B2)
3		1.2.2004	február-2004	Q1_2004	
4		1.3.2004	marec-2004	Q1_2004	

YEAR(sériové_číslo)

Obrázok 11 Formátovanie pre výber roku z dátumu

Tvorba a mapovanie tabuľky faktov

Vytvoríme nový hárok, napr. s názvom **fakt**. Stĺpec A bude reprezentovať identifikátory transakcií. Do stĺpca B vložíme merateľný fakt. V našom prípade to je stĺpec s názvom CENA. Ďalšie stĺpce pomenujeme rovnako ako stĺpce identifikátorov v našich dimenziách. Predpripravená tabuľka by mala vyzerať ako na obrázku nižšie.

	A	B	C	D	E	F
1	ID_FAKT	CENA	ID_MIESTA	ID_DATUM	ID_PREDAJNE	ID_PRODUKTU
2		6,8				
3		32,8				
4		154,4				
5		1,9				
6		4,6				
7		66,6				
8		74,2				
9		6,3				
10		8,7				

Obrázok 12 Medzivýsledok tabuľky faktov

Pre nájdenie cudzích kľúčov z dimenzií použijeme funkciu LOOKUP v tvare:

- =LOOKUP(transakcie!G2;miesta!B:B;miesta!A:A)

Kde hodnoty oddeľujeme bodkočiarkou. Prvá hodnota je názov miesta, ktorého **ID chceme nájsť**, druhá hodnota je zoznam usporiadaných miest (celý stĺpec **B** hárku **miesta**) a tretia hodnota je ID, ktoré chceme vložiť to tabuľky faktov ak sa táto hodnota nájde (stĺpec **A** hárku **miesta**).

Prvá hodnota v transakciách v bunke G2 je Detva. Funkcia tento názov nájde v hárku miesta a priradí mu hodnotu 7 z vygenerovaných identifikátorov. Táto funkcia by mohla vrátiť aj samotný názov, ak by sme tretiu hodnotu zmenili z A na stĺpec B. Zvyšné hodnoty doplníme dvojklikom na symbol **+** označenej bunky.

Rovnako budeme postupovať aj pri iných dimenziách. Vyhľadávame najmenšiu úroveň v rámci hierarchie. **Označovanie celých stĺpcov** aj s hlavičkou **nie je problém**, keďže v transakciách nemáme, napr. produkt s názvom PRODUKT tak mu nebude priradená hodnota ID_PRODUKTU. Výsledná tabuľka faktu je zobrazená nižšie.

F4 =LOOKUP(transakcie!B4;produkty!B:B;produkty!A:A)

	A	B	C	D	E	F
1	ID_FAKT	CENA	ID_MIESTA	ID_DATUM	ID_PREDAJNE	ID_PRODUKTU
2	1	6,8	7	1	9	32
3	2	32,8	7	1	9	10
4	3	154,4	7	1	9	1
5	4	1,9	7	1	9	29
6	5	4,6	7	1	9	48
7	6	66,6	7	1	9	25
8	7	74,2	7	1	9	26
9	8	6,3	7	1	9	36
10	9	8,7	7	1	9	45
11	10	19,7	7	1	9	42
12	11	29	7	1	9	44
13	12	29,8	7	1	9	41
14	13	2	7	1	9	16

Obrázok 13 Výsledok predspracovania tabuľky faktu

Záver

Takto upravené dáta vieme importovať do databázy. Nie všetky databázy môžu podporovať prácu so súbormi programu Excel s viacerými hárkami. Niekedy je preto potrebné hárky rozdeliť do samostatných súborov a následne ich importovať do databázy. Pri kopírovaní dát, je potrebné používať vlastnosť **kopírovať hodnoty buniek**. Je to z toho dôvodu, že prednastavené kopírovanie kopíruje funkciu v danej bunke a to by vyvolalo nesprávne hodnoty alebo chyby (#ODKAZ!) v novom súbore.