Možnosti analýzy podnikových dat

Business Data Analyses Possibilities

Bc. Dagmar Pokorná

Diplomová práce
2010

Univerzita Tomáše Bati ve Zlíně
Fakulta aplikované informatiky

Univerzita Tomáše Bati ve Zlíně
Fakulta aplikované informatiky
akademický rok: 2009/2010

ZADÁNÍ DIPLOMOVÉ PRÁCE

(PROJEKTU, UMĚLECKÉHO DÍLA, UMĚLECKÉHO VÝKONU)

Jméno a příjmení:	**Bc. Dagmar POKORNÁ**
Studijní program:	**N 3902 Inženýrská informatika**
Studijní obor:	**Informační technologie**
Téma práce:	**Možnosti analýzy podnikových dat**

Zásady pro vypracování:

1. **Rozbor problematiky a vypracování literární rešerše na dané téma.**
2. **Navrhněte ukázkový příklad pro analýzu podnikových dat.**
3. **Realizujte navržený příklad pomocí nástrojů MS SQL Serveru.**
4. **Zhodnoťte řešení, uveďte jeho výhody a nevýhody.**

Rozsah práce:

Rozsah příloh:

Forma zpracování diplomové práce: **tištěná/elektronická**

Seznam odborné literatury:

1. **NOVOTNÝ, Ota, POUR, Jan, SLÁNSKÝ, David. Business Intelligence : Jak využít bohatství ve vašich datech. Praha : Grada Publishing, 2004. 256 s. ISBN 80-247-1094-3.**
2. **POUR, Jan, GÁLA, Libor, ŠEDIVÁ, Zuzana. Podniková informatika. 2. přeprac. vyd. Praha : GRADA Publishing, 2009. 496 s. ISBN 978-80-247-2615-1.**
3. **TVRDÍKOVÁ, Milena. Aplikace moderních informačních technologií v řízení firmy. Praha : GRADA Publishing, 2008. 176 s. ISBN 978-80-247-2728-8.**
4. **SODOMKA, Petr. Informační systémy v podnikové praxi. Praha : Computer Press, 2006. 352 s. ISBN 80-251-1200-4.**
5. **LACKO, Luboslav. Databáze: datové sklady, OLAP a dolování dat. Praha : Computer Press, 2003. 488 s. ISBN 80-7226-969-0.**
6. **LACKO, Luboslav. Business Inteligence v SQL Serveru 2005 : Reportovací, analytické a další datové služby. Praha : Computer Press, 2006. 392 s. ISBN 80-251-1110-5.**
7. **BRUST, Andrew, FORTE, Stephen. Mistrovství v programování SQL Serveru 2005 . Praha : Computer Press, 2007. 848 s. ISBN 978-80-251-1607-4.**

Vedoucí diplomové práce: **doc. Ing. Zdenka Prokopová, CSc.**
Ústav počítačových a komunikačních systémů

Datum zadání diplomové práce: **19. února 2010**

Termín odevzdání diplomové práce: **8. června 2010**

Ve Zlíně dne 19. února 2010

prof. Ing. Vladimír Vašek, CSc.
děkan

prof. Ing. Vladimír Vašek, CSc.
ředitel ústavu

ABSTRAKT

Dnešní doba je charakterizována expanzí dat. Téměř všechny podnikové procesy jsou nějakým způsobem uloženy v databázových strukturách, do kterých denně přitékají další tisíce záznamů. Uvnitř té směsi bitů je ukryto množství cenných informací, vypovídajících o skrytém potenciálu podniku. Úkolem analytických nástrojů je tyto informace získat a převést je do podoby, využitelné pro další řízení a rozhodování. Cílem mé diplomové práce bylo ověření možností analýz podnikových dat vzhledem k dostupnosti vhodných nástrojů určených koncovým uživatelům.

Klíčová slova:

Analýza vícerozměrných dat, Business Inteligence, dolování dat, OLAP, datový sklad

ABSTRACT

Data expansion is specific for these days. Almost every business processes are somehow stored in database structures, into which thousands of new records are daily loaded. Inside this data structure is a great deal of valuable information which predict hidden potential of a company. Function of analytical instruments is to gain this information and transform it into form that can be used for another management. The main aim of my diploma thesis was to verify possibilities of business data analysis with a view on availability of appropriate tools used by end users.

Keywords:

Multivariable Data Analysis, Business Inteligence, Data Mining, OLAP, Datawarehouse

Na tomto místě bych chtěla vyjádřit své poděkování doc. Ing. Zdeňce Prokopové, CSc. za cenné informace a odborné rady, kterými přispěla k vypracování této diplomové práce.

Prohlašuji, že

- beru na vědomí, že odevzdáním diplomové/bakalářské práce souhlasím se zveřejněním své práce podle zákona č. 111/1998 Sb. o vysokých školách a o změně a doplnění dalších zákonů (zákon o vysokých školách), ve znění pozdějších právních předpisů, bez ohledu na výsledek obhajoby;
- beru na vědomí, že diplomová/bakalářská práce bude uložena v elektronické podobě v univerzitním informačním systému dostupná k prezenčnímu nahlédnutí, že jeden výtisk diplomové/bakalářské práce bude uložen v příruční knihovně Fakulty aplikované informatiky Univerzity Tomáše Bati ve Zlíně a jeden výtisk bude uložen u vedoucího práce;
- byl/a jsem seznámen/a s tím, že na moji diplomovou/bakalářskou práci se plně vztahuje zákon č. 121/2000 Sb. o právu autorském, o právech souvisejících s právem autorským a o změně některých zákonů (autorský zákon) ve znění pozdějších právních předpisů, zejm. § 35 odst. 3;
- beru na vědomí, že podle § 60 odst. 1 autorského zákona má UTB ve Zlíně právo na uzavření licenční smlouvy o užití školního díla v rozsahu § 12 odst. 4 autorského zákona;
- beru na vědomí, že podle § 60 odst. 2 a 3 autorského zákona mohu užít své dílo – diplomovou/bakalářskou práci nebo poskytnout licenci k jejímu využití jen s předchozím písemným souhlasem Univerzity Tomáše Bati ve Zlíně, která je oprávněna v takovém případě ode mne požadovat přiměřený příspěvek na úhradu nákladů, které byly Univerzitou Tomáše Bati ve Zlíně na vytvoření díla vynaloženy (až do jejich skutečné výše);
- beru na vědomí, že pokud bylo k vypracování diplomové/bakalářské práce využito softwaru poskytnutého Univerzitou Tomáše Bati ve Zlíně nebo jinými subjekty pouze ke studijním a výzkumným účelům (tedy pouze k nekomerčnímu využití), nelze výsledky diplomové/bakalářské práce využít ke komerčním účelům;
- beru na vědomí, že pokud je výstupem diplomové/bakalářské práce jakýkoliv softwarový produkt, považují se za součást práce rovněž i zdrojové kódy, popř. soubory, ze kterých se projekt skládá. Neodevzdání této součásti může být důvodem k neobhájení práce.

Prohlašuji,

- že jsem na diplomové práci pracoval samostatně a použitou literaturu jsem citoval. V případě publikace výsledků budu uveden jako spoluautor.
- že odevzdaná verze diplomové práce a verze elektronická nahraná do IS/STAG jsou totožné.

Ve Zlíně …….

podpis diplomanta

OBSAH

ÚVOD

Informační technologie, které nás dnes doprovázejí doslova na každém kroku, sebou nesou obrovské objemy shromažďovaných dat, v nichž je ukryto velké množství informací. Data a informace nejsou synonyma, data jsou pouze klíčem k bohatství informací.

Kvalitní analýza dat a úroveň získaných informací stojí na pozadí všech správných manažerských rozhodnutí. Dobří manažeři je dokáží využít ke zvýšení výkonnosti a konkurenceschopnosti podniku, protože dovedou nejenom rozebrat události minulé, ale také predikovat trendy a možné budoucí směry vývoje a tomu pak podřídit další kroky řízení. Dokáží odhalit anomálie trhu, zaměřit se na vhodné zájmové skupiny zákazníků, měřit a porovnávat výkonnost. Dobré řídící pracovníky dokáží vhodně použité analytické nástroje podpořit v jejich krocích. A ty špatné, snad, alespoň občas od některých chybných rozhodnutí odvrátit. Analýza dat se už dávno netýká pouze vrcholových manažerů a potřeb jejich strategických rozhodování. Nachází uplatnění na všech úrovních řízení. Snad právě proto jsem se ve své praxi často, mnohdy už na prvních obchodních prezentacích, setkávala s dotazy, které se týkaly možností výstupů a analýz nabízených informačních systémů. Zjistila jsem, že sestavy, jejich tvorba, rozsah a dostupnost, se velmi výrazně podílí na celkovém vnímání a hodnocení systému. Poznala jsem informační systémy, které poskytovaly pružné parametrické reporty s širokými možnostmi exportů dat, ale také systémy, které měly integrovány pouze sestavy s pevně definovanou strukturou nebo systémy, jejichž jedinou možností exportu byl výstup do formátu pdf. A i když dodavatelé systémů ve většině případů nabízejí jako službu vlastní tvorbu sestav podle požadavků zákazníků nebo prodávají komerční moduly pro vytváření reportů, každý zkušenější zákazník ví, že ceny sestav se mohou časem (v případě zakázkové tvorby sestav) nebo jednorázově (u komerčních modulů pro tvorbu sestav) velmi výrazně promítnout do nákladů spojených s informačním systémem. Protože sestavy a následné rozbory podnikových dat patří k tomu nejcennějšímu, co informační systémy manažerům a vedoucím pracovníkům na všech stupních řízení nabízejí, zaměřila jsem se ve své diplomové práci na běžně dostupné způsoby získávání dat a možnosti jejich analýz.

Nástroje, dovolující hloubkové analýzy podnikových dat, jsou dnes již integrovány do databázových strojů a jsou tak přímou součástí podnikových informačních systémů. Stojí nad transakčními databázemi a komunikují s běžnými kancelářskými programy. Tím

poskytují možnost přístupu k podnikovým datům i těm firemním uživatelům, na něž by se v případě nákupu specializovaného software nejspíš nedostalo.

I. TEORETICKÁ ČÁST

1 VÝVOJ ANALYTICKÝCH SYSTÉMŮ

Se společností IBM je spojeno mnoho technologických prvenství z oblasti informatiky a výpočetní techniky. Patří k nim také první definice principů Business Intelligence. U jejich zrodu stál výzkumný pracovník německého původu Hans Peter Luhn, který v roce 1958 publikoval v IBM Journal článek s názvem "A Business Intelligence System". V něm formuloval hlavní myšlenky této filozofie, která vychází z toho, že obchodní cíle společností by měly být stanoveny na základě vyhodnocení existujících faktů. [1]

I přesto, že koncepce Business Inteligence byla vytvořena v době, kdy výpočetní technologie byly dostupné pouze úzkému okruhu odborníků, intelektuální předpoklady stanovené v roce 1958 se o několik desítek let později promítly do softwarových programů, určených k poskytování manažerských informací.

Do povědomí širší veřejnosti uvedl termín Business Intelligence až v roce 1989 analytik společnosti Gartner Group Howard J. Dresner. Popsal jej jako „sadu konceptů a metod určených pro zkvalitnění analytických a rozhodovacích procesů v organizacích“. Zaměřil se na význam datové analýzy, reportingu a dotazovacích nástrojů, které nabízejí uživateli množství dat a pomáhají mu se syntézou hodnotných a užitečných informací. [2]

Zajímavé je, že koncepce H.P.Luhna se objevila v době, ve které výpočetní systémy začínaly teprve velmi zvolna pronikat z oblastí armádních a vládních výzkumů do oblastí komerčních. Výraznější rozšíření výpočetní techniky do těchto sfér umožnily až integrované obvody, které znamenaly jak postupnou miniaturizaci tak také snižování cen počítačů. Zlom představovaly zejména legendární počítače řady IBM System 360, které byly oficiálně představené 7.dubna 1964. Byly to první počítače, které těžily z vynálezu integrovaných obvodů[1], které používaly vzájemně kompatibilní periferie a jejichž architektura byla postavená na modulární struktuře[2]. Revoluční byla také, do té doby nevídaná, přenositelnost a zpětná kompatibilita software. [3] [4]

---

[1] Tyto počítače ještě nebyly osazeny mikročipy přímo. Byly vybaveny hybridními obvody, které používaly čipová lůžka, na které byly napájené externí součástky. Šlo o kompromis mezi druhou (tranzistorovou) generací a třetí (čipovou) generací počítačů.

[2] Při zvýšení nároků mohli zákazníci vyměnit nebo rozšířit jenom potřebné části hardware. Nemuseli již investovat do zakoupení zcela nového stroje.

Právě na těchto počítačích byly v šedesátých letech minulého století provozovány první informační systémy ve velkých podnicích a bankách. Přestože nesly označení Management Information Systems (systémy určené k podpoře řízení), jednalo se o běžné rutinní agendy zaměřené především na zpracování dat z oblasti účetnictví. Potřebám strategického rozhodování a řízení vycházely vstříc pouze nepřímo poskytováním předem definovaných periodických reportů.

Proto se počátkem sedmdesátých let, většinou ještě jako součást Management Information Systems objevují systémy, určené lidem zabývajícím se nejenom každodenním provozním řízením, ale i řízením strategickým. Světlo světa tak spatřily aplikace souhrnně označované jako Decision Support Systems (systémy pro podporu rozhodování). Jejich základním úkolem bylo poskytování informací a nástrojů, pomocí nichž mohly být modelované a vyhodnocované různé podnikové alternativy a strategie.

O rozvoj systémů pro podporu rozhodování se však nezasloužily pouze potřeby manažerů, ale také další vývoj v oblasti hardware a software. Za klíčové pro DSS lze přitom považovat zejména dvě oblasti - změnu rychlosti přístupu k datům (náhrada magnetických disků se sekvenčním přístupem za magnetické disky s přímým přístupem) a převratný návrh relačního datového modelu dr. Codda, založeného na matematické teorii množin.

Tyto systémy byly již u řídících pracovníků úspěšnější než mnohem šířeji zaměřené systémy pro podporu řízení (MIS). Dobře se prosadily u středních řídících kádrů, ale u nejvyšších řídících složek (výkonných ředitelů, prezidentů společností apod.) příliš neuspěly.

Koncem osmdesátých let, s nástupem graficky orientovaných uživatelských rozhraní, s nástupem myší a s dalších pokroků na poli software i hardware, se pak objevuje třetí vlna prostředků, vydávajících se na pomoc řídícím procesům. Jedná se o prostředky označované jako Executive Information Systems (někdy též Executive Support Systems), které již míří k manažerům nejvyšším a snaží se jim nabídnout přímý (on-line) přístup k aktuálním informacím o stavu řízených organizací a to s maximální možnou mírou intuitivnosti, využívajíc "uživatelské přítulnosti" grafických uživatelských rozhraní. První aplikace tohoto typu pracovaly přímo na pořízených datech. Zatěžovaly však primární systémy, a proto došlo k oddělování provozních dat a dat pro analýzy. [5] [6] [7]

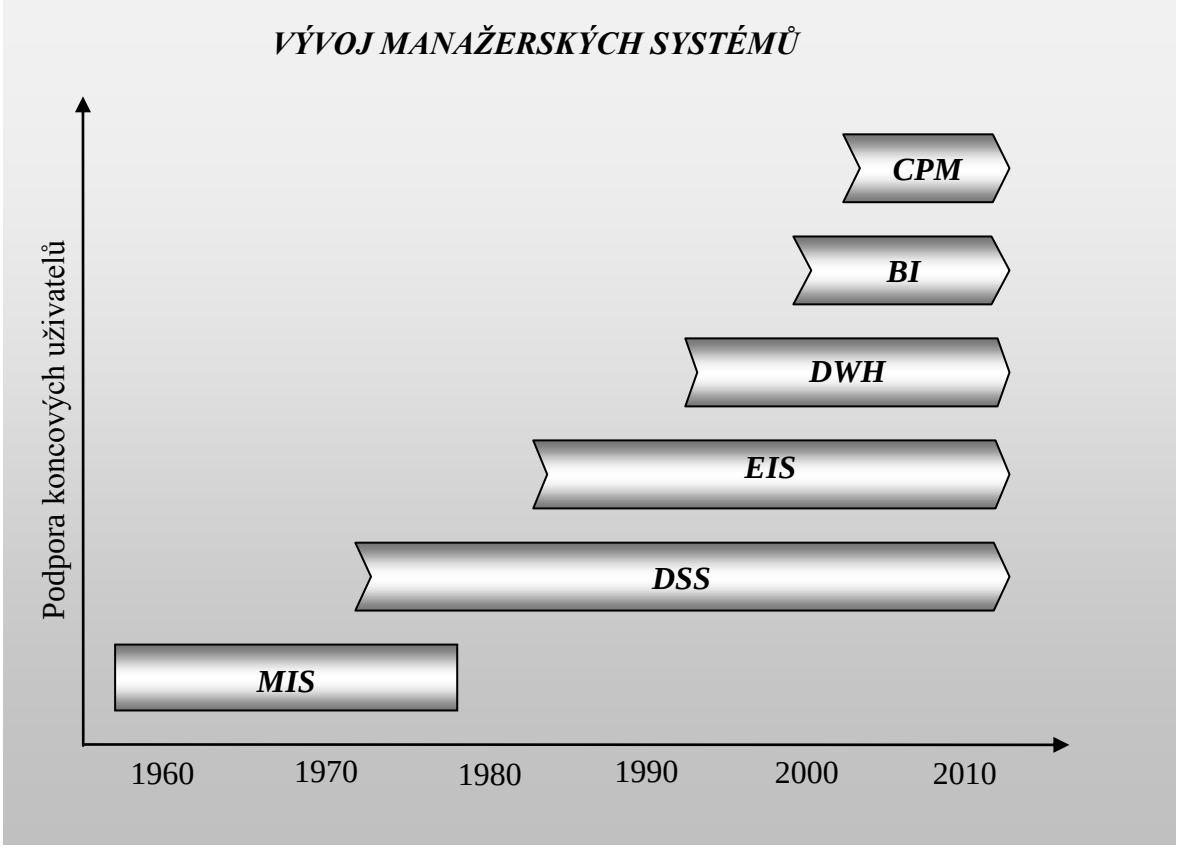
To byl počátek datový skladů, jejichž filozofii poprvé publikoval v knize Building the Data Warehouse v roce 1991 Bill Inmon, dnes světově uznávaný odborník na danou problematiku. Jeho definice datových skladů zní:

„Datový sklad je podnikový strukturovaný depozitář předmětově (subjektově) orientovaných, vzájemně provázaných, nepodléhající změnám, časově proměnných, historických dat, používaných na získávání informací a pro podporu rozhodování. V datovém skladu jsou uloženy data detailní (atomická) i sumární.“ [8]

Pravá příčina vzniku datových skladů však souvisí zejména s masivním nasazováním serverových podnikových systémů, ke kterému došlo na konci osmdesátých let minulého století a s jejich koncepcí samostatných, vzájemně nezávislých, aplikací. V podnicích vznikala potřeba komplexních analýz a hledání souvislostí mezi oddělenými daty, jejichž objem navíc znepokojivě narůstal[3]. Vznik datových skladů tak byl logickým vyřešením tohoto problému. Datové sklady vznikly jako samostatné informační systémy, postavené nad podnikovými daty. [9] A tuto funkci plní i dnes, jen s tím rozdílem, že už se nejedná o nespojitá data jednoho systému, ale o oddělená data různých nezávislých podnikových ERP systémů, specializovaných systémů a jiných externích zdrojů dat. Kořeny Inmonova datového skladu, budovaného na principu entitně-relační databáze, ležely ve správě dat. To mělo řadu výhod, ale také řadu nevýhod a omezení. S novým pohledem na danou problematiku přišel o tři roky později, v roce 1994, Ralph Kimball. Jeho filozofie byla založena na tzv. datových tržištích. Jednalo se v podstatě o vyčlení vybraných dat datového skladu do okruhů, určených omezeným skupinám uživatelů (např. data vybraná pro pracovníky marketingu …). Zatímco datové sklady jsou předmětově orientované (data jsou rozdělována podle typu), datová tržiště jsou orientovaná problémově. K ukládání dat sloužil nový multidimenzionální databázový model, který umožňoval snadno a rychle vytvářet nejrůznější pohledy na data pomocí různých řezů datovou kostkou. Tato technologie je základem dnešních analytických nástrojů Business Inteligence a je v této diplomové práci podrobněji rozpracována dále. [8] [10]

[3] Každá část těchto systémů byla orientována na jednu vybranou oblast (výroba, sklady …) a chovala se jako samostatný systém. Pracovala s vlastními, vzájemně oddělenými daty.

Propojením BI s nástroji podnikového plánování vznikl nový typ aplikací, kterým se dnes říká Corporate Performance Management (CPM) – Řízení podnikového výkonu.



Obr. 1 Vývoj manažerských systémů. Podle [11]

2 PRINCIPY A NÁSTROJE ANALYTICKÝCH SYSTÉMŮ

2.1 Principy

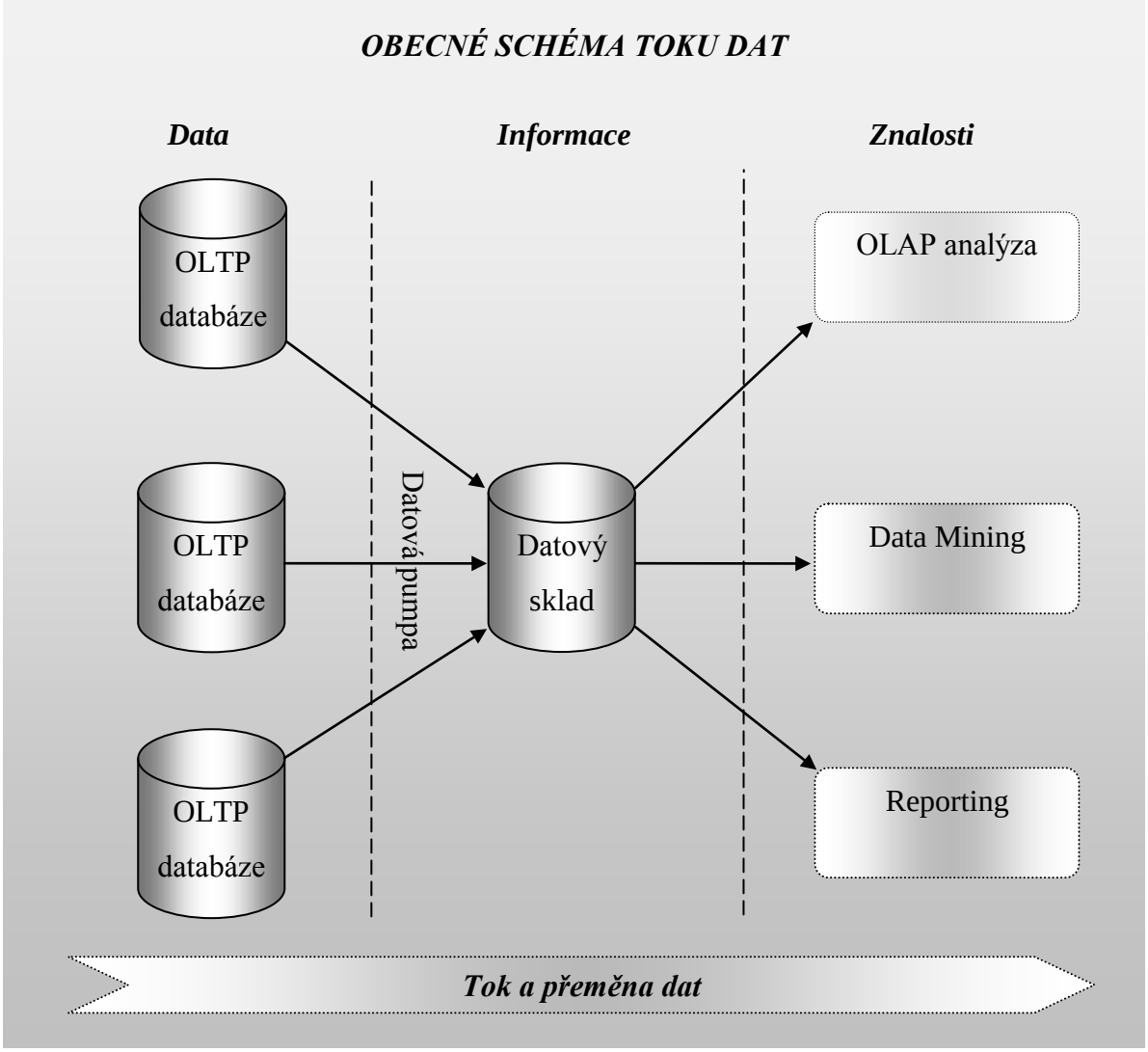
Hloubková analýza dat podnikových informačních systémů a jejich následné využití při řízení firmy spadá pod obecné označení Business Inteligence. K českému ekvivalentu tohoto pojmu má asi nejblíže termín „manažerské rozhodování“. Přesná definice, zatím neexistuje, Česká společnost pro systémovou integraci používá následující:

“Business Inteligence (BI) je sada procesů, aplikací a technologií, jejichž cílem je účinně a účelně podporovat řídící aktivity ve firmě. Podporují analytické a plánovací činnosti organizací a jsou postaveny na principech multidimenzionálních pohledů na podniková data. Aplikace BI pokrývají analytické a plánovací funkce většiny oblastí podnikového řízení., tj. prodeje nákupu, marketingu, finančního řízení, controllingu, majetku, řízení lidských zdrojů, výroby.“ [2]

Z definice vyplývá analytický a plánovací charakter BI aplikací, které se od běžných provozních systémů liší v uživatelském pohledu na data. Zatímco provozní úlohy pracují s detailními informacemi, analytické úlohy pracují s agregovanými daty. V praxi si to můžeme představit takto:

- Obchodník, pracující s **provozními daty**, bude mít k dispozici podrobné informace jednotlivých obchodních případů. Bude vědět kdy a co který zákazník nakoupil, ale nebude schopen rychle a jednoduše vyhodnotit jeho chování během libovolného časového období, případně predikovat další vývoj.
- Obchodník, pracující s **analytickými agregovanými daty**, bude mít k dispozici informace obecného charakteru. Bude znát celkový prodej dle zákazníků, produktů, prodejců. Obratem zjistí marži nebo obrat jakéhokoliv časového období. Na základě historických faktů snadno vysleduje trend a tím pádem lépe a efektivněji zacílí marketingové aktivity.

Aby bylo možné analytický pohled realizovat, bylo nutné změnit technologii přístupu k datům. Zatímco provozní systémy pracují s transakčními entitě-relačními databázemi, analytické systémy pracují s datovými sklady a multidimenzionálními databázemi, coby poskytovatelem informací. V té souvislosti se také často mluví o transformaci dat na informace a na znalosti.



Obr. 2 Schéma toku dat v analytických systémech. Volně podle [22]

Obrázek znázorňuje základní schéma toku dat a jejich přeměny z dat na informace a znalosti. Data, pocházející z různých provozních systémů, textových a XML souborů, tabulek, webových aplikací jsou posbírány, očištěny, upraveny, sjednoceny a pomocí datových pump „pumpovány“ do datového skladu“, odkud jsou pomocí analytických nástrojů zpřístupněny koncovým uživatelům. Mohlo by se zdát, že celý tento složitý proces je zbytečný, protože uživatelé by mohli analyzovat zdrojová data přímo. To by ale bylo možné pouze v malých organizacích, provozujících jeden málo vytížený informační systém. V podnicích, ve kterých se analyzují data ze systémů, v nichž mohou probíhat až tisíce transakcí za vteřinu nebo ze systémů korporací, provozujících své pobočky a závody po celém světě, je takové řešení zatím organizačně i technicky nemožné. [14]

2.2 Hlavní komponenty

Obecnou koncepci analytických řešení tvoří čtyři základní komponenty, z nichž každá poskytuje několik nástrojů [12]:

- **Komponenty datové transformace**

  slouží ke sběru a přenosu dat ze zdrojových systémů do datových skladů a úložišť. Zahrnují:

  - ETL systémy pro extrakci, transformaci a přenos dat
  - EAI systémy pro integraci aplikací

- **Databázové komponenty**

  zajišťují procesy uklání, aktualizace a správy dat.

  Zahrnují:

  - Datové sklady - Data Warehouse (DWH)
    - centrální úložiště podnikových dat
  - Datová tržiště - Data Marts (DMA)
    - subjektově orientované analytické databáze
  - Operativní datová úložiště - Operational Data Store (ODS)
    - podpůrné analytické databáze
  - Dočasná datová úložiště - Data Staging Areas (DSA)
    - databáze pro dočasné uložení dat před jejich vlastním zpracováním

- **Analytické komponenty**

  pokrývají činnosti spojené s vlastním zpřístupněním a analýzou dat

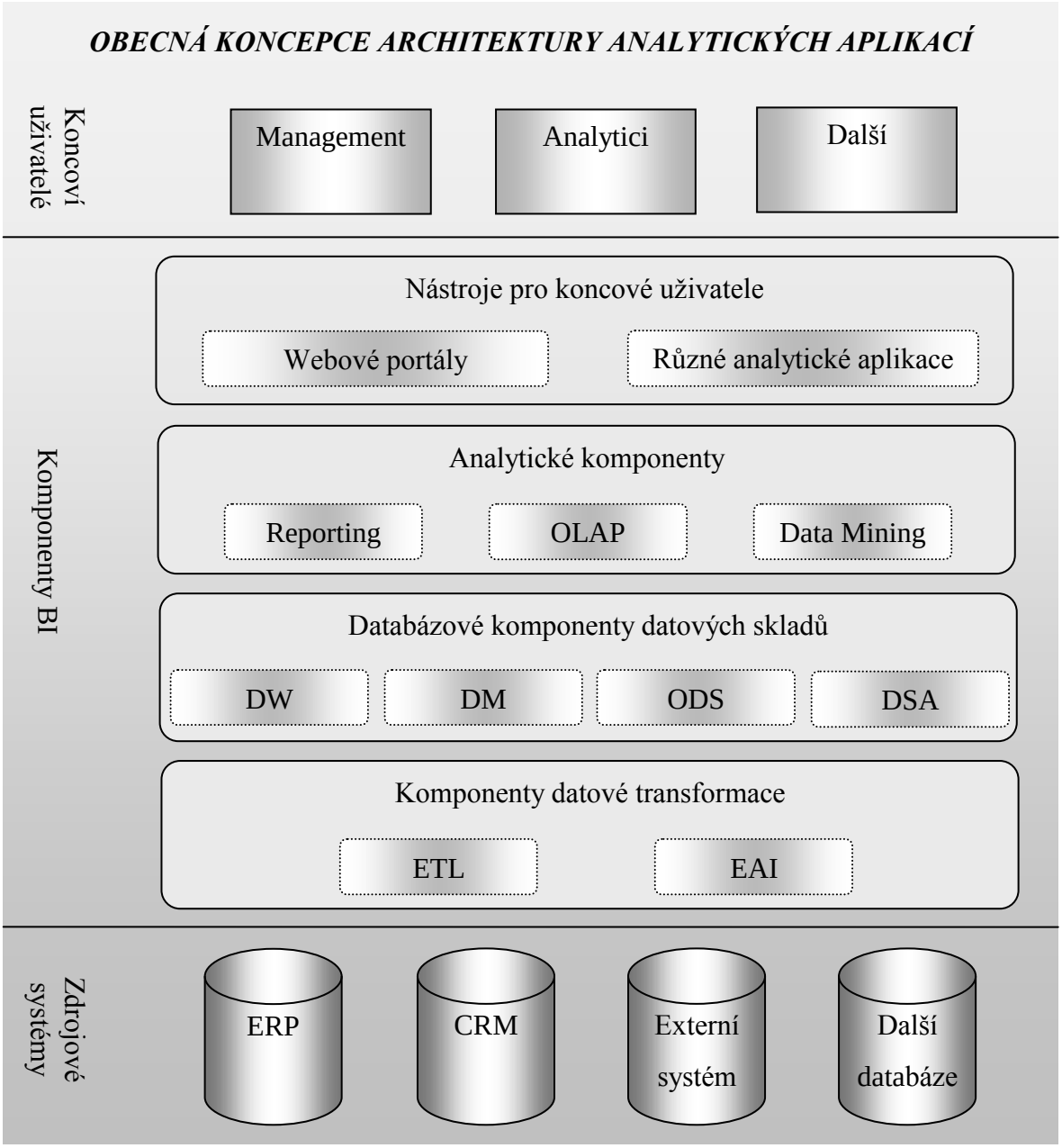
  Zahrnují:

  - Reporting
    - analytická vrstva, zaměřená na standardní nebo ad hoc dotazovací proces do databázových komponent

- On-Line Analytical Processing (OLAP)
  - vrstva zaměřená na pokročilé dynamické analytické úlohy
- Dolování dat (Data mining)
  - systémy zaměřené na sofistikovanou analýzu velkého množství dat

- **Nástroje pro koncové uživatele**

  zajišťují komunikaci koncových uživatelů s ostatními komponentami řešení, tedy zejména sběr požadavků na analytické operace a následnou prezentaci výsledků.



Obr. 3 Obecná koncepce architektury analytických aplikací. Podle [12]

3 KOMPONENTY DATOVÉ TRANSFORMACE - DATOVÉ PUMPY

Abychom získali data, vhodná pro další analýzu, musíme je z provozních (primárních, transakčních, OLTP či „legacy) systémů vytáhnout a převést do datového skladu. Pak můžeme provádět analýzy pomocí OLAP technologií (základní analýzy dat), Data Mining technologií (pokročilé analýzy dat) nebo pomocí reportovacích služeb vytvářet sestavy.

Obr. 4 Transformace dat v analytických systémech

Tento proces je při budování datových skladů nejdůležitější a také nejnáročnější, protože je při něm nutné zajistit analýzu obsahově i technologicky nehomogenních datových zdrojů, pocházejících z různých interních i externích systémů (například souborových databází, relačních databází, flat file souborů, XML struktur nebo dalších) a potom podle potřeb řízení firmy vybrat relevantní data, vzájemně je centralizovat, integrovat a agregovat. Klíčovými nástroji, používanými v této technicky obtížné fázi jsou:

- ETL (Extract, Transform and Load)
- EAI (Enterprise Application Itegration)

3.1 ETL (Extract, Transform and Load)

Plnění datového skladu (ETL proces) začíná extrakcí dat z primárních zdrojů (Extraction). Během této fáze jsou vyhledávány a odstraňovány různé datové nekonzistence. Extrahovaná data mohou být před jejich transformací do datových schémat uložena v dočasných úložištích. Komponenta dočasného úložiště dat (Data Staging Area - DSA) bývá nejčastěji součástí těch řešení datových skladů, jejichž zdrojem jsou velmi vytížené transakční systémy. Nasazením DSA se sníží potřeba využití transakčních systémů při procesu ETL a ty se pak mohou plně věnovat obsluze podnikových procesů. DSA je možné využít také v případě, kdy je nutné data převést do požadovaného databázového formátu například z textových souborů. [13]

Po extrakci následuje transformace dat (Transformation), která převede data, získaná z jednotlivých datových zdrojů, do unifikovaného datového modelu, nad nímž je možné vytvářet agregace a seskupování. [13]

Závěrečnou fází ETL je přenos údajů z paměti zdrojových dat nebo z dočasného úložiště do databázových tabulek datového skladu. Při prvotním naplnění může jít o obrovské množství dat. Protože ETL pracuje v dávkovém (batch) režimu, další pravidelné aktualizace, už přinášejí jen takové množství dat, jaké v odpovídajícím časovém období (den, týden, měsíc) v OLPT databázích vznikne.

3.2 EAI (Enterprise Application Itegration)

Nástroje EAI jsou využívány ve vrstvě zdrojových systémů. Jejich cílem je integrace primárních podnikových systémů a razantní redukce počtu jejich vzájemných rozhraní. Tyto nástroje pracují na dvou úrovních:

- na úrovni datové integrace, kde jsou použity pro integraci a distribuci dat
- na úrovni aplikační integrace, kde jsou využity nejenom pro integraci a distribuci dat, ale především pro sdílení vybraných funkcí informačních systémů.

EAI platformy pracují, na rozdíl od nástrojů ETL, v reálném čase. [12]

4 DATABÁZOVÉ KOMPONENTY - DATOVÉ SKLADY

Datové sklady, v anglicky psaných dokumentech „Data Warehouse“ (DW, DWH), jsou zvláštním typem podnikových databází, které obsahují konsolidovaná data ze všech dostupných provozních systémů. Nejsou optimalizované pro rychlé zpracování transakcí, nýbrž pro rychlé poskytování analytických informací získaných z velkého množství dat. Transakční databáze provozních systémů nejsou pro tento účel vhodné z několika důvodů:

- **Kapacita zdrojové databáze**

  V transakčních systémech jsou z kapacitních důvodů zpravidla udržována pouze aktuální provozní data. Starší údaje bývají určitou dobu archivovány, ale přístup k nim je komplikovaný a omezený.

- **Vytížení zdrojové databáze**

  Provozní systémy, které tvoří hlavní zdroj dat, bývají dostatečně vytížené vlastní činností. Jejich jediným úkolem je zajištění operativního fungování firmy. Zatěžování výpočetního výkonu prováděním složitých analýz a agregací by zbytečně omezovalo primární informační systém.

- **Struktura dat ve zdrojových databázích**

  Data podnikových aplikací bývají uloženy v entitě-relačních databázích, které umožňují rychlé provádění datových transakcí, zajišťují integritu dat, starají se o bezpečnostní přístup k datům. Tato struktura ale neposkytuje okamžitý přístup ke kumulovaným datům a to navíc na různých úrovních agregace (za podnik, za útvar, za skupinu zákazníků, za vybrané období apod.) a už vůbec nedovoluje pružně měnit kritéria pohledů.

- **Zdroje dat**

  Data se nenacházejí pouze na jednom místě, ale jsou roztříštěny pod různým označením v mnoha různých podnikových systémech a databázích. K jejich analýze se tak přistupuje napříč informačními systémy. Patří k nim jak interní aplikace, tak i různé externí databáze.

Pojetí datového skladu se ve srovnání s běžným podnikovým skladem značně liší. Zatímco materiály, součástky, polotovary nebo hotové výrobky, držené ve skladu, máme zájem co

nejrychleji vyexpedovat (a nikoli dlouhodobě skladovat), v datovém skladu chceme shromažďovat a uchovávat informační bohatství firmy po co nejdelší časové období. Spíše než ke klasickým skladům tak lze datové sklady přirovnat k muzejním depozitářům. I v nich se snažíme exponáty shromažďovat a časově, geograficky nebo jinak třídit. A stejně tak je tomu i v případě dat a datových skladů. [14]

Podle definice je datový sklad podnikový strukturovaný depozitář subjektově orientovaných, integrovaných, nepodléhající změnám, časově proměnlivých historických dat použitých na získání informací. V tomto složitém vyjádření znamená [14]:

- **Subjektová orientace**

  Údaje jsou do datového skladu zapisovány podle předmětu zájmu, ne podle aplikace, ve které byly vytvořené. To znamená, že se k sobě sbíhají například všechny dostupné informace o zákaznících, bez ohledu na to, ze které aplikace pocházejí.

- **Integrovatelnost**

  Datový sklad musí být jednotný. Údaje, týkající se jednoho předmětu zájmu, se do něj ukládají pouze jednou. Proto se musí zavést jednotná terminologie a vybrat vhodné jednotky veličin. To je velmi důležité, protože data, vstupující do datového skladu pocházejí z různých, vzájemně nesouvisejících systémů.

- **Časová variabilnost**

  Klíčovým atributem v datových skladech je čas. Spojují a vyhodnocují se v nich data, se stejnou časovou periodou (např. prodej produktů podle měsíců). Zatímco v provozních systémech jsou data platná v okamžiku přístupu a mohou se rychle měnit, data uložená v datových skladech jsou historická. V provozních databázích jsou držena data kratších časových období (max. měsíce), v datových skladech se jedná o údaje delších časových období (roky).

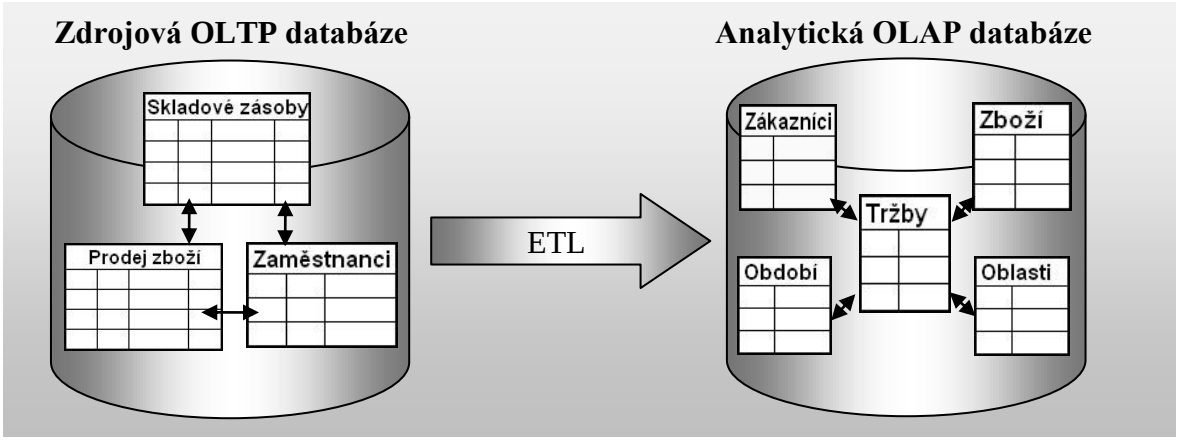
- **Neměnnost**

  U dat v provozních systémech dochází k průběžným změnám. Data do nich přibývají (insert), ubývají (delete) nebo se mění (update). Naproti tomu data v datových skladech se obvykle nemění ani neodstraňují, jen do nich v pravidelných intervalech přibývají. Provádějí se s nimi v podstatě jen dvě databázové operace –

vkládání (insert) a čtení (select). Z toho vyplývá, že metody pro optimalizaci a normalizaci dat a transakční přístup k nim nejsou v datovém skladě potřebné. (8)

4.1 Porovnání provozní relační databáze a datového skladu

Datový sklad bývá fyzicky i logicky oddělen od provozní databáze. U běžné relační databáze je obvyklá snaha o co nejmenší redundanci a maximální integritu uložených dat, které je dosahováno jejich normalizací a vnitřním provázáním jednotlivých logických funkčních celků. V datovém skladu, který je určen výhradně ke čtení, je naproti tomu řešení vždy vedeno snahou o jasnou vnitřní separaci jednotlivých funkčních celků a to i za cenu zvýšených nároků na paměťový prostor. Při popisu struktury datového skladu mluvíme o multidimenzionální (vícerozměrné) struktuře uložených dat. Provozní databáze bývají optimalizovány pro rychlé zpracování velkého množství malých transakcí a data se zpracovávají okamžitě při vzniku požadavku. Aktualizace datového skladu, tj. přidávání nových datových agregátů, probíhá obvykle periodicky jednou denně, jednou týdně, jednou měsíčně v závislosti na interních potřebách. Tyto akce je ale možné považovat za součást údržby datového skladu, která probíhá ve speciálním režimu při momentálním vyloučení zpracování OLAP požadavků uživatelů datového skladu. V běžném režimu práce (tzn. při provádění dotazů a analýz) není obsah datového skladu modifikován. Provozní aplikace nad relační databází řeší určitý okruh úloh nad „svými“ specifickými daty. V datovém skladu se naproti tomu shromažďují informace z mnoha různých zdrojů a seskupují se nikoliv podle původu, ale podle logického významu. Relační databáze je navrhována pomocí ERD (entitě-relační modelování), datový sklad je navrhován pomocí dimenzionálního modelování. [8]



Obr. 5 Srovnání struktur OLTP a OLAP databáze [15]

4.2 Typy datových skladů

Rozlišujeme dva základní typy datových skladů a dva typy pomocných skladů:

- **Základní datové sklady**
    - Datový sklad
    - Datové tržiště
- **Pomocné datové sklady**
    - Operativní datová úložiště
    - Dočasná datová úložiště

4.2.1 Datový sklad - Data Warehouse – DWH

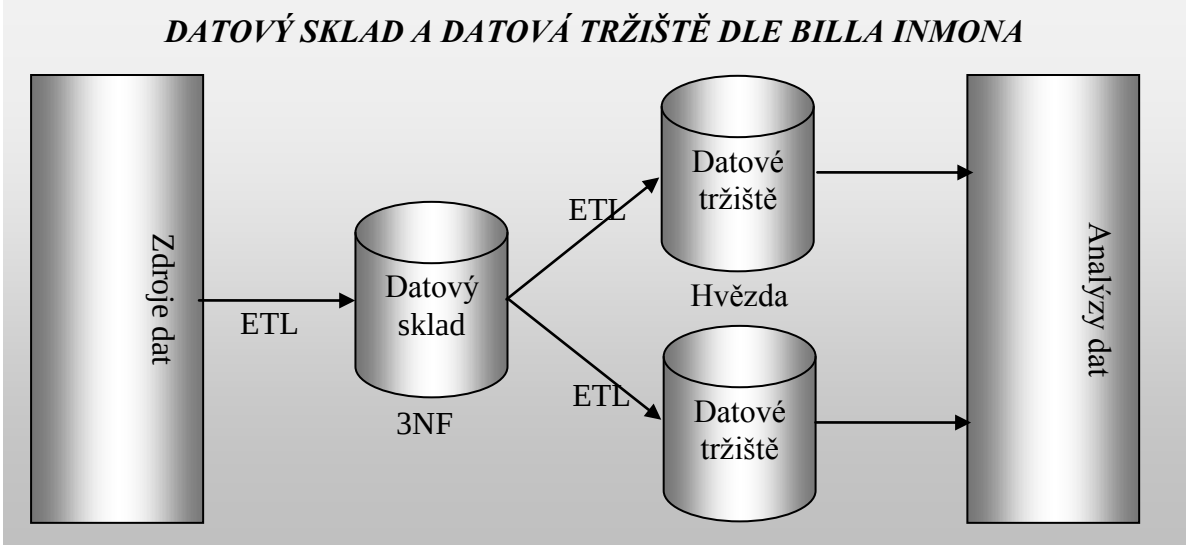
Datový sklad je rozsáhlá centrální podniková databáze, ve které jsou uložena transformovaná data pocházející z různých provozních systémů a externích databází. Tyto data jsou určené k dalším analýzám.

4.2.2 Datové tržiště - Data Marts – DMA

Princip datových tržišť je podobný jako princip datových skladů. Rozdíl je pouze v tom, že datové tržiště jsou decentralizované a tematicky orientované. Analytické informace, které poskytují, jsou zacílené na určitou skupinu uživatelů (marketing, prodej apod.). [15] Existují dvě koncepce budování datových skladů.

4.2.2.1 Koncepce Billa Inmona

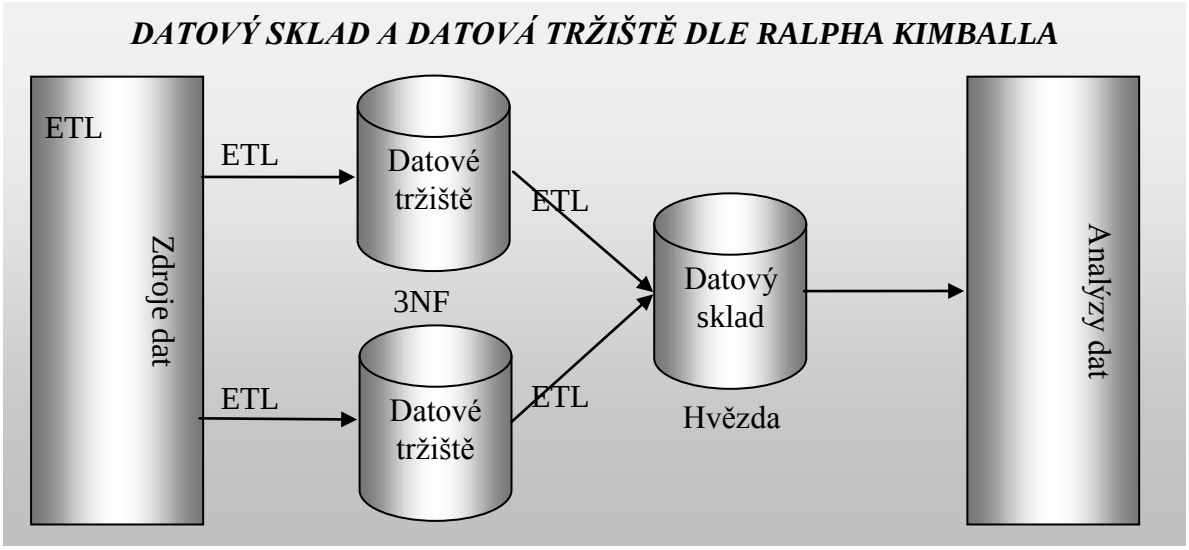
Při tomto konceptu se nejprve buduje centrální datový sklad a teprve z něj se vytvářejí problémově orientovaná datová tržiště. Toto řešení vyžaduje vyšší počáteční náklady na analýzu a dlouhou dobu na úplnou realizaci, na druhou stranu se ale jedná o nejčistější řešení s relativně jednoduchou správou. [8]



Obr. 6 Datový sklad a datová tržiště podle Billa Inmona

4.2.2.2 Koncepce Rapha Kimballa

Výrok Ralpha Kimballa: „Datový sklad není nic jiného než sjednocení datových tržišť“ charakterizuje jeho pojetí datových skladů. Podle něho vede cesta k centrálnímu skladu sjednocením tematicky orientovaných datových tržišť. Výhodou je postupné budování datového skladu a tím i pozvolné rozložení finančních nákladů a poměrně rychlé poskytování prvních analytických výstupů, nevýhodou je, že datová tržiště jsou budována jen na základě požadavků jednotlivých útvarů a do doby vytvoření centrálního datového skladu neexistuje celopodnikový pohled na data. Při centralizaci datových tržišť je navíc nutné ještě jednou podstoupit složitý a náročný proces ETL. [15]



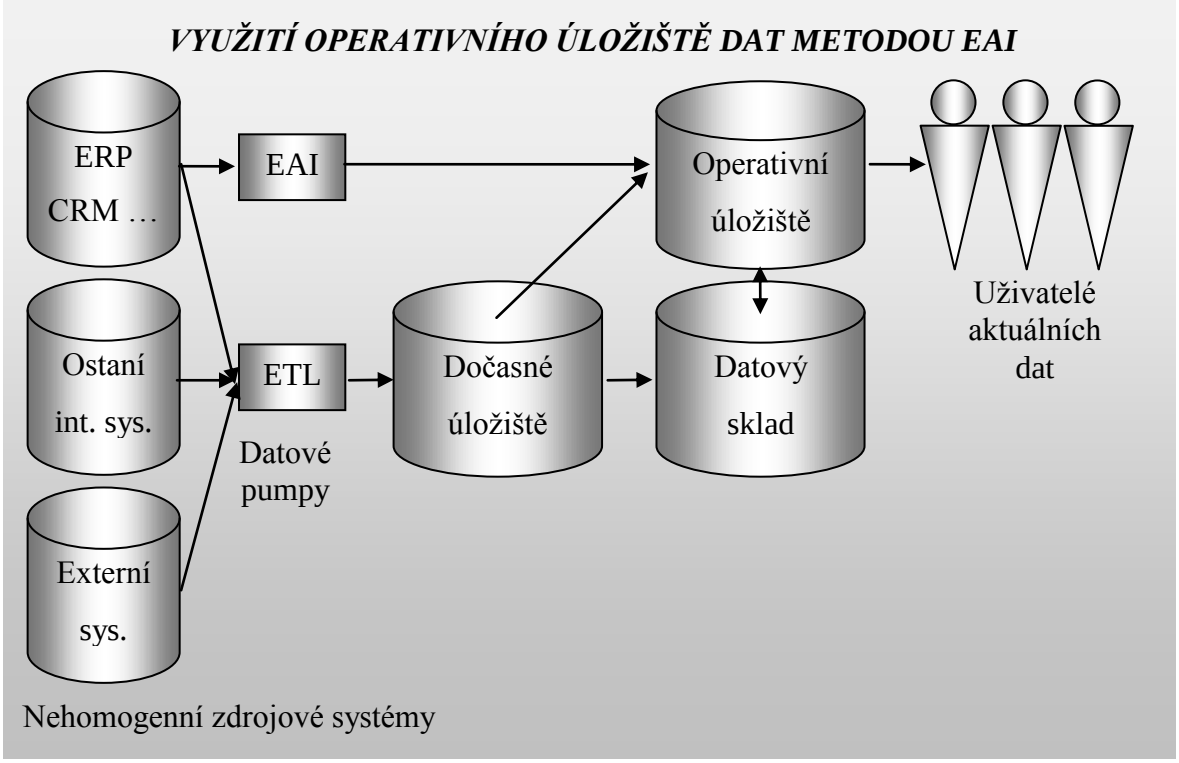
Obr. 7 Datový sklad a datová tržiště podle Ralpha Kimballa

4.2.3 Dočasné datové úložiště - Data Staging Areas – DSA

Používají se v procesu ETL pro dočasné uložení nekonzistentních neagregovaných netransformovaných dat, extrahovaných z produkčních systémů. Jedná se o nepovinnou komponentu, která nachází své hlavní uplatnění u vytížených produkčních systémů, kde je potřeba transferovat data s minimálním dopadem na jejich výkonnost nebo u systémů, jejichž data je potřeba před zpracováním konvertovat do databázového formátu. [14]

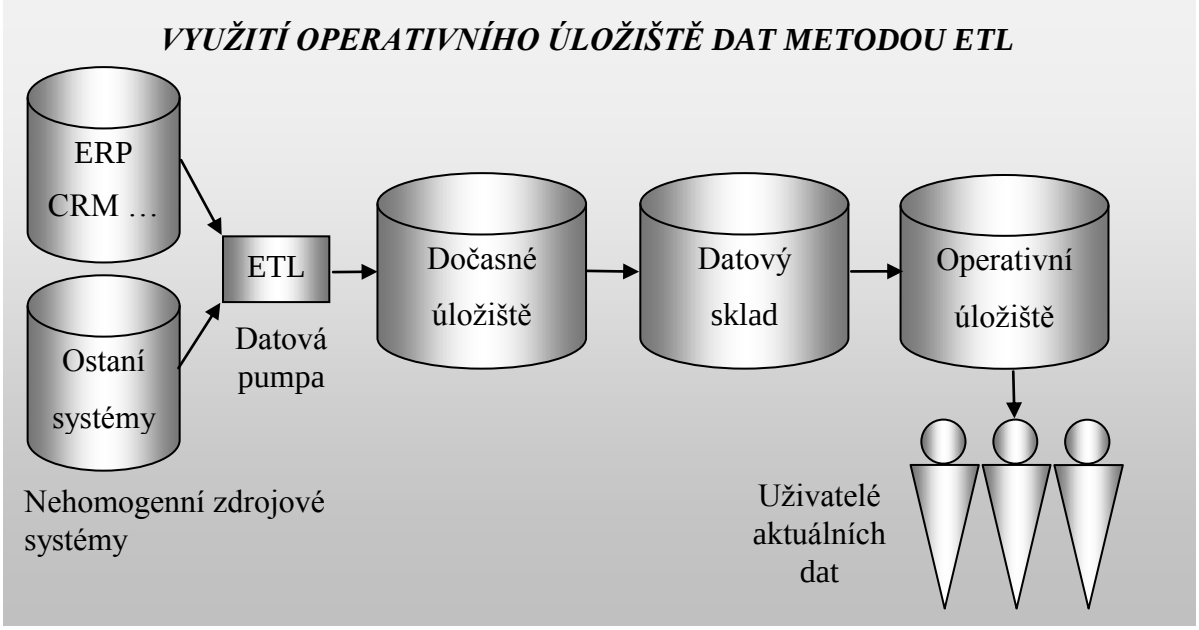
4.2.4 Operativní datové úložiště - Operational Data Store - ODS

Operativní datové úložiště je definováno jako jednotné místo datové integrace aktuálních dat z primárních systémů s minimální dobou odezvy po zpracování (tedy sledování téměř v reálném čase). Můžeme je nalézt např. v call centrech, kde je potřeba u každého zákazníka rychle znát jeho aktuální profil, aktivované nebo objednané produkty či zařazení do segmentu marketingových nabídek. Pro účely maximální „operativnosti“ bývá operativní úložiště napojeno na datové zdroje prostřednictvím EAI platforem. Ty umožňují vzájemnou komunikaci mezi libovolnými dvěmi aplikacemi v reálném čase a nevyžadují přitom jejich přímé propojení. Na rozdíl od ETL platforem, které zpracovávají události dávkově, EAI platformy reagují na jednotlivé události okamžitě. [12] [14]



Obr. 8 Využití operativního úložiště dat metodou EAI. Podle [15]

V některých případech slouží ODS jako mezisklad k provádění rychlých dotazů nad malým množstvím aktuálních analytických dat. V tomto případě nepracuje s provozními daty.



Obr. 9 Využití operativního úložiště dat metodou EAI. Podle [15]

4.3 Typy schémat datových skladů

Datové modely provozních systémů bývají velmi složité, protože obsahují mnoho tabulek a vazeb a stávají se tak pro běžného uživatele obtížně pochopitelné. Z toho důvodu se objevily snahy o zjednodušení ERD diagramů a jejich přizpůsobení potřebám datových skladů. Vznikly dva typy relačních dimenzionálních modelů, které určují strukturu datového skladu. Rozlišují se podle napojení dimenzí na tabulku faktů:

- Hvězdicové schéma (Star Schema)
- Schéma sněhové vločky (Snowflake Schema)

Základem každého schématu je jedna nebo více tabulek faktů, v nichž jsou uložena vlastní analyzovaná data (tj. sledované veličiny – hodnoty použité k agregovaným výpočtům). Tyto tabulky obsahují detailní údaje ze všech zdrojů a cizí klíče, pomocí nichž jsou spojeny s tabulkami dimenzí. Dimenzionální tabulky slouží k uložení popisných informací tabulek faktů (viz obrázky schémat). K samotným agregacím či filtrům nejsou tyto údaje v podstatě potřebné, těžko bychom ale dokázali analyzovat produkt nebo zákazníka podle jeho ID, navíc pokud jsou takových zákazníků či produktů stovky nebo tisíce. Dimenze určují

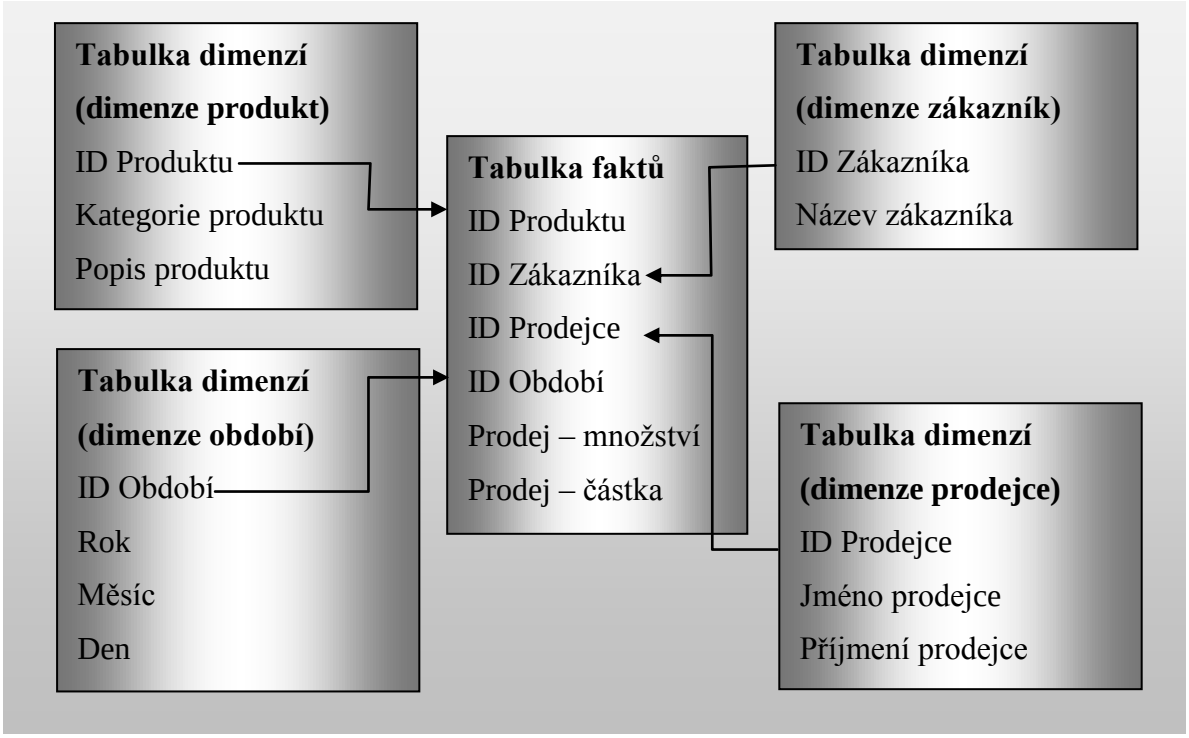
způsob pohledu (produkt, zákazník …) a hierarchii (skupina produktů, produkt …). S tabulkami faktů je spojen ještě termín granulita. Ta určuje úroveň podrobností v tabulce faktů. Čím nižší je úroveň granulity v tabulce faktů, tím detailnější jsou data, určená k provádění matematických operací. [8] [12]

V hvězdicovém schématu je každá dimenze reprezentována právě jednou dimenzionální tabulkou. Toto schéma poskytuje vysoký dotazovací výkon, je jednodušší, a proto také častější.

U schématu sněhové vločky jsou dimenzionální tabulky normalizovány. Dimenzionální tabulky se rozdělí podle hierarchických úrovní dimenzí do dalších tabulek. Omezí se sice redundance dat, ale z důvodu většího množství spojení mezi tabulkami se také sníží dotazovací výkon.

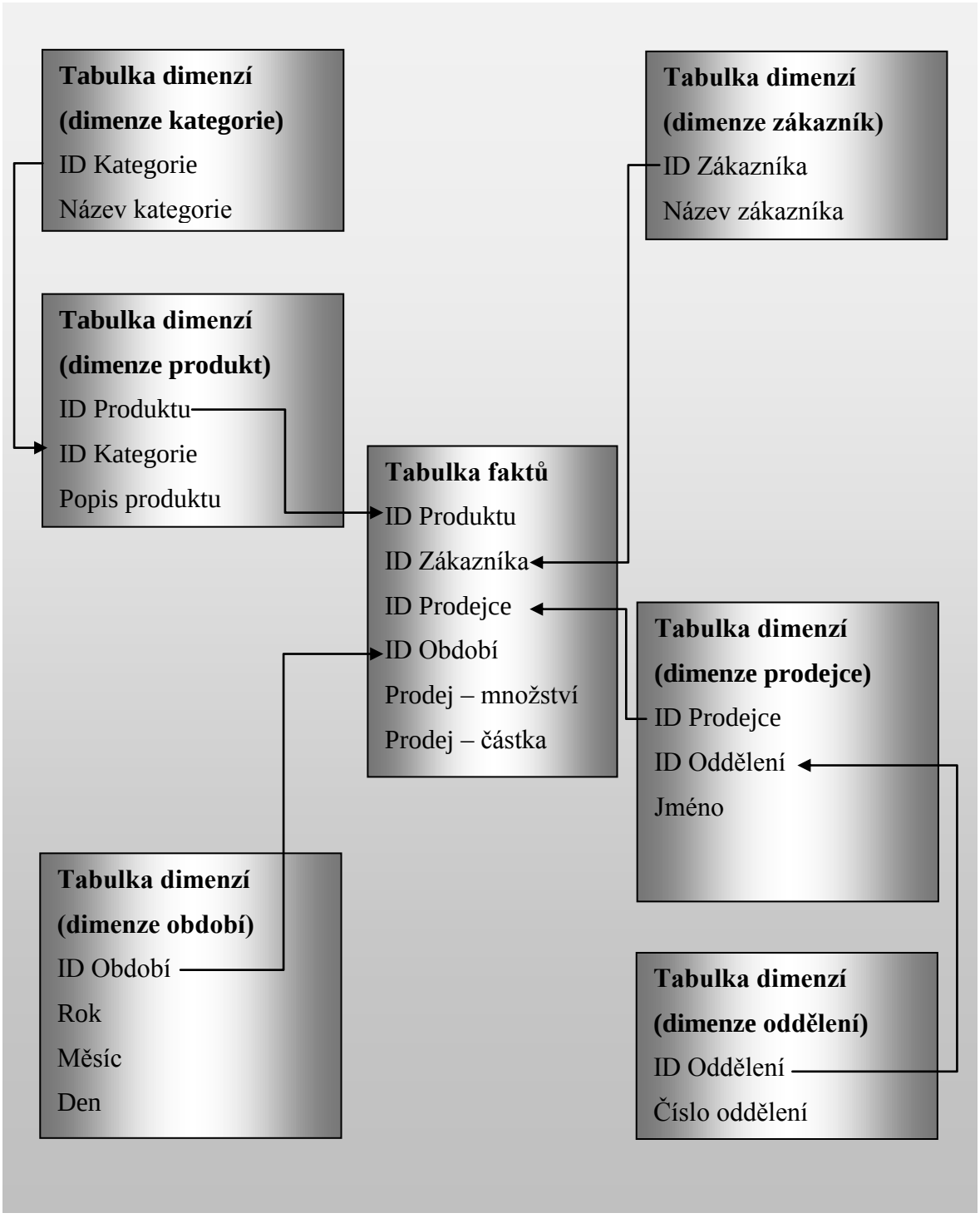
Schéma constellation (souhvězdí) je soubor schémat typu „hvězda“ (více tabulek faktů se sdílenými dimenzemi). [8]

4.3.1 Hvězdicové schéma



Obr. 10 Hvězdicové schéma datového skladu

4.3.2 Schéma sněhové vločky



Obr. 11 Schéma sněhové vločky

5 ANALYTICKÉ KOMPONENTY

5.1 Analýza vícerozměrných dat

Data v datovém skladu jsou sice vyčištěná a integrovaná, ale často také hodně objemná. Pro jejich analýzu se používají speciální datové struktury a technologie, které se označují jako OLAP (On-line Analytical Processing). K jednoduchým běžně dostupným a mnoha manažery velmi oblíbeným OLAP nástrojům, umožňujícím rychlé a pružné provádění vícerozměrných analýz, patří kontingenční tabulky MS Excel.

Termín OLAP poprvé zavedl dr. E.F.Codd k popsání technologie, která měla překlenout rozdíl mezi využitím osobních počítačů a řízením podnikových dat. Jeho definice zní:

„OLAP je volně definovaný řád principů, které poskytují dimenzionální rámec pro podporu rozhodování.“ [19]

Dr. Codd spolu se svými spolupracovníky definoval v roce 1993 dvanáct pravidel OLAP [14]:

- **Multidimenzionální konceptuální pohled**

  Systém by měl poskytovat multidimenzionální model odpovídající podnikatelským potřebám a měl by umožňovat intuitivní manipulaci a analýzu získaných údajů.

- **Transparentnost**

  Systém by měl být propojen na front-end systémy

- **Dostupnost**

  Systém by měl poskytovat pouze data potřebná k analýze. Uživatele nezajímá, jak systém k heterogenním zdrojům přistupuje.

- **Konzistentní výkon**

  Výkon systému nesmí záviset na počtu dimenzí. Ani při rostoucí velikosti databáze by se neměl výkon snížit.

- **Architektura klient-server**

  Systém OLAP musí být typu klient-server.

- **Generická dimenzionalita**

Každá dimenze údajů musí být ekvivalentní ve struktuře i operačních schopnostech.

- **Dynamické ošetření řídkých matic**

  Systém by měl být schopen adaptovat své fyzické schéma na analytický model, optimalizující ošetření řídkých matic.

- **Podpora pro více uživatelů**

  Systém by měl podporovat týmovou práci uživatelů a souběžné zpracování údajů.

- **Neomezené křížové dimenzionální operace**

  Systém musí dokázat rozeznat dimenzionální hierarchie a automaticky vykonat asociované kumulované kalkulace v rámci dimenzí i mezi nimi.

- **Intuitivní manipulace s údaji**

  Uživatelské rozhraní by mělo být intuitivní. Mělo by umožňovat rychlé zobrazení nebo skrytí detailů (drill-down, drill-up).

- **Flexibilní vykazování**

  Systém by měl umožňovat změnu uspořádání řádků a sloupců podle potřeb analýzy.

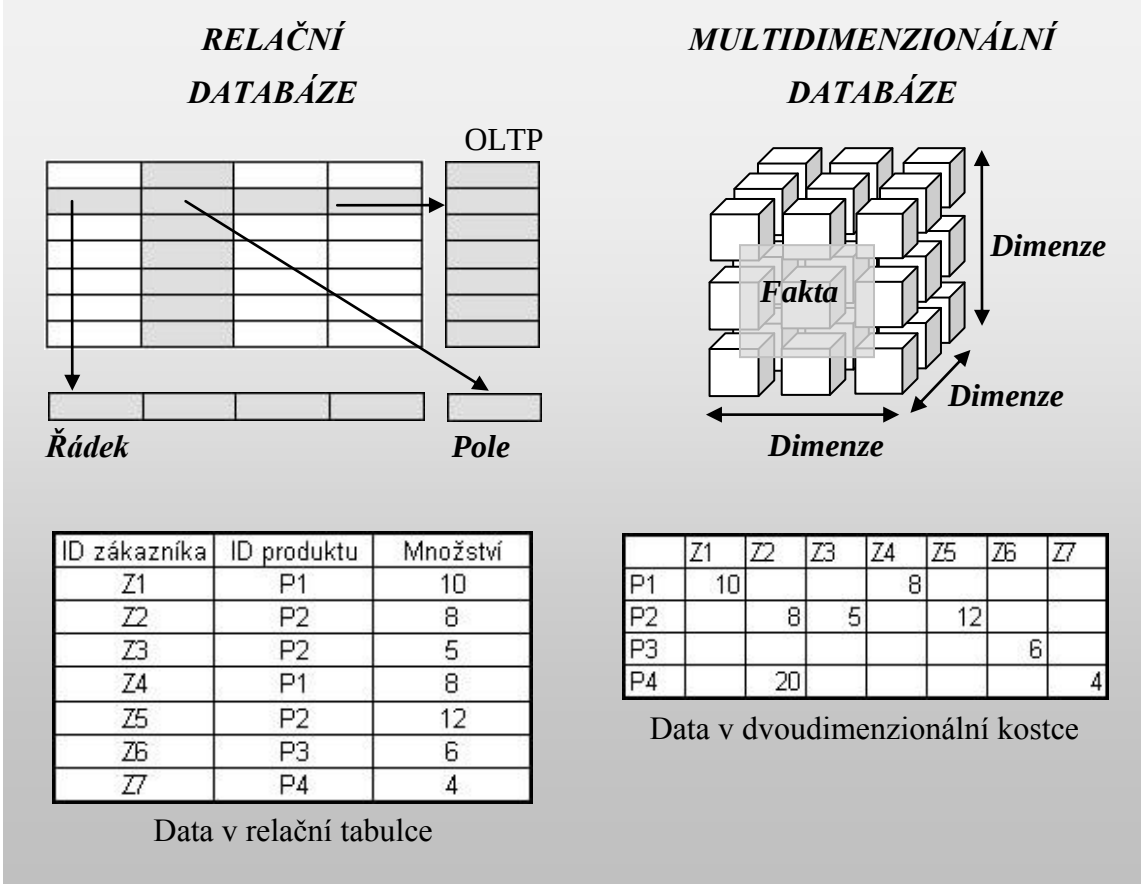
- **Neomezený počet dimenzí a úrovní agregace**

  Systém OLAP by neměl zavádět žádné umělé omezení počtu dimenzí nebo úrovní agregace. Měl by podporovat vícenásobné hierarchie. [14]

V roce 1995 bylo přijato praktičtější pojetí téhož - FASMI (Fast Analysis of Shared Multidimensional Information – rychlý, analytický, sdílený, mnohorozměrný, informační).

5.1.1 Popis OLAP technologie

OLAP technologie pracuje s tzv. multidimenzionálními daty. Na rozdíl od dvourozměrného uložení dat v relačních databázích (sloupec, řádek), je zde obdobou tabulky n-rozměrná datová kostka (Cube, Data Cube, OLAP). Můžeme si ji představit jako obyčejnou prostorovou kostku, která ale na rozdíl od geometrické kostky může mít více než tři rozměry (dimenze). Lze ji také přirovnat k vícerozměrnému poli, představitelnému každému programátorovi. Multidimenzionální databáze není normalizovaná, používány jsou převážně nenormalizované tabulky. Tvoří je tabulky dimenzí a faktů (measures) uspořádaných do schémat.



ID zákazníka	ID produktu	Množství
Z1	P1	10
Z2	P2	8
Z3	P2	5
Z4	P1	8
Z5	P2	12
Z6	P3	6
Z7	P4	4

Data v relační tabulce

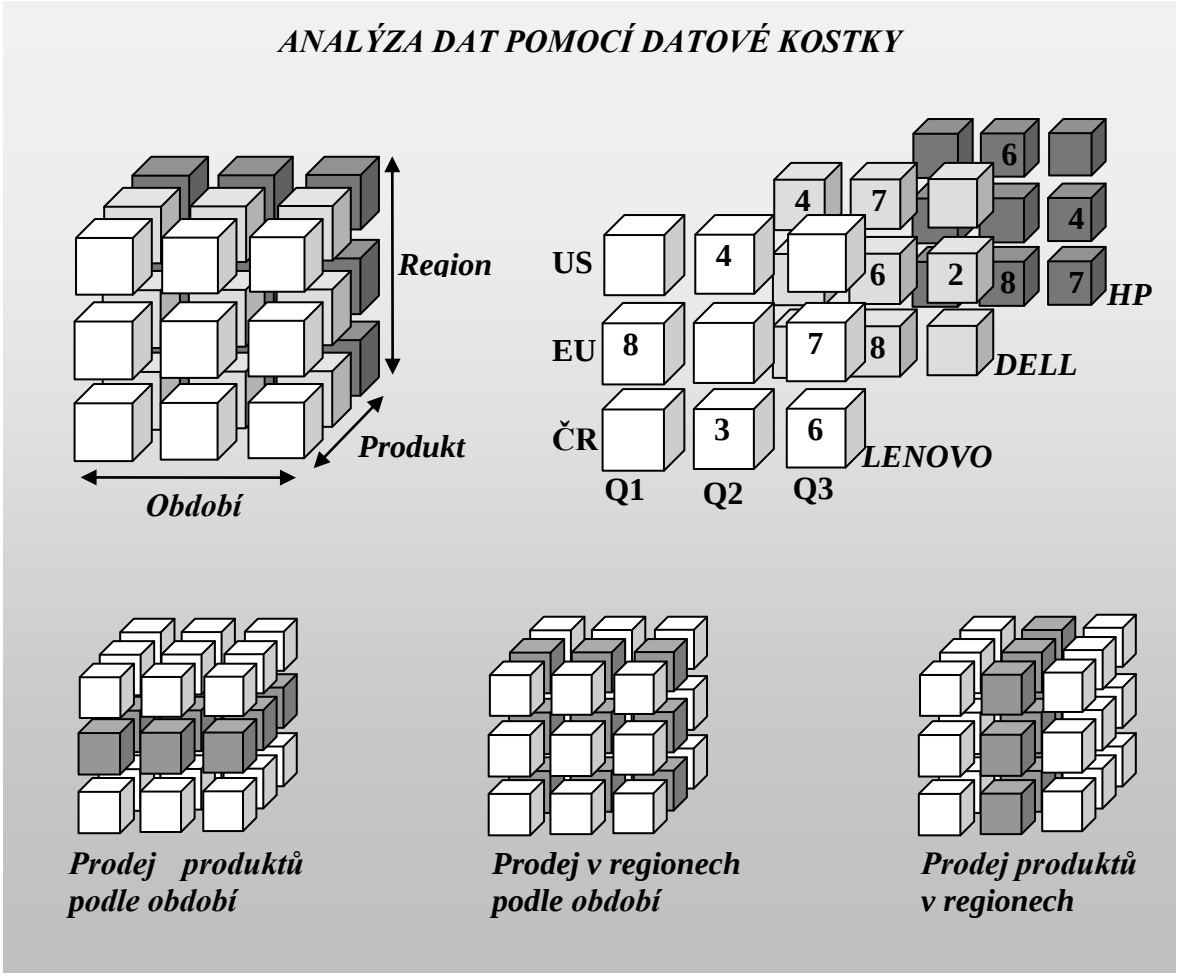
	Z1	Z2	Z3	Z4	Z5	Z6	Z7
P1	10			8			
P2		8	5		12		
P3						6	
P4		20					4

Data v dvoudimenzionální kostce

Obr. 12 Srovnání struktur relační a multidimenzionální databáze

Každá dimenze reprezentuje jiný úhel pohledu na data. Obsahuje jeden nebo více popisných atributů. Ty mohou být uspořádány nejenom logicky, ale také hierarchicky. Příkladem běžné dimenze je například dimenze geografická, ve které jsou uloženy dimenze zákazníků (může být hierarchicky členěná na země – regiony – města – adresy), dimenze časová (roky – měsíce - dny) nebo dimenze produktová (kategorie produktu - produkt). Z hlediska realizace dimenze v hierarchii existuje dimenze s implicitní hierarchií a dimenze s explicitní hierarchií. Implicitní hierarchie je vyjádřená zařazením potřebných atributů přímo v tabulce dimenze (zákazník – město - kraj). Explicitní hierarchie vytváří pro danou dimenzi řetěz tabulek (zákazník – ID_města – ID_kraje). V ekonomických aplikacích bývá přítomna jedna ekonomická dimenze a čas. Ostatní dimenze se navrhují s ohledem na požadavky uživatelů. Číselné údaje, pocházející z obchodních nebo jiných ekonomických činností podniku (množství prodaného zboží, zisky, tržby, náklady …), se nacházejí v tabulce faktů. Sledovanými vlastnostmi faktů jsou granulita, která určuje úroveň podrobností faktů (vysoká granulita znamená uložení dat v nízkém stupni agregace - velkým detailem dat) a aditiva, která určuje, zda je možné fakta sumarizovat podle

dimenzí. Atributy, do kterých se ukládají fakta, mohou být aditivní (lze agregovat podle všech dimezí), semiaditivní (lze agregovat jenom podle některých dimenzí) a neaditivní. Dimenze i fakta se do datového skladu, který je zdrojem analytických informací, dostávají z externích zdrojů. V datové kostce se ale nacházejí také různé typy agregací (průměry, počty, součty …). K jejich výpočtu dochází až v datovém skladu. [16]



Obr. 13 Analýza dat pomocí datové krychle

5.1.2 Fyzická realizace multidimenzionálního datového modelu

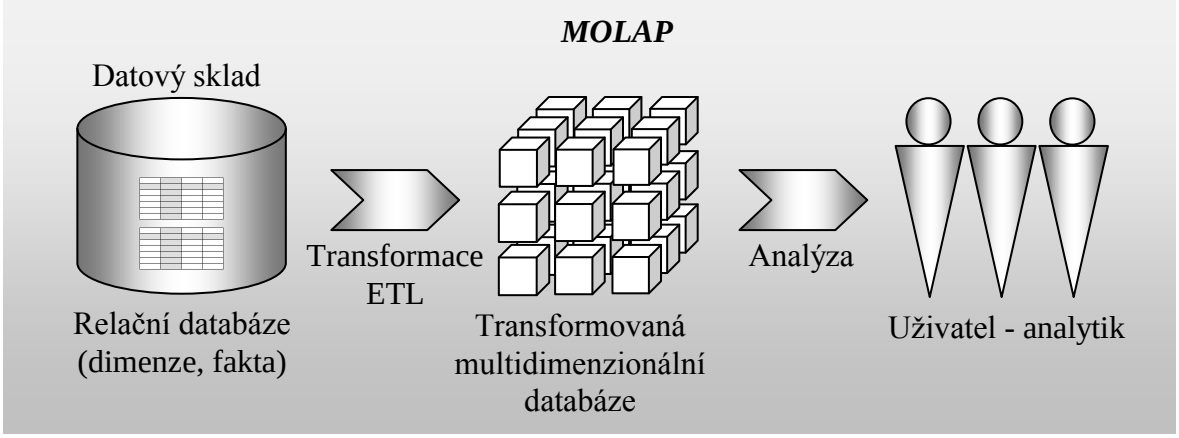
5.1.2.1 MOLAP – Multidimenzionální OLAP

Logická struktura – multidimenzionální datový model (hvězda / vločka)

Fyzické řešení – multidimenzionální databáze

MOLAP vyžaduje ke své práci vedle vlastního datového skladu ještě speciální multidimenzionální databázi. Ta je v pravidelných intervalech aktualizována daty

z datového skladu. [17] V datovém skladu jsou data uložena v relační databázi v tabulkách dimenzí a faktů, v multidimenzionální databázi jsou uložena nejenom data z datových skladů, ale také všechny možné agregace, které se vypočítají při aktualizaci datové krychle. Nevýhodou tohoto uložení je vysoká redundance dat (data jsou uložena na dvou místech) a nutnost transformace z RDB do MDB, výhodou je rychlost přístupu k potřebným datům. Je vhodný pro malé až středně velké objemy dat, kdy kopírování všech dat do multidimenzionálního formátu nevyžaduje výrazně dlouhou dobu ani nespotřebovává příliš mnoho diskového prostoru.



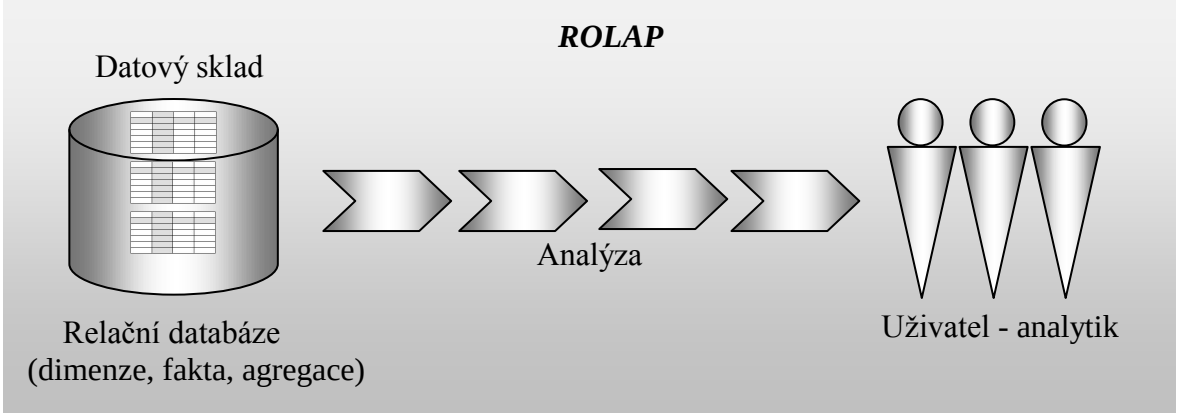
Obr. 14 MOLAP

5.1.2.2 ROLAP – Relační databázový OLAP

Logická struktura – multidimenzionální datový model (hvězda / vločka)

Fyzické řešení – relační databáze

ROLAP odráží multidimenzionální souvislosti, aniž by k uložení dat používal multidimenzionální strukturu. Pracuje nad relační databází datového skladu nebo datového tržiště. Multidimenzionální dotazy automaticky překládá na odpovídající SQL příkazy SELECT. Databáze obsahuje jak tabulky dimenzí a faktů, tak také oddělenou sadu pomocných tabulek, sloužících k ukládání agregací. Výhodu tohoto typu je okamžitý přístup k aktuálním analytickým datům, nevýhodou pomalejší odezva ve srovnání s MOLAP. Je vhodný i pro rozsáhlé databáze.



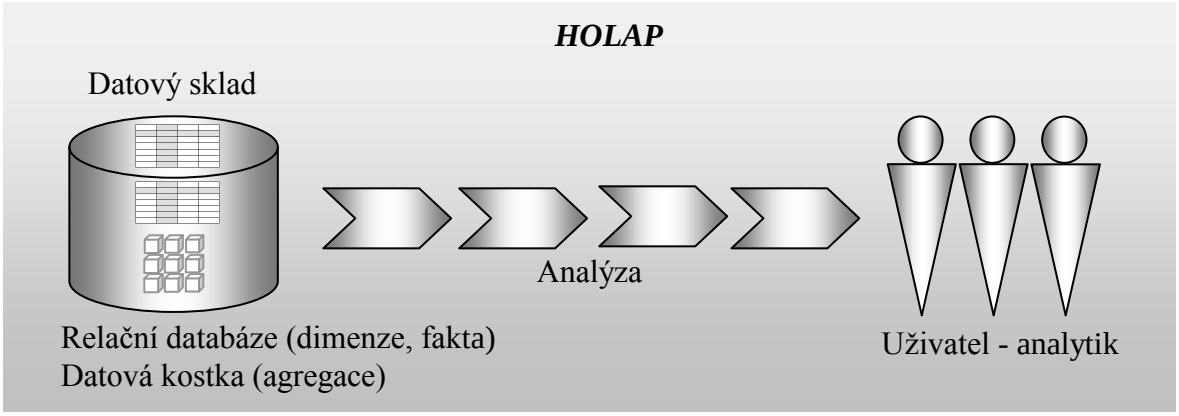
Obr. 15 ROLAP

5.1.2.3 HOLAP – Hybridní OLAP

Logická struktura – multidimenzionální datový model

Fyzické řešení – relační databáze s multidimenzionálními agregacemi

HOLAP je specifickým případem kombinace obou přístupů. Analýza dat probíhá nad relační databází, ale sumarizované hodnoty, se kterými se pracuje nejčastěji, jsou uloženy v multidimenzionální struktuře v datovém skladu, což zajišťuje rychlý přístup k těmto datům. Údaje nejnižší úrovně zůstávají uložené v relační databázi. Oddělení nejnižší úrovně od ostatních zajišťuje rychlejší přístup k nejčastěji používaným údajům. Výhodou tohoto řešení je rozsáhlý přístup k datům při současně rychlé agregaci, nevýhodou je nutnost udržování dat na dvou místech. [18]



Obr. 16 HOLAP

5.1.2.4 DOLAP – Dynamický OLAP

DOLAP je speciálním typem OLAP, při kterém je multidimenzionální matice (kostka) budována virtuálně v RAM paměti. Z toho plane základní výhoda tohoto řešení, kterým je neomezená flexibilita. Nevýhodou jsou vysoké nároky na RAM paměť a nutnost budovat kostku pokaždé znovu. [18]

5.1.3 Operace s daty v OLAP analýze

Typicky se s multidimenzionálními daty provádí tyto základní operace:

- **Zanoření - Drill-down**

  Posun v hierarchii dimenze směrem k nižší úrovni. – detailnější pohled na data.

- **Vynoření - Roll-up**

  Posun v hierarchii dimenze směrem k vyšší úrovni. – obecnější detailnější pohled na data.

- **Zanoření na nejnižší úroveň – Drill-across**

  Posun na nejnižší úroveň – zobrazení dostupných neagregovaných údajů.

- **Rotace – Pivot / Rotate**

  Změna os datové kostky – změna úhlu pohledu. V kontingenční tabulce to znamená otočení dimenzí. Dimenze umístěné do sloupců se přemístí do řádků a naopak.

- **Projekce - Slice**

  Řez datovou kostkou a aplikace filtru.

- **Selekce - Dice**

  Filtr pro více dimenzí. Vybere pouze ty údaje, které splňují zadanou podmínku. Úroveň dimenzí ani jejich počet se nemění.

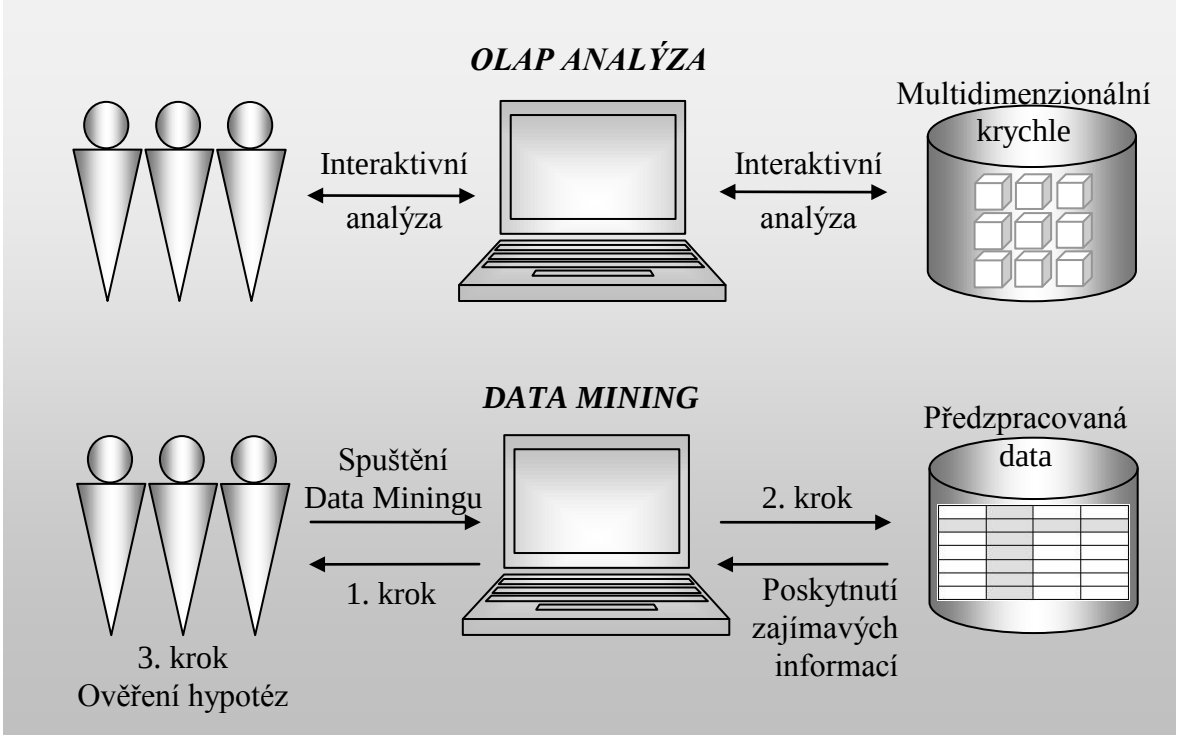
5.2 Dobývání znalostí z dat

Dolování dat (Data Mining) je proces hledání informací a skrytých, předem neznámých, souvislostí a vztahů ve velkých objemech dat. Usama Fayyad, jeden z nejuznávanějších expertů na tuto problematiku, definoval data mining takto:

„Data Mining je proces výběru, hledávání a modelování ve velkých objemech dat, sloužící k odhalení dříve neznámých vztahů mezi daty za účelem získání obchodní výhody.“ [20]

Rozvoj a postupné rozšiřování této analytické metody souvisí s obrovským nárustem dat, uložených v podnikových databázích. Přímou úměrou k jejich objemu roste ale i množství chyb v datech a obtížnost získávání smysluplných informací. [21]

Na rozdíl od OLAP analýzy, při které se používá deduktivní způsob práce a hledání odpovědí na předem známé otázky, v případě Data Miningu se jedná o induktivní přístup, při němž se teprve na základě skutečných dat vytvářejí možné hypotézy, které je pak potřeba na vybraných vzorcích ověřit a na základě výsledků průzkumů buď přijmout nebo nepřijmout. OLAP i Data Mining spojuje práce s historickými daty. Ale zatímco OLAP analýza pracuje se sumačními daty a interaktivně ji řídí analytik, Data Mining pracuje s detailními záznamy a je řízený systémem.

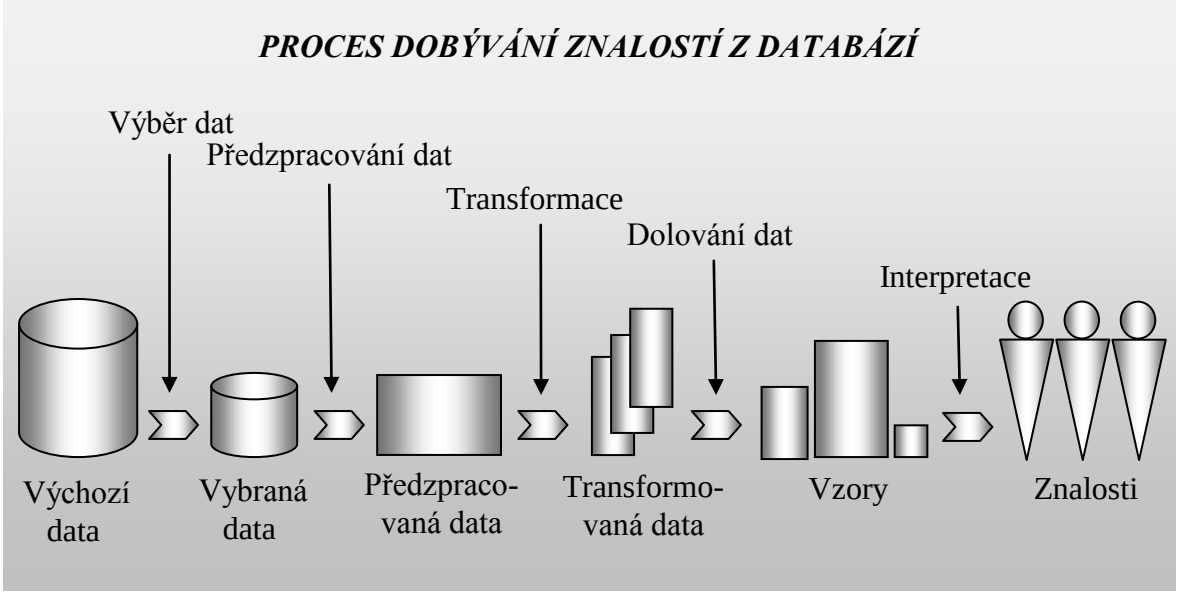


Obr. 17 Srovnání OLAP analýzy a Data Miningu

Data Mining vznikl spojením databázových a statistických disciplín, které se po léta vyvíjely nezávisle na sobě. Databázové technologie představují osvědčený prostředek k tomu, jak uchovávat rozsáhlá data a vyhledávat v nich informace, data představují obrovský potenciál skrytých informací, využitelných při řízení podniku. Statistika poskytuje nástroje k modelování a analyzování závislostí. Data Mining využívá různé složité algoritmy, pomocí nichž je možné predikovat vývoj nebo segmentovat či shlukovat příbuzné údaje. Z hlediska matematické a statistické teorie je založený hlavně na hledání korelací a testování hypotéz. [14]

5.2.1 Proces dobývání znalostí

Pro dolování dat je velice důležitá kvalita vstupních dat. Pokud v datech není podchycen některý důležitý údaj, může být výsledek analýzy nesprávný (například analýza denního prodeje nápojů bez zahrnutí vnější teploty apod). Proto se klade důraz už na přípravu dat, určených pro analýzu. Při přípravě se obvykle z datového skladu, obsahujícího relevantní údaje, vytváří jedna tabulka, které obsahuje předzpracované a očištěné data. Při interpretaci se nalezené znalosti obvykle hodnotí z pohledu koncového uživatele. [23]



Obr. 18 Proces dobývání znalostí z databází [23]

- **Stanovení cílů**

  Impulsem k zahájení procesu dobývání znalostí je nějaký reálný problém. Na jeho konci by mělo být získání co největšího množství informací, vhodných k řešení

daného problému. Asi největšího uplatnění nachází Data Mining v oblasti marketingu.

- **Výběr dat**

  V této fázi je nutné vytipovat data pro data Mining a to jak z hlediska zaměření (demografická, behaviorální, psychografická získaná průzkumem veřejného mínění) tak podle zdrojových databází. Data jsou obvykle extrahována ze zdrojových systémů na zvláštní server.

- **Předzpracování dat**

  Příprava dat je nejnáročnější a nejkritičtější fází procesu. Závisí na ní výsledek analýzy. Z objemných databází je nutné vybrat odpovídající informace tak, aby mohly být uloženy do jednoduché tabulky. [12]

  „Výsledné modely jsou tak dobré, jak dobré jsou data, použité na jejich vytvoření.“

  Předzpracování dat se skládá z těchto kroků:

  - Čištění dat

    Je nutné řešit problém chybějících, neúplných hodnot nebo nekonzistentních hodnot. Neúplné hodnoty je možné zanedbat nebo doplnit průměrnou hodnotou či konstantou „unknown“. Chybějící hodnoty mohou být vyhlazeny podle sousedních hodnot (binding), mohou být přiřazeny podle odpovídající skupiny apod.

  - Integrace dat

    Při integraci dat z různých zdrojů do jedné databáze je nutné řešit redundanci dat, různé názvosloví, různé vyjádření hodnoty apod.

  - Transformace dat

    Data se musí transformovat do formátu vhodného pro dolování dat. Data se sumují a zobecňují, přidávají se nové odvozené atributy, numerické hodnoty v intervalech se diskreditují.

  - Redukce dat

    Odstraňují se nadbytečné a nepoužívané atributy, komprimují se a nahrazují alternativní menší reprezentací dat.

- **Dolování dat**

  V této fázi se na předzpracovaná data aplikují vybrané algoritmy a vytvářejí se matematické modely. Tato fáze je nejkratší a nejjednodušší. Po ní už následuje využití modelů a převedení jejich výsledků do konkrétních podnikových úkolů a plánů.

5.2.2 Modely dolování dat

V případě dobývání znalostí existuje několik typových úloh [12]:

- **Explorační analýza dat**

  Nezávislé zkoumání dat bez předchozích znalostí, které by mohly hledání ovlivnit.

- **Deskripce**

  Popisuje celou datovou množinu. Podle projevů chování se vytvářejí skupiny, do kterých se dají projevy v datech rozdělit.

- **Predikce**

  Snaží se předpovídat hodnotu určité veličiny na základě znalostí hodnot ostatních veličin. Z hlediska statistiky je takovou metodou regresivní analýza.

- **Hledání vzorů a pravidel (nugettů)**

  Podstatou je hledání vztahů a vzorů chování. Ke klasickým úlohám patří analýza nákupního košíku, rozkrývající druhy zboží, které zákazníci nakupují současně.

- **Hledání podle vzorů**

  Při řešení těchto úkolů má analytik k dispozici určité vzory a jeho cílem je nalézt takové nebo podobné vzory i v datech.

5.2.3 Metodiky dolování dat

S rozvojem technologie začaly vznikat metodiky, které si kladou za cíl poskytnutí jednotného rámce pro řešení různých úloh z oblasti Data Miningu. Tyto metodiky umožňují sdílet a přenášet zkušenosti z úspěšných projektů. Definují vhodné postupy a fáze procesu získávání znalostí [24]:

- **Metodika „5A“**
    - Assess – posouzení potřeb projektu, stanovení cílů a strategií
    - Access – sběr a příprava dat
    - Analyze – provádění analýz a hledání odpovědí na otázky z prvního bodu
    - Akt – přeměna získaných znalostí na akční znalosti
    - Automate – převedení výsledků analýzy do praxe
- **Metodika SEMMA**
    - Sample – vybírání vhodných objektů
    - Explore – vizuální explorace a redukce dat
    - Modify – seskupování objektů a hodnot atributů, datové transformace
    - Model – analýza dat
    - Assess – porovnání modelů a interpretace
- **CRISP-DM**
    - Business understanding – porozumění problematice, formulování úlohy
    - Data understanding – výběr dat
    - Data preparation – příprava dat
    - Evaluation – porozumění výsledkům
    - Deployment – využití výsledků

5.2.4 Metody dolování dat

- regresní metody (lineární regresní analýza, nelineární regresní analýza, neuronové sítě)
- klasifikace (diskriminační analýza, logistická regresní analýza, rozhodovací stromy, neuronové sítě),
- segmentace – shlukování (shluková analýza, genetické algoritmy, neuronové shlukování – Kohonenovy mapy)

- analýza vztahů (asociační algoritmus pro odvozování pravidel typu „ if X then Y“)
- predikce v časových řadách (Boxova-Jenkinsonova metoda, neuronové sítě, autoregresní modely, ARIMA)
- detekce odchylek [24]

II. PRAKTICKÁ ČÁST

6 ANALYTICKÉ NÁSTROJE MS SQL SERVERU 2008

Od počátku služby OLAP v MS SQL Serveru 7.0 se Microsoft postupně snažil o model samoobslužných analytických nástrojů. Myšlenkou bylo, aby uživatelé sami mohli vytvářet analýzy, aniž by při tom museli projít různými vrstvami správy databáze. A tak se postupně vytvářely a zdokonalovaly nástroje, které toto chování podporují. Ještě ve verzi MS SQL Server 2000 byly mezi jednotlivými analytickými nástroji poměrně striktní hranice. To způsobovalo velké množství redundantních dat (redundantních z pohledu datových modelů, nikoli z pohledu obsahu datových skladů). Ve verzi MS SQL Server 2005 byly všechny analytické vrstvy sjednocené do jednotného Unified Dimesional Model. UDM vychází z Data Driven modelu, který v současné době patří k nejpreferovanějším trendům vývoje podnikových projektů. Podle něho je budování projektu zahájeno fází modelování a teprve na základě vytvořeného modelu se vygeneruje struktura datového skladu a vytvoří se úlohy pro jeho naplnění. Budování projektu pokračuje návrhem měřítek, dimenzí, kostek a reportů. V MS SQL Serveru 2008 jsou jádrem analytických nástrojů Analysis Services SQL Serveru 2008. Jejich součástí jsou technologie OLAP, dolování dat a patří k nim také služby Reporting Services a Integration Services. [14] [28]

- **Integration Services**

  Služba SQL Server Integration Services (SSIS) plní funkci datové pumpy ETL (Extrakt, Transform, Load). Kromě toho ale dovoluje také vytvářet aplikace pro správu databáze a systémových prostředků, umožňuje manipulovat se soubory v adresářích, importovat a exportovat data. Ve verzi MS SQL Server 2005 nahradila službu MS SQL Serveru 2000 Data Transformation Services (DTS).

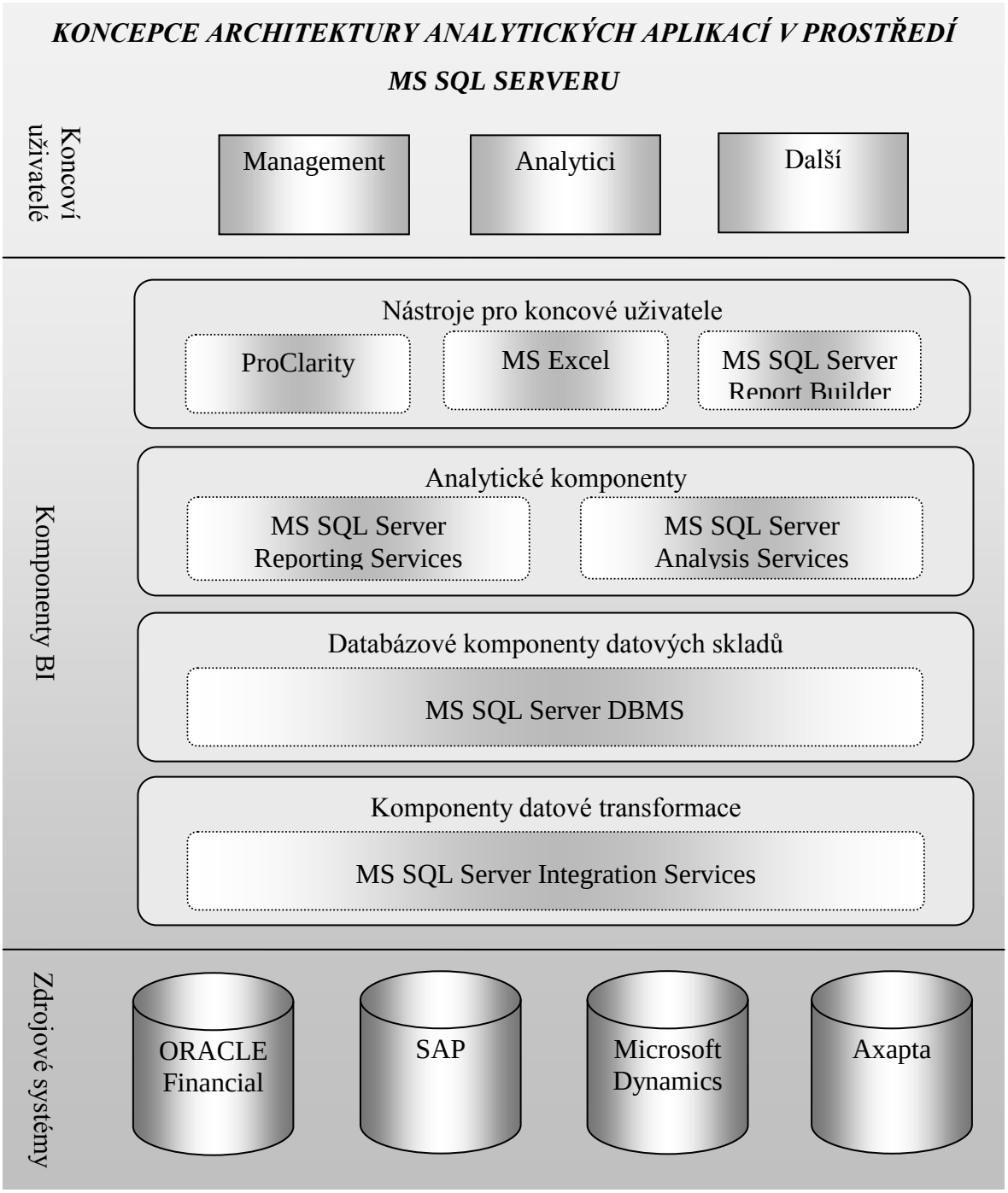
- **Reporting Services**

  Služba SQL Server Reporting Services (SSRS) poskytuje pružnou platformu pro tvorbu a distribuci sestav. Spolupracuje s klientským nástrojem MS SQL Server Report Builder, který je jako doplněk MS SQL Serveru koncovým uživatelům k dispozici zcela zdarma.

- **Analysis Services**

  Služba SQL Server Analysis Services (SSAS) je klíčovou komponentou analýz podnikových dat. Obsahuje dvě komponenty:

- Modul OLAP pro analýzu vícerozměrných dat umožňuje zavádění, dotazování a správu datových krychlí, vytvořených v Business Inteligence Development Studio (BIDS)
- Modul Data Mining rozšiřuje možnosti podnikových analýz o hledání vzorů a predikci vývoje.



Obr. 19 Koncepce architektury analytických aplikací v MS SQL Serveru

7 UŽIVATELSKÉ NÁSTROJE ANALÝZY DAT

7.1 Analýza dat pomocí MS Excel

Nejjednodušší a nejdostupnější způsob analýzy podnikových dat nabízí program MS Excel. Jedná se určitě také o způsob nejlevnější, protože myslím, že bez ohledu na provozovaný informační systém, neexistuje dnes manažer nebo jiný vedoucí pracovník, který by neměl tento program na svém notebooku nebo PC nainstalovaný a který by s ním neuměl pracovat. Není proto nutné utrácet peníze za nákup licencí specializovaného software nebo je investovat je do produktových školení. Uživatelé mohou začít tvořit analytické sestavy a grafy ihned. Analýzy dat pomocí MS Excel jsou velmi dynamické a výkonné, umožňují mnoho různých pohledů i grafických reprezentací. Data do MS Exel jsou získávána několika způsoby. Nejčastěji se jedná o manuální vyplňování tabulek vlastními daty uživatelů tak, jak je získávají z podnikových sestav nebo jak si je pořizují během výkonu své práce. Druhý méně pracný způsob představuje import dat z podnikových informačních systémů. Data z IS jsou nejprve exportována do souborů s oddělovači a z nich jsou následně nahrána do MS Excel. Třetí, nejoperativnější způsob představuje přímé napojení na databázi podnikového informačního systému.

7.1.1 Analýza dat pomocí kontingenčních tabulek a grafů

Kontingenční tabulky představují jeden z nejmocnějších nástrojů MS Excel. Díky nim lze data snadno sumarizovat, filtrovat a třídit. Ze stejných zdrojových dat je možné vytvořit mnoho různých pohledů, sestav a grafů. Data přitom mohou pocházet z jiného listu, databáze, datové krychle, textového souboru nebo jiného zdroje dat. Vytvořenou kontingenční tabulku mohou uživatelé jednoduchým přetahováním polí snadno měnit, přidávat do ní nebo ubírat z ní data, sloupce, řádky nebo měnit typy souhrnů a výpočtů bez toho, aby ovlivnili zdrojové data. Kontingenční tabulky jsou výkonným nástrojem vícerozměrné analýzy, velmi často se používají také jako klientský nástroj pro přístup k datovým krychlím MS SQL Serveru.

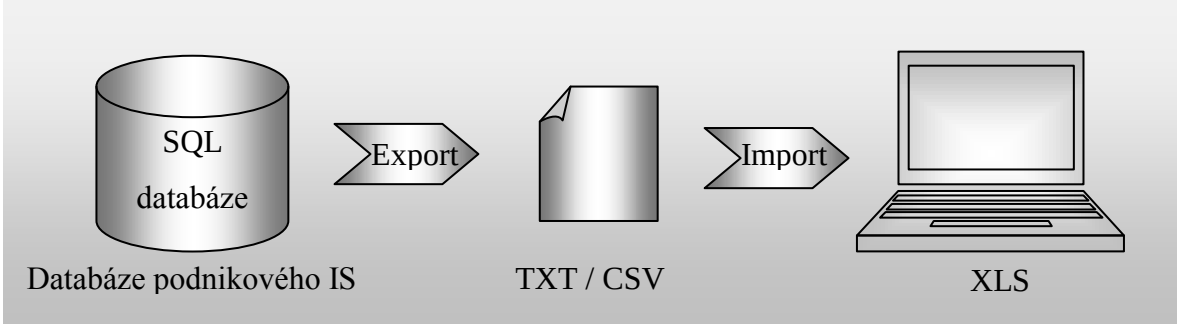
7.1.1.1 Ručně zadávaná data

Analýza dat pomocí kontingenčních tabulek z ručně zadávaných dat představuje patrně nejběžnější a nejrozšířenější způsob analýzy podnikových dat. Je dostupná všem

uživatelům bez ohledu na jejich přístup k datům podnikového systému. Nevyžaduje od nich znalosti jazyka SQL ani jiné programátorské dovednosti. Uživatelé, kteří tímto způsobem analyzují data, svá zdrojová data dobře znají a již při pořizování dat vědí, jaké výstupy je zajímají. To jim tvorbu kontingenčních tabulek značně usnadňuje. Zdrojové tabulky je možné kdykoliv rozšířit o další sloupce, takže uživatelé nejsou nijak svazováni ani strukturou dat. Na druhou stranu se kvůli pracnému pořizování dat jedná o způsob nejméně efektivní.

7.1.1.2 Importovaná data

Součásti mnoha informačních systémů jsou nástroje pro export dat. Často je možné podnikové tabulkové data exportovat přímo do xls formátu nebo alespoň do textového souboru, který je možné následně do MS Excel importovat a tam s ním dále pracovat. Pole v textovém souboru mohou být odděleny speciálním oddělovačem (čárka, středník, tabelátor) nebo mohou mít pevnou šířku sloupců, definovanou relativní pozicí v řádku.

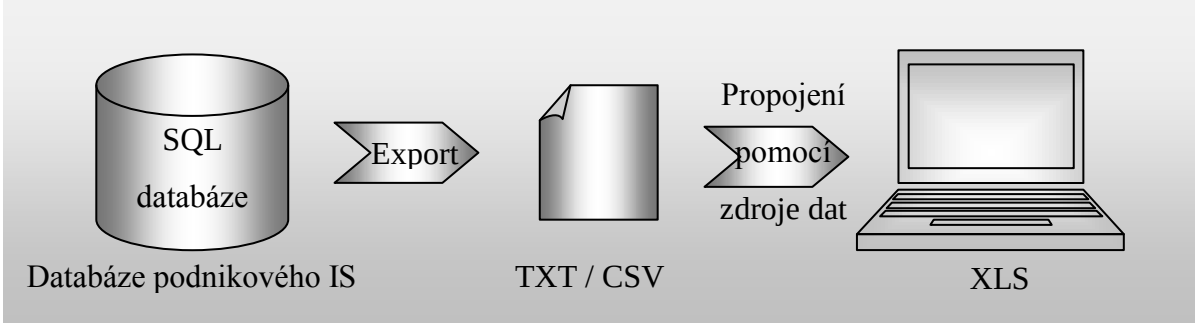


Obr. 20 Proces importu dat z datového souboru do MS Excel

7.1.1.3 Datové soubory připojené pomocí zdroje dat

Pokud jsou datové soubory z podnikových systémů pravidelně exportovány a aktualizovány, nemusíme jejich import do MS Excel provádět pokaždé ručně. Stačí když vytvoříme propojení se zdrojem dat. Tak budeme přistupovat vždy k aktuálnímu souboru bez nutnosti opakovaných importů. Když v MS Excel vytváříme propojení s datovým souborem, vznikne nám v adresáři datového souboru soubor Schema.ini a v adresáři C:\Program Files\Common Files\ODBC\Data Source soubor jmeno_zdroje_dat.dsn. Soubor Schema.ini obsahuje informace o struktuře importovaných dat, soubor jmeno_zdroje_dat.dsn obsahuje spojovací informace. Oba soubory můžeme otevřít

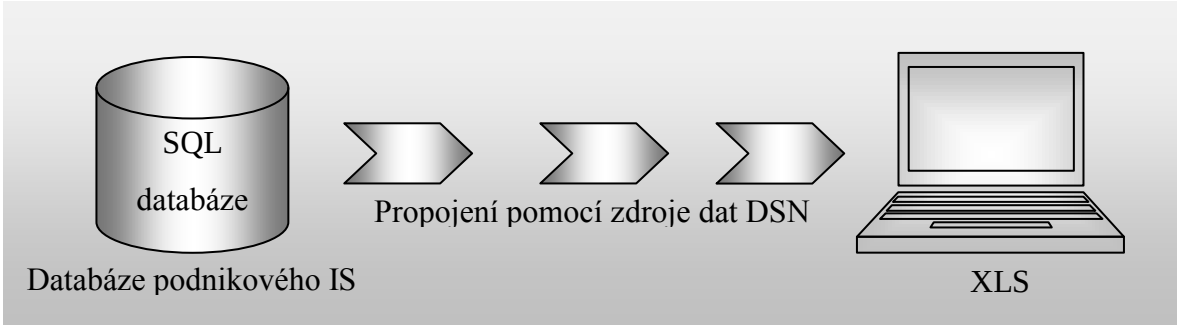
v běžném editoru (Poznámkový blok, PSPad …) a podle potřeby upravit. Při dalším připojení ke zdroji dat budou použity nově nastavené podmínky. [25]



Obr. 21 Proces připojení datového souboru do MS Excel

7.1.1.4 Databáze připojené pomocí zdroje dat

Nejenom exportované TXT/CSV soubory, ale také datové komponenty SQL databáze je možné propojit pomocí zdroje dat přímo s MS Excel. Takto můžeme přistupovat nejenom k uloženým tabulkám a pohledům, ale také k OLAP krychlím.

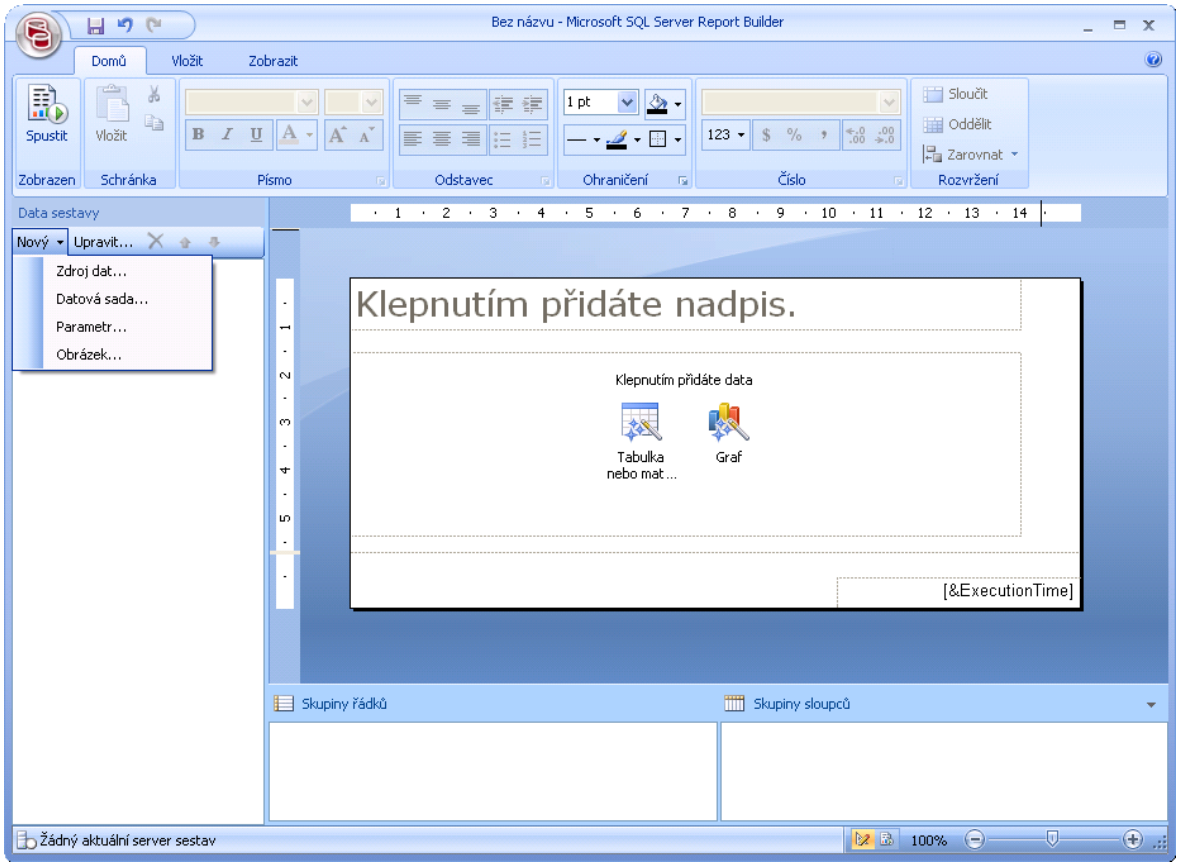


Obr. 22 Proces připojení datového souboru do MS Excel

Pomocí průvodce dotazem mohou i uživatelé, neznalí jazyka SQL, filtrovat data, třídit data a spojovat tabulky.

7.2 Analýza dat a tvorba sestav pomocí Report Builderu

Vynikajícím, ale přitom málo známým nástrojem, určeným koncovým uživatelům analytických služeb MS SQL Serveru 2008 je aplikace Report Builder. Jedná se o bezplatné lokalizované rozšíření služeb Reporting Services. Má příjemné grafické rozhraní, je velmi intuitivní a od uživatelů vyžaduje minimální znalosti. Instaluje se na klientské stanice, odkud se ke zdrojům dat připojuje pomocí definovaného zdroje dat. K vlastním datům se pak přistupuje pomocí definované sady dat. Z Report Builderu je možné přistupovat nejenom k datům MS SQL Serveru, ale i jiným OLTP a OLAP databázím.



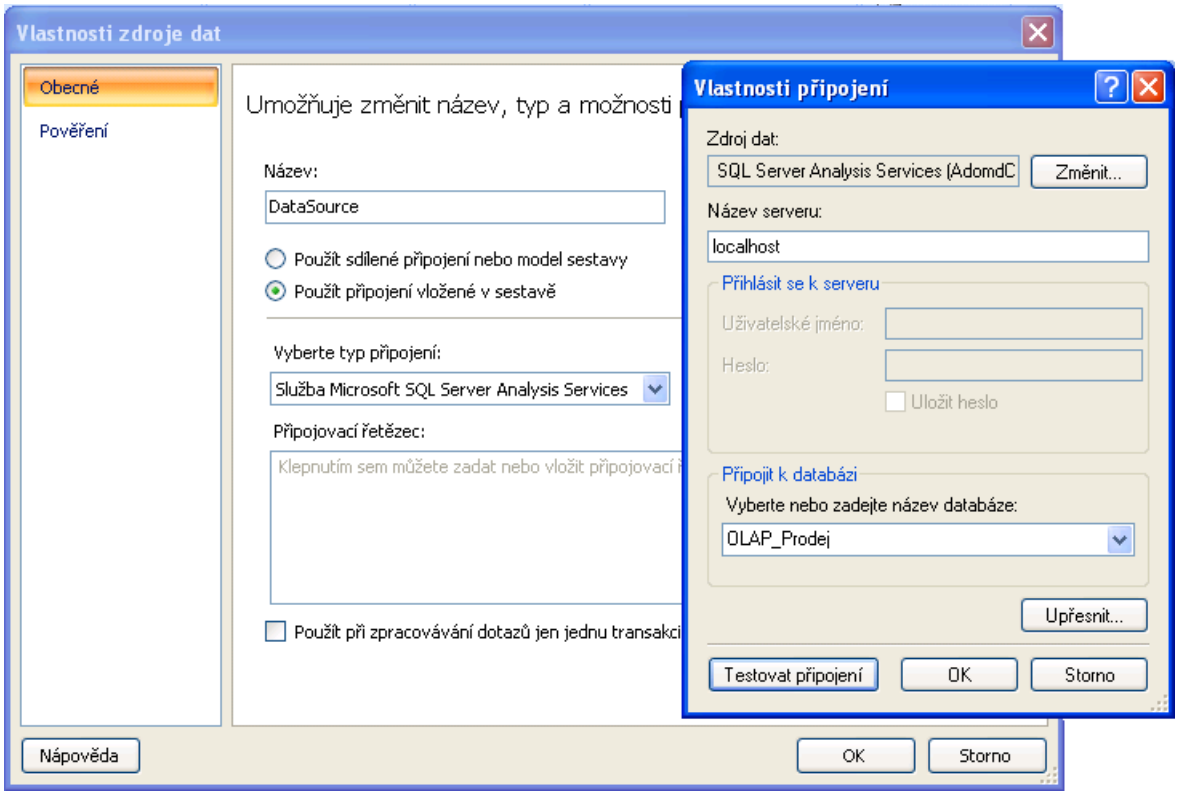
Obr. 23 Hlavní okno aplikace Report Builder

Tvorba sestavy se skládá ze tří kroků:

- Definice zdroje dat
- Definice datové sady
- Definice sestavy nebo grafu

7.2.1 Definice zdroje dat

Zdroje dat i datových sad se definují pomocí průvodce, spustitelného z hlavního okna aplikace ve volbě Nový. V prvním okně se zadává název zdroje dat, tedy název datového spojení, a jeho typ. Protože se jedná o analytickou sestavu, je nutné zvolit typ připojení Služba Microsoft Server Analysis Services. V dalším kroku následuje zadání názvu serveru a výběr analytické databáze.

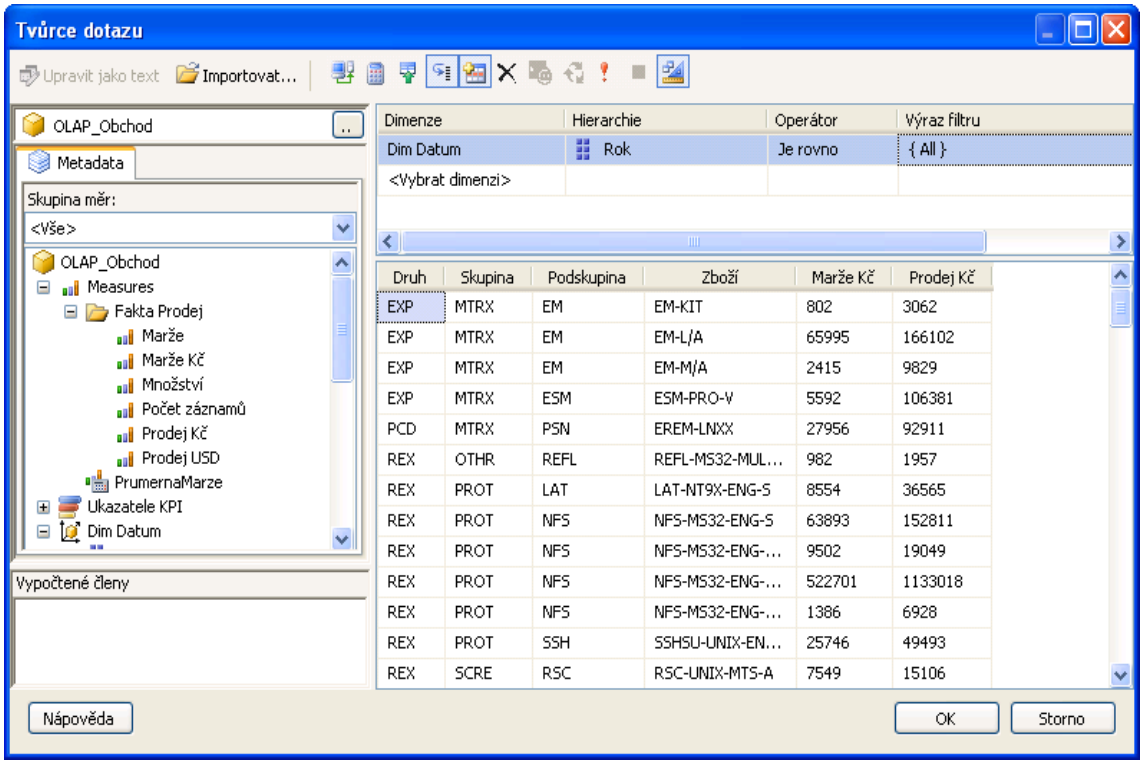


Obr. 24 Definice zdroje dat v Report Builderu

7.2.2 Definice sady dat

V definici sady dat se vybírají vlastní údaje, zahrnuté do sestavy. Opět se postupuje se pomocí srozumitelného průvodce. Tvorba analytických sestav, postavených na datových krychlích, je velmi podobná tvorbě kontingenčních tabulek. Stejným způsobem se přetahují pole do oblasti sloupců, řádků a dat, podobně se pracuje i s filtry. V náhledu se okamžitě zobrazují vybraná data.

U analytických sestav, vytvářených pomocí Report Builderu, je jediným nutným předpokladem, že jejich tvůrci rozumí pojmům měřítka a dimenze.



Obr. 25 Definice sady dat v Report Builderu

Prodej produktů

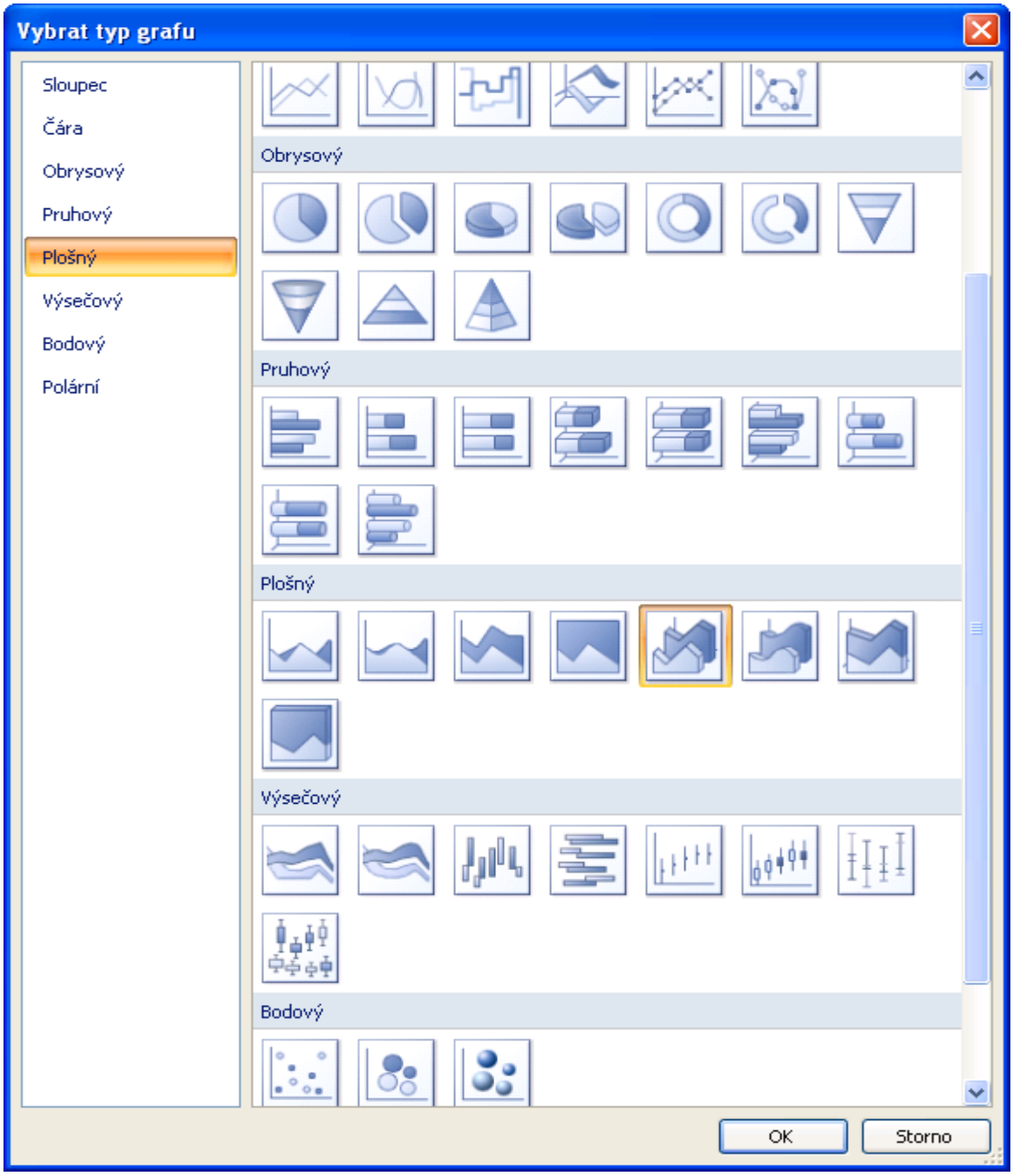
Druh	Skupina	Podskupina	Zboží	Marže Kč	Prodej Kč
⊟ EXP	⊞ MTRX	Celkem		5592	106381
	Celkem			5592	106381
⊟ PCD	⊞ MTRX	Celkem		27956	92911
	Celkem			27956	92911
⊟ REX	⊟ PROT	⊞ NFS	Celkem	106515	212011
		⊞ SSH	Celkem	25746	49493
		Celkem		132261	261504
	⊟ SCRE	⊟ RSCU	RSCU-ALL-MTV/A	3 062,00 Kč	11748
			RSCU-UNIX-ENG-VPKUP	1 284,00 Kč	1284
			RSCU-UNIX-ENG-VPLUP	0,00 Kč	0
			Celkem	4346	13032
		⊞ RSCW	Celkem	29121	102601
		⊞ RSSU	Celkem	72385	260601
		⊞ RSSW	Celkem	2298	8424
		Celkem		108150	384658
	⊞ WNBE	Celkem		317778	651299
	⊞ XSRV	Celkem		2312907	7110779
	Celkem			2871096	8408240
Celkem				2904644	8607532

Obr. 26 Ukázka interaktivní tabulkové sestavy s hierarchickými dimenzemi

7.2.3 Definice sestavy a grafů

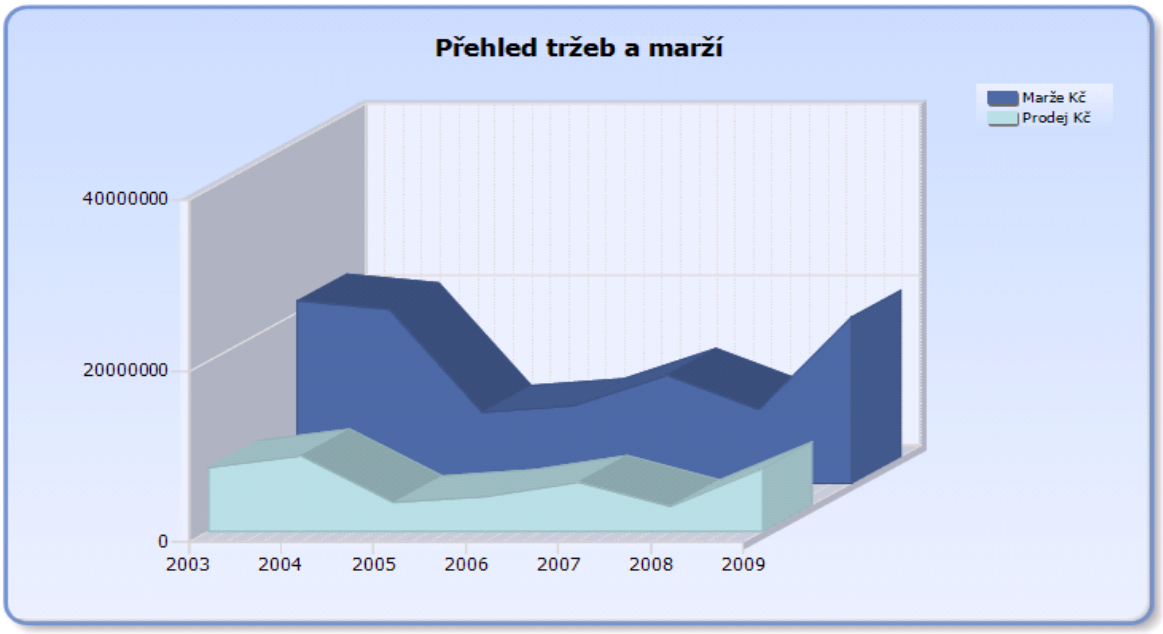
Vlastní definice sestav, postavená na datech, definovaných v sadách dat se spouští z hlavního okna aplikace výběrem typu sestavy (tabulka, graf).

Tvorba je jednoduchá, opět probíhá pomocí průvodce. Během něho je na výběr jen několik základních předdefinovaných sestav a typů grafů. Ty je ale možné po dokončení průvodce upravit nebo výrazně změnit.



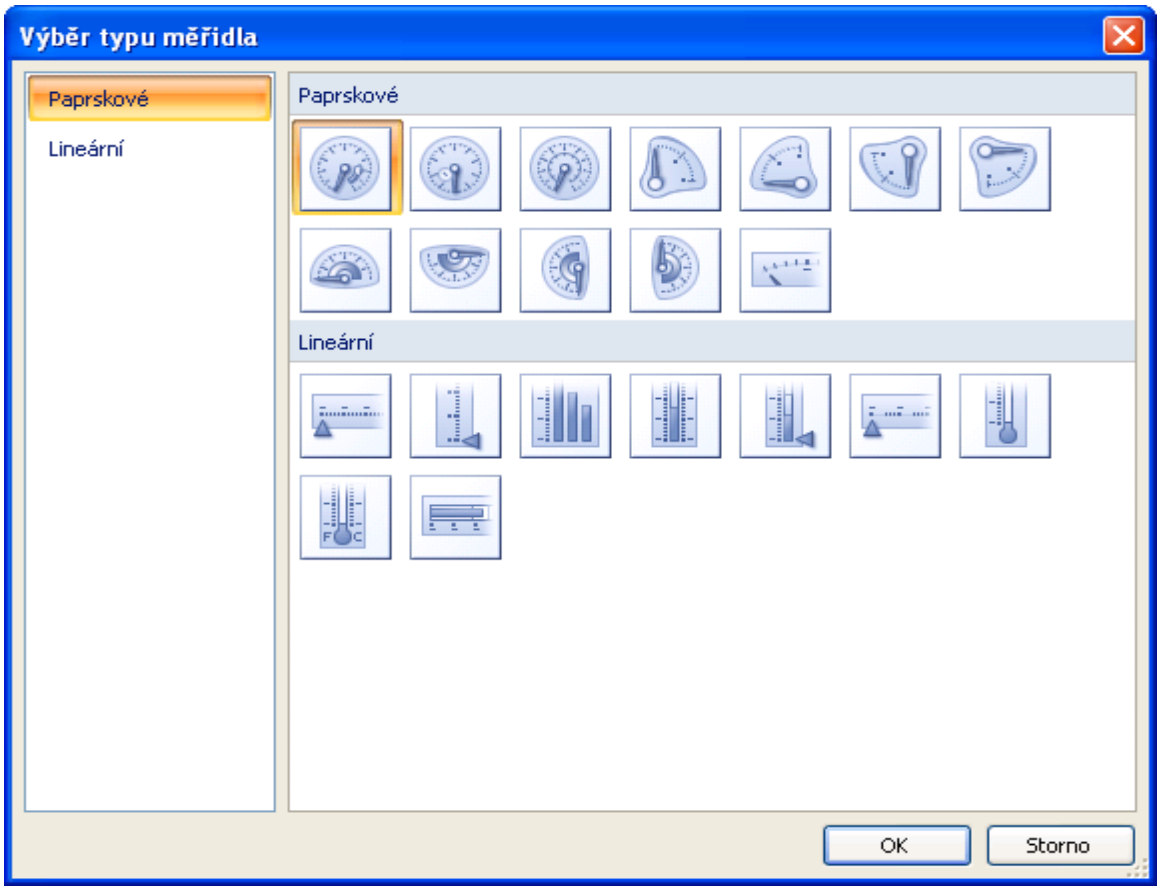
Obr. 27 Možnosti grafů v Report Builderu

Změny jednotlivých vlastností grafu se provádějí podobně jako v MS Excel.



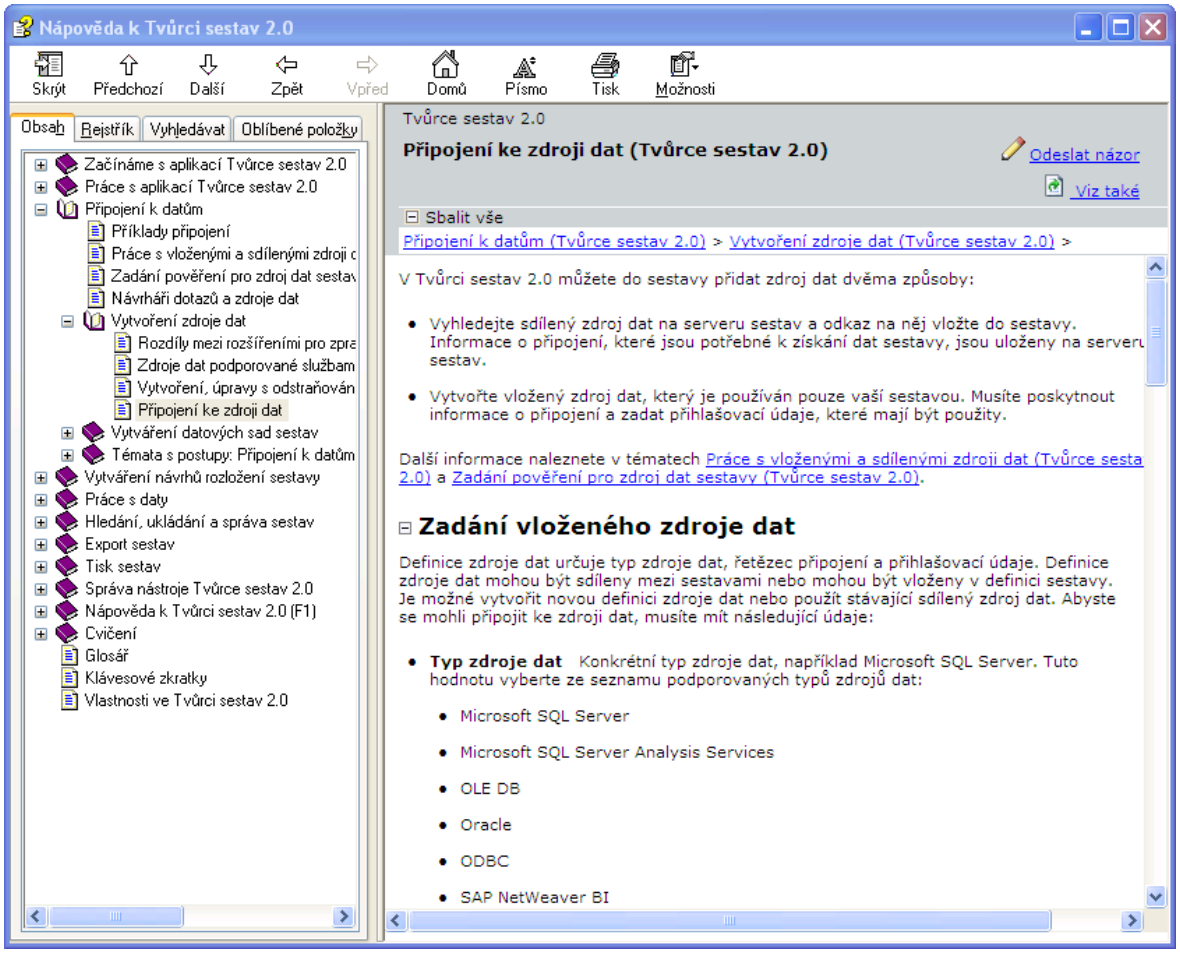
Obr. 28 Ukázka grafu vytvořeného v Report Builderu

Sestavy je možné obohatit o celou řadu měřidel.



Obr. 29 Možnosti měřidel v Report Builderu

Neocenitelným pomocníkem pro spoustu českých uživatelů je nejenom lokalizace produktu, ale také jeho nápovědy.



Obr. 30 Ukázka lokalizované nápovědy v Report Builderu

Vytvořené sestavy je možné publikovat na podnikovém portálu, zasílat e-mailem, exportovat do mnoha různých formátů. Jeho interaktivní možnosti a efektní výstupy jsou v kombinaci s datovými krychlemi, vytvořenými pomocí MS SQL Serveru 2008, vynikajícím nástrojem pro analýzu podnikových dat.

8 ANALÝZA OBCHODNÍCH ČINNOSTÍ DISTRIBUTORA SW POMOCÍ NÁSTROJŮ MS SQL SERVER 2008

Společnost, zabývající se poskytováním služeb v oblasti software a jeho prodejem, potřebuje analyzovat své obchodní aktivity. Stávající způsob, při kterém si každý obchodník vede svoji individuální evidenci v MS Excel je nevyhovující. Neexistují společné data, ze kterých by bylo možné vysledovat vývoj prodeje jednotlivých produktů a predikovat další vývoj. Chybí hodnocení segmentace trhu. Není možné posoudit celkovou úspěšnost marketingových aktivit. Obchodní aktivity jsou málo transparentní. Protože každý z obchodníků má jiné zkušenosti a také jiné preference, má management před sebou řadu různých názorů, ale pro rozhodování v podstatě nedostatek informací. Z toho důvodu se firma rozhodla ověřit možnosti, které v tomto směru nabízí MS SQL Server 2008.

8.1 Příprava dat

K dispozici jsem měla data jednotlivých obchodníků. Jednalo se různě formátované tabulky MS Excel. Aby bylo možné s analýzou vůbec začít, bylo nutné data uspořádat a navrhnou jejich jednotnou strukturu. Poté bylo nutné data očistit, sjednotit a v nově definovaném formátu uložit. Jako úložiště posloužil opět MS Excel. Jednalo se vůbec o nejnáročnější část práce. Ale na jejím konci byly podklady, pomocí nichž už mohli manažeři začít smysluplně analyzovat své obchodní aktivity.

Protože cílem bylo ověření analytických nástrojů MS SQL Serveru 2008, následovala tvorba a naplnění malého datového skladu. Datový sklad tvoří dvě tabulky faktů:

- **FaktaProdej**

  Tabulka faktů pro analýzu prodeje

- **FaktaNakup**

  Tabulka faktů pro analýzu nákupu

A následující tabulky dimenzí:

- **DimDatum**

  Tabulka společná pro FaktaNakup i FaktaProdej. Tabulka časových dimenzí.

- **DimZSkupinaA**

Tabulka společná pro FaktaNakup i FaktaProdej. Dimenze druhů zboží.

- **DimZSkupinaB**

  Tabulka společná pro FaktaNakup i FaktaProdej. Dimenze skupin zboží.

- **DimZSkupinaC**

  Tabulka společná pro FaktaNakup i FaktaProdej. Dimenze podskupin zboží.

- **DimZbozi**

  Tabulka společná pro FaktaNakup i FaktaProdej. Dimenze zboží. Umožňuje nadefinovat hierarchickou strukturu dimenze (Druh → Skupina → Podskupina → Zboží).

- **DimNFaktura**

  Tabulka pouze pro FaktaNakup. Přijaté dodavatelské faktury Analýza je zaměřená na skutečně realizované dodávky, proto pomíjí nákupní objednávky, které nemusely být plně vykryté nebo mohly být stornované. Nákupní faktury představují skutečně dodané a přijaté zboží.

- **DimNDodavatel**

  Tabulka pouze pro FaktaNakup. Dimenze dodavatelů.

- **DimPFaktura**

  Tabulka pouze pro FaktaProdej. Vystavené odběratelské faktury. Analýza je zaměřená na skutečně realizovaný prodej, proto pomíjí objednávky zákazníků. Vystavené faktury představují skutečně prodané zboží.

- **DimProdejce**

  Tabulka pouze pro FaktaProdej. Dimenze prodejců.

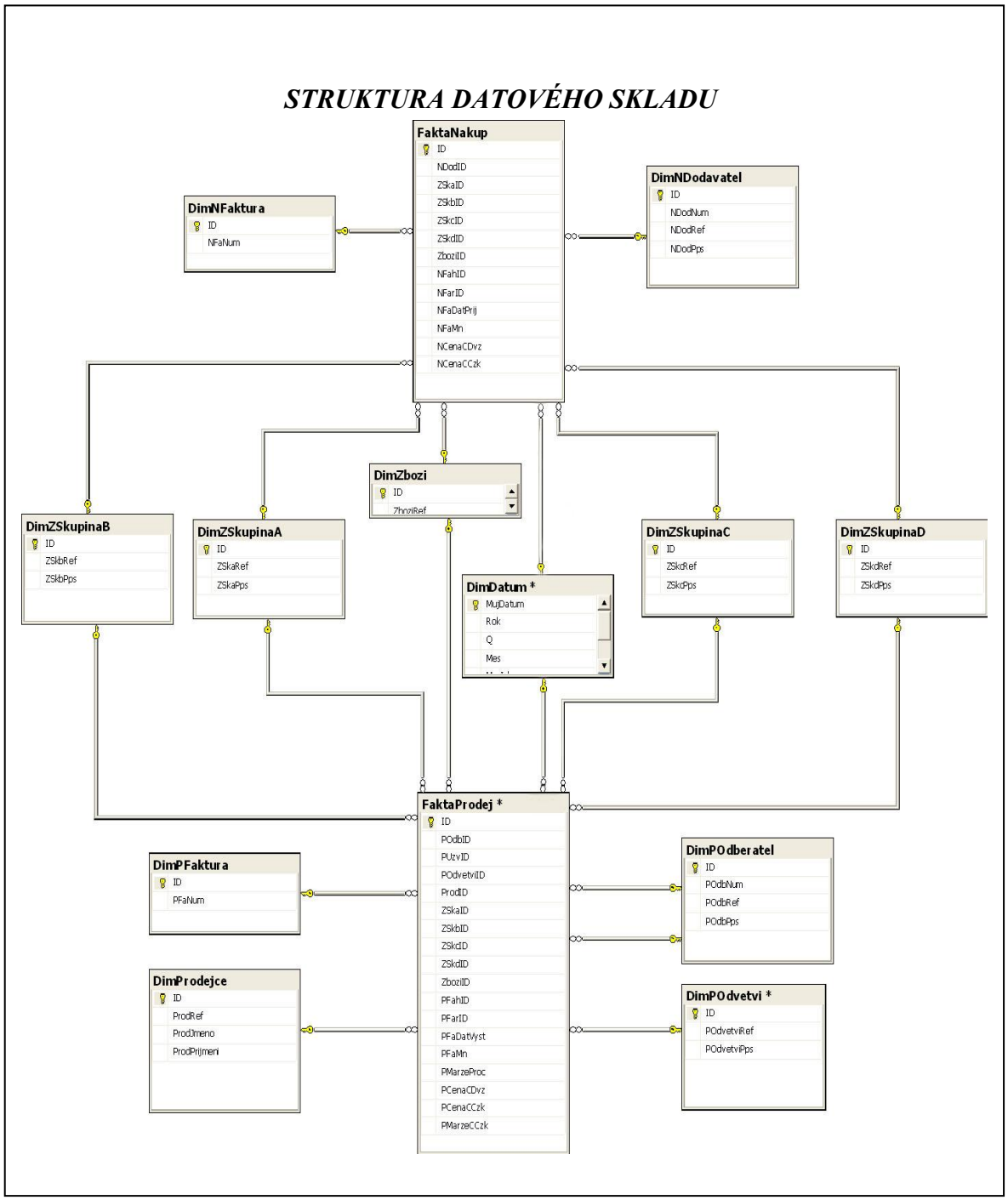
- **DimPOberatel**

  Tabulka pouze pro FaktaProdej. Dimenze odběratelů. Obsahuje jak informace o zákaznících, nakupujících zboží, tak i o koncových uživatelích, kteří mohou software nakupovat prostřednictvím svých IT dodavatelů. Zpravidla se to týká uživatelů, kterým podnikové informační technologie zajišťuje externí firma prostřednictvím outsourcingu. Z hlediska analýzy prodeje je důležitější znalost

struktury koncových uživatelů než znalost zákazníků, objednávajících zboží. Umožní lépe zacílit marketingové aktivity na výběr potenciálních uživatelů. Umožňuje nadefinovat hierarchickou dimenzi (Odvětví → Odběratel)

- **DimPOdvetvi**

  Tabulka pouze pro FaktaProdej. Segmentace průmyslových odvětví dle EU.



Obr. 31 Struktura datového skladu analyzovaných dat

Z obrázku je zcela zřetelná struktura datového skladu typu souhvězdí. Je to dáno sdílenými tabulkami produktových a časových dimenzí.

Před plněním datového skladu bylo nutné připravit také všechny potřebné číselníky, tak aby byly zajištěny vazby mezi tabulkami dimenzí a tabulkami faktů (tj. vazby mezi primárními klíči tabulek dimenzí a cizími klíči tabulek faktů). K importu dat jsem použila vestavěné funkce MS SQL Serveru 2008.

Vzhledem k malému objemu vstupních dat, jsem zachovala nízkou granulitu.

8.2 Vytvoření datové krychle

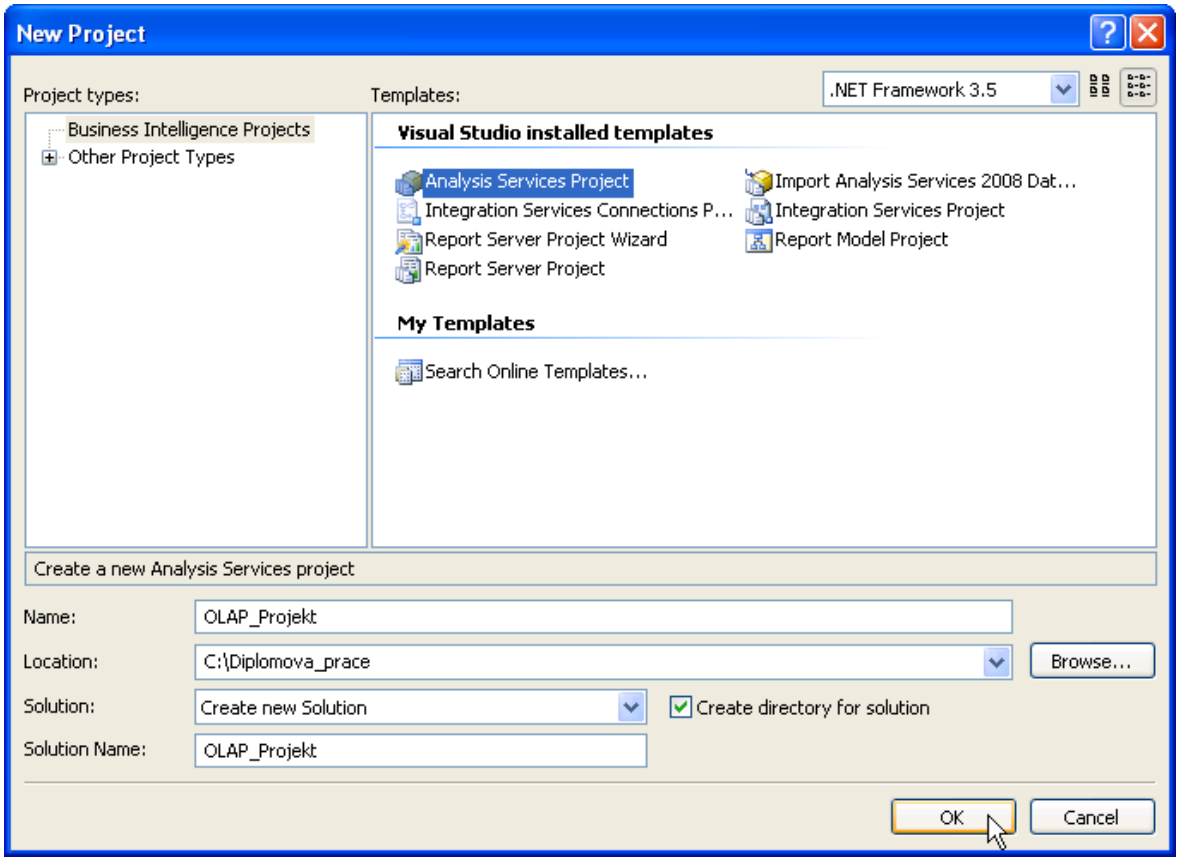
Vícerozměrná analýza dat se provádí pomocí tzv. datové krychle (OLAP cube). Vývojovým prostředím pro tvorbu datových OLAP krychlí v MS SQL Serveru 2008 je Business Inteligence Development Studio (BIDS). Jedná se upravenou verzi MS Visual Studia 2008, které je bezúplatně dodáváno s MS SQL Serverem. BIDS podporuje offline vývoj projektů, protože pracuje se snímky schémat zpřístupněných datových zdrojů. [27] To je velká výhoda, protože se tím omezuje také zatížení serveru a komunikace v síti.

Vytvoření datové krychle se skládá z těchto kroků:

- Definování datových zdrojů
- Definování pohledu na datové zdroje
- Návrh datové krychle
- Konfigurace dimenzí
- Zpřístupnění datové krychle

8.2.1 Vytvoření projektu

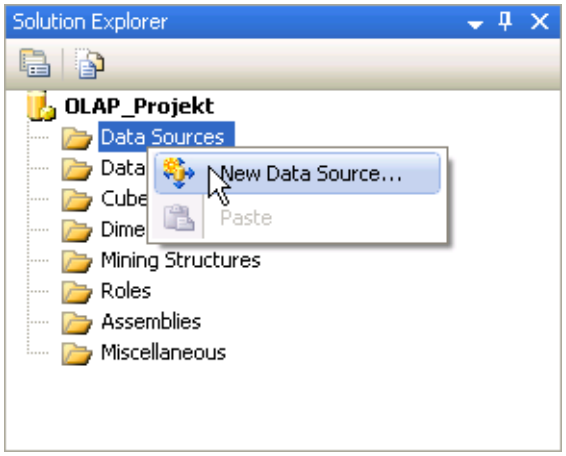
Nový projekt Analysis Services začíná spuštěním vývojového prostředí BIDS a volbou File → New → Project. Analytické projekty vznikají na základě šablony Analysis Services Project. Po zadání názvu a umístění projektu je se otevře okno vývojového prostředí ve kterém je možné začít modelovat datovou krychli.



Obr. 32 Vytvoření projektu v BIDS

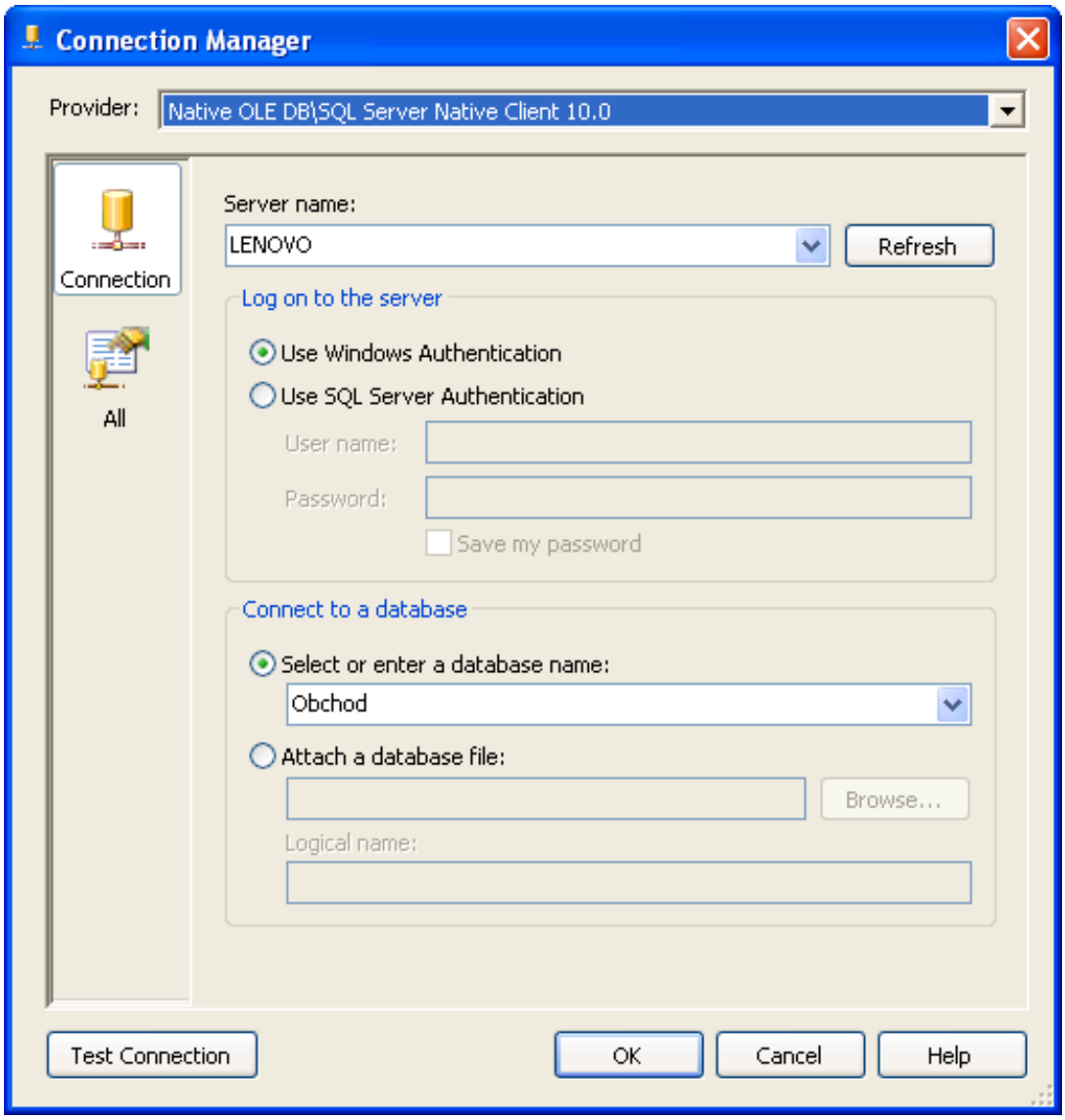
8.2.2 Definování datových zdrojů

Prvním krokem při budování krychle je vytvoření datového zdroje. Datový zdroj odkazuje na databázi, ze které projekt přistupuje k datům. Datový zdroj nemusí být propojen pouze s databází MS SQL Serveru, ale s jakoukoliv jinou databází, přístupnou pomocí OLE DB nebo ODBC.



Obr. 33 Definování datových zdroje

Datový zdroj se definuje v okně Solution Explorer (pokud není zobrazeno, zpřístupní se volbou View → Data Source). Kliknutím pravým tlačítkem myši na složku Data Source → New Data Source se spustí průvodce Data Source Wizard. Průvodce nabídne v prvním kroku dostupné datové zdroje. Pokud žádný z nich nevyhovuje, je nutné vytvořit datový zdroj nový.

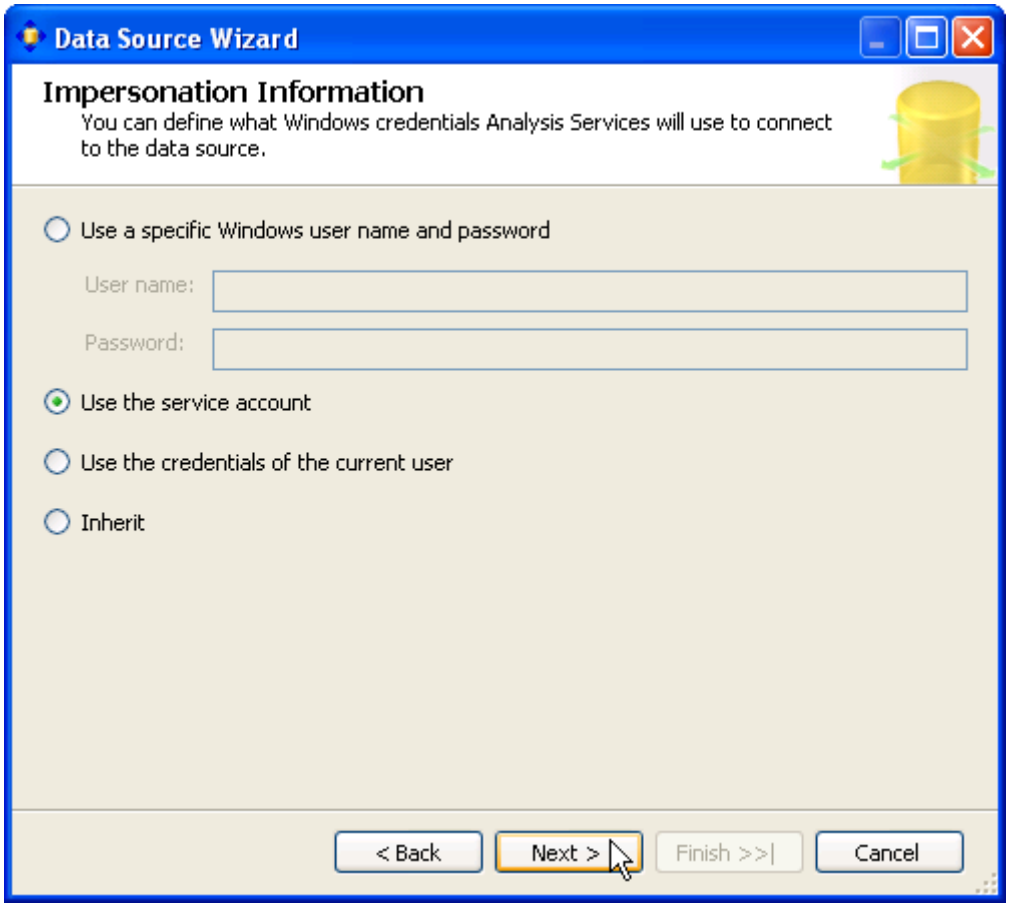


Obr. 34 Definování nového datového spojení.

Při tvorbě nového datového zdroje se vybírá server, databáze a typ požadovaného zabezpečení.

Následně se definuje způsob připojení serveru služeb Analysis Services (nikoli uživatel pracující v prostředí BIDS). Volba aktuálně přihlášeného uživatele (Use the Credentials of the Current User) se zásadně nedoporučuje. Nejvhodnější volba je přihlášení účtem služby

(Use the Service Account). Tato volba předpokládá, že účet, pod kterým služby Analysis Services běží, mají přístup k datovému zdroji.



Obr. 35 Definování způsobu připojení serveru služeb Analysis Services

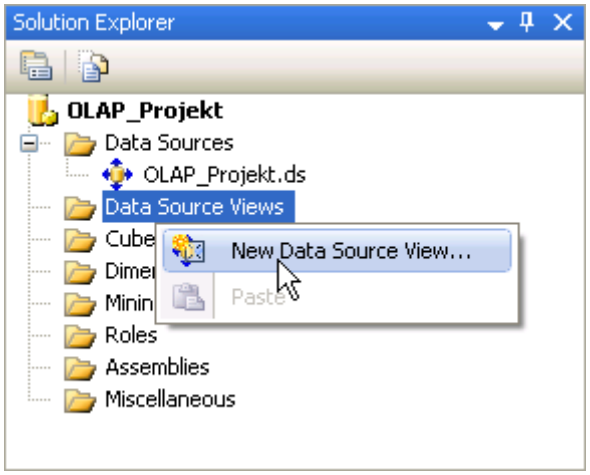
Dokončením průvodce se vytvoří nový zdroj dat. Ten je možné podle potřeby, dvojitým kliknutím myši na název zdroje dat, upravit. Vytvořený datový zdroj se jako soubor s příponou ds automaticky umístí do adresáře projektu. Soubor je možné zobrazit v poznámkovém bloku, PSPadu či podobném editoru.

8.2.3 Definování pohledů na datové zdroje

Po vytvoření zdroje dat následuje vytvoření pohledu na datové zdroje. Jedná se specifickou záležitost, která úzce souvisí s offline způsobem práce v Business Inteligence Developer Studio. Pohledy na datové zdroje jsou souborem metadat, asociovaných s tabulkami a pohledy, které projekt Analysis Services používá.

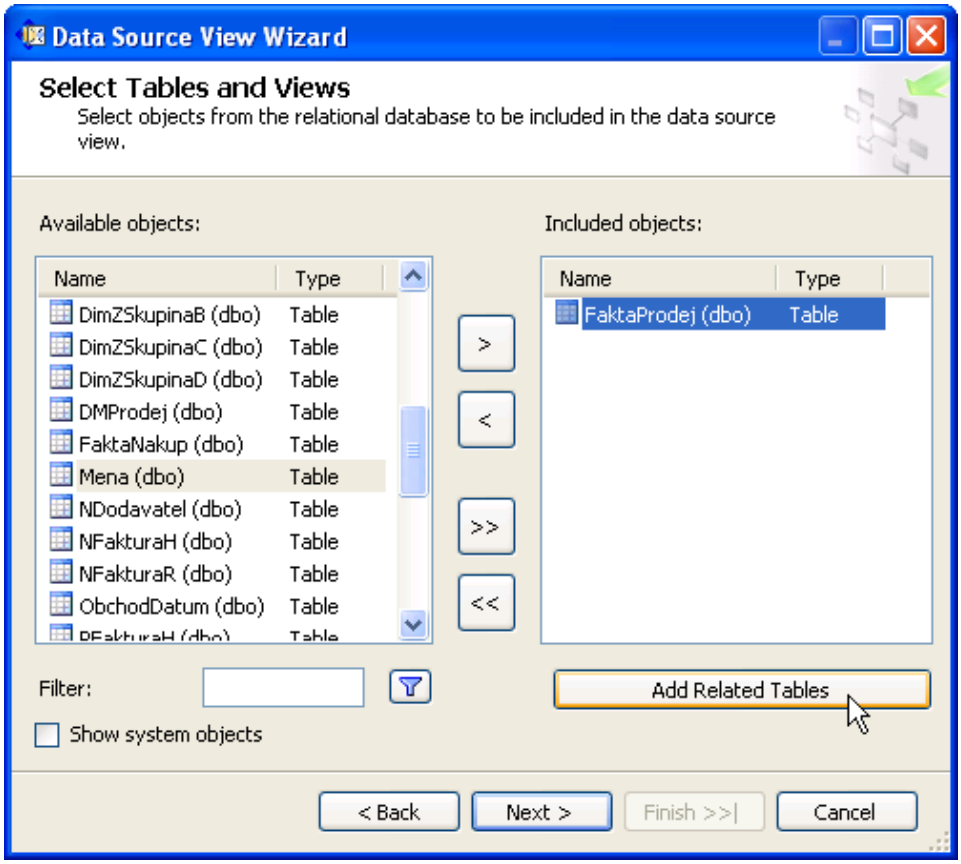
Není to tedy pohled na celou připojenou databázi, ale pouze na její vybranou část, zvolenou pro stavbu datové krychle.

Průvodce tvorbou pohledu na datové zdroje se spouští podobně jako průvodce tvorbou datového zdroje, stisknutím pravého tlačítka myši na složku Data Source View a volbou New Data Source View.



Obr. 36 Definování pohledu na datové zdroje

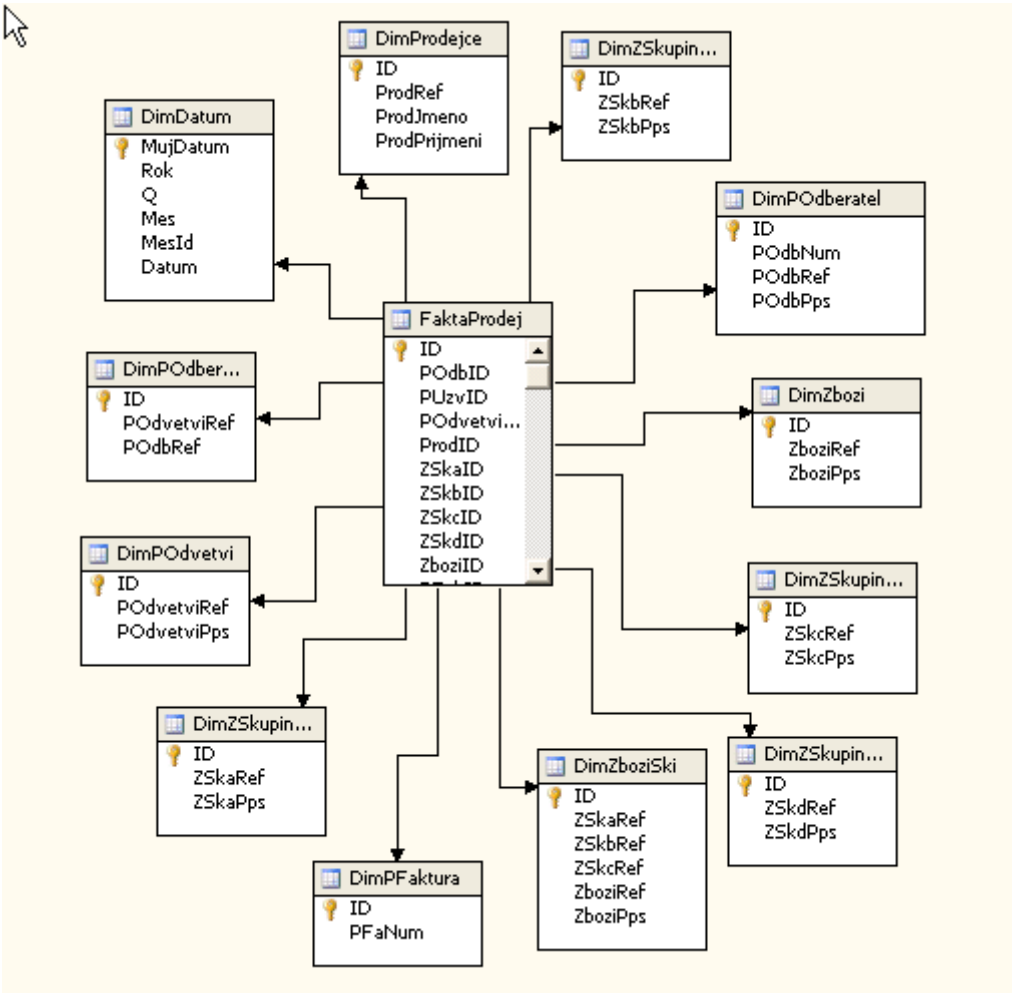
V prvním kroku průvodce se vybírá odpovídající datový zdroj. Následuje výběr potřebných tabulek, zahrnutých do datového zdroje.



Obr. 37 Výběr tabulek, které budou zahrnuté do datového zdroje

Je rozumné, jako první vybrat tabulku faktů a kliknutím na tlačítko Add Related Tables přidat související tabulky. Systém je vybere automaticky na základě definovaných primárních a cizích klíčů.

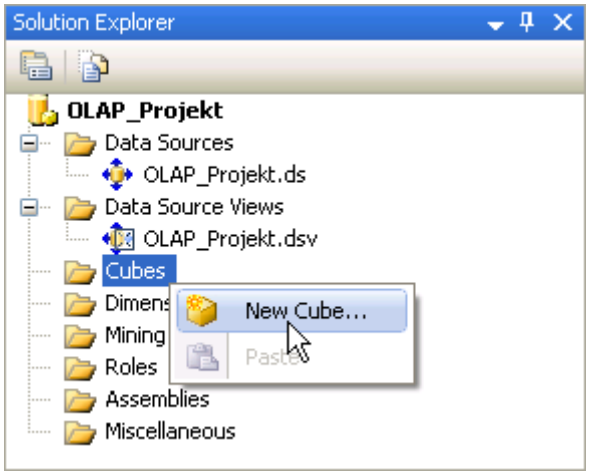
Po dokončení průvodce se vytvoří nový pohled na datový zdroj a jako soubor s příponou dsv se automaticky umístí do adresáře projektu, odkud ho je možné zobrazit v poznámkovém bloku, PSPadu nebo jiném editoru. Současně se v okně zobrazení návrhu objeví struktura datového diagramu. Jedná se zobrazení vazeb tabulek dimenzí s tabulkou faktů. Srovnáním se strukturou datového skladu je vidět, že datovou krychli tvoří pouze omezený výběr tabulek dimenzí a faktů.



Obr. 38 Struktura pohledu na datové zdroje

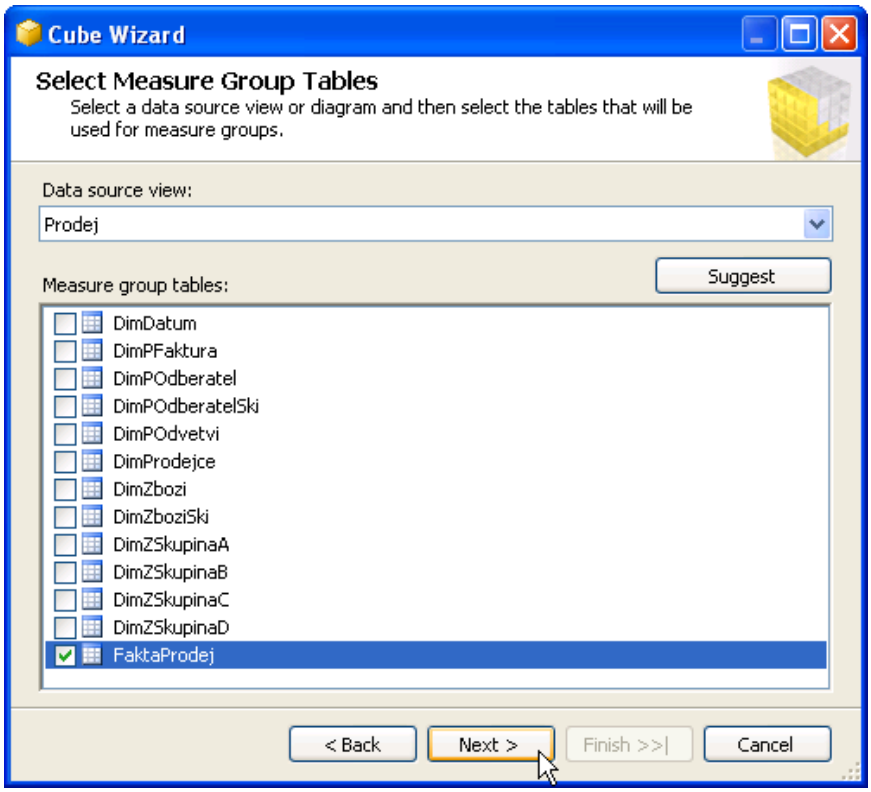
8.2.4 Definice datové krychle

Po vytvoření datových zdrojů a pohledů na datové zdroje následuje samotné vytvoření datové krychle. Spouští se pomocí průvodce.



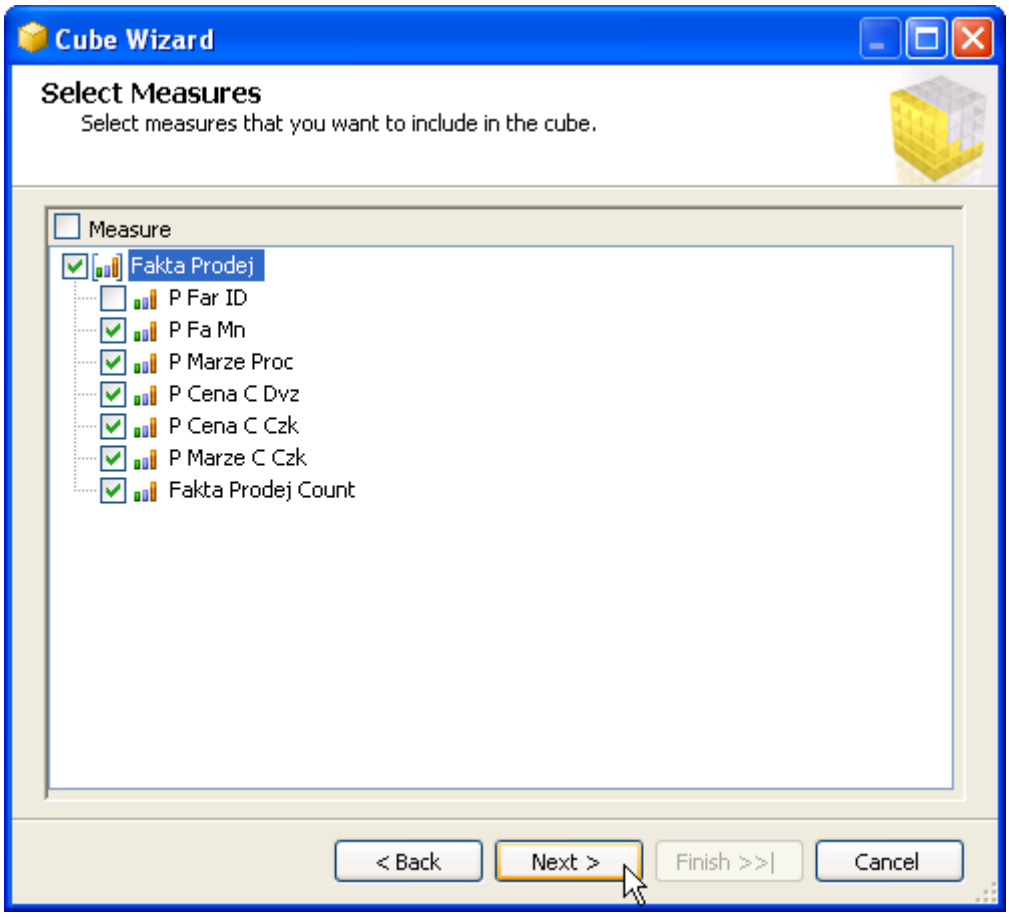
Obr. 39 Definice datové krychle

V prvním kroku je nutné vybrat způsob vytvoření datové krychle. Můžeme vytvořit prázdnou datovou kostku, vytvořit ji pomocí zdroje dat nebo, nejjednodušeji, pomocí existujících tabulek. Po přesunu na další obrazovku nastává výběr tabulek do skupiny měřítek. Měřítka (measures) jsou sledované číselné údaje obchodních aktivit (množství prodaného zboží, zisky, tržby, náklady …), umístěné v tabulkách faktů. Podle potřeby je možné vybrat jednu nebo více tabulek faktů. Protože jsem měla pro potřeby analýzy prodeje všechna data umístěna v tabulce FaktaProdej, vybrala jsem tuto tabulku.



Obr. 40 Výběr tabulky faktů do datové krychle

Po výběru tabulky faktů je nutné zvolit sledovaná měřítka.

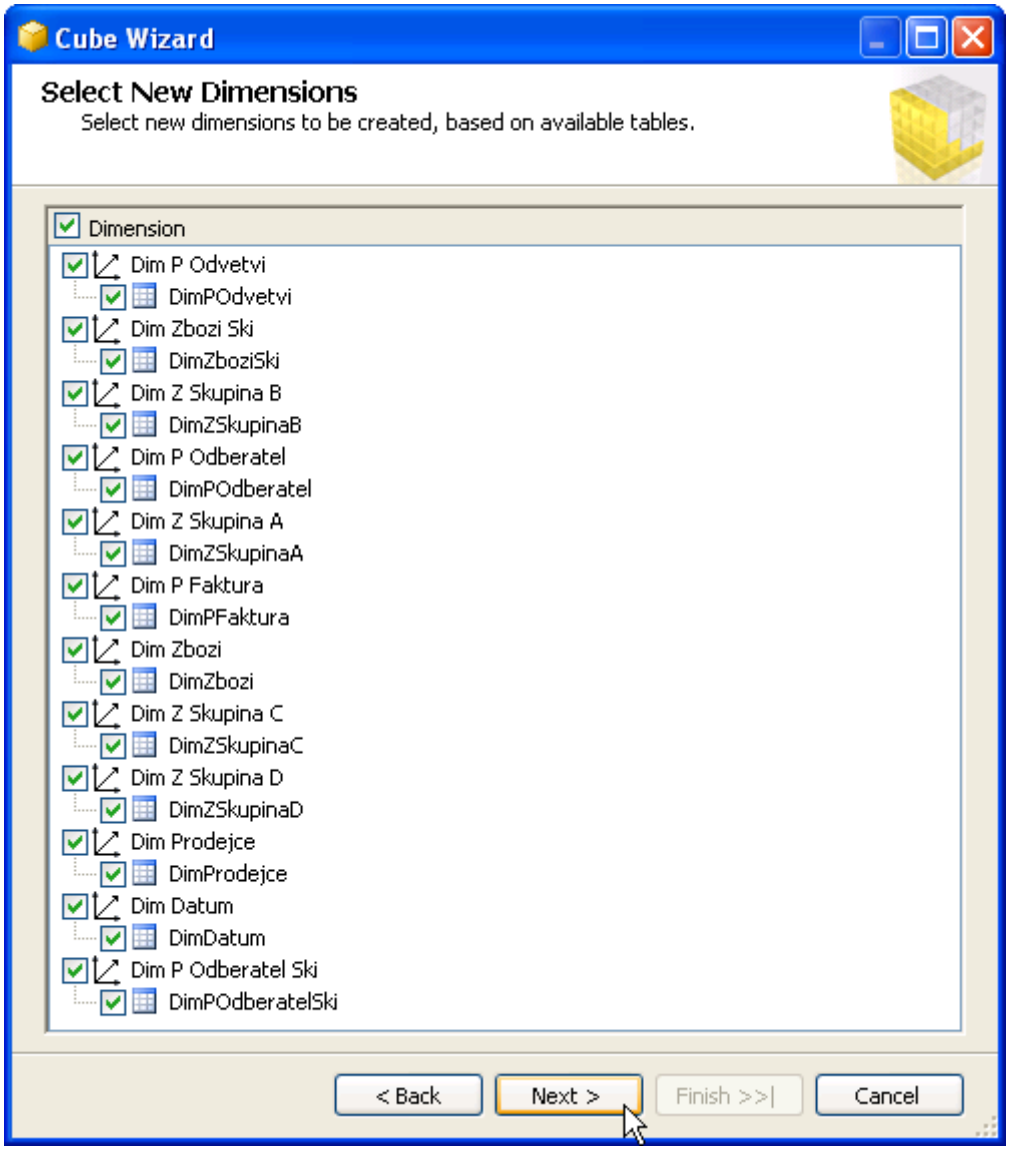


Obr. 41 Výběr měřítek do datové krychle

Protože sloupec PFaiID je definován jako primární klíč tabulky FaktaProdej a není tedy sledovanou hodnou, je potřeba jeho zaškrtnutí zrušit. Ostatní údaje jsou pro analýzu zajímavé, a proto zůstávají zachovány. (PFaMn – celkové množství zboží, PMarzeProc – marže vybrané položky v procentech, PCenaCDvz – celková cena prodeje v cizí měně[4], PCenaCCzk – celková cena prodeje v Kč, PMarzeCCzk – celková marže v Kč). Měřítko FaktaProdejCount uvádí celkový počet záznamů a nabízí ho systém. Analýzu zisků není možné provést, protože údaje o vynaložených nákladech nejsou k dispozici.

V dalším kroku následuje výběr dimenzí, které budou použity k analýze.

[4] Prodejní cena software je uváděna v USD a teprve při fakturaci dochází ke stanovení ceny v Kč, přepočtem aktuálním kurzem. Cena v USD je z hlediska porovnání prodávaných objemů důležitá, protože není zkreslená vývojem kurzu.



Obr. 42 Výběr dimenzí do datové krychle

Po dokončení průvodce implicitně zvolí název krychle podle použitého Data Source View.

8.2.5 Konfigurace dimenzí

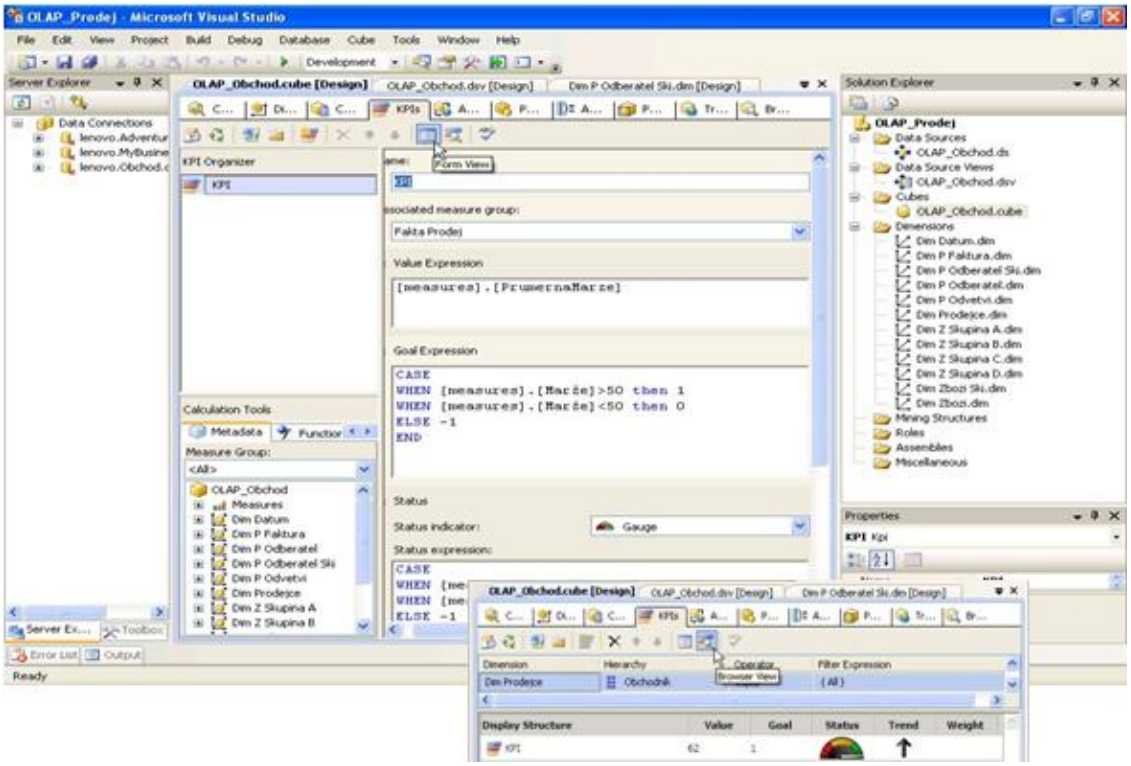
Posledním krokem před zveřejněním projektu je konfigurace dimenzí. V této fázi se nastavují vlastnosti, definují hierarchie, způsoby zobrazení hodnot nebo přidávají vypočítaná pole. Úprava se provádí v návrháři Dimension Designer, který se aktivuje dvojitým kliknutím na vybranou dimenzi a slouží především k úpravě zobrazení koncovým uživatelům.

8.2.6 Sledování klíčových indikátorů výkonnosti KPI

Z pohledu analýzy podnikových dat je velmi zajímavá možnost sledování klíčových indikátorů výkonnosti. KPI (Key Performance Indicator). Jedná se o ukazatele, které pomáhají organizaci dosáhnout stanovených cílů pomocí jejich definování a měření průběhu jejich plnění. Podle KPI poznáme, zda se k cíli blížíme, či nikoli. Zároveň vidíme, jakou cestu jsme již urazili a kolik nám k dosažení cíle ještě chybí. Výsledkem těchto měření je nalezení a eliminování slabých bodů sledovaného procesu, a tedy i zefektivnění cesty k dosažení definovaného cíle [28].

Sledování KPI je možné k projektu datové krychle připojit. Slouží k tomu záložka KPIs v okně projektu. KPI se definují pomocí MDX výrazu [27]:

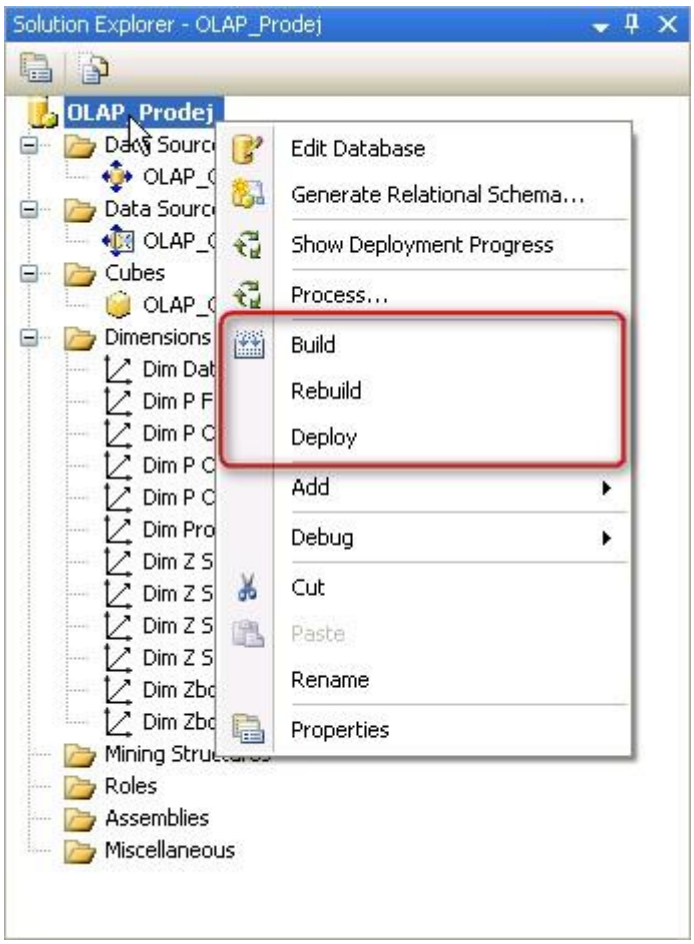
- Hodnota (MDX výraz)
- Cílová hodnota (MDX výraz)
- Stav (MDX výraz v rozsahu od -1 do 1)
- Trend (MDX výraz v rozsahu od -1 do 1)
- Typ vizualizace (semafor, ukazatel …)



Obr. 43 Definice a zobrazení KPI v projektu datové krychle

8.2.7 Zveřejnění projektu

Závěrečná fáze projektu se skládá ze dvou částí. Build zahrnuje vývoj a předběžné ladění na lokálním vývojářském počítači, Deploy zavádí odladěný proces na analytický server a předává ho k dispozici koncovým uživatelům. V rutinní praxi pak dochází k pravidelné aktualizaci datových krychlí v předem stanovených periodách. Ty závisí na objemu dat a na potřebě jejich aktuálnosti.



Obr. 44 Zveřejnění projektu

8.3 Klientský přístup k datovým krychlím – výhody a nevýhody

Pro práci s datovými krychlemi slouží vývojářům prostředí BIDS. To je pro koncové uživatele nejenom nevhodné, ale také nedostupné.

Koncoví uživatelé mohou volit mezi přístupem tlustého klienta, kdy klientská aplikace běží na jejich lokálním počítači (MS Excel, Report Builder) a tenkého klienta, který pro přístup k serverovým aplikací a datům používá internetový prohlížeč. V případě tlustého klienta

nedochází k tak velkému zatížení serveru, jako v případě tenkého klienta, protože výpočetní výkon se rozděluje mezi lokální počítač a mezi server. U tenkého klienta leží celá aplikační logika ne serveru, výpočetní možnosti klienta se nevyužívají. [27]

Já jsem se ve své práci zaměřila na dva klientské nástroje, shodu okolností oba patří do skupiny tzv.tlustých klientů.

MS Excel je vynikajícím tabulkovým procesorem o jehož kvalitách dnes už asi nikdo nepochybuje. Většina uživatelů zdaleka nevyužívá možností, které jim nabízí. Je to výborný klient pro přístup k datovým krychlím, který dokáže pokrýt potřeby podnikových analytiků. Nevýhodou je, že práce s ním vyžaduje interaktivní přístup a není proto asi nejvhodnějším nástrojem top manažerů na úrovni generálních ředitelů nadnárodních korporací. Tito lidé potřebují hotové výstupy na jejichž základě musejí provádět rychlá strategická rozhodování. Pro velkou většinu ostatních manažerů je plnohodnotným klientem analytických aplikací.

Zajímavým programem je i Report Builder. Dokáže tvořit velmi zajímavé grafy a interaktivní reporty doplněné o možnost vizualizace pomocí měřidel výkonů. Umožňuje sestavy sdílet na podnikových portálech. Podporuje export do mnoha různých formátů (XML, TIFF, PDF, XLX, DOC …). Formátování sestav a grafů nabízí stejné možnosti jako MS Excel. Tvorba sestav je postavena na principu kontingenčních tabulek a tak uživatelům, kteří je v MS Excel používají, nebude činit potíže. Dokonce bych řekla, že díky jasné struktuře dimenzí a faktů, je tato práce ještě jednodušší než tvorba kontingenčních tabulek z rozsáhlých xls souborů Na ovládání ve stylu Office 2007 nejsou asi zatím všichni uživatelé zcela zvyklí, přesto myslím, že by jim práce s tímto programem připadala intuitivní a pohodlná. U této aplikace totiž odpadá zatížení její předchozí verzí. Report Builder je výborná aplikace, která by si zasloužila mnohem větší pozornost.

ZÁVĚR

Cílem mé práce bylo ověření možností analýz podnikových dat vzhledem k dostupnosti vhodných nástrojů. Protože tyto analýzy se už dávno netýkají pouze vrcholových manažerů a potřeb jejich strategických rozhodování, zaměřila jsem se zejména na možnosti běžných koncových uživatelů, které dostupnost informací ve formě vhodných analytických výstupů trápí nejvíce. I přes množství reportů, tabulek a grafů, které jim jejich informační systémy nabízejí, se vždy objeví takový výstup, který některému z uživatelů chybí. Někdy se může jednat o tvrdošíjnost uživatele, trvajícího za každou cenu na „své“ sestavě a to i v případě, když jiné reporty tytéž informace nabízejí také. Tato varianta je celkem jednoduše řešitelná. Záleží pouze na rozhodovacích pravomocích nespokojeného uživatele. Horší případ ale nastává, pokud systém požadované informace ve formě jakéhokoliv výstupu neposkytuje vůbec. V té situaci zpravidla přicházejí na řadu investice do zákaznických řešení. Ty ale ve velkých společnostech, v nichž jakékoli náklady podléhají složitým schvalovacím procedurám, mohou představovat poměrně náročný proces. Přitom požadované analytické výstupy nemusí být jenom produktem informačních systémů. Často stačí používání nástrojů, které už podnik vlastní. Jedná se zejména o MS Excel či bezplatný lokalizovaný doplněk MS SQL Serveru Report Builder. Propojením těchto aplikací s podnikovými databázemi získají koncoví uživatelé okamžitý přístup k aktuálním datům bez jakýchkoliv dalších nákladů. Nevýhodou jsou pouze zvýšené nároky na zabezpečení databáze, které musí zajistit databázoví administrátoři. Přínosem jsou podnikové informace, on-line dostupné všem zainteresovaným uživatelům a navíc ve formě, kterou si uživatelé v jim oblíbeném a blízkém prostředí nadefinují sami a navíc ve tvaru, který je pro ně nejvhodnější. V MS Excel jsou to zejména dynamické kontingenční tabulky a grafy, pracující jak s provozními OLTP tak analytickými OLAP daty, v případě Report Builderu se jedná o různé typy tradičních i interaktivních sestav, publikovaných na webových portálech, v intranetu, zasílaných e-mailem či jinak zveřejňovaných. Nezanedbatelná je také skutečnost, že mezi návrhem sestavy a vlastní analýzou dat uplyne pouze krátký čas. Zkušený uživatel totiž dokáže ve známém prostředí vytvořit sestavu poměrně rychle. Oba uvedené nástroje, které byly pro analýzu dat zvoleny (MS Excel kvůli jeho kvalitám a obrovské rozšířenosti, Report Builder kvůli jeho ceně a možnostem) jsou výborným doplňkem analytických možností MS SQL Serveru 2008 mezi něž patří jak analýza vícerozměrných dat (OLAP), tak také technologie dolování dat (Data Mining).

CONCLUSION

The main aim of my thesis was verification of business intelligence possibilities with regard to availability of convenient instrument. Business intelligence is not nowadays related only to the top management and their needs of strategic decision. Therefore did I focus on possibilities of common end users which information availability in the form of suitable analytical outputs bothers the most. In spite of amount of reports, tables and diagrams, which are offered by their information systems, every time appears such an output that some user misses. Sometimes can it be user stubbornness, who insists on his arrangement, in spite of the fact that other reports offer this information as well. This option is simply answerable. It only depends on power to take decisions of dissatisfied user. What is even worse is if the system demanded information in the form of any output do not offer at all. In this situation come into customer resolution investments. These can represent relatively demanding process in big companies, where all costs undergo to complex assent procedure. Demanded analytical outputs do not have to be only a product of information systems. Frequently is enough usage of instruments that the company once owns. It goes especially about MS Excel cost-free localized add-on of MS SQL Server Report Builder. End users profit by connection those applications with business database immediate access to current data without other costs. The only drawback is increased requirement on database security that needs to be ensured by administrators. What is a benefit is business information available to all involved users in the form, which they define in their favorite setting. In MS Excel it is related especially with dynamic pivot tables and diagrams, which work with operational OLTP and analytical OLAP data. In case of Report Builder is typically used for different types of traditional and interactive reports published on web sites, in intranet, sent by e-mails or advertised in other ways. Indispensable is the fact that between report layout and own data analysis passed only a short time. Experienced user can create report in familiar environment relatively fast. Both mentioned instruments, which were for data analysis selected (MS Excel due to its qualities and familiarity, Report Builder due to its price and possibilities), are excellent addition to analytical possibilities of MS SQL Server 2008. Among those belong multidimensional data analysis (OLAP) and also technology of data mining.

SEZNAM POUŽITÉ LITERATURY

[1] POWER, Dan. DSSResources.com [online]. 2005 [cit. 2010-06-07]. What is business intelligence?. Dostupné z WWW: <http://dssresources.com/faq/index.php?action=artikel&id=4>.

[2] POUR, Jan; SLÁNSKÝ, David. cssi.cz [online]. 2000 [cit. 2010-06-07]. Efekty a rizika Business Intelligence . Dostupné z WWW: <http://www.cssi.cz/cssi/system/files/all/SI_04_2_pour.pdf.>

[3] BOČEK, Jan. Extrahardware.cnews.cz [online]. 2009 [cit. 2010-06-07]. Historie počítačů VII. – Integrovaná generace. Dostupné z WWW: <http://extrahardware.cnews.cz/historie-pocitacu-vii-%E2%80%93-integrovana-generace>.

[4] HALCIN, Jakub. Galaxie.name [online]. 2005 [cit. 2010-06-07]. Příběh počítače (4.díl). Dostupné z WWW: <http://www.galaxie.name/index.php?clanek=pribeh-pocitace-4-dil&k=1>.

[5] POWER, Dan. Dssresources.com [online]. 2007 [cit. 2010-06-07]. A Brief History of Decision Support Systems. Dostupné z WWW: <http://dssresources.com/history/dsshistory.html>.

[6] VOCHOZKA, Josef. Ics.muni.cz [online]. 1999 [cit. 2010-06-07]. DSS a Internetovské dokumenty. Dostupné z WWW: <http://www.ics.muni.cz/zpravodaj/articles/154.html>.

[7] PETERKA, Jiří. Earchiv.cz [online]. 2005 [cit. 2010-06-07]. MIS, DSS, EIS. Dostupné z WWW: <hhttp://earchiv.cz/a94/a442c120.php3>.

[8] DANEL, Roman. Homel.vsb.cz [online]. 2005 [cit. 2010-06-07]. Homel . Dostupné z WWW: <http://homel.vsb.cz/~dan11/isys/Danel%20-%20IS%20-%20Datovy%20sklad.pdf>.

[9] HINCA, P. Použití datových skladů v pojistné matematice [online]. 2003 [cit. 2010-06-07]. Actuaria.cz. Dostupné z WWW: <http://www.actuaria.cz/upload/prednDW.doc>.

[10] Itnirvanas.com [online]. 2009 [cit. 2010-06-08]. ITNirvanas. Dostupné z WWW: <http://www.itnirvanas.com/2009/02/kimball-vs-inmon.html>.

[11] KALPAKIDIS, Kosmas. Braincourt [online]. 2010 [cit. 2010-06-07]. A historical summary. Dostupné z WWW: <http://braincourt.de/MIS-FIS-DSS-EIS-etc.116.0.html?&L=1&sjxvuuvej>.

[12] NOVOTNÝ, Ota; POUR, Jan; SLÁNSKÝ, David. Business Inteligence : Jak využít bohatství ve vašich datech. Praha : Grada, 2005. 254 s. ISBN 80-247-1094-3.

[13] SODOMKA, Petr. Informační systémy v podnikové praxi. . Brno : Computer Press, 2005. 352 s. ISBN 80-251-1200-4.

[14] LACKO, Lubomír . Datové sklady, analýza OLAP a dolování dat. Brno : Computer Press, 2003. 487 s. ISBN 80-7226-969-0.

[15] GÁLA, Libor; POUR, Jan; ŠEDIVÁ, Zuzana. Podniková informatika. 2., přepracované a aktualizované vydání. Praha : Grada, 2010. 496 s. ISBN 978-80-247-2615-1.

[16] HORÁK, Jiří ; HORÁKOVÁC, Bronislava. Datové sklady a využití datové struktury typu hvězda pro prostorová data [online]. 2004 [cit. 2010-06-08]. Gis.vsb.cz. Dostupné z WWW: <http://gis.vsb.cz/GIS_Ostrava/GIS_Ova_2007/sbornik/Referaty/Sekce3/hvezdaF4.pdf>.

[17] TVRDÍKOVÁ, Milena. Cssi.cz [online]. 2005 [cit. 2010-06-08]. Nástroje business inteligence - struktura a integrační charakter. Dostupné z WWW: http://www.cssi.cz/cssi/system/files/all/SI_05_2_tvrdikova.pdf.

[18] DANEL, Roman. Homel.vsb.cz [online]. 2005 [cit. 2010-06-07]. Homel . Dostupné z WWW: http://homel.vsb.cz/~dan11/isys/Danel%20-%20IS%20-%20OLAP.pdf>.

[19] JAŠA, Petr. Common.cz [online]. 2005 [cit. 2010-06-08]. Datové sklady. Dostupné z WWW: <www.common.cz/attachments/118_petr_jasa_datove_sklady.pdf>.

[20] PŮLPÁN, Jaroslav. Systemonline.cz [online]. 2001 [cit. 2010-06-08]. Dolování dat aneb Hledání skrytých souvislostí. Dostupné z WWW: <http://www.systemonline.cz/clanky/dolovani-dat-aneb-hledani-skrytych-souvislosti.htm>.

[21] VÍTEK, J. Datamining.xf.cz [online]. 2002 [cit. 2010-06-08]. Pojem Data Mining. Dostupné z WWW: http://datamining.xf.cz/view.php?cisloclanku=2002102702

[22] VÍTEK, J. Datamining.xf.cz [online]. 2002 [cit. 2010-06-08]. Pojem Data Mining. Dostupné z WWW: http://datamining.xf.cz/view.php?cisloclanku=2002102808

[23] BERKA, Petr. Sorry.vse.cz [online]. 2005 [cit. 2010-06-08]. Dobývání znalostí z databází. Dostupné z WWW: <http://sorry.vse.cz/~berka/docs/izi456/kap_1.pdf.>

[24] DANEL, Roman. Homel.vsb.cz [online]. 2005 [cit. 2010-06-07]. Homel . Dostupné z WWW: DANEL, Roman. Homel.vsb.cz [online]. 2005 [cit. 2010-06-07]. Homel . Dostupné z WWW: http://homel.vsb.cz/~dan11/isys/Danel%20-%20IS%20-%20OLAP.pdf>.

[25] ZAPAWA , Timothy . Microsoft Excel -- Získávání, analýza a prezentace dat . Computer Press : Brno, 2007. 430 s. ISBN 978-80-251-1535-0.

[26] HOTEK, Mike . Zvětšit obálku Hodnocení: průměrné hodnocení: 5,00 ohodnotit počet hodnocení: 1 Microsoft SQL Server 2008 Krok za krokem. Brno : Computer Press, 2009. 488 s. ISBN 978-80-251-2466-6.

[27] WALTERS, Rober, et al. Mistrovství v Microsoft SQL Server 2008 . Brno : Computer Press, 2009. 864 s. ISBN 978-80-251-2329-4.

[28] BAŠTÝŘ, Petr. Realit.cz [online]. 2010 [cit. 2010-06-08]. KPI jako nástroj snížení nákladů. Dostupné z WWW: <http://realit.cz/clanek/kpi-jako-nastroj-snizeni-nakladu>.

SEZNAM POUŽITÝCH SYMBOLŮ A ZKRATEK

MIS Management Information Systems (systémy určené k podpoře řízení)

DSS Decision Support Systems (systémy pro podporu rozhodování)

EIS Executive Information Systems (systémy pro vrcholové řízení podniku)

ERP Enterprise Resource Planning

CPM Corporate Performance Management (systémy určené k řízení podnikového výkonu)

BI Business Inteligence (manažerské rozhodování)

OLAP Online Analytical Processing (technologie uložení dat v databázi, která umožňuje uspořádat velké objemy dat tak, aby byla data přístupná a srozumitelná uživatelům zabývajícím se analýzou obchodních trendů a výsledků)

OLTP Online Transaction Processing (technologie uložení dat v databázi, která umožňuje jejich co nejsnadnější a nejbezpečnější modifikaci v mnohauživatelském prostředí)

ETL Extract, transform, and load (technologie plnění datového skladu)

EAI Extract, transform, and load (technologie plnění datového skladu)

DWH Data Warehouse (datový sklad)

DMA Data Marts (datové tržiště)

DSA Data Staging Areas (dočasné datové úložiště)

ODS Operational Data Store (operativní datové úložiště)

RDB Relační databáze

MDB Multidimenzionální databáze

ERP Enterprise Resource Planning (ERP) je informační systém, který integruje a automatizuje velké množství procesů souvisejících s produkčními činnostmi podniku.

SEZNAM OBRÁZKŮ

SEZNAM PŘÍLOH

PŘÍLOHA P I: KONCEPCE BI DEFINOVANÁ H.P.LUHNEM 1958

H. P. Luhn

A Business Intelligence System

Abstract: An automatic system is being developed to disseminate information to the various sections of any industrial, scientific or government organization. This intelligence system will utilize data-processing machines for auto-abstracting and auto-encoding of documents and for creating interest profiles for each of the "action points" in an organization. Both incoming and internally generated documents are automatically abstracted, characterized by a word pattern, and sent automatically to appropriate action points. This paper shows the flexibility of such a system in identifying known information, in finding who needs to know it and in disseminating it efficiently either in abstract form or as a complete document.

Introduction

Efficient communication is a key to progress in all fields of human endeavor. It has become evident in recent years that present communication methods are totally inadequate for future requirements. Information is now being generated and utilized at an ever-increasing rate because of the accelerated pace and scope of human activities and the steady rise in the average level of education. At the same time the growth of organizations and increased specialization and divisionalization have created new barriers to the flow of information. There is also a growing need for more prompt decisions at levels of responsibility far below those customary in the past. Undoubtedly the most formidable communications problem is the sheer bulk of information that has to be dealt with. In view of the present growth trends, automation appears to offer the most efficient methods for retrieval and dissemination of this information.

During the past decade significant progress has been made in applying machines to the processes of information retrieval. Automatic dissemination has so far been given little consideration; however, unless substantial portions of human effort in this area can be replaced by automatic operations, no significant over-all improvement will be achieved. Even the information retrieval processes mechanized so far still require appreciable human effort to organize the information before it is entered into machines.

It is believed that techniques now being developed will greatly contribute to the solution of the problem by extending automatic processes to the preparatory phases of mechanical information-retrieval systems, to the area of dissemination and to associated functions. Ideally, an automatic system is needed which can accept information in its original form, disseminate the data promptly to the proper places and furnish information on demand.

The techniques proposed here to make these things possible are:

1. Auto-abstracting of documents;
2. Auto-encoding of documents;
3. Automatic creation and updating of *action-point* profiles.

All of these techniques are based on statistical procedures which can be performed on present-day data processing machines. Together with proper communication facilities and input-output equipment a comprehensive system may be assembled to accommodate all information problems of an organization. We call this a *Business Intelligence System.*

Objectives and principles

Before the system operation is described, the term *Business Intelligence System* should be defined and the objectives and principles stated.

In this paper, *business* is a collection of activities carried on for whatever purpose, be it science, technology, commerce, industry, law, government, defense, et cetera. The communication facility serving the conduct of a business (in the broad sense) may be referred to as an *intelligence system.* The notion of *intelligence* is also defined here, in a more general sense, as "the ability to apprehend the interrelationships of presented facts in such a way as to guide action towards a desired goal."[1]

The term *document* is used to designate a block of information confined physically in a medium such as a

letter, report, paper or book. The term may also include the medium itself.

The objective of the system is to supply suitable information to support specific activities carried out by individuals, groups, departments, divisions, or even larger units. These are the *action points* previously referred to. To this end the system concerns itself with the admission or acquisition of new information, its dissemination, storage, retrieval and transmittal to the action points it serves.

More particularly the object of the system is to perform these functions speedily and efficiently, taking advantage of novel procedures which utilize the inherent capabilities of electronic devices.

One of the most crucial problems in communication is that of channeling a given item of information to those who need to know it. Present methods of accomplishing this are inadequate and the general practice is to disseminate information rather broadly to be on the safe side. Since this method tends to swamp the recipients with paper, the probability of not communicating at all becomes great. The Business Intelligence System provides means for selective dissemination to each of its action points in accordance with their current requirements or desires. This is accomplished by the mechanical creation of *profiles* reflecting the sphere of interest of each point and by updating these profiles as dictated by changes in the attitude of the respective action points and as recorded by the system on the basis of certain transactions.

Another problem in communication is to discover the person or section within an organization whose interests or activities coincide most closely with a given situation. Presently, the difficulty of finding such relationships often results in improper decisions, wrong actions, inaction, or duplication. An objective of the Business Intelligence System is to identify related interests by use of profiles of action points.

The problem of discovering information which has a bearing on a given situation has probably received the most attention in recent years, and various mechanical systems have been developed and put into operation. This phase of communication is commonly referred to as *information retrieval* or, more broadly, as the *library problem*. Information retrieval is necessarily a major function of the Business Intelligence System. Means are provided not only to integrate this function with the rest of the system but also to produce additional useful functions, as will be described later.

The achievement of these objectives is governed by principles essential to effective service and convenience of the user. Some of these are listed below:

1. Information admitted to the system includes communications, addressed to action points individually, which contain information of potential interest to other action points.
2. New information which is pertinent or useful to certain action points is selectively disseminated to such points without delay. A function of the system is to present this information to the action point in such a manner that its existence will be readily recognized.
3. Transmittal of information either as a result of dissemination or of retrieval is to be guided by progressive stages of acceptance by an action point. This procedure saves the recipient's time by reducing the amount of material to be transmitted and eliminating the non-pertinent material.
4. The system is to provide means for quickly discovering similarity of interests and activities that might exist amongst action points so that subjects and problems of common concern may be discussed and advanced through direct interchange of ideas between such points, if so desired.
5. The system is not to impose conditions on its user which require special training to obtain its services. Instead the system is to be operated by experienced library workers. Thus, in the case of an inquiry, the user will be required only to call the librarian, who will accept the query and will ask for any amplification which, in accordance with his experience, will be most helpful in securing the desired information.
6. Similarly, information lingering at an action point but of potential value to other action points is *mobilized* for efficient communication through inquiries of skilled reporters.

Description of the Business Intelligence System

The following description is given in rather general terms, and references to any specific type of *business* have been substantially avoided. Furthermore, the fact that certain devices are being referred to as implementation of the system, should not be interpreted as implying a specific size of the operation.

The description is given in accordance with main functional sections of the system, each illustrated by the diagram. Our assembly of these functional sections into a complete system is shown in Fig. 1.

Document input

Each document entering the system shown in Fig. 1 is assigned a serial number and is photographically reproduced on some medium such as microfilm. In those cases where the document has been addressed specifically to an action point, the original is promptly transmitted to the addressee. In all other cases the original is stored in a file for a reasonably short time and thereafter destroyed, unless there are reasons for preserving it for longer periods.

The microfilm copy of the document is transcribed onto magnetic tape by a human transcriber or a print-reading device. In those cases where the original document is available in machine-readable form, the transcription is done mechanically. The document is now available both as a microfilm copy and a magnetic tape record.

The microfilm copy is then recopied onto the storage medium of a *document microcopy* storage device. The microfilm record is stored elsewhere to constitute a microfilm master file which may serve to regenerate records in cases of emergency.

The magnetic tape record is now introduced into the auto-abstracting and encoding device. This device submits the document to a statistical analysis based on the physical properties of the text, and data are derived on word frequency and distribution. From these data the device then selects certain sentences of the document to produce an auto-abstract.[2] This is printed out, together with the title, author, and document serial number. This printout is photographically transferred onto the storage medium of the *auto-abstract microcopy storage* device.

The process of creating auto-abstracts consists of ascertaining the frequency of word occurrences in a document. A predetermined portion of the words of highest frequency is then given the status of *significant words* and an analysis is made of all the sentences in the text containing such words. A relative value of sentence significance is then established by a formula which reflects the number of significant words contained in a sentence and the proximity of these words to each other within this sentence. Several sentences which rank highest in value of significance are then extracted from the text to constitute the auto-abstract.

As soon as the auto-abstract has been created, the statistical data are further processed to derive an information pattern which characterizes the document. This process of *encoding* constitutes a further abstraction and involves procedures such as the categorization of words by means of a thesaurus.[3]

Useful patterns may be derived by listing a given portion of the words of highest frequency together with a selection of specific words. The interrelationship of words may also be indicated and certain frequently occurring combinations of words may be noted. Because of variation of word usage amongst authors the normalization of such words becomes an important function of encoding. Index lookup in a thesaurus-like dictionary will replace words, including those of foreign languages, by a notional family designation. The selection of specific words may also be accomplished by index lookup.

The document pattern derived by the above process is then transferred into a special pattern-storage device together with the title, author, and document serial number. This information is stored in coded form on a medium that may be subjected to serial scanning. As an alternative the resulting pattern may be rearranged and be distributed over a storage array to permit random access according to characteristics.

The tape or film transcript of the document may be stored in a library for reference if it later becomes necessary to change the method or scope of encoding.

Action-point profiles

As indicated earlier, one of the basic requirements of the system is the ability to recognize by mechanical means the sphere of interest and the type of activities that characterize each of the action points the system is to serve. This is accomplished by means of an information pattern similar to that of the documents.

Initially, the creation of these action-point profiles is best accomplished by having each action point create a document describing the various aspects of its activities and enumerating the types of information needed. Such documents are then introduced at the input of the system and are identified by action-point designation. The machine-readable transcripts of these documents are then described in connection with the document input. The resulting patterns are then stored in the Pattern Storage area in a special profile-storage device. Also stored, with each of these profile patterns, is the date of entry.

Selective dissemination of new information

Based on the document-input operation and the creation of profiles, the system is ready to perform the service function of selective dissemination of new information.

As soon as a new document has been entered into the system and its pattern developed, this pattern is set up in a comparison device which has access to all of the action-point profiles. The comparisons are carried out on the basis of degree of similarity, expressed in terms of a fraction, for each of the profile patterns. This fraction is subject to change as time goes on, depending upon conditions to be explained later.

Whenever a profile agrees to a given extent with a given document pattern, the serial number, title, and author of the affected document, together with the action-point profile designation, are transferred and stored in a monitoring device. This procedure is repeated for any subsequent similar occasion. The monitor is substantially a random-access storage device and has the functional capabilities of performing inventory operations. In this capacity it will transmit the serial number, title and author of the document in question to the desk printer at the selected action point and keep a record of this transaction.

Of the various ways in which such an announcement may be transmitted to the affected action points, the most effective one is by means of a printing device at each action-point location. An objective of the system is to command attention of the recipient. The use of individual printing devices is more effective than are centrally located devices serving several action points.

Selective acceptance of disseminated information

The dissemination of information so far has consisted in furnishing the action point with the serial number, title, and author of documents selected for it. This selection, however, is considered to be a provisional one, and the system withholds any further information if the action point can determine, on the basis of information given so far, that certain of the selected subjects are not of sufficient interest. If an announcement is of interest, and more detailed information on the subject is desired, the system will produce such information on demand. This step is initiated when the action point connects itself by telephone to the monitor and dials the serial numbers of the documents affected. Upon receipt of this message the monitor will relay an instruction to the microcopy storage

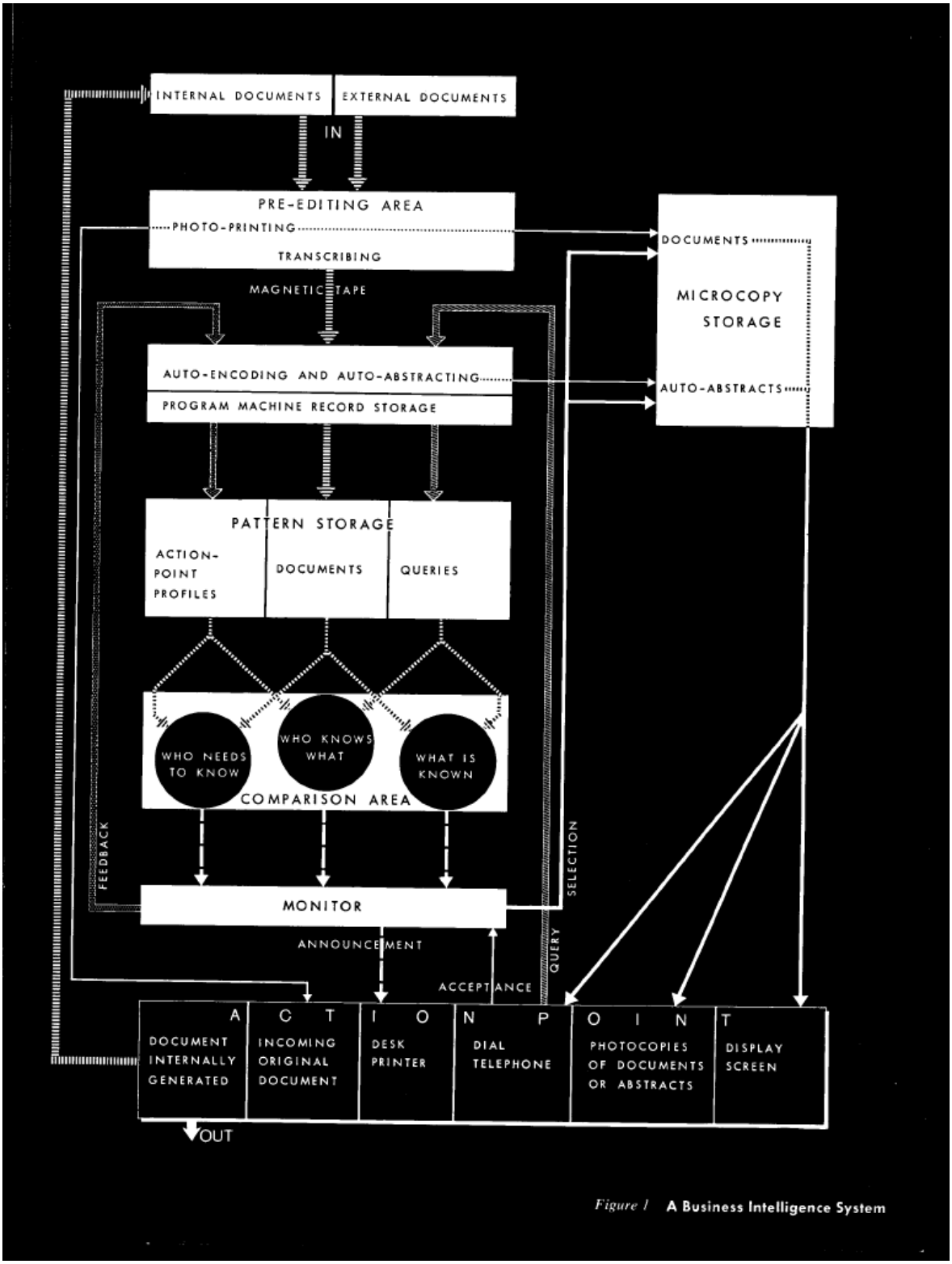


Figure 1 **A Business Intelligence System**

device to produce photoprints of the auto-abstracts of these documents and to mark them with the action-point designation. The auto-abstracts are then transmitted to the action point either in the form of a paper copy or by speedier means, such as Telefax or TV display.

The action point may now peruse the abstracts to determine which of the documents are desired in their entirety. These decisions are then entered into the system in the form of *acceptances*. An acceptance is made at an action point by dialing the document number, prefixed by a code symbol, whereupon the monitor will instruct the microcopy storage device to produce a photocopy of the complete document, properly marked with the action-point designation. These photocopies are then delivered to the action point.

The monitor will record the incidence of acceptance by modifying the affected records contained in its storage. At the same time the monitor will also instruct the auto-encoding device to transfer copies of the code patterns of the affected documents to the profile section of pattern storage, together with the identification of the action point involved and the date of transferral.

As a result of these operations the profile of a given action point has been updated to reflect interest in a currently communicated subject. As time goes on there is the probability that an increasing number of new documents will be announced to an action point because of possible shift of interests. In order to avoid such cumulative effects, the system is so arranged that the response to past interests is gradually relaxed. This relaxation is related to the date affixed to each new pattern that is superimposed on an action point's profile. Depending on the age of each of these patterns, an adjustment is made on the fraction of similarity that must be met in the comparison process of new documents. The older the profile pattern, the closer an agreement is needed for selection for dissemination, and consequently the fewer documents are selected. On the other hand those documents selected are more closely related to the original subject.

Information retrieval

This phase of the system concerns itself with the retrieval of those stored documents which might be relevant to a topic under consideration by an action point. The information to be discovered may vary widely and may consist of anything ranging from factual data to an extensive bibliography on a broad subject. Under the supervision of an experienced librarian the process of information retrieval is performed in the following way.

An action point telephones the librarian and states the information wanted. The librarian will then interpret the inquiry and will solicit sufficient background information from the action point in order to provide a document similar in format to that of documents normally entering the system. This query document is transmitted to the auto-encoding device in machine-readable form. An information pattern is then derived from the query document in a manner similar to that used for normal documents.

The resulting query pattern, together with a serial number and designation of the originating action point, is then sent to the queries section of the pattern-storage device. Subsequently, a copy of this query pattern is set up in the comparison device and is compared with all of the document patterns stored in the document-pattern storage device. This operation is similar to the one described in connection with selective dissemination. In the present case, the query pattern replaces the profile pattern.

Whenever similar patterns are detected by this means, the document designation is transmitted to the monitor, where it is registered and then announced to the action point.

Although the service of a librarian is considered a convenience to the action point, in certain cases, means may be provided at the action-point location to permit direct access to the system. This would be justified where many of the inquiries concern lookup-type retrieval of data.

When an action point desires information relative to a given document, the number of the document at hand would be dialed and instructions for search given to the monitor. Thereupon the monitor would select the corresponding pattern from document-pattern storage and provide instruction for use as a query pattern in the ensuing comparison operation.

Selective acceptance of retrieved information

The considerations which prompted the step-by-step acceptance of documents in the dissemination process are also applied to information retrieval. The processes employed, therefore, are identical.

The function of information retrieval, however, differs from that of dissemination in that the choice is not that of accepting or rejecting one document, but rather a selection of one or several from a special group of potentially relevant documents. Although in some cases a first search may have produced satisfactory references, in other cases the material produced may not be satisfactory. The action point must then relay this fact to the librarian and discuss with him how the searching procedure or the query should be modified so as to improve the probability of getting relevant material.

In those cases where pertinent information has been discovered, the acceptance of the complete documents of such information will cause the updating of the action-point profile, as was the case in dissemination. The query pattern will be impressed on the profile as a matter of course, whether or not the inquiry has been satisfied, so that new documents relevant to the subject of the inquiry will be made known subsequently.

Detection of an action point having given characteristics

In the process of transacting business it is often desired to determine who concerns himself with a given subject. The usual type of question asked is: "Who does or knows a certain thing?" A function of the Business Intelligence System is to answer questions of this type.

The manner in which this function is performed by the system is similar to the information retrieval procedure. However, instead of simulating a document pattern, a profile pattern is developed which represents most closely the characteristics of an action point sought. This synthetic profile is then compared with those in the profile storage and when a given degree of similarity is discovered, the identification of the affected action point is transferred to the monitor, together with the identification of the inquirer point. Thereafter the identities are announced by the tape-printing device at the inquiring action point so that personal contacts may be made.

Document output

The functions described so far have concerned themselves with documents admitted or acquired by the system from the outside. The document-output phase deals with internally generated documents. This type of document is essentially the product of action points and may be addressed to other action points within the organization or to external points. An objective of the system is to facilitate selective dissemination and retrieval of such documents in substantially the same way as for outside documents.

When a document has been created at an action point, a copy is produced, preferably in machinable form. This copy is then dispatched for processing to the input point of the system and the original is sent to the addressee.

Since this type of document is an indication of the interest of the originating action point, the information pattern derived by the auto-encoding process is not only stored in document-pattern storage but also is impressed on the profile of its originator, thereby updating it.

In the dissemination process this internally created document is announced to other action points in the same fashion as were outside documents.

Miscellaneous functions of the system

The comprehensive system for the various functions so far described is illustrated by Fig. 1. A number of additional useful functions which may be derived from the system are briefly described here.

It might be desirable to check each new document for duplication by comparing it with all of the documents in storage. Similarly a list of related documents may be prepared to serve as references applying to a new document.

When retrieving information it might be found advantageous to compare a query first with all the queries stored, in order to discover whether similar queries have been submitted in the past. If a list of the documents retrieved is available, the process of retrieval may be greatly simplified. This method may also be used to bring together the respective inquirers to furnish an opportunity to discuss the problems which apparently brought about similar inquiries. Periodic analysis of the profiles may also furnish valuable information on trends and possible overlapping of activities or interests.

Since a history of the usage of the system is stored in the monitor, an analysis of its records will disclose the efficiency of system operation. The findings may serve to adjust the system for optimum efficiency.

There are many details which might have to be provided to adjust the general form of the system to specific applications. One such requirement might be classification, by an editor, of documents with regard to security, proprietary interests and proper utilization of information.

A plurality of systems may be organized in hierarchical fashion, in which a first system would serve a number of more specialized systems. In this case the specialized system would each assume the role of an action point in the mother system.

It also appears quite feasible to share the system equipment among a number of organizations.

Prospects for establishing a Business Intelligence System

The system described here employs rather advanced design techniques and the question arises as to how far away such systems may be from realization. It may therefore be of interest to review the state of system and machine development.

The availability of documents in machine-readable form is a basic requirement of the system. Typewriters with paper-tape punching attachments are already used extensively in information processing and communication operations. Their use as standard equipment in the future would provide machine-readable records of new information. The transcription of old records would pose a problem, since in most cases it would be uneconomical to perform this job by hand. The mechanization of this operation will therefore have to wait until print-reading devices have been perfected.

The type of equipment required for processing information in accordance with the system is presently available as far as the functions are concerned. It is safe to assume that special equipment will eventually be required to optimize the operation.

The auto-abstracting and auto-encoding systems are in their early stage of development and a great deal of research has yet to be done to perfect them. Perhaps the techniques which ultimately find greatest use will bear little resemblance to those now visualized, but some form of automation will ultimately provide an effective answer to business intelligence problems.

References

1. *Webster's New Collegiate Dictionary*, G. & C. Merriam Co., Springfield, Mass.
2. H. P. Luhn, "The Automatic Creation of Literature Abstracts," *IBM Journal of Research and Development*, **2**, No. 2, 159 (April 1958).
3. H. P. Luhn, "A Statistical Approach to Mechanized Encoding and Searching of Literary Information," *IBM Journal of Research and Development*, **1**, No. 4, 309 (October 1957).

Received July 1, 1958