OLAP PRO SOCIO-EKONOMICKÉ PRŮZKUMNÉ ANALÝZY A PREVENCI KRIMINALITY

Jiří HORÁK[1], Igor IVAN[2], Bronislava HORÁKOVÁ[3], Michala DROZDOVÁ[4], Petr BALA[5]

[1,2,4]Institut geoinformatiky, Hornicko-geologická fakulta, Vysoká škola báňská – Technická univerzita Ostrava, 17. listopadu 15, 708 33 Ostrava, Česká republika

jiri.horak@vsb.cz; igor.ivan@vsb.cz; michala.drozdova@vsb.cz

[3,5]Hexagon Safety & Infrastructure, Prosek Point - budova A, Prosecká 851/64, 190 00 Praha 9 – Prosek, Česká republika

bronislava.horakova@hexagonsi.com; petr.bala@hexagonsi.com

Abstrakt

OLAP (On-line Analytical Processing) je považováno za součást Business Intelligence (BI), která umožňuje efektivní a intuitivní přístup ke konsolidovaným datům uloženým v multidimenzionální databázi. Na rozdíl od běžných komerčních aplikací BI řešení pro socioekonomická data vyžaduje odlišný přístup. Datový sklad integruje data o obyvatelstvu, nezaměstnanosti, kriminalitě, bydlení a službách v území a čase. Pro OLAP řešení byly vyzkoušeny možnosti MS SQL Server 2014, SPSS a Intergraph Thin Client (ITC) s modulem NGIS. Pilotní řešení v ITC demonstruje možnosti interaktivního dotazování a výběrů dat pomocí grafických objektů v panelu v souladu s drill-down přístupem.

Abstract

OLAP for Socio-Economic Exploratory Analysis and Crime Prevention: OLAP (On-line Analytical Processing) is considered as a part of Business Intelligence (BI) which enables effective and intuitive access to consolidated data stored in a multidimensional database. Despite usual commercial BI application the solutions for socio-economic data (usually a public sector domain) require a different approach. The data warehouse integrates data about population, unemployment, crime, housing and facilities in the territory and time. Following OLAP solutions has been tested - MS SQL Server 2014, SPSS and Intergraph Thin Client (ITC) with the NGIS module. A pilot ITC based solution demonstrates possibilities of interactive querying and data selection using graphical objects in a panel and effectiveness of drill-down OLAP approach.

Klíčová slova: OLAP, socioekonomická data, kriminalita, business intelligence, GIS

Keywords: OLAP, human geography, crime, business intelligence, GIS

1. ÚVOD

On-line analytical processing (OLAP) je zpravidla chápán jako součást součást Business Intelligence (BI), která umožňuje efektivní a intuitivní přístup ke konsolidovaným (harmonizovaným a agregovaným) datům uloženým v multidimenzionální databázi.

Základní vlastností OLAP je multidimenzionalita uložení dat v takzvané multidimenzionální kostce. OLAP podporuje agregace kumulativních dat (Loshin, 2012) a specifické analytické operace jako drill-down, roll-up, drill-across, slice-and-dice a pivoting.

Dvojice operací označovaných jako drill-up a drill-down označuje posun v hierarchii dané dimenze buď výše nebo níže a poskytuje tedy buď generalizovanější, nebo detailnější pohled na data (Lacko, 2003).

Pivotování (pivot) označuje operaci otáčení kostkou, při níž se získává jiný pohled na data. Typicky například záměna sloupců za řádky při pohledu na řez kostky, a nebo výměna zobrazované dimenze v řádcích nebo sloupcích. Při reportování tak lze snadno změnit způsob agregace dat a členění výsledků.

Drill-Across spojuje několik tabulek faktů za podmínky stejné granularity dat. Umožňuje tak na základě územních a časových (případně i jiných) jednotek propojovat předmětově různá data.

Slice-and-Dice provádí řezy multidimenzionální kostkou, tedy výběry podle jednoho zvoleného kritéria (slice) nebo podle více kritérií současně, čímž vzniká menší datová kostka (dice).

Prostorová extenze OLAP, SOLAP, může být definována podle (Bédard, 1999, Rivest et al., 2001) jako vizuální platforma určená pro podporu snadných časoprostorových analýz a explorační analýzy multidimenzionální dat na jednotlivých agregačních úrovních.

Datové sklady a multidimenzionální modelování byly původně navrženy pro ekonomické aplikace. Mohou však být velmi užitečné i pro jiné oblasti. Ve veřejném sektoru se často integrují data z různých zdrojů, které zachovávají po jistou dobu stejnou strukturu a dochází u nich k často malým změnám v případě časté aktualizace při současně velkém celkovém objemu dat. Multidimenzionální modelování a OLAP proto může představovat u nich vhodný efektivní nástroj. Jednou z takových aplikací je i databáze pro prevenci kriminality vyvinutá v rámci projektu „Geoinformatika jako nástroj pro podporu integrované činnosti bezpečnostních a záchranných složek státu“.

2. TESTOVÁNÍ MOŽNOSTÍ OLAP

Možnosti OLAP byly testovány ve 3 prostředích – MS SQL Server 2014, SPSS v.18 a Integraph Thin Client (ITC).

2.1 MS SQL SERVER 2014

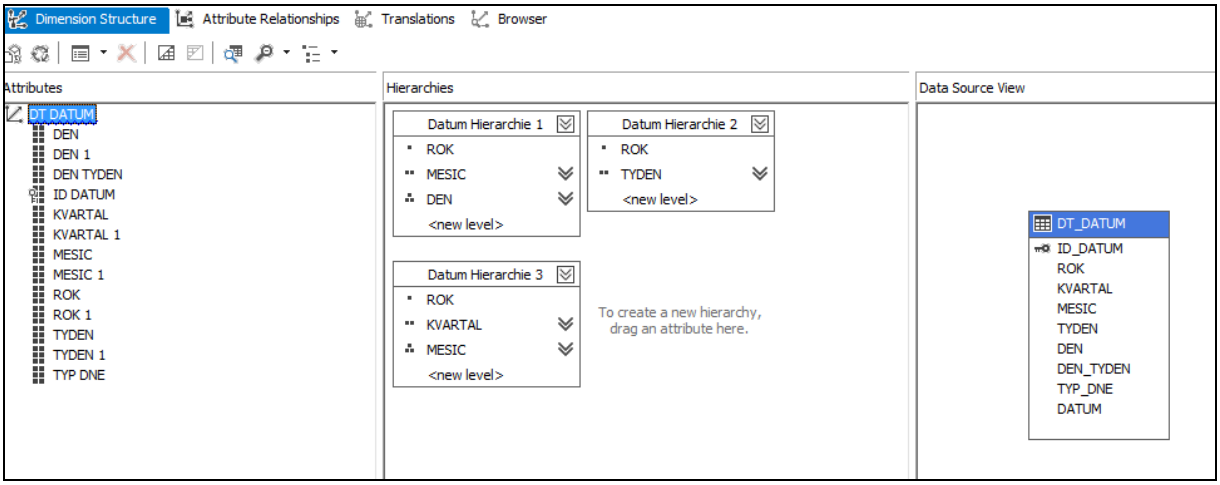
Byla použita verze Microsoft SQL Server 2014 v edici Enterprise, která umožňuje plně využít funkce a služby podporující Business Intelligence.

Při tvorbě multidimenzionální datové kostky byla použita sada nástrojů Sql Server Data Tools, integrovaná do Visual Studia. Byl vytvořen projekt typu Analysis Services Multidimensional and Data Mining Project, který je typu Bussiness Intelligence z kategorie Analysis Services.

Sql Server Data Tools (SSDT) je sada nástrojů umožňující návrh a tvorbu databází a aplikací ve výsledku nasazených pod správu SQL Serveru nebo SQL Azure, přímo z vývojového prostředí Microsoft Visual Studia. Po nainstalování SSDT, které nejsou součástí samotné instalace SQL Serveru, jsou do Visual Studia přidány šablony projektů souhrnně nazvané jako Business Intelligence.

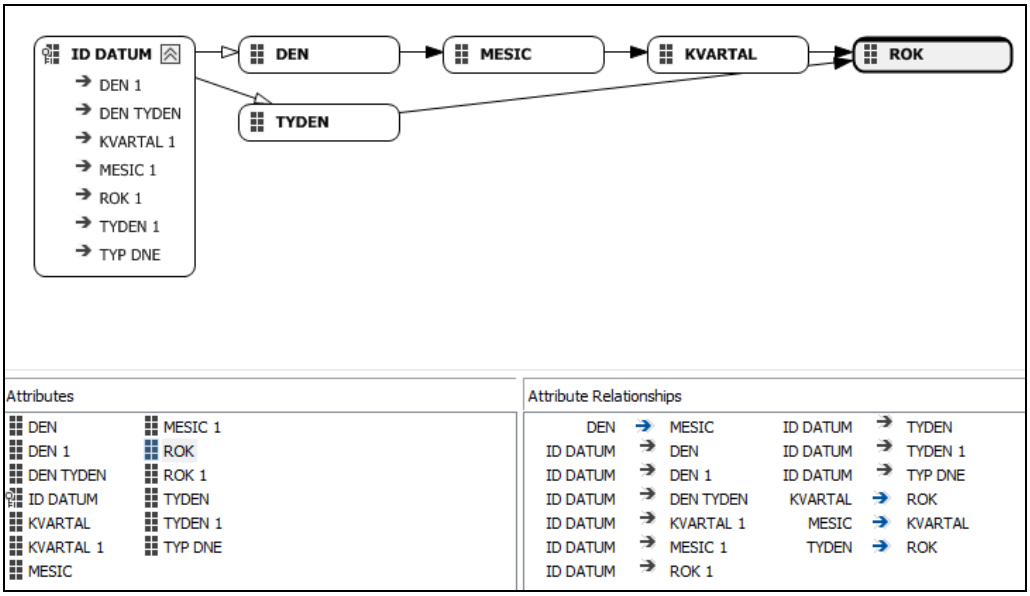
Po určení datového zdroje přichází na řadu definice dimenzí a specifikace jejich atributů. Následně volba tabulky faktů a měrných jednotek, přičemž měrné jednotky mohou být i agregované.

V dalších krocích realizace následují úpravy definovaných dimenzí, mezi které je možné zařadit např. definici hierarchií, díky nimž můžeme v rámci dimenze jednoduše sestoupit na nižší úroveň agregace (drill-down) nebo naopak přejít na vyšší agregační úroveň (roll-up). Na následujícím obrázku jsou ukázány 3 hierarchie v rámci dimenze DATUM.



Obr. 1. Hierarchie v dimenzi DATUM.

Aby byla hierarchie správně specifikována, je potřeba u každého atributu vytvořit kolekce klíčových sloupců. V rámci námi vytvořených hierarchií nad dimenzí DATUM byly vytvořeny následující: den-měsíc-kvartál-rok u atributu DEN, měsíc-kvartál-rok u atributu MESIC, týden-rok u atributu TYDEN, a kvartál-rok u atributu KVARTAL. Další důležitou věcí je doplnění všech potřebných vazeb mezi atributy participujícími v hierarchii.



Obr. 2. Doplněné vazby mezi atributy v dimenzi DATUM.

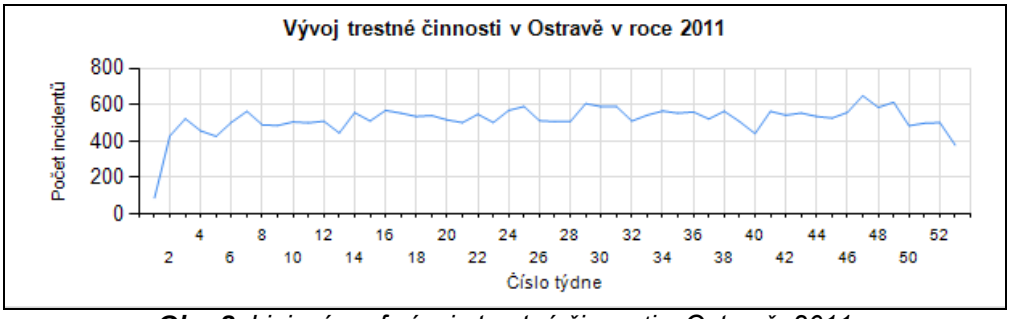
Posledním krokem před vlastním vytvořením datové kostky bývá definice nových vypočítaných členů a klíčových ukazatelů výkonnosti v záložkách Calculations a KPIs (Key Performance Indicators). Calculations slouží k vypočtení nových hodnot odvozených z kombinace agregovaných měrných jednotek nebo dat v dimenzích. Tyto hodnoty fungují na principu měrných jednotek, nejsou však definovány přímo v tabulce faktů, protože jsou vypočítány až na základě výsledků, které jsou poskytnuty datovou kostkou. KPIs jsou určeny pro stanovení cílů, které by měly být splněny, díky čemuž je možné kontrolovat, jak vypadá současný vývoj dané hodnoty v porovnání s naším očekáváním.

Datovou kostku, která je již ve správě analytických služeb, je možné prohlížet v Microsoft Visual Studiu, v Excelu nebo v SQS Server Management Studiu (SSMS).

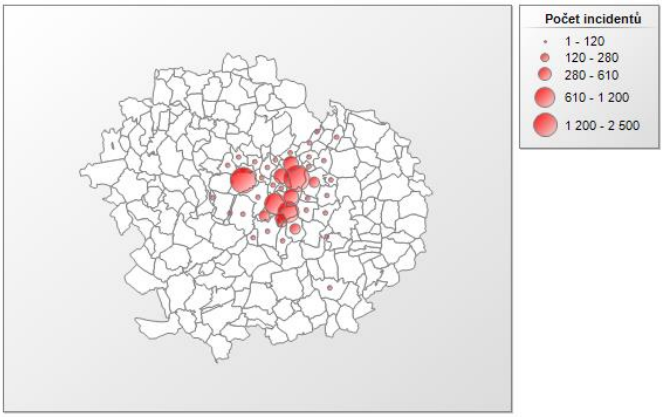
V prostředí Sql Server Management Studio (SSMS) je možné se nad datovou kostkou dotazovat jazykem MultiDimensional Expression (MDX) Language a získávat tak různé pohledy na data v kostce uložená.

Pro realizace reportů a vizualizaci výsledků byla využita služba SQL Server Reporting Services, k čemuž byl vytvořen projekt ve vývojovém Prostředí Visual Studio projekt. Tento projekt umožňuje vytvářet nad datovou kostkou libovolné pohledy a vhodně je graficky vizualizovat.

Služba Reporting Services nástroje Microsoft SQL Server umožňuje vytváření komplexních reportů včetně celé řady objektů. Mezi ně patří tabulky, matice, grafy celé řady typů (obr. 3), obrázky, datové sloupce (data bars), sparklines a 3 typy „map“ – základní, kartogram (colour analytical) a bodový kartodiagram (bubble maps). Podle očekávání je realizace map příliš jednoduchá a mapovou kompozici nelze nastavit ani upravit do potřebné podoby (obr. 4).



Obr. 3. *Liniový graf vývoje trestné činnosti v Ostravě, 2011*



Obr. 4. Bodový kartodiagram četností výskytu TSK kategorie E.

2.2 SPSS

OLAP v SPSS je podporován pomocí definice OLAP kostky, kde je možné specifikovat různé statistické indikátory, výpočty a srovnání. Příklad sumární statistické tabulky počtů a podílů krádeží podle typů objektů pro pachatele ve věkové skupině 20-25 let z Ostravy z let 2010-2011 je uveden na obr. 5. Manipulace s daty je podpořena především pomocí pivotingu. Mapové výstupy však nejsou dostatečně podpořeny.

OLAP Cubes

vek_p 20-25

typ_objekt		Sum	N	Mean	Std. Deviation	% of Total Sum	% of Total N
pocet_udal	BYT	12	12	1,00	,000	,1%	,1%
	čerpací stanice	7	5	1,40	,548	,1%	,0%
	DOPRAVNÍ PROSTŘEDEK - PHM	2	2	1,00	,000	,0%	,0%
	DOPRAVNÍ PROSTŘEDEK - součástky motorových vozidel	27	27	1,00	,000	,2%	,3%
	DOPRAVNÍ PROSTŘEDEK - věci uložené v motorovém vozidle	70	56	1,25	,548	,6%	,5%
	MIMO OBJEKT - zastávka ČD	4	2	2,00	,000	,0%	,0%
	MIMO OBJEKT - zastávka MHD	1	1	1,00	.	,0%	,0%
	neurčeno	571	441	1,29	,884	5,1%	4,1%
	PODNIK, OBJEKTY VÝROBY A SLUŽEB	20	12	1,67	,492	,2%	,1%
	PRODEJNA - supermarket (hypermarket, obchodní dům)	63	61	1,03	,180	,6%	,6%
	RESTAURACE - bar	5	5	1,00	,000	,0%	,0%
	RESTAURACE - restaurace	4	4	1,00	,000	,0%	,0%
	ZDRAVOTNICKÝ OBJEKT - nemocnice	1	1	1,00	.	,0%	,0%
	Total	787	629	1,25	,774	7,1%	5,9%

Obr. 5. Evidované objekty krádeží pro věkovou skupinu pachatelů 20-25 let (Ostrava, 2010-2011)

2.3 ITC (INTERGRAPH THIN CLIENT)

ITC (Intergraph Thin Client) je modulární systém tenkého GIS klienta, který je koncipován a navržen s ohledem na jednoduchost ovládání pro koncového uživatele. Tento nástroj byl vybrán pro pilotní řešení s ohledem na potřebu zpřístupnit a vizualizovat zpracovaná data v datovém modelu multidimenzionální databáze uživatelsky jednoduchou formou s využitím prostorových informací v datech.

Aplikace používá knihovny OpenLayers a jQuery a umožňuje integrovat různé nástroje a služby. Pro implementaci byly využity pouze funkcionality a některé moduly ITC, konkrétně mapové okno, NGIS modul a „vrstvy".

Mapové okno je určeno pro zobrazování prostorových dat z mapových služeb standardu OGC WMS a WMTS.

NGIS modul poskytuje specifické služby pro interaktivní manipulaci s daty z multidimenzionální databáze a jejich vizualizaci. NGIS modul je představuje z pohledu uživatele panel v aplikaci ITC, který zobrazuje určitá vybraná data formou grafů (popř. i tabulek) a umožňuje jejich zobrazení formou kartogramu nad mapovým podkladem. Prostřednictvím panelu s grafy lze provádět i dotazování - filtrování dat. Filtrování je interaktivní a uživatel jej provádí přímo prostřednictvím grafů, kdy výběrem určité hodnoty v grafu dojde k jejímu odfiltrování (či odfiltrování ostatních hodnot) a hodnoty v grafu jsou ihned přepočteny.

NGIS modul využívá JavaScript knihovny DataDrivenDocuments (také známé jako D3) a Crossfilter.

D3.js (http://d3js.org/) je knihovna publikována pod BSD licencí a využitelná především pro tvorbu dynamických a interaktivních vizualizací dat v prostředí webových prohlížečů. Knihovna je rychlá, podporuje velké datové soubory a dynamické chování pro zajištění interakce a animací. Pomocí Document Object Model (DOM) umožňuje připojit data a transformovat je do potřebných cílových dokumentů (HTML tabulky, interaktivní SVG grafy apod.).

Crossfilter je knihovna určena pro zpracovávání multidimenzionálních dat v prostředí webových prohlížečů (http://square.github.io/crossfilter/). Podporuje vybrané OLAP analytické metody, zejména drill-down přístup.

Využitím obou těchto knihoven lze vytvářet aplikace pro tzv. „koordinovanou vizualizaci".

NGIS modul využívá CSV formát pro analytická, tj. multidimensionální data a JSON nebo GeoJSON pro prostorová referenční data. Tematické mapy se zobrazují v mapovém okně ITC (často s využitím webových mapových služeb) nebo v mapovém okně umístěném jako objekt na panelu NGIS.

Konfigurace NGIS modulu zahrnuje definici obsahu (např. jaké grafy, jaké popisy) v HTML souborech, nastavení D3 a Crossfilter objektů v js souborech (zejména typ a vzhled grafů, zdroje dat, způsob vizalizace, vazební atribut mezi analytickými a referenčními daty) a konfiguraci serveru v XML souborech (mapové pozadí pro mapové okno ITC, výběr dalších konfiguračních souborů, pravidla autorizace atd.). Konfigurace určuje instanci NGIS modulu pro určitá multidimensionální data a území.

Aplikace ITC na začátku vyžaduje autentizaci a autorizaci uživatele. Ve výchozím zobrazení je načten plný rozsah analytických dat a v mapovém okně ITC jsou zobrazeny výchozí (default) prostorové jednotky, zpravidla grid 1x1km, u jiných konfigurací než „kriminalita" jsou použity části obcí. Změnit prostorovou jednotku je možné volbou ve výběrovém seznamu „Typ vizualizace" v NGIS panelu. Při změně měřítka mapy (přiblížení) na menší než 1:15 000 dojde k automatickému přepnutí aktuálně použitých prostorových jednotek na čtverce 100x100m.

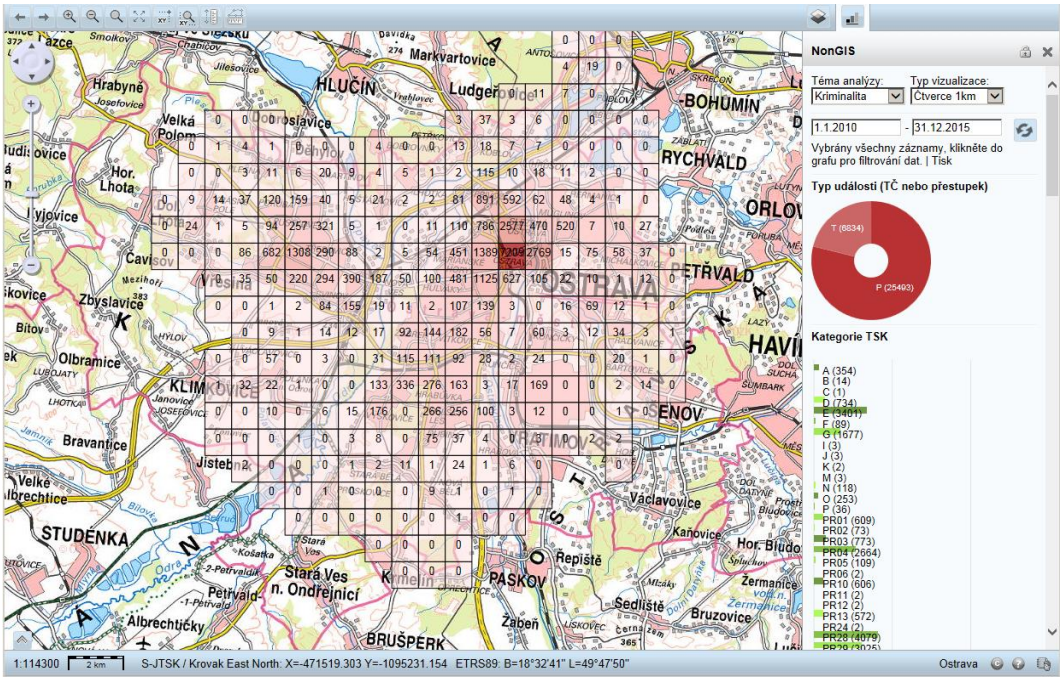
Grafy v NGIS panelu představují vždy klasifikaci dle určité dimenze dat. Graf zobrazuje vždy počty unikátních hodnot v příslušné dimenzi jak graficky (podíl výseče u koláčového grafu, velikost sloupce u sloupcového grafu), tak v absolutních číslech (popisek v závorce na grafu, případně i jako plovoucí text při najetí kurzorem myši nad příslušnou část grafu).

Grafy slouží také pro filtrování zobrazovaných dat. Omezit rozsah zobrazovaných dat je možné kliknutím na konkrétní kategorii (výseč či sloupec) v grafu.

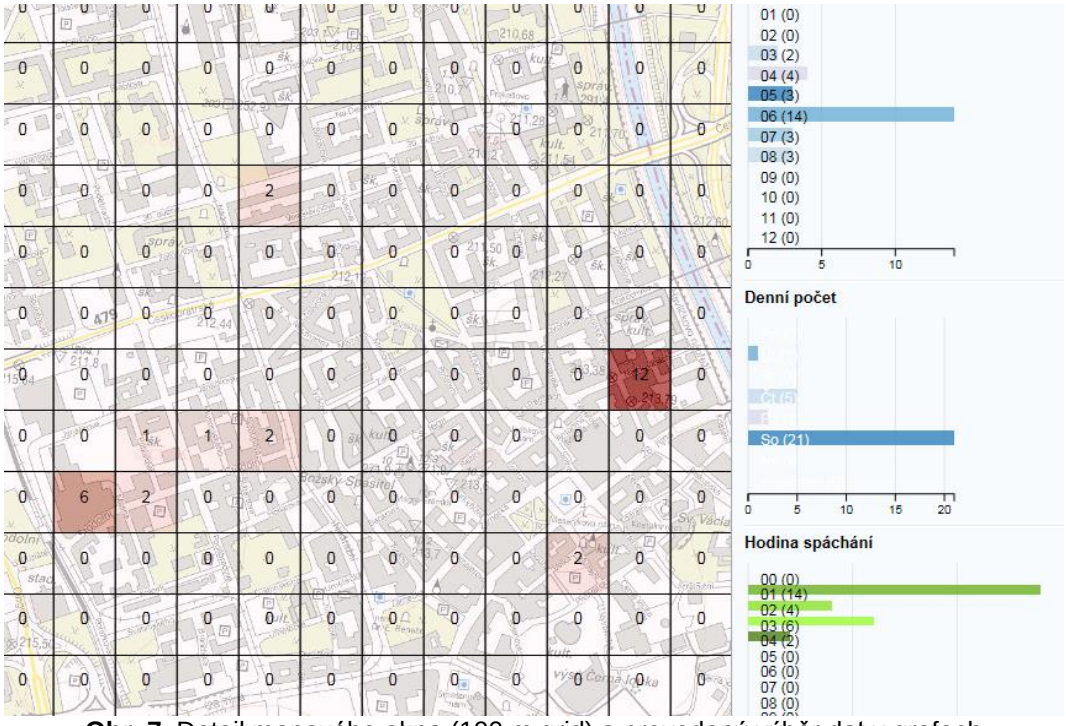
Zrušit filtry a obnovit výchozí plný rozsah dat je možné pomocí:

- „reset všech filtrů“ – zruší všechny aplikované filtry,
- „reset“ – zruší pouze filtr dimenze příslušného grafu.

Po výběru části dat jsou grafy a mapa odpovídajícím způsobem překresleny (obr. 7). Filtry (podmínky dotazu) lze snadno kombinovat opakovaným kurzorovým výběrem.



Obr. 6. Náhled na ITC aplikaci: mapové okno 1 km sítí a NGIS panel s interaktivním grafy



Obr. 7. Detail mapového okna (100 m grid) a provedený výběr dat v grafech

V rámci pilotního řešení byly připraveny instance pro několik typů dat pro Ostravu, Kolín a České Budějovice.

3. ZÁVĚR

Business Intelligence umožňuje vytvářet efektivní a uživatelsky jednoduché aplikace nejen pro ekonomické aplikace, ale i pro veřejný sektor. Hlavní výhody lze najít v dobře připravených ETL procesech a v nabídce analytických nástrojů pro jednoduché časoprostorové analýzy multidimenzionálních dat.

Možnosti OLAP nástrojů byly testovány ve 3 prostředích – MS SQL Server 2014, SPSS v.18 a Intergraph Thin Client (ITC). V rámci MS SQL Server 2014 byly aplikovány nástroje SQL Server Data Tools, Analysis Services a SQL Server Reporting Services. Prostředí lze doporučit zejména pro aplikace, kde je kladen důraz na dotazování dat a výpočty odvozených dat a nových indikátorů. Využití SPSS pro OLAP je vhodné pro statistické analýzy, zejména pro multivariační analýzy a testování hypotéz nad multidimenzionálními daty. ITC (Intergraph Thin Client) představuje tenkého GIS klienta v kombinaci s NGIS modulem (využití DataDrivenDocuments a Crossfilter JavaScript knihoven), který umožňuje drill-down přístup. Dovoluje zkoumat data podle jednotlivých dimenzí a postupně modifikovat datové pohledy pomocí přírůstkového filtrování a tak vytvářet více specifické a detailní pohledy. Aplikace je uživatelsky přátelská díky své interaktivitě – uživatel vytváří dotazy prostřednictvím interaktivních NGIS objektů. Postupný výběr jedné nebo více částí dat v grafech a propojených tabulkách a mapách provádí přepočty všech relevantních hodnot ve výběrech a statistikách a poskytuje vhodný vizuálně-explorační nástroj pro uživatele.

Poděkování

Příspěvek vznikl v rámci projektu VF20142015034 „Geoinformatika jako nástroj pro podporu integrované činnosti bezpečnostních a záchranných složek státu“. Data poskytly následující subjekty: Policie ČR, Městská policie Ostrava, Městská policie Kolín, Městská policie České Budějovice, ČSÚ, MV ČR, MF ČR, MZdr ČR.

LITERATURA

Bédard Y. (1999): Visual modelling of spatial databases: towards spatial PVL and UML. Geomatica 12/1999; 53(2):169-186

Horák, J.: Informační systémy. VŠB-TU Ostrava, výukový text. 2013.

Lacko, L.: Databáze: datové sklady, OLAP a dolování dat s příklady v Microsoft SQL Serveru a Oracle. Computer Press, Brno. 2003.

Loshin D. (2012): Business Intelligence: The Savvy Managers Guide. MORGAN Kaufmann, Newnes, USA. Pp.370. ISBN 978-0-12-385889-4.

Rivest S., Bédard Y., Marchand P. (2001): Towards better support for spatial decision-making: Defining the characteristics of Spatial On-Line Analytical Processing (SOLAP). Geomatica 12/2001; 55(4):539-555.