Microsoft® SQL Server™ 2008

Business Intelligence

Ľuboslav Lacko

2. aktualizované vydanie (SQL Server 2008 – finálna verzia)

Microsoft®

Microsoft SQL Server 2008

Business Intelligence na platforme Microsoft SQL Server 2008

Obsah:

Pre praktické zoznámenie sa s črtami Business Intelligence na platforme nového databázového servera MS SQL Server 2008 odporúčame ako prvý krok jeho inštaláciu. Popis je v prílohe č. 1. Ako veľmi zaujímavú alternatívu pre testovanie a zoznamovanie sa so serverom SQL Server 2008 odporúčame nainštalovať ho na virtuálny počítač vytvorený pomocou nástroja Microsoft Virtual PC 2007. Ako operačný systém odporúčame pre vývojárov a študentov Windows Vista, a pre migrujúcich administrátorov Windows Server 2008.

Kapitola 1: Microsoft SQL Server 2008 ako platforma pre BI

Termín Business Intelligence (BI) prvýkrát v roku 1989 definoval Howard Dresner zo spoločnosti Gartner Group ako „množinu konceptov a metodík, ktoré zlepšujú rozhodovací proces za použitia metrík alebo systémov založených na metrikách." Je to proces transformácie údajov na informácie a prevod týchto informácii na poznatky prostredníctvom objavovania. Účelom procesu je konvertovať veľké objemy údajov na poznatky, ktoré sú potrebné pre koncových používateľov. Tieto poznatky môžeme potom efektívne využiť napríklad v procese rozhodovania. S problematikou BI sú úzko spojené údajové sklady. **Údajový sklad** je podnikovo štruktúrovaný depozitár subjektovo orientovaných, integrovaných, časovo premenných, historických údajov použitých na získavanie informácií a podporu rozhodovania. V údajovom sklade sú uložené atomické a sumárne údaje. Údaje sa získavajú a ukladajú do produkčných (operačných) databáz, ktoré môžu byť v rôznych oddeleniach firiem, alebo dokonca v rozličných geografických lokalitách.

Tieto údaje v pravidelných intervaloch zozbierame, predspracujeme a zavedieme do údajového skladu. Údajový sklad je v podstate tiež databáza, len je organizovaná podľa trochu iných pravidiel, tabuľky napríklad nemusia byť normalizované.

Údaje sa do údajového skladu zapisujú skôr podľa predmetu záujmu, než podľa aplikácie, v ktorej boli vytvorené. Pri orientácii na subjekt sú údaje v údajovom sklade kategorizované podľa subjektu, ktorým môže byť napr. zákazník, dodávateľ, zamestnanec, výrobok a podobne.
Údajový sklad musí byť jednotný a integrovaný. To znamená, že údaje týkajúce sa konkrétneho predmetu sa do údajového skladu ukladajú len raz. Preto musíme zaviesť jednotnú terminológiu, jednotné a konzistentné jednotky veličín. Nie je to úloha jednoduchá, pretože údaje prichádzajú do údajového skladu z nekonzistentného a neintegrovaného operačného prostredia. Preto musia byť údaje v etape prípravy a zavedenia upravené, vyčistené a zjednotené. Ak údaje nie sú konzistentné a dôveryhodné, údajový sklad stráca význam. Údaje sa ukladajú do údajového skladu ako séria snímok, z ktorých každá reprezentuje určitý časový úsek. Na rozdiel od operačného prostredia, kde sú údaje platné v okamihu prístupu, v údajových skladoch sú údaje platné pre určitý časový moment, časový snímok. Zatiaľ čo v operačnom databázovom prostredí sa údaje ukladajú za kratšie časové obdobie dní, maximálne mesiacov, v údajovom sklade sú údaje za dlhšie časové obdobie, typicky niekoľko rokov. Kľúčové atribúty v údajovom sklade obsahujú čas, ktorý v operačných databázach nemusí byť uvádzaný. Len čo je v údajovom sklade zaznamenaná konkrétna snímka údajov z operatívnej databázy, nemôžu byť už tieto údaje v údajovom sklade modifikované.

SQL 2008 borí mýtus o nevhodnosti transakčných databáz pre analýzy

Transakčné OLTP databázy sú určené na ukladanie operačných údajov. Výsledkom dopytovania sú databázové tabuľky, súhrny získané pomocou agregačných funkcií, rôzne zostavy a podobne. OLTP databázy sú z dôvodu jednoduchého dopytovania a vylúčenia redundancie spravidla normalizované. Takéto systémy dosahujú vysoké výkony skôr pri online transakciách než pri zložitých analýzach, ktoré sú veľmi náročné na výpočtovú kapacitu. Komplexná analýza vyžaduje iné techniky návrhu databáz, napríklad použitie multidimenzionálnych a hviezdicových schém s tabuľkami faktov, ktoré obsahujú merné jednotky obchodovania a vysoko nenormalizované tabuľky dimenzií. Azda najväčšou prekážkou použitia databázových systémov OLTP pre analýzy je skutočnosť, že tieto systémy nemajú k dispozícii integrovaný zdroj údajov zo všetkých operačných systémov v rámci podniku tak, aby umožnili tvorbu komplexných analýz, to znamená, že potrebné údaje, alebo údaje ktoré by mali slúžiť ako podklady pre analýzy sú roztrúsené v rôznych spravidla heterogénnych OLTP systémoch a musia sa zakaždým prácne integrovať skôr, ako je možné získať požadované informácie. Časová náročnosť prípadných analýz, hoci nemusí ísť ani o príliš zložité ani príliš komplexné analýzy je preto pomerne vysoká. Niekedy sa dokonca ani nepodarí koordinovať údaje medzi jednotlivými systémami, takže vlastne ani nemôžeme získať globálny obraz o stave podnikania.

Nevýhody použitia transakčných databáz pre analytické účely by sme mohli zhrnúť do niekoľkých bodov:
- nie je jednoduché nájsť príčiny a vysvetlenia problémov,
- zložité hľadanie závislosti jednotlivých veličín,
- príliš rozsiahle výstupy z transakčných systémov,
- dlhý čas výpočtu a degradácia výpočtového výkonu databázového stroja transakčnej databázy,
- transakčný systém neuchováva historické údaje,

– nehomogénna štruktúra údajov,
– dlhý čas prípravy údajov.

Výhody
Na druhej strane je snaha o komplexný prístup k údajom v údajovom sklade pri akceptovaní hlavnej požiadavky, ktorú prináša dnešná hektická globalizovaná ekonomika – „všetko v reálnom čase". Kým v minulosti sa údaje z údajových skladov spracovávali v dávkach pre nejaké obdobie z blízkej minulosti (včerajší deň, minulý týždeň), v súčasnosti manažéri a analytici požadujú prístup k výsledkom analýz v reálnom čase. V nadväznosti na túto požiadavku už začína byť minulosťou aj oddelenie transakčných systémov a údajových skladov. Súčasným trendom vyplývajúcim z citovaných požiadaviek je viacúrovňový podnikový údajový sklad s možnosťou prístupu pomocou integrovaného aplikačného prostredia. Vznikne tak akýsi komplexný BI ekosystém, do ktorého sú na strane vstupov spravidla zahrnuté aj externé zdroje údajov a ktorý na výstupe poskytuje informácie nielen pre firmu, ktorá je vlastníkom, budovateľom a prevádzkovateľom BI ekosystému, ale aj pre jej partnerov, dodávateľov a zákazníkov.

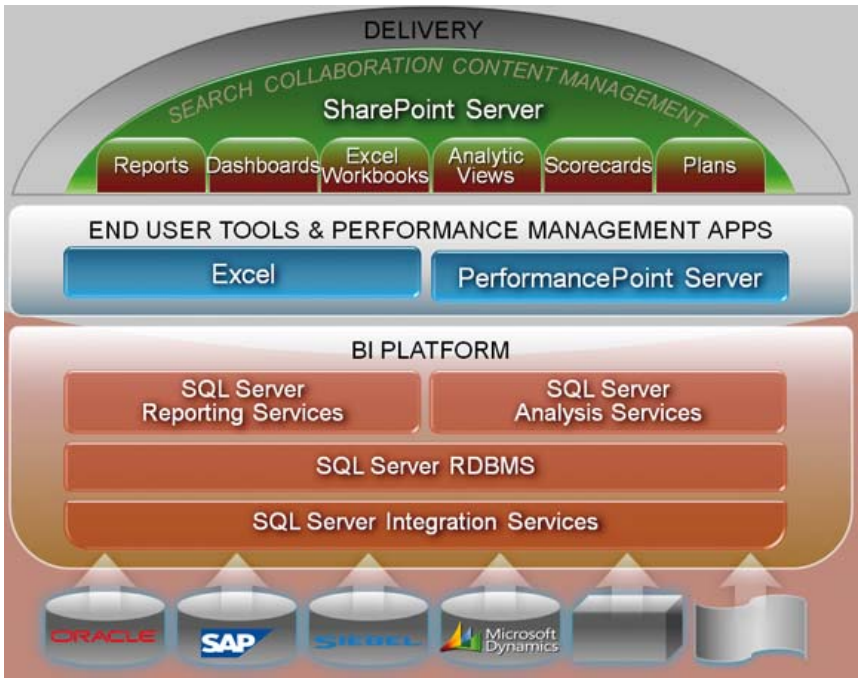
BI systémy pracujúce v reálnom čase

BI systémy pracujúce v reálnom čase sú založené na troch primárnych faktoroch:
– paralelný priebeh etapy ETL, teda extrakcie, transformácie a zavedenia údajov, pričom všetky ETL procesy musia byť dostatočne nadimenzované na predpokladané objemy spracovávaných údajov a to aj v dobe špičiek, ako sú napríklad predvianočný predaj, koncoročné uzávierky alebo marketingové kampane,
– relačné systémy musia okrem operačnej prevádzky, teda spracovávania transakcií zvládnuť aj zložité a kapacitne náročné BI dopyty,
– doručovanie výsledkov BI procesov, teda reportov a výsledkov analýz a to buď periodicky, alebo na vyžiadanie.

Charakteristickou črtou údajových skladov je veľké množstvo údajov v niektorých databázových tabuľkách. V takýchto prípadoch systémov reálneho času je výhodné rozdeliť veľkú tabuľku na niekoľko logických oddielov (partícií), podľa nejakého pravidla vyplývajúceho z povahy údajov, napríklad jeden logický oddiel môže obsahovať údaje za nejaké časové obdobie (mesiac, rok), prípadne údaje podľa iného pravidla, napríklad v zozname osôb, osoby, ktorých priezvisko začína príslušným písmenom a podobne. Takéto delenie je potom logické aj z hľadiska následného spracovania údajov, pretože trebárs reklamácie a fakturácie sa obvykle vykonávajú za príslušné kalendárne obdobie.

Pri analýze možností integrácie jednotlivých podnikových systémov a procesov, či už ekonomických, alebo výrobných, nachádzame ako hlavný integračný prvok databázu, kam všetky integrované a konsolidované systémy ukladajú svoje údaje. Firma Microsoft v tejto oblasti profituje z tradície jedného z najvýznamnejších dodávateľov databázových systémov. Ich nový databázový produkt SQL Server 2008 je moderná komplexná serverová platforma pre ukladanie a správu údajov v databázach a údajových skladoch a balíky nástrojov pre Business Intelligence vrátane pokročilého reportovania. IT Ekosystém podnikových aplikácií využívajúci SQL Server 2008 by mal byť koncipovaný tak, že na najvyššej úrovni architektúry je SharePoint Server pre spoluprácu, a vrstva Content management sa využíva na vyhľadávanie a správu obsahu.

SQL Server 2008 je vyšším evolučným stupňom databázových serverov. Umožňuje prelínanie a integráciu podnikových procesov aj na úrovni analytických služieb. Analytické služby implementované priamo v databázovom serveri sa tak môžu využívať buď izolovane v jednotlivých odvetviach alebo komplexne. Nemusí to byť len podpora rozhodovania na strategickej úrovni týkajúca sa top manažmentu. Na jednej strane analytických procesov sú podklady pre analýzy uložené v databázach alebo údajových skladoch. Rovnako dôležitý je na druhej strane aj spôsob prezentovania výsledkov analýz. Pracovníci pred špecializovanými aplikáciami preferujú používanie nástrojov na ktoré sú zvyknutí zo svojej bežnej práce, napríklad programy kancelárskych balíkov.



Business Intelligence na platforme Microsoft

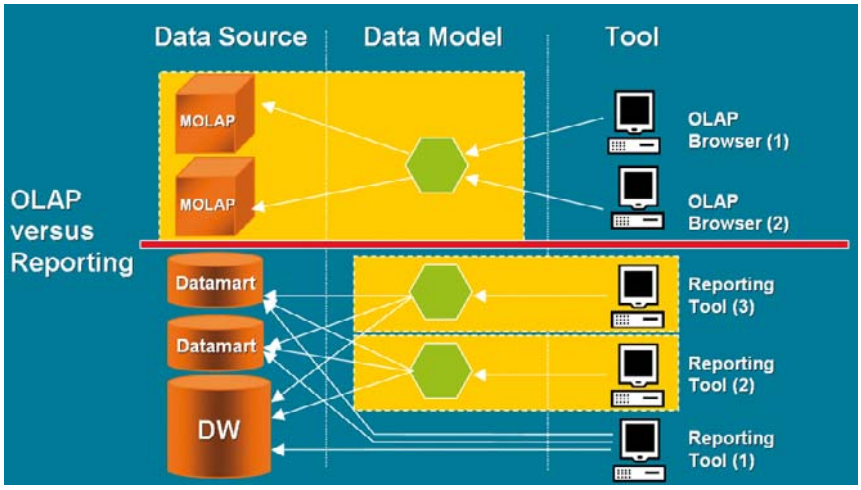
Kapitola 2: Modelovanie Business Intelligence projektov

Jedným z preferovaných trendov pre budovanie podnikových projektov a aplikácií je model driven development, kedy sa budovanie projektu začne fázou modelovania. Túto filozofiu podporuje aj SQL Server 2008, ktorý pre modelovanie projektov typu Business Intelligence využíva technológiu UDM (Unified Dimension Model). Táto technológia bola po prvý krát použitá vo verzii SQL Server 2005. Ťažisko vytvorenia projektu spočíva v modelovaní a následne na základe vytvoreného modelu BI aplikácie bude vygenerovaná štruktúra pre údajový sklad, vytvorené dávky pre jeho plnenie prostredníctvom možností Integračných služieb. Budovanie BI projektu pokračuje návrhom mierok, dimenzií, kociek a podobne.

Prvý predpoklad pre unifikovaný prístup k modelovaniu je unifikácia technológií. Tento predpoklad je splnený integráciou databázových analytických a reportovacích služieb do jedného balíka. Druhým predpokladom je unifikované používateľské rozhranie, čo v prípade databázového servera predstavujú nástroje pre administráciu, dopytovanie a návrh štruktúr a čiastkových modelov. Aj táto požiadavka je v prípade SQL Servera 2008 splnená. K dispozícii je komfortný nástroj **SQL Server Business Intelligence Development Studio**.

Základné princípy UDM

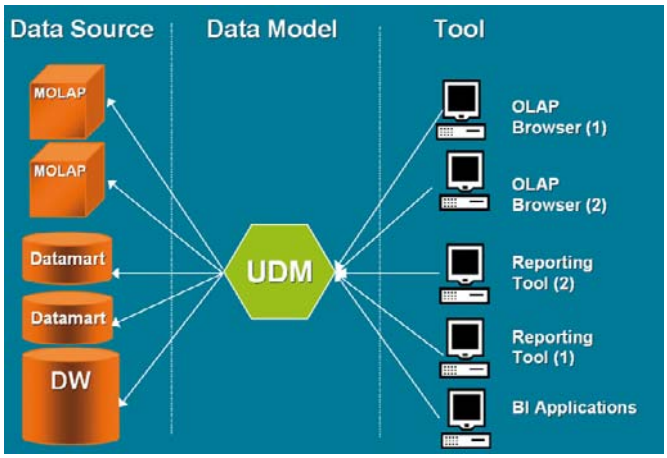
Do verzie servera SQL Server 2005 boli vrstvy databázových, analytických a reportovacích služieb pomerne striktne oddelené, aj keď aj tu by sme už našli množstvo zjednocujúcich prvkov – je možné analyzovať údaje z relačných databáz, prípadne generovať reporty z relačných aj analytických databáz.



Oddelenie vrstvy OLAP a reportovacích služieb v serveri SQL Server 2000

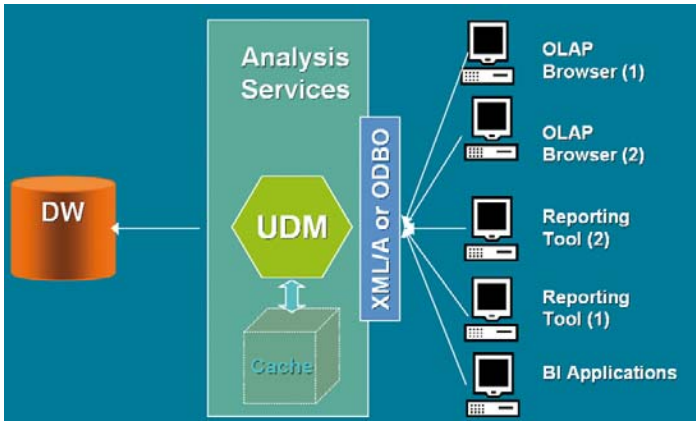
Hlavná nevýhoda architektúry predchádzajúcej verzie – redundancia údajov a modelov je zrejmá už na prvý pohľad. Na obrázku v ľavej časti vidíme už na prvý pohľad, že dochádza k redundancii údajov, pretože tie isté údaje sú jednak v relačných a jednak v multidimenzionálnych databázach. Pri hlbšej analýze údajových tokov zistíme, že situácia z hľadiska redundancie je ešte komplikovanejšia. Tie isté údaje sa prenášajú z relačných databáz do údajových skladov, odtiaľ v mnohých prípadoch do údajových tržníc a odtiaľ do multidimenzionálnych OLAP kociek. Taktiež nezanedbateľná redundancia a s ňou spojené množstvo práce vysokokvalifikovaných špecialistov je vo vrstve modelov. Iným problémom takejto oddelenej architektúry sú rozdielne návyky pracovníkov čo sa týka práce s údajmi.

Modelovanie s využitím UDM umožňuje využitie výhod relačných aj multidimenzionálnych databáz a do istej miery eliminuje ich nevýhody. Od verzie SQL Servera 2005 sú vrstvy Business Intelligence zjednotené do jednotného modelu Unified Dimensional Model (UDM). Tento model prevzal to najlepšie z reportovania a OLAP analýz. K zjednoteniu dochádza jednak na úrovni modelov, kedy je potrebný len jeden dimenzionálny model pre generovanie reportov aj OLAP kociek.



Unified Dimensional Model – zjednotenie na úrovni modelov

Klasický postup analýzy začínal výberom tabuliek z relačných databáz alebo z údajového skladu. Tieto tabuľky slúžili na vytvorenie faktov a dimenzií. Podobne je to aj pri UDM, ale údaje zostávajú v pôvodných úložiskách Mechanizmus MOLAP (multidimenzionálne online analytické spracovanie) potom uloží analytické údaje vo vlastných údajových štruktúrach a sumároch. Počas tohto procesu sa napočíta toľko predbežných výsledkov, koľko je z technického a časového hľadiska možné. Údaje v úložisku typu MOLAP sa teda budú ukladať ako vopred vypočítané pole. Hodnoty údajov aj indexov sa uchovávajú v jednotlivých poliach multidimenzionálnej databázy. Databáza je organizovaná tak, aby umožnila rýchle získavanie príslušných údajov z viacerých dimenzií. Preto je hlavnou výhodou MOLAP maximálny výkon vzhľadom na dopyty používateľa, nevýhodou je redundancia údajov, nakoľko tieto sú uložené jednak v relačnej databáze, jednak v multidimenzionálnej databáze. Požiadavky na úložnú kapacitu môžu v prípade použitia viacerých dimenzií extrémne narastať. Východiskom z tejto situácie pri serveri SQL Server 2005 je vyrovnávacia MOLAP cache pamäť, kam sa ukladajú najčastejšie používané alebo očakávané výsledky agregácií. Navyše toto proaktívne cachovanie je pri serveri SQL Server 2005 plne automatizované. Údaje teda zostávajú v relačných databázach a napočítané agregácie sa ukladajú do multidimenzionálnych štruktúr. Pri dopytovaní sa údaje vyberajú do multidimenzionálnej pamäte cache.



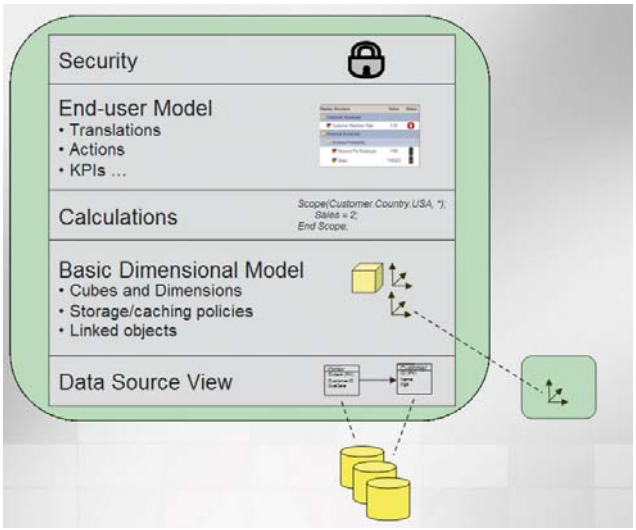
Unified Dimensional Model – cache

Významnou novinkou je zavedenie novej skupiny služieb s názvom Integration Services. Na úrovni tejto vrstvy sa budú integrovať údaje z rôznych údajových zdrojov vrátane ich požadovaných transformácií.

Ak by sme mali zhrnúť dosiaľ prezentované fakty do jednoduchej definície UDM, dospejeme k záveru, že UDM je akýmsi pomysleným mostom medzi používateľom a jeho údajmi.

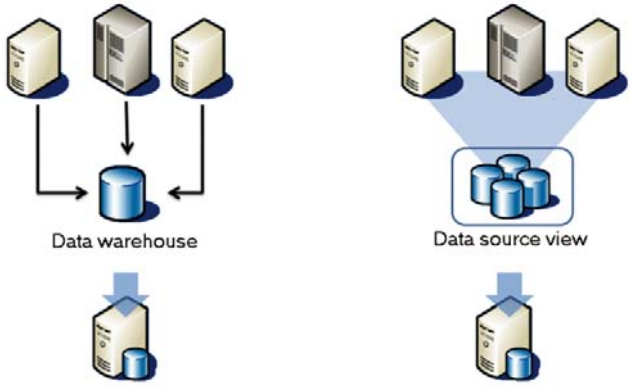
Implementácia UDM na platforme SQL Server 2008

Azda najlepšie pochopíme filozofiu modelovania na hierarchickej schéme. Celá vrstva slúžiaca na vytvorenie výstupu (reportu), je z architektonického hľadiska postavená nad údajmi v databázach. Samozrejme je výhodné, aby tieto údaje boli čo najviac vyčistené, prípadne do rozumnej miery normalizované bez zbytočných redundancií a neprehľadných komplikovaných relačných vzťahov. Samozrejme, nie každý deň sú Vianoce a tak často pri budovaní BI vrstvy musíme vychádzať z údajov aké sú k dispozícii, prípadne aké vyhovujú logike predmetu podnikania, alebo k akej štruktúre údajov sme sa dopracovali postupným budovaním informačných systémov.



Unified Dimensional Model

Samotná BI nadstavba nad databázou je rozdelená do štyroch vrstiev. Nad databázami je vrstva údajových pohľadov, v ktorej už presne vyšpecifikujeme tabuľky, pohľady a to až na úroveň jednotlivých atribútov a relačných väzieb. Modelovaním vrstvy **Data Source View** do určitej miery prispejeme síce nie k fyzickému vyčisteniu údajov, no k nadefinovaniu priamejšieho prístupu k nim. Zjednodušene povedané – na tejto úrovni vytvárame relačné schémy, na základe ktorých budú vytvárané dimenzie a kocky. Veľmi dôležitou úlohou úrovne Data Source View je zjednotenie údajov z heterogénnych zdrojov.



Data source view abstrahuje údaje pre analýzy od údajových zdrojov pri analýzach

Nasleduje úroveň dimenzií kde špecifikujeme, ktoré atribúty sa viažu k merným jednotkám obchodovania. Na úrovni **základného dimenzionálneho modelu** do značnej miery ovplyvňujeme aj spôsob ukladania a cachovania multidimenzionálnych údajov. Okrem faktov a dimenzii vstupujú do hry aj **kalkulácie**, to znamená údaje vypočítané podľa jednoduchých, prípadne aj zložitých vzťahov z hodnôt iných atribútov. Napríklad, ak by sme mali v tabuľke ako atribút rodné číslo, dokážeme z neho určiť dátum narodenia a teda aj vek príslušnej osoby.