20 Možnosti využití dat z informačního systému nemocnice – zkušenosti Masarykova onkologického ústavu.

Brabec P., Andres P., Komínek J., Šebesta R., Klimeš D., Vyzula R., Žaloudík J., Dušek L.
Masarykův onkologický ústav, Brno
Institut biostatistiky a analýz, Masarykova univerzita, Brno

Souhrn

Informační systémy v prostředí nemocničního zařízení procházejí neustálým vývojem, který ale mnohdy nereaguje efektivně na aktuální potřeby nemocnice. To je způsobeno složitostí problematiky, dlouhodobým sběrem neparametrických dat a implementací dodatečných softwarových modulů, které nakonec vytvářejí z nemocničních informačních systémů značně heterogenní prostředí, které neumožňuje nad daty provádět jednoduchou analýzu a reporting. Velice častým problémem, se kterým je možno se běžně setkat na úrovni managementu nemocnice i na úrovni běžného provozního personálu, je nepochopení požadavků uživatelů na straně nasazených konzultantů. Zadání, které pak dostává tvůrce softwaru, nerespektují informační systém nemocnice jako celek a ústí pouze v parciální řešení, která mohou v budoucnosti způsobit velké problémy. Těmto problémům je lépe předcházet, a to co nejdříve. Pokud totiž nebudou data jednotlivých pracovišť alespoň částečně parametrizovatelná, nebude v budoucnu možné je navzájem propojovat a provádět nad nimi analýzu. Nebude tak možné rozhodovat o efektivním řízení léčby, vyhodnocovat ekonomickou situaci nemocnice a provádět benchmarking mezi pracovišti. Tvrzení, že dané pracoviště léčí správně a současně hospodárně, tak nebude ve skutečnosti ničím podloženo. V současné době je v ČR velice málo nemocnic, které by skutečně měly svá data „pod kontrolou“ a jednoznačně dokázaly pohotově argumentovat na dotazy. Jedním ze skutečně efektivních řešení je datový sklad (Data Ware House) a sofistikovaný reportingový systém. Jeden takovýto sklad je právě vyvíjen v Masarykově onkologickém ústavu a první výsledky ukazují, že by to mohla být za současné situace rozumná cesta, jak konečně dostat „data pod kontrolu“.

Popis současného stavu

Problémem dnešních nemocnic není nedostatek informací, ale jejich **využitelnost**, **nákladovost** a možnost **integrace**. Každého z nás napadne jistě napadne že když již existují nějaká data a jsou využitelná i jinde měla by se využít ty již existující. Jedním z podobných problémů může být hlášení NOR, které je dáno zákonem a nařízeno ministerstvem zdravotnictví. Teď nediskutuji o metodice ale o možnosti využitelnosti. ÚZIS v letošním roce zavedl pilotní projekt ve kterém je hlavním cílem integrovat data z hlášení NOR do jednotných datových struktur v NIS u onkologických pracovišť. To je jistě rozumná myšlenka, která ovšem nese řadu otázek. Pokud se nad tímto problém chvíli pozastavíme, napadne nás spousta otázek typu „Jak to asi bude drahé? Má to vůbec smysl? Je to reálné? Co to pro nás bude znamenat? Vyjděme z předpokladu, že hlášení NOR má jednotnou logickou strukturu a unifikovanou metodiku. Pak nastává problém kde získat data a jak je napojit do centrální databáze. V případě, že všechny uvedené problémy jsou vyřešeny, pak problém vlastně neexistuje a řeší se pouze technologická otázka jak data vyexportovat a z jakých zdrojů.

Dalším podobným problémem může být otázka provázanosti parametrů mezi systémy. Příkladem může být složitější otázka typu „Kolik máme v každém měsíci unikátních pacientů ve třetím klinickém stadiu, s jakou diagnózou, jakou mají nejčastější léčbu, jaké jsou jejich náklady a jaký je průměrný počet ‘lůžko-dnů‘ u těchto pacientů?“ Toto je velice komplikovaný dotaz na který dokáže málo kdo odpovědět během krátké chvíle. Z praxe víme, že pokud systémy a to nejen NIS, ale i ty ostatní nejsou provázané, nejsme schopni adekvátně odpovědět aniž bychom zadali tento požadavek do analytického oddělení a čekali 14 dní na výsledek analýz. V rámci tohoto dotazu můžeme narazit i na problém, že některý parametr vyskytující se v otázce není v NIS či jiném systému veden parametricky a stádium prostě nezjistíme, aniž bychom procházeli jednotlivé textové lékařské zprávy.

Dle zkušeností různých pracovišť v ČR patří mezi nečastější problémy:

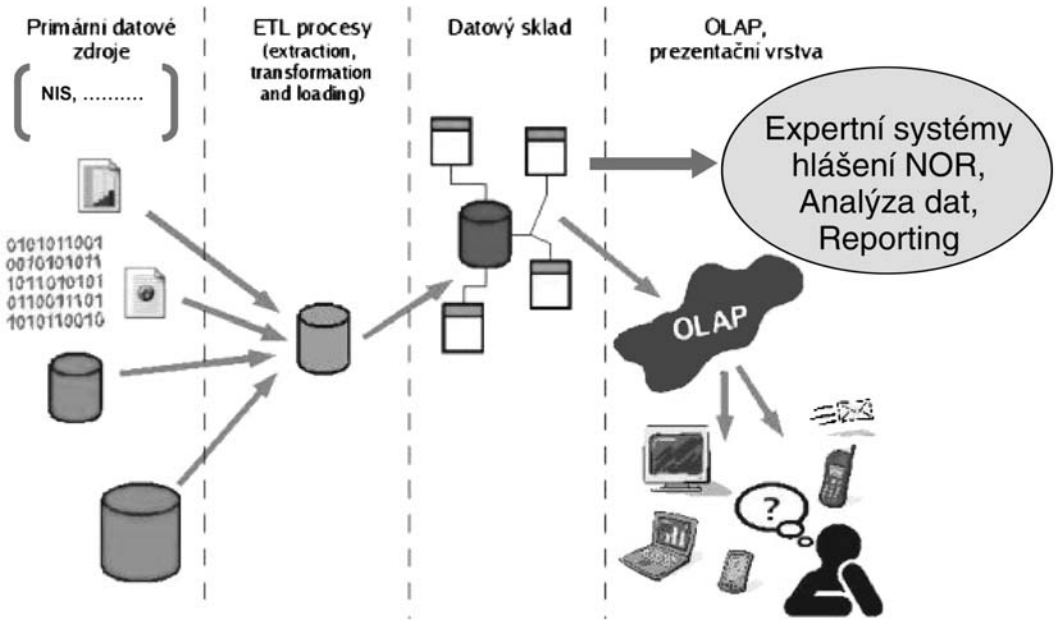
- Výrazná heterogenita informačních systémů
- Nejednoznačná identifikace provázanosti parametrů mezi systémy
- Chybějící dokumentace
- Nedostatečná definice finálních výstupů
- Záznamy parametrů v textové podobě
- Roztříštěnost modulů
- Špatná implementace
- Systém nerespektuje procesní chod nemocnice
- Složitá integrace

Jaká jsou možná řešení ?

Z výše popsaných problémů by se mohlo zdát že splnit předpoklad využitelnosti dat je prostě nemožné, nebo příliš komplikované. Není tomu tak. Když si uvědomíme, kolik informací již v současné době máme v různých systémech v nemocnici tak nám jednoznačně vyplívá že musí existovat jednoduché řešení. Osobně se domnívám, že není možné aby existoval jednotný

systém, který bude naprosto variabilní, flexibilní, parametrický a přitom měl nízké provozní náklady. Dnešní nemocnice se neobejdou bez IT či ICT oddělení, neustálého doplňování parametrů, rozšiřování datových struktur. Vznikají nové přístroje, diagnostické metody, mění se legislativa atd. to je potřeba všechno respektovat a přitom je nutné data umět integrovat. Jedním z možných východisek za daných podmínek může být datový sklad (Data Ware House DWH).

Datový sklad si můžeme pro jednoduchost představit jako velkou databázi, která obsahuje kromě dat samotných i vazby a definice vazeb mezi jednotlivými systémy. Jakmile jsou data uloženy, můžeme na data nahlížet různými zobrazovacími nástroji a dál data zpracovávat do dalších struktur jako jsou například OLAP kostky, rovnou vytvářet reporty, napojovat další aplikace apod. Příklad jak vypadá proces získávání dat do DWH ilustruje obrázek 1.



Obrázek 1

Implementace DWH na MOÚ

Masarykův onkologický ústav jako jedno z progresivních pracovišť se rozhodl technologie DWH využít a před 2 lety se rozběhla implementace zahrnující nejrozsáhlejší informační systémy nemocnice. Z pohledu navrhovaných a možných řešení se rozhodovalo na základě definice požadavků na DWH a dospělo se k následujícím závěrům: Návrh řešení datového skladu, musí být dostatečně robustní nejen z hlediska spolehlivosti a bezpečnosti, ale musí být také schopné plnit analytické funkce manažerského informačního systému s požadovanými výkonovými a ekonomickými parametry. Řešení je charakterizováno zejména následujícími vlastnostmi:

1. Použitá metodologie relačního on-line analytického zpracování umožňuje provádět **komplexní analýzu dat** (včetně adhoc analýzy, analýzy jednotlivých transakcí, tzv. analýzy nákupního koše, implementace data mining metod, atd.) a **minimalizuje dodatečné investice** při dalším rozvoji systému (rozšiřování počtu uživatelů, objemu analyzovaných dat, jejich granularity, počtu dimenzí a aplikace komplexnějších analytických metod).
2. Použitá technologie tzv. ETL procesů (zajišťujících přenos dat z provozních systémů do datového skladu a kontrolu jejich kvality) využívá distribuovaného zpracování pro zajištění **vysoké spolehlivosti** procesu a nesymetrického šifrování a systému certifikátů pro zajištění **bezpečnosti a autenticity dat**.
3. Systém využívá open source technologií, což kromě **nízkých celkových nákladů na vlastnictví** znamená i vyšší bezpečnost, úroveň **ochrany investice** a transparentnost celého řešení v důsledku přístupu ke zdrojovým kódům všech použitých komponent.
4. technologie musí obsahovat prezentační vrstvu, která bude splňovat předpoklad jednoduchosti a získatelnosti informace v reálném čase.

Cílem prezentační vrstvy je zpřístupnit analytickou informaci uživateli tak, aby pro něj byla **snadno vyhodnotitelná,** snadno **dostupná** (tj. informace „ve správný čas na správném místě“) a **důvěryhodná**. Samozřejmým požadavkem je potřeba volit takové technologie, které jsou optimální z hlediska **celkových nákladů na vlastnictví** systému (tedy nízké náklady na údržbu a administraci, na provoz hardware a síťové infrastruktury, nízké náklady na rozšiřování počtu uživatelů). Splnění výše uvedených požadavků je zajištěno vhodným výběrem použitých technologií a jejich promyšlenou implementací:

- **Snadné a jednoznačné vyhodnocení informace:** Kromě standardních výstupů ve formě vhodně formátovaných a parametrizovatelných tabulek je možné analytickou informaci prezentovat ve formě 2D grafů, map nebo přímo exportovat do

souborů tabulkového procesoru. Je také možné vytvářet prezentace dat v 3D prostoru s možností interaktivní manipulace a tak umožnit uživateli snadno získat přehled o struktuře dat a trendech jejího vývoje.

- **Snadná dostupnost informace:** Analytická informace musí být snadno dostupná jak uvnitř firemní sítě, tak i pro mobilní uživatele. Kromě interaktivního přístupu (např. přes web portál) je vhodné mít možnost doručit potřebnou informaci uživateli i proaktivně pomocí nejrůznějších komunikačních kanálů (např. e-mail, SMS, fax).
- **Nízké celkové náklady na vlastnictví systému:** Pro implementaci je využito open source technologií, jejichž licenční model zaručuje uživateli jejich nízkou pořizovací cenu, vysokou míru ochrany investic a díky důslednému používání otevřených standardů i malou závislost na jednom dodavateli. Kritické části systému jsou implementovány na platformě operačního systému GNU/Linux, díky jehož vysoké stabilitě a snadné údržbě je možné udržet velmi nízké náklady na provoz řešení.

Konkrétní požadavky a návrh datové struktury je součástí detailní Projektové dokumentace, která jasně specifikuje, struktury do kterých mají dodavatelé systémů exportovat data, včetně popisu jednotlivých parametrů.

Obecně lze říci že řešené požadavky se vyskytují z následujících oblastí:

- Klinická (pacient, hodnocení pacienta v čase, čistě klinická záležitost)
- Ekonomická (náklady, výnosy, hospodaření nemocnice…)
- Multilaterální (náklady na pacienta, obecně data agregovaná z více systémů)

Přístup prostřednictvím www portálu je vhodný pro širokou škálu uživatelů, kterým může poskytnout různou míru funkcionality a interaktivity podle jejich konkrétních potřeb. Vývojáři (tzv. „power-user“) mohou prostřednictvím portálu vyvíjet nové reporty a analytické objekty (metriky, filtry, atd.), které pak mohou využít ostatní uživatelé na různých stupních rozhodování ve firmě. Předem připravené reporty jsou snadno začlenitelné např. do firemního portálu, kde pak mohou poskytovat potřebné informace pro širokou komunitu uživatelů, a to i těch, kteří nejsou pro práci s OLAP technologií nijak speciálně školeni.

Uživatel může vytvářet reporty velmi jednoduchým způsobem přenášením (tzv. **„drag and drop“**) **atributů** a **metrik** do předem připravené mřížky modelu tabulky. Uživatel si tedy sám určí rozložení tabulky a které metriky chce sledovat v návaznosti na vybrané atributy. Součástí definice reportu je i možnost definice filtrace reportu a jeho formátování. Formát zobrazení lze definovat přímo v prostředí tenkého klienta (web browser) interaktivním způsobem využívajícím obvyklých ovládacích prvků pro definici barvy výstupu, typu a velikosti písma, zarovnání atd.

Pro formátování reportu lze také využít tzv. podmíněného formátování. Je možné definovat podmínku (např. obrat je větší než…) a na její splnění vázat určitý způsob formátování výstupní buňky reportu (barva, typ písma, buňku je možné doplnit o různé ikony, poznámky atd.).

V průběhu životního cyklu systému roste počet vytvořených objektů (atributů, faktů, metrik, filtrů, reportů, mřížek atd.) a je třeba přesně kontrolovat přístupová práva uživatelů k reportům a k jednotlivým analytickým objektům. Lze si např. představit situaci, kdy primář má být schopen vytvářet reporty zobrazující výkony jednotlivých lékařů jeho kliniky, ale neměl by mít možnost sledovat výkony ostatních klinik. Pro řízení přístupu k datům je použit systém tzv. **„uživatelských rolí“**. Každý uživatel pak může v systému vystupovat v jedné nebo více rolích a každý objekt (např. report, metrika, atribut) má pak svůj vlastní tzv. **„Access Control List“**. Snadno lze pak definovat, které operace s objektem jsou povoleny a které zakázány pro konkrétní roli.

Pro práci ve vícejazyčném prostředí je výhodná důsledná internacionalizace použité technologie a lokalizace prostředí aplikace podle potřeby konkrétní organizace. Nastavením uživatelských preferencí se **lokalizuje** nejen prostředí www portálu, ale také **názvy** všech vytvořených **objektů** (metrik, filtrů, reportů) a **obsah reportů** (např. názvy diagnóz a formát data). Alternativou zobrazení dat je grafická reprezentace pomocí **2D grafů** a **3D prezentací**, která umožní uživateli snadno a rychle se zorientovat v rozsáhlých reportech a nacházet různé spojitosti a trendy v datech, které nejsou z tabulkové prezentace na první pohled patrné. Grafická prezentace výsledků je přístupná prostřednictvím tenkého klienta (web browser), a to i včetně možnosti interaktivní manipulace.

Užitečnou pomocnou metodou pro analýzu vícerozměrných dat je **interaktivní zobrazení dat v třírozměrném prostoru** (tyto metody bývají zahrnovány do oblasti vizuálního data miningu). Možnost interaktivní manipulace s 3D prezentací usnadňuje uživateli pochopení struktury dat, popř. trendů jejich vývoje. Vyžití statistického software projektu „R“ umožňuje analytikům použít širokou škálu pokročilých statistických metod a metod pro dolování dat (**data mining**), vyvíjených na nejvýznamnějších světových akademických pracovištích. Grafické knihovny projektu „R“ lze také využít pro generování kvalitní tzv. „statistické“ a „obchodní“ grafiky.

Využití kontingenčních tabulek tabulkového procesoru. Po přenosu dat z datového tržiště do tabulkového procesoru je možné velice efektivně provádět tzv. „řezy“ vícerozměrným datovým prostorem, včetně snadného drillování, pivotování a filtrování. Výsledky je možné snadno formátovat, zobrazovat v podobě 2D a 3D grafů a publikovat (ve formě PDF dokumentů nebo XLS souborů). Tento způsob práce s daty je vhodný pro menší objemy zpracovávané informace a menší počty sledovaných atributů a metrik. Data konkrétního reportu je také možno exportovat přímo z rozhraní tenkého klienta do tabulkového procesoru (MS Office, OpenOffice.org, StarOffice) ve formě tabulky včetně formátování. Jako vhodný tabulkový procesor je možné jme-

novat produkty OpenOffice.org nebo StarOffice firmy Sun, zejména pro dobře propracovaný a pružný systém definování kontingenčních tabulek a kvalitní grafický výstup (i ve formě 3D grafů).

Použité technologie

Navrhované řešení manažerského informačního systému je založeno na koncepci relační on-line analýzy dat, (tzv. **ROLAP** – Relational On-Line Analytical Processing), která poskytuje vynikající možnosti analýzy jak agregovaných ukazatelů, tak i atomických dat (např. dat jednotlivých obchodních transakcí). Využitím těchto prostředků je možné efektivně připravovat širokou škálu výstupů, od jednoduchých reportů až po komplexní analýzy, jinými technologiemi nerealizovatelnými. Systém umožňuje zpřístupnit výstupy uživatelům jak prostřednictvím tenkých klientů (WEB technologie), tak i formou přímé analýzy prostřednictvím kontingenčních tabulek v tabulkovém procesoru nebo je automaticky vhodně zformátované distribuovat uživatelům prostřednictvím faxu, elektronické pošty nebo SMS zpráv.

Použitá platforma díky své modulárnosti a škálovatelnosti umožňuje efektivně rozvrhnout projekt do jednotlivých etap a tak optimálně rozložit a zhodnotit nutné investice i zapojení vnitřních zdrojů odběratele. Otevřenost systému a využití standardních technologií (**XML, SOAP API**) zaručuje jeho snadnou integraci s technologiemi třetích stran. Flexibilita relačního datového modelování dává možnost optimálně navrhnout datový sklad a pružně reagovat na jeho budoucí modifikace a optimalizovat jej tak, aby bylo zaručeno efektivní fungování systému i při jeho dalším rozvoji. Podobně jako obecný systém pro podporu rozhodování, tak i řešení využívající technologie projektu MOÚ je rozděleno do tří základních vrstev: ETL procesy, datový sklad a prezentační vrstva. Tzv. ETL procesy zajišťují extrakci dat z primárních datových zdrojů, jejich transformaci a zavedení do datového skladu (z angl. „Extraction, Transformation and Loading“). Podle určení systému tato vrstva může být reprezentována datovým tržištěm nebo datovým skladem.

Vlastní úložiště datového skladu (ev. datového tržiště) může být zabezpečeno obecným relačním systémem řízení báze dat. V současné době se jeví jako výhodné pro uvažované objemy využít technologie databáze MySQL (http://www.mysql.com), která disponuje vynikajícími vlastnostmi pro využití v režimu tzv. RO (read-only) databáze při jedné z nejlepších propustností na trhu. Pro využití v oblasti datových skladů stojí za zmínku zejména následující vlastnosti:

- možnost uzamčení tabulek v paměti a využití hash indexů
- možnost komprimace tabulek a indexů
- možnost uložení a rychlého čtení dat přímo z indexových B-stromů
- tzv. „Merge“ tabulky umožňující partikulární komprimaci virtuální datové tabulky
- tzv. stripping tabulek umožňující jejich rozložení přes více diskových svazků a tím zvýšení I/O propustnosti
- možnost replikace databáze na více fyzických systémů jak pro zvýšení spolehlivosti, tak pro rozložení zátěže při čtecích operacích

Systém řízení báze dat MySQL je produktem švédské firmy MySQL AB. Jedná se o Open Source technologii publikovanou jak pod GPL, tak pod komerční licencí. Se svými více než 4 miliony aktivních instalací se řadí mezi celosvětově nejrozšířenější databázové platformy. Díky tak širokému nasazení a dostupnosti zdrojových kódů se vyznačuje výbornou stabilitou. V oblasti datových skladů firma MySQL AB disponuje referencí databáze o velikosti tabulek faktů 50GB, celkové velikosti 200 GB (7 mil. záznamů/měsíc) udržující historii až 10 let.