**Univerzita Hradec Králové**
**Fakulta informatiky a managementu**
**Katedra informatiky a kvantitativních metod**

# OPEN SOURCE V ÚLOHÁCH BUSINESS INTELLIGENCE

Bakalářská práce

Autor: Jan Poisl
Studijní obor: IM-3

Vedoucí práce: prof. RNDr. Hana Skalská, CSc.

Hradec Králové Duben 2017

Prohlášení:

Prohlašuji, že jsem bakalářskou práci zpracoval samostatně a s použitím uvedené literatury.

V Hradci Králové dne 28.4.2017 Jan Poisl

Poděkování:

Chci poděkovat všem, kteří mě podporovali a pomáhali jak při studiu, tak při tvorbě bakalářské práce. Obzvláště děkuji mé vedoucí bakalářské práce prof. RNDr. Haně Skalské, CSc. za metodické vedení práce, praktické rady při vedení této práce a odbornou pomoc.

Jan Poisl

# Anotace

Tato bakalářská práce je zaměřena na téma open source Business Intelligence systémů. Cílem práce je vysvětlení pojmů software, open source, informační systém, Business Intelligence a srovnání trhu Business Intelligence systémů a produktů. Neopomenutelnou součástí práce je charakterizace informačních systémů a typů úloh, kterým se tyto informační systémy věnují, přičemž pozornost je zaměřena především na charakterizaci úloh Business Intelligence. Praktická část této práce je věnována zpracování přehledu o vybraném open source Business Intelligence systému Pentaho. Tento systém byl vyzkoušen, podrobně popsán, byly vymezen úlohy, kterým se věnuje, provedlo se zhodnocení jeho schopností, doporučení pro typ podniku, kterému by mohl přinést největší prospěch a bylo provedeno celkové porovnání nabízených funkcí dle určitých kritérií.



# Annotation

## Title: Open source in Business Intelligence

This bachelor work is focused on the topic of open source Business Intelligence systems. The focus of this work is the explanation of terms software, open source, information system, Business Intelligence and comparison of market of Business Intelligence systems and products. The crucial part of this work is the characterization of information systems and types of tasks they are performing, but the special focus is characterization of Business Intelligence tasks. Practical part of this work is dedicated to creating summary of chosen open source Business Intelligence system Pentaho. This system had been tested, described in great detail, its important tasks were defined, enumeration of its possibilities was made, recommendation for types of business it could be the most beneficial was made and system was evaluated based on certain criteria.

# Obsah

# Seznam obrázků



# Seznam tabulek

# 1 Úvod

Informační systémy jsou nezbytnou součástí všech organizací, které pracují s daty. Vzhledem k neustále rostoucímu počtu lidí využívajících internet, internetové obchody, služby, cloudová řešení, zaměstnanecké či školní informační systémy a nespočet dalších systémů či služeb využívajících data, vzrůstá i celkový počet dat a informací, které je nutno organizacemi sbírat, skladovat, a především z nich vyvozovat správné důsledky a předvídání trendů. Existuje mnoho typů informačních systémů, přičemž každý typ systému je zaměřen na vykonávání specifické činnosti. V ideálním případě přispívají ke zvyšování produktivity organizace či společnosti, zjednodušují práci, vnitřní či vnější organizaci, zpřehledňují data, kterými organizace či společnost disponují.

Vzhledem k často enormnímu množství dat, kterými organizace či společnosti disponují, dochází k problému, kdy je těžké z tohoto enormního množství dat získat relevantní informace. Z tohoto důvodu je velmi zajímavé téma Business Intelligence systémů, jelikož cílem těchto systémů je sběr, analýza a interpretace těchto dat. Práce s daty ale není omezena pouze na Business Intelligence. Práci s velkým počtem dat se věnují i další obory jako například v současné době velmi populární big data. Big data se snaží řešit situaci, kdy je k dispozici takové enormní množství dat, že je nelze běžnými softwarovými prostředky zpracovávat. Téma této bakalářské práce ale předpokládá situaci, kdy je množství dat zpracovatelné.

Proprietární informační systémy se vyznačují svou vysokou cenou a uzavřeností kódu. Částečnou odpovědí na tento problém jsou open source informační systémy, které jsou často v základní verzi dostupné zdarma, popřípadě v různých placených schopnějších verzích. Z tohoto důvodu se mohou zdát open source informační systémy pro určité organizace a společnosti výhodně, ať už je důvodem cena či otevřenost a tím pádem upravitelnost kódu. Otázkou ale je, zdali tyto systémy nabízejí plnohodnotná Business Intelligence řešení.

# 2 Cíl práce

Práce je rozdělena na teoretickou část a praktickou část, přičemž každá z těchto částí má jiný cíl. V teoretické části je hlavním cílem rozšíření teoretických znalostí ohledně open source softwaru a Business Intelligence. Tyto teoretické znalosti budou poté dále využity v praktické části, která se už věnuje konkrétnímu Business Intelligence řešení a práci v něm.

Prvním cílem je tedy vymezení důležitých pojmů, které je nutné znát pro plné pochopení tématu této práce. Je nutné, aby byl vysvětlen pojem zdrojový kód a popsán, jaký je jeho smysl a z čeho je složen. Se zdrojovým kódem dále souvisejí pojmy jako open source, free software a proprietární software. Ke každému typu softwaru také náleží určitá licence, která upravuje podmínky používání tohoto softwaru. Každý, kdo by projevil zájem pracovat s nebo využívat open source Business Intelligence systémy, musí být obeznámen s různými typy licencí, které se mohou na tento software vztahovat a jejich legálními důsledky. Vzhledem k vysokému typu licencí a zaměření práce na open source Business Intelligence systémy, budou popsány a vysvětleny licence, které se váží k open source softwaru, jeho používání, upravování a dalšímu šíření. Další teoretickou součástí je obeznámení s pojmem *informační systém*, které druhy informačních systémů existují a k jaké práci reálně slouží. Poslední a největší téma, kterému se teoretická část práce věnuje, je vysvětlení Business Intelligence a charakterizace Business Intelligence úloh.

Druhým cílem této práce je praktické vyzkoušení vybraného open source Business Inteligence systému – Pentaho. K provedení tohoto cíle byl vybraný systém nainstalován, bylo zjištěno, které Business Intelligence úlohy je schopen vykonávat a bylo provedeno zhodnocení. Hlavním zaměřením praktické zkoušky systémů byly Business Intelligence úlohy popsané v teoretické části. Při hodnocení byla brána v potaz schopnost zpracování jednotlivých úloh. Při celkovém hodnocení byla ale brána v úvahu i uživatelská přívětivost, rozšiřitelnost daného softwaru, přidaná hodnota a potenciální unikátní vlastnosti, kterými systém může disponovat a tím pádem se lišit od ostatních.

# 3 Metodika zpracování

## 3.1 Hypotézy/výzkumné otázky

Vzhledem k tomu, že Business Intelligence není pouze jeden konkrétní proces, ale jedná se o souhrn procesů, je možné od sebe tyto procesy oddělit a konkrétně je definovat. Oddělením procesů lze získat seznam procesů, které jsou od Business Intelligence systému očekávány. V testovaném systému lze poté zjistit, které z těchto procesů obsahuje a zda se tím pádem dá označit za plnohodnotné Business Intelligence řešení.

Vzhledem k předchozím profesním zkušenostem byl vybrán BI systém Pentaho. Autorova hypotéza je, že Pentaho obsahuje všechny nezbytné BI procesy, a je možné jej označit za plnohodnotné Business Intelligence řešení.

## 3.2 Metodika

Metodika teoretické části práce spočívá především v práci s odbornou literaturou a internetovými zdroji.

Metodika praktické části práce spočívá v instalaci konkrétního open source Business Intelligence systému, ve vytvoření specifik a požadavků, podle kterých bude možné daný systém objektivně zhodnotit. Z tohoto zhodnocení lze potvrdit či vyvrátit položenou hypotézu.

# 4 Teoretická východiska

Vzhledem k vysoké specializaci určitých termínů, které tato práce používá, je ke správnému pochopení tématu nezbytně nutné znát důležité pojmy a termíny, které se budou dále v práci vyskytovat.

Z důvodu zaměření práce na *open source* systémy je nutno vysvětlit, co je *zdrojový kód* a jaké jsou výhody jeho volné dostupnosti. Z tohoto důvodu je nutné vysvětlit také pojem *free software* a jeho opak – *proprietární software*. Samotný fakt, že software má otevřený zdrojový kód ovšem neznamená, že není vázán podmínkami užití a omezeními. Z tohoto důvodu je nutné být obeznámen s *licencemi* a jejich omezeními. Vzhledem k tématu Business Intelligence je kromě něj také nutné vysvětlení a popis infomačních systémů, jejich typů a určení.

## *4.1 Zdrojový kód*

*„Forma, ve které jsou napsané softwarové programy. Většina softwaru je napsána v lidsky čitelné formě zvané zdrojový kód. Tato forma je poté procesována dávkovým počítačovým programem zvaným překladač, linker nebo builder, který vytvoří spustitelný program. Výsledek je nazván binární nebo objektový kód.“ (1)*

Každý program si lze představit jako set instrukcí. Když na počítači běží program jako například internetový prohlížeč či emailový klient, je počítačem volán soubor, který obsahuje spustitelné instrukce. Spustitelné instrukce jsou instrukce, které dokáže počítač přeložit a provádět dle nich určité akce. Programátoři ovšem nepíšou spustitelný kód, ale zdrojový kód (1).

Zdrojový kód si lze představit jako sadu instrukcí, zpravidla s využitím slov z anglického jazyka – tím více, pokud je celý zdrojový kód psaný v angličtině, což je častou konvencí i u organizací, které operují v zemích, kde není angličtina hlavním jazykem. Zdrojový kód je složen z instrukcí, které počítači „nařizují“, co má dělat. Pouhý zdrojový kód ale počítači k funkčnosti nestačí. Zdrojový kód musí být vložen do programu, kterému se říká překladač. Překladač přeloží zdrojový kód do objektového kódu, což je soubor instrukcí, které jsou srozumitelné pro procesor, kterým disponuje počítač.

## 4.2 Free software

*„Free software hnutí bylo přímou reakcí na privatizaci systému UNIX. Počítačoví odborníci, především akademici, se domnívali, že operační systémy potřebují být volně přístupné ve formě zdrojového kódu. Tudíž „free software" odkazuje na volnou přístupnost zdrojového kódu, nikoliv cenu." (1) [free může také znamenat zdarma – poznámka autora]*

Free Software Foundation (2) používá o free softwaru následující přirovnání:

„Myslete o něm jako o svobodě slova, ne jako o pivu zdarma." *(2)*

Meeker (1) vyjmenovává následující podmínky, které musí software splňovat, aby mohl být označen jako free software:

1) Volnost využívat program, pro jakékoliv účely.
2) Volnost studovat, jak program pracuje a volnost adaptovat tuto vědomost pro své potřeby.
3) Volnost redistribuovat kopie pro pomoc svému bližnímu.
4) Volnost vylepšit program a volnost vydání vylepšení pro veřejnost, aby měla z vylepšení užitek celá komunita.

## 4.3 Open source

*„Vývojářský model a druh licenční smlouvy. Vývojářský model referuje k modelu, pod kterým mnoho přispěvatelů z komunity může přispívat do kódové základny. Licenční model referuje k softwaru licencovanému pod dědičnými či nedědičnými dohodami, které dělají zdrojový kód volně přístupný." (1)*

Meeker (1) tvrdí, že k tomu, aby se software mohl nazývat open source, musí bezpodmínečně splňovat následující podmínky:

1) Volná redistribuce
2) Přístupný zdrojový kód
3) Povolená odvozená díla

4) Integrita autorova zdroje

### 4.3.1 Open source licence

Přestože open source software disponuje otevřenou kódovou základnou, může podléhat různým licencím, které upravují způsob, jakým mohou přispěvovatelé upravovat kód, specifické požadavky souvisejí s úpravou kódu, distribuci upraveného kódu a kombinování kódové základny s jinou kódovou základnou. V open source se lze nejčastěji setkat s následujícími licencemi:

## GPL

GPL je nejpoužívanější licence používaná u open source softwaru. GLP je hlídaná Free Software Foundation (FSF). Ten je také odpovědný za vydávání nových verzí. Jako první licence sama říká, že text licence nemůže být modifikován. Právní důsledek tohoto omezení není znám, jelikož legální dokumenty zpravidla nejsou předmětem ochrany autorských práv. Licence samotná nemůže být modifikována držiteli licence, na které je vázaná (1).

## GPL + Exception

GLP + Exception obsahuje speciální výjimku, která povoluje, aby jiný, proprietární kód mohl být zkombinován to stejného spustitelného souboru do kódu, který spadá pod tuto verzi licence. Tato verze platí pro standardní C knihovnu, která je asociovaná s GPL C překladačem. Pro veškerý kód generovaný tímto překladačem také platí GPL (1).

## GPL + FLOSS Exception

GPL + FLOSS Exception je variací GPL + Exception. Tuto licenci využívá MySQL AB. Tato výjimka umožňuje linkování do softwaru, který je pokrytý ne-GPL open source licencí, ale ne proprietárním kódem (1).

## LGPL

LGPL je známo jako „Lesser“ (menší, pozn. autora) GPL. Kód pokrytý LGPL může být použit jako dynamicky linkovaná knihovna kódem pokrytým jakoukoliv jinou licencí (1).

### *Omezení*

LGPL má dvě omezení. Prvním omezením je limitace použití pouze 10 řádků, pokud jej používá proprietární software. Druhým omezením je zákaz zpětného inženýrství. Z těchto důvodů je nutno dbát při implementaci LGPL knihoven zvýšené obezřetnosti (1).

## Korporátní dědičné softwarové licence

Korporátní dědičné licence byly pokusem o to, zachytit základní licenční paradigma GPL. První z těchto licencí byla Mozilla Public Licence. Pod tuto licenci spadal internetový prohlížeč Netscape po tom, co byl přetransformován na open source projekt. Mozilla Public Licence byl i základ pro prohlížeč Firefox. Tato licence je ovšem napsána tak, že ji nelze aplikovat na jiný kód bez modifikace, protože obsahuje výslovná ustanovení ohledně ochranné známky Mozilla a obsahuje explicitní reference na Mozilla kód. Licence byla z těchto důvodů kritizována pro nedostatek adaptability (1).

## Ostatní korporátní dědičné softwarové licence

Tyto licence nejsou založeny na předchozích korporátních dědičných softwarových licencích. Nejznámějším představitelem ostatních korporátních dědičných softwarových licencí je The Open Software Licence. Tato licence byla napsána právníkem jménem Lawrence Rosen. Jedná se o dědičnou licenci. Tato licence je volnější než GPL. Zajímavostí této licence je zahrnutí limitované původní záruky. Většina open source licencí se zříkávají jakýchkoliv záruk (1).

## Permisivní licence

V podstatě existují tři typy permisivních licencí. Keener (1) uvádí následující licence a jejich následující popis:

### *Berkeley Software Distribution*

Tato licence je nejpoužívanější permisivní open source licence. Skládá se z prohlášení o oprávnění, oznámení a upozornění. Někteří právníci znepokojeni tím, že „oprávnění“ není licence. V reálném použití ovšem k problémům nedochází. BSD licence v praxi obvykle není vynucována.

### *Apache 2.0*

Licence Apache 2.0 byla vydána roku 2004. Jedná se o permisivní licenci s velmi nenáročnými podmínkami.

### *Artistic licence*

Aristic licence (umělecká licence) je open source licence, jejíž autor je Larry Wall a jejím cílem bylo pokrytí PERL překladače. Tato licence obsahuje ustanovení jako například povinnost zahrnutí oznámení v každém změněném souboru, které konstatuje, jak a kdy byl soubor změněn.

## *4.4 Proprietární software*

*„Termín používaný ve free software komunitě pro software, který je licencovaný v binární formě pod komerčními licenčními podmínkami. Toto je chybné označení, jehož důsledkem je, že někteří lidé si myslí, že open source software není proprietární. Nicméně, oba open source a proprietární software jsou subjekty proprietárních autorských práv.“ (1)*

## *4.5 Systém*

*„Entity, které intuitivně uznáváme jako systémy, spolu sdílejí podobné charakteristiky, které adoptujeme jako abstraktní definici pojmu systém: každý [systém] je tvořen souborem komponent, které jsou nějakým způsobem propojeny tak, aby poskytovaly jeden kolektivní výsledek a obecný účel.“ (3)*

Každý systém může být charakterizován hranicemi, které oddělují jeho vnitřní komponenty od externího prostředí. Pokud je možno překračovat tuto

hranici oběma směry, systém je označen jako otevřený. Pokud hranici překračovat nelze, systém je označen jako uzavřený (3).

Pro technické systémy je často nezbytné posuzování výkonu systému. Výkon systému lze hodnotit dle následujících dvou kritérií:

### 4.5.1 Efektivita

Měřítka efektivity vyjadřují úroveň shody konkrétního systému úkolům, pro jejich řešení byl navrhnut. Ukazatelé výkonnosti jsou spojeny s výstupy systému. Jako výstupy systému si lze pro názornost představit například objem výroby, týdenní prodeje a výtěžek z prodeje. Zjednodušeně lze říci, že efektivita se posuzuje z toho, zda systém provedl korektní rozhodnutí (3).

### 4.5.2 Účinnost

Měření účinnosti vyzdvihují vztahy mezi vstupy a výstupy systému. Tato měření se zaměřují kvalitu transformace vstupních informací. Měření mohou například vyjadřovat množství zdrojů nezbytných k dosažení určitého objemu prodeje. Zjednodušeně lze říci, že účinnost se posuzuje podle toho, zda je akce prováděna nejlepším možným způsobem (3).

## *4.6 Informační systémy*

### Manažerské informační systémy

Tyto informační systémy se začaly objevovat na přelomu 60. a 70. let. Manažerské informační systémy zpracovávají data a poskytují finální výsledky ve formě takzvaných reportů. Tyto systémy však často nedosáhly výsledků, ať už z objektivních či subjektivních důvodů. I přes tyto nedostatky se ale staly východiskem a inspirací pro další, tentokrát už dokonalejší informační systémy (4).

### Systémy pro podporu rozhodování

Systémy pro podporu rozhodování jsou zaměřené na hlubší porozumění firemních dat a jejich následnému dalšímu využití pro manažerské účely. Oproti manažerským informačním systémům jsou schopny analyzovat data a zkoumat

z nich budoucí možný vývoj. Další schopností, kterou oplývají, je schopnost modelování situací a následně vyhodnocování možných dopadů těchto situací na firmu. Přibyla také takzvaná „*what-if*“ analýza.

U systémů pro podporu rozhodování lze říci, že se na rozdíl od manažerských informačních systémů soustředí na budoucnost – co bude; nikoliv na to, co už bylo. Zvýšená komplexita těchto systémů znamenala i vyšší uživatelské nároky. Tudíž se systémy začali pracovat analytici a databázoví specialisté místo původních samotných manažerů (4).

## Systémy pro podporu skupinové práce

Tento typ systému se začal využívat v 70. a 80. letech ve Spojených státech. Jeho tehdejší práce spočívala v koordinování 200–300 odborníků, kteří byli z různých částí zemí, a jejich úkolem bylo řešení krizových situací. Rozhodování se dělo na jednom místě – jednalo se zpravidla o místnost, která byla vybavená terminály, centrálním zobrazováním a samozřejmě programovým vybavením. Používá se dodnes. (4).

Systémy pro podporu skupinové práce zachovávají všechny výhody vyplývající ze skupinového rozhodování a zároveň potlačují negativní stránky skupinové práce (například obavy z prosazování vlastního názoru). Nevýhodou těchto systémů ale je to, že díky anonymitě účastníci mohou podceňovat rizika prosazovaného řešení situace (4).

## Exekutivní informační systémy

U exekutivních informačních systémů se dá mluvit o systémech, které jsou ušity na míru. Jejich další charakteristikou je zaměření na uživatelskou přívětivost a komfort a jejich intuitivní ovládání. Informace se zobrazují formou zpravodajství. I přes uživatelský komfort je od těchto systémů očekávaná schopnost podrobnější analýzy (4).

## Inteligentní systémy pro podporu managementu

Inteligentní systémy jsou dalším přiblížením k lidskému vnímání a uvažování. Systémy ze svého principu nemohou dokonale napodobit člověka, ale

mohou napodobit výsledky člověka a lidské zdůvodnění výsledků. Dobrým příkladem inteligentních systémů jsou expertní systémy. Expertní systémy fungují jako virtuální experti. Nevýhodou těchto systémů je jejich velmi úzké zaměření a také potřeba lidských expertů, kteří jsou ochotni podělit se o své znalosti a zkušenosti (4).

## 4.7 Business Intelligence

*„Business Intelligence (BI) představuje komplex procesů, aplikací a technologií IS/ICT, které téměř výlučně podporují analytické a plánovací činnosti podniků a organizací a jsou postaveny na principu multidimenzionality, kterým zde rozumíme možnost nahlížet na realitu z několika možných úhlů pohledu.“ (5)*

Business Intelligence je tedy kombinace propojených prvků – především datových zdrojů a proaktivních nástrojů. Toto spojení prvků je schopno reagovat na měnící se požadavky. Business Intelligence dále sjednocuje jednotlivé datové zdroje v organizaci či společnosti, a to z toho důvodu, aby nedocházelo k různé interpretaci stejných dat, jako tomu docházelo v dřívějších manažerských systémech (4).

Business Intelligence s sebou přináší několik prvotních problémů. Nejdříve je nutno najít cestu, jak reprezentovat nashromážděná data takovým způsobem, aby byla užitečná. Přičemž je také nutné postarání se o to, aby tato data pokrývala celou společnost či organizaci, popřípadě alespoň její velkou část. K tomu, aby tento proces byl proveditelný, je potřeba další administrativa (6).

Dalším problémem může být saturace dat. Mnoho datových skladišť roste neúměrně rychle z důvodů neschopnosti rozlišení důležitých a nedůležitých dat a neschopnosti říci „ne“ (7).

### 4.7.1 Reporting

Reporting je klíčovou aktivitou Business Intelligence. Reporting je obvykle podstatnou částí z řady aktivit, které organizace a společnosti provádějí se svými daty (6).

Janert (6) tvrdí, že reporting dat v současné době trpí dvěma velkými specifickými problémy:

1) Z důvodů hospodaření se zdroji jsou reportovací řešení vybudována genericky – jako jeden reporting systém, který podporuje všechny uživatelské potřeby. Z čehož vyplývá, že ve skutečnosti systém neslouží dobře nikomu.
2) Většina reportingu zaměňuje „*aktuální*“ s „*v reálném čase*“, což je důsledek toho, že data pro reporting jsou brána z databáze. Data z databáze zaručují aktuálnost, ovšem při jejich zpracování dojde ke zpoždění, které ve výsledku znamená fakt, že konečný report z konkrétních dat ztratí na své aktuálnosti.

### Reprezentativní reporty

Reprezentativní reporty jsou určené pro externí uživatele. Ať už se jedná o data pro čtvrtletní zhodnocování či pro zákazníky. Ve zkratce, reprezentativní reporty jsou všechny reporty, které jsou publikovány (6).

### Operační reporty

Na rozdíl od reprezentativních reportů jsou tyto reporty používány manažery při chodu společnosti. Jedná se o vnitřní informace, které nejsou specifikované. Může jít o informace od počtu objednávek, velikost katalogu, skladových zásob do vytíženosti procesorů na serverech.

## 4.7.2 Data mining

*„Data mining je cesta k vytvoření business intelligence z dat, která organizace sbírá, organizuje a skladuje. Široký rozsah data mining technik je využíván organizacemi k získání lepšího porozumění o jejich zákaznících a jejich operacích a k řešení komplexních organizačních problémů (8).“*

Turban, Sharda a Delen (8) tvrdí, že ačkoliv data mining stojí na základech tradiční statistické analýzy a umělé inteligence, získal si pozornost společností a uvádějí následující důvody k adopci data mining technik společnostmi:

- Intenzivnější konkurence v globálním měřítku, způsobená neustále se měnícími přáními a potřebami ve stále více saturovaném konkurenčním prostředí.
- Všeobecné uznání nevyužité hodnoty schované ve velkých zdrojích dat
- Konsolidace a integrace databázových záznamů, které zpřístupňují jednotný pohled zákazníků, prodejců, transakcí a podobně.
- Konsolidace databází a dalších zdrojů dat do jednoho místa ve formě skladiště dat.
- Exponenciální růst zpracovávání dat a technologií uchovávání dat.
- Významné snížení ceny hardwaru a softwaru pro skladování dat a jejich zpracování.
- Hnutí pro demasifikaci obchodních praktik.

## Využití

Důvodem, proč využívat data miningu může být velké množství dat generované internetem. Vzrůstá počet těchto dat a jejich komplexita. Velké množství dat může být také generováno různými vědními disciplínami jako například astrologie nebo atomová fyzika.

Za generování velkého množství dat mohou být zodpovědné také vědecké výzkumy. Za příklad vědeckých výzkumů, které často využívají data mining mohou posloužit lékařské výzkumy. Tyto výzkumy mohou používat data mining k identifikaci přesnějších způsobů diagnostiky a léčení nemocí, popřípadě jej mohou využít k objevování nových nebo zdokonalených léků (8).

Mezi komerční sektory využívající data mining technik patří finanční, obchodní a lékařské sektory. Turban, Sharda a Delen uvádí několik příkladů komerčního využití data miningu:

- Detekce a redukování podvodných aktivit
- Identifikace nákupních vzorů konkrétního zákazníka
- Znovuzískávání výnosných zákazníků
- Identifikace obchodních pravidel z historických dat
- Jako pomůcka ke zvyšování ziskovosti (analýza tržních košů)

- Zdokonalení zaměřování služeb na klienty

## Charakteristiky a výhody

*„Technicky řečeno, data mining je proces, který využívá statistické, matematické a umělou inteligenci využívající techniky pro extrahování a identifikaci užitečných informací a dalších vědomostí (nebo vzorů) z velkých souborů dat.“ (8)*

Nejedná se o novou disciplínu, jedná se především o novou definici pro využívání mnoha disciplín. Jako příklad disciplín, které data mining využívá, mohou sloužit disciplíny jako statistika, umělá inteligence, strojové učení, informační systémy a databáze (8).

Turban, Sharda a Delen (8) uvádí hlavní charakteristiky a prováděné úkoly data miningu jako:

- Data jsou často hluboce zahrabána uvnitř velmi velkých databází, které často obsahují data z několika let. V mnoha případech jsou data čištěna a konsolidovaná do datových skladů.
- Data mining prostředí je většinou klient-server architekturou (konkrétní uživatel používá aplikaci, která se připojuje na server) nebo webovou architekturou informačního systému.
- Sofistikované nástroje včetně pokročilých vizualizačních nástrojů pomáhají odstraňovat informace schované v korporátních souborech nebo archivovaných veřejných záznamech. Nalézání takových informací vyžaduje zasílání informací a synchronizaci dat k tomu, aby uživatel dostal správné výsledky.
- Miner (horník) je často koncový uživatel disponující data „vrtačkami“ a schopnými dotazovacími nástroji k tomu, aby se ptal ad-hoc otázky a obdržel odpovědi rychle, bez minimálních či žádných programovacích dovedností.
- Správný výsledek často spočívá nalezení nečekaného výsledku a vyžaduje, aby koncový uživatel myslel kreativně skrz celý proces, včetně interpretace nalezených výsledků.

- Data mining nástroje jsou snadno kombinovány s tabulkovými procesory a dalšími vývojářskými nástroji, což znamená, že nalezená data mohou být snadno analyzována a nasazena.
- Vzhledem k velkým počtům dat a masivním úsilí při hledání dat, je někdy nezbytné využívat paralelního zpracování.

## Data pro data mining

*„Pouze malá část času je věnována samotné analýze. Často je větší část času a snahy vynaložená na rozmanitých úkolech, které se mohou v porovnání s analýzou zdát triviální, ale jsou kriticky nezbytné: získávání dat; ověřování, čištění a formátování; pracování s aktualizacemi, místem a archivováním.“ (6)*

Mezi nejčastější zdroje dat ve firemním prostředí patří databáze a soubory se záznamy (takzvané logy). Tyto zdroje adresují jiné požadavky. Databáze obvykle obsahují data související s firmou a jejími obchody, zatímco logy jsou zdrojem „operačních“ dat. Pokud se uživatel ptá „co jsme prodali a komu?“, je k odpovědi využívaná databáze. Pokud se uživatel ptá „co jsme udělali a kdy?“, využívá „operačních“ dat z logů (6).

### *Čištění dat*

Janert (6) uvádí, že nezpracovaná data, ať už získaná z databází či ze záznamových souborů, typicky potřebují vyčištění a uvádí následující oblasti, které často vyžadují pozornost:

#### Chybějící hodnoty

Pokud individuální atributy nebo celé datové body chybí, je nutno rozhodnout o dalším naložení s těmito daty. Jako řešení se nabízí smazání celého záznamu, označení chybějící informace jako chybějící, doplnění dat. Volba naložení záleží na konkrétních specifických situacích a cílech.

**Neobvyklé hodnoty**

Obecně je nutno dbát zvýšené opatrnosti při odstraňování neobvyklých hodnot – existuje možnost, že neobvyklá hodnota je správná hodnota, kterou se hledá. Tyto hodnoty by se nikdy neměly odstraňovat bez vědomí oprávněných osob.

**Neužitečná data**

Data, která přicházejí přes síť, mohou obsahovat tzv. „non-printable“ znaky, popřípadě další neužitečná data podobného typu. Tato data jsou neužitečná a mohou v některých případech zmást aplikace, které se snaží procesovat data.

**Formátování**

Individuální hodnoty mohou být nevhodně naformátovány. Příkladem nevhodného formátování je text, který je celý psaný velkými či malými písmeny; mazání mezer mezi slovy nebo nahrazování tečkami; nahrazování časových údajů Juliánským kalendářem; nahrazování číselných datových typů znakovými a naopak.

**Duplikované záznamy**

Sety dat často obsahují duplikované záznamy, které je nutno správně identifikovat a odstranit.

**Spojování datových sad**

Potřeba spojování datových sad vyvstává často – například pokud data pochází z různých databází. Při spojování těchto dat je nutno dbát na to, aby data byla skutečně kompatibilní, zvláště pokud jsou databáze umístěny v různých částech světa. V takovém případě je nutno rozlišovat mezi časovými zónami, ale nesmí být opomenuty ani věci jako jiná měna. V případě různých jazyků je nutno být si vědom teoretických problémů s lokalizací, odlišné kódování použitých fontů, popřípadě jiný formát data.

#### Písařské chyby

Jako písařské chyby se dají označit data, která byla nesprávně zadána. Může se jednat o hodnotu 0.1 místo 0.01, nebo data vložená do nesprávných řádků, popřípadě sloupců.

### *Typy dat*

Při zpracování dat je nutno brát v ohledu fakt, že existuje více typů dat. Turban, Sharda a Delen (8) uvádí následující typy dat a jejich popis:

#### Kategorická data

Tato data reprezentují označení používaná při rozdělování do specifických skupin. Může se jednat o rasu, pohlaví, věkovou skupinu, popřípadě úroveň vzdělání. Přestože je možno reprezentovat některé skupiny stejným typem dat, například reprezentace věku a roku narození použitím čísla, je vhodné od sebe tyto skupiny rozdělit, aby byla zachována informativnost.

Kategorická data lze nazývat daty diskrétními, jelikož nejdou rozdělit na části, a to ani v případě, že jsou použita čísla. Čísla nejsou nic jiného než symboly a nenaznačují možnost kalkulace zlomkových hodnot.

#### Nominální data

Jako nominální data si lze představit jednoduché kódy přiřazené objektům jako označení, které nejsou měřením. Nominální data mohou být reprezentována binomickými hodnotami (ano/ne, pravda/lež), popřípadě hodnotami, ve kterých je na výběr z více než dvou možností (svobodný/ženatý/rozvedený).

#### Ordinální data

Ordinální data obsahují kódy přiřazené objektům nebo událostem, které zároveň reprezentují hodnocení mezi nimi. Může se jednat o dosažené vzdělání – základní, středoškolské, vysokoškolské.

**Numerická data**

Reprezentují číselné hodnoty specifických proměnných. Mezi typické příklady numerických dat patří věk, počet dětí, příjem, teplota. Numerická data mohou být nazvaná nepřerušovanými daty, pokud proměnná obsahuje nepřerušované hodnoty na specifické stupnici, která umožňuje vložení dílčích hodnot.

**Intervalová data**

Intervalová data jsou hodnoty, které mohou být měřeny na intervalové stupnici. Často používaným příkladem intervalového data je teplota.

**Poměrová data**

Zahrnují měřicí proměnné běžně se vyskytující ve fyzických vědách a inženýrství. Jako příklad poměrových data může sloužit hmotnost, délka, čas, úhel, energie a elektrický náboj.

**Ostatní data**

Mezi ostatní data lze řadit data, čas, nestrukturovaný text, obrázky a audio. Před tím, než je možno s těmito typy pracovat, je nutné provést jejich konverzi do kategorické nebo numerické reprezentace. Tato data lze také rozdělovat na statická či dynamická.

Některé metody data miningu mohou být specifické ohledně toho, jaké typy dat jsou schopné zpracovat. Dodání nekompatibilních datových typů může vést k nekorektním modelům, popřípadě zastavení vývoje modelu. Většina implementací algoritmů v široce dostupných softwarových nástrojích je schopna akceptovat kombinaci numerických a nominálních hodnot a poté interně provést nezbytnou konverzi před procesováním dat (8).

## Text mining

*„Text mining je polo-automatizovaný proces extrahování vzorců (užitečných informací a znalostí) z velkých objemů nestrukturovaných zdrojů dat.“ (8)*

Text mining má stejný smysl a používá stejné procesy jako data mining. Od data miningu se liší tím, že text mining pro vstup využívá nestrukturované (nebo jen velmi sporadicky strukturované) textové soubory jako jsou Microsoft Word dokumenty, PDF soubory, výňatky textu, XML soubory a další (8).

Výhody text miningu se prokazují především v oblastech, které generují velké obsahy dat. Mohou to být oblasti jako právo, akademický výzkum, finance, medicína, biologie, technologie a marketing (8).

Turban, Sharda a Delen (8) uvádí následující nejpopulárnější oblasti aplikace text miningu:

### *Extrakce informací*

Extrakce informací se skládá z identifikace klíčových frází a vztahů uvnitř textu vyhledáváním předdefinovaných vět v textu pomocí vyhledávání vzorů.

### *Sledování témat*

Sledování témat je založeno na uživatelském profilu a dokumentech, které uživatel sleduje, text mining dokáže předpovídat další dokumenty, které by mohly uživatele zajímat.

### *Sumarizace*

Sumarizací dokumentu se rozumí shrnutí obsahu dokumentu z důvodu ušetření času na čtenářově straně.

### *Kategorizace*

Kategorizací se rozumí identifikace hlavních témat dokumentu a správné zařazení dokumentu do předdefinovaných souborů kategorií.

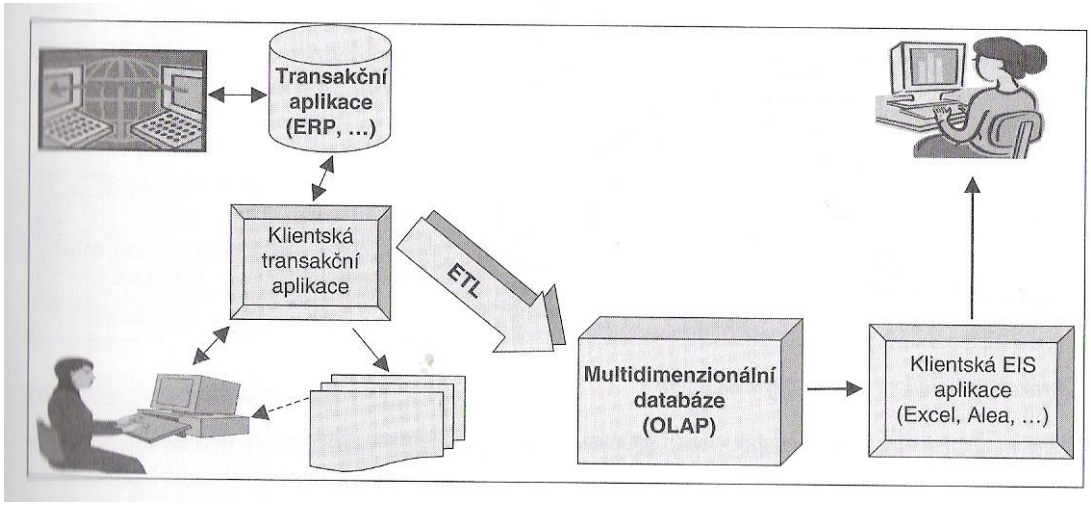
### *Shlukování*

Shlukování dokumentů se skládá ze shromažďování tematicky nebo typově podobných dokumentů bez předdefinovaného souboru kategorií.

## 4.8 Principy BI řešení

Vzhledem k tomu, že BI se využívá v mnoha různých variantách a podobách, neexistuje jednotná varianta řešení. Lze si ale představit základní řešení a komplexní řešení, které využívá další podpůrné BI produkty a technologie (5).

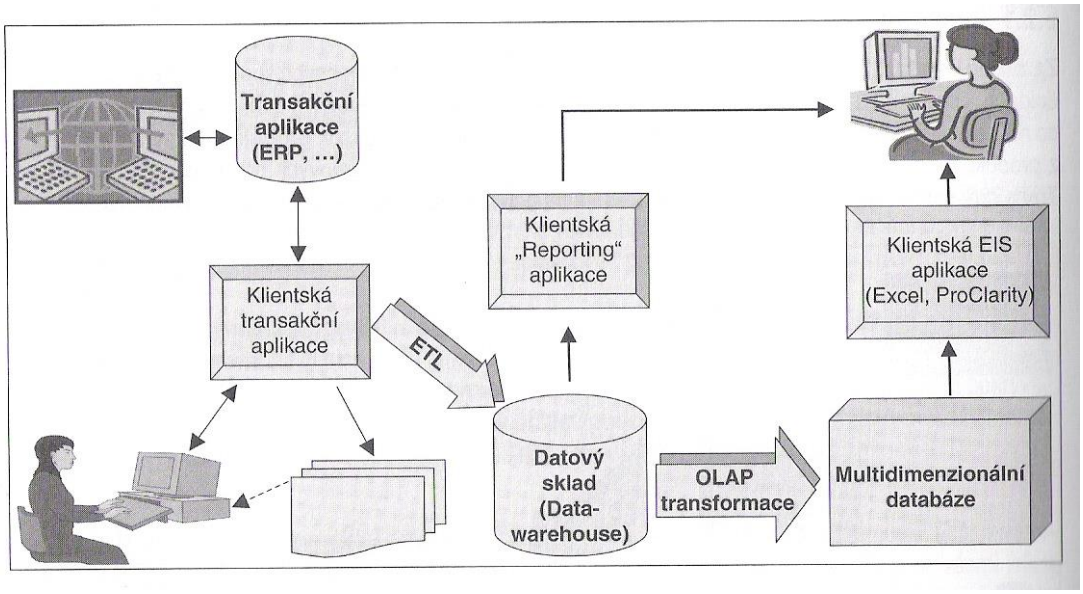
### 4.8.1 Základní BI řešení




**Obrázek 1 Základní řešení BI s multidimenzionální databází (5).**

Základním řešením lze rozumět řešení, které je postaveno na společných principech Business Intelligence. Řešení na obrázku dokumentuje řešení Business Intelligence aplikací, které využívají OLAP databázi. Data jsou uživateli zpřístupněna pomocí klientských aplikací (5).

### 4.8.2 Základní BI řešení s datovým skladem




**Obrázek 2 Řešení BI s datovým skladem (5).**

Toto základní řešení je doplněno o datový sklad, multidimenzionální databázi a reporting aplikace.

K plnému porozumění předchozích obrázků je nutné být obeznámen s prvky systémů, jejich smyslem a činnostmi, které provozují.

## Produkční (zdrojové) systémy

Produkční systémy jsou takové systémy, ze kterých BI aplikace čerpají data a zároveň nepatří do skupiny BI aplikací. Jejich architektura zpravidla podporuje ukládání a modifikaci dat v reálném čase. Produkční systémy nejsou navrženy pro vykonávání analytických úloh. Jedná se o hlavní vstup do BI, často jediný. V praxi jsou ovšem obvykle doplněny spektrem dalších vstupních systémů (5).

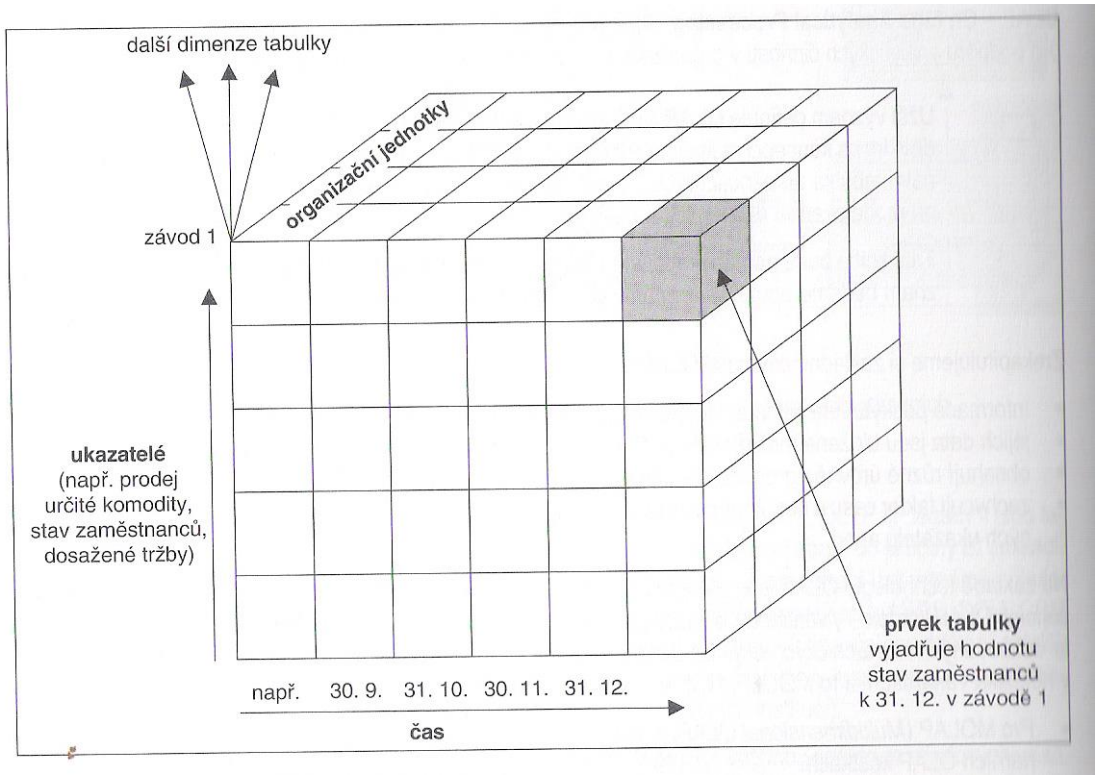
## ETL

ETL je zkratkou *Extraction, Transformation and Loading*, přeloženo na Extrakce, Transformace a Nahrávání. Někdy se ETL také říká *datová pumpa*. Stará se o získání a vybrání dat ze zdrojových systémů, dále upravení do požadované formy, vyčištění a nahrání do specifických datových struktur. Používají se na přenos dat mezi dvěma nebo více systémy. ELT nástroje pracují v dávkovém

režimu. Dávkový režim lze porozumět jako režim, kdy jsou data přenášena v určitých časových intervalech. V praxi se využívají denní, týdenní a někdy také měsíční intervaly (5).

## Multidimenzionální databáze

Multidimenzionální databáze jsou optimalizované pro ukládání a interaktivní využívání multidimenzionálních dat. Výhodou je rychlost zpracování a efektivní analýzy těchto dat (5).




**Obrázek 3 Princip multidimenzionální databáze (5).**

## *4.9 Skladiště dat*

*„Skladiště dat je nejpřednější sklad pro data dostupná pro vývoj Business Intelligence architektur a systémů pro podporu rozhodování. Termín skladování dat indikuje set aktivit ve vzájemném vztahu, které jsou zapojeny při designování, implementaci a užívání datového skladu.“ (3)*

Další definice skladiště dat může znít následovně:

*„Skladiště dat je subjektově-orientovaná, integrovaná, časově variabilní a stálá kolekce dat organizovaná pro podporu rozhodování v managementu. Skladiště dat od operačních databází odlišuje několik faktorů. Protože oba systémy poskytují zcela odlišné funkcionality a vyžadují jiné typy dat, je nezbytné udržovat skladiště dat oddělená od operačních databází.“ (9)*

### 4.9.1 Typy skladovaných dat

Existuje několik typů dat, která lze skladovat. Vercellis (3) uvádí následující tři typy dat, která se ukládají ve skladištích dat:

## Interní data

Vnitřní data se obvykle skladují na databázích, které se označují jako transakční systémy nebo operační systémy. Tyto systémy jsou páteří podnikových systémů. Tato data jsou shromažďována transakčními aplikacemi, které řídí některé operace společnosti, mezi něž patří administrace, účetnictví, produkce a logistika. Této kolekci aplikací se říká *enterprise resource planning* (ERP), česky plánování podnikových zdrojů.

## Externí data

Jako externí data se rozumí data, která nejsou shromažďována společností jako interní data, ale jsou zpravidla kupována od jiných společností, například od těch, která sbírají data o obchodování, tržním podílu, či předvídají trendy, popřípadě ekonomické a finanční indikátory. Tato data rozšiřují bohatost interních dat.

## Osobní data

Lidé provádějící Business Intelligence analýzu ve většině případů také spoléhají na informace, které jsou uloženy v tabulkových procesorech nebo lokálních databázích, které se nachází v jejich počítačích. Tato data jsou nazývána jako osobní data a je důležitá jejich integrace se strukturovanými daty z interních a externích zdrojů.

### 4.9.2 OLPT/OLAP

Aplikace, které jsou srdcem operačních systémů, jsou nazývány on-line transaction processing (OLTP). Celým kolekcím nástrojů zaměřených na provádění Business Intelligence analýz a podporu rozhodovacích procesů, se říká on-line analytical processing (OLAP) (3).

Vercellis (3) uvádí následující důvody pro implementaci skladiště dat odděleně od databází podporujících OLPT a jejich popis:

#### Integrace

V mnoha případech pracují systémy s informacemi, které pocházejí z mnoha zdrojů, distribuovaných po různých částech organizace, popřípadě pocházejí z externích zdrojů. Z tohoto důvodu je nutná data integrovat. Tuto integraci je možno dosáhnout různými metodami – jednotné kódovací metody, konverze do standardních měřících jednotek, dosahování sémantické homogenity informací.

#### Kvalita

Data převáděná z operačních systémů do skladišť dat jsou z důvodu dosažení spolehlivých a bezchybných informací zkoumána a korigována. Zkoumání a korigování dat zvyšuje praktickou hodnotu Business Intelligence systémů.

#### Efektivita

Některé dotazy zaměřené na extrakci dat souvisejících s Business Intelligence analýzou mohou způsobovat vysoké zatěžování výpočetních zdrojů a času potřebného na jejich procesování. Pokud by takový dotaz byl nasměrovaný do transakčních systémů, byla by riskována efektivnost systému, na které závisí plánovací aplikace a rutinní činnosti organizace. Z tohoto důvodu je vhodnějším řešením směřování těchto dotazů pro OLAP analýzu do datového skladiště.

#### Rozšiřitelnost

Data uložená v transakčních systémech mají z důvodů paměťové kapacity rozsah pouze v limitovaném časovém horizontu. Data vztahující se k minulým obdobím jsou archivována a poté mazána z OLPT.

Toto chování přímo koliduje s Business Intelligence systémy, které zpravidla potřebují historické informace z důvodů předvídání trendů a detekce opakujících se vzorců chování. Z tohoto důvodů je vhodnější řešení užití datového skladiště, která uchovávají historické informace.

## Zaměření na entity

Data uložená v datových skladištích jsou primárně zaměřena na hlavní entity, pro které se provádí analýza – entity jako produkty, zákazníci, objednávky a prodeje.

Na druhou stranu, transakční systémy jsou zaměřeny na operační aktivity. Orientace na entity během Business Intelligence analýzy umožňuje snadnější vyhodnocení výsledků a jednodušší detekci zdrojů neefektivnosti.

## Integrace

Všechna data pocházející z různých zdrojů jsou před nahráním do datového skladu homogenizována. Pod homogenizací si lze například představit harmonizaci měřících jednotek.

## Časové varianty

Všechna data vstupující do datového skladiště jsou označena časovým úsekem, na který odkazují. Díky tomu může Business Intelligence analýza zjišťovat historické trendy a trendy předvídat.

## Trvalost

Poté, co jsou data nahrána do datového skladiště, už obvykle nejsou modifikována a jsou skladována permanentně. Tato vlastnost umožňuje snadnější organizaci read-only (pouze pro čtení) přístupu pro uživatele. Také zjednodušuje proces aktualizace.

## Konsolidace

Některá data uložená v datových skladech jsou obdržena jako souhrn primárních dat patřících do operačních systémů, ze kterých pocházejí. Příkladem

může být mobilní operátor, který pro každého zákazníka skladuje týdenní celkovou cenu za volání, místo toho, aby skladoval každý jednotlivý hovor. Důsledkem je úspora místa potřebného na skladování takovýchto informací.

## Denormalizace

Na rozdíl od operačních databází nejsou data skladována v datových skladištích strukturována v normální formě. Důvodem je zkrácení operačního času pro komplexní dotazy.

### 4.9.3 Srovnání charakteristik OLPT a OLAP

Vercellis (3) uvádí následující srovnání charakteristik OLPT a OLAP:

| Charakteristika | OLTP | OLAP |
|---|---|---|
| Typ dat | dynamická data | statická data |
| Čas | pouze aktuální data | aktuální a historická data |
| Časová dimenze | implicitní a aktuální | explicitní a variabilní |
| Detailnost | detailní data | agregovaná a konsolidovaná data |
| Aktualizování | neustálé a nepravidelné | periodické a pravidelné |
| Aktivity | opakující se | nepředvídatelné |
| Flexibilita | nízká | vysoká |
| Výkon | vysoký, několik sekund pro dotaz | může být nízká pro komplexní dotazy |
| Uživatelé | zaměstnanci | znalostní pracovníci |
| Funkce | operační | analytická |
| Důvod užití | transakce | komplexní dotazy, podpora rozhodování |
| Priorita | vysoký výkon | vysoká flexibilita |
| Měřítko | míra transakcí | efektivní odpověď |
| Velikost | MB až GB | GB až TB |

**Tabulka 1 Srovnání charakteristik OLPT a OLAP. Přeloženo.**

### 4.9.4 Kvalita dat

Kvalita dat je nikdy nekončící záležitostí pro pracovníky, kteří jsou zodpovědní za design a aktualizace skladišť dat. Vercellis (3) uvádí následující faktory, které mohou mít vliv na kvalitu dat:

## Přesnost

Data musí být vysoce přesná. Například je bezpodmínečně nutné kontrolovat správnost jmen a kódování – zda jsou správně uvedeny a zda jsou v korektních rozsazích.

## Úplnost

Aby byla data použitelná, neměla by obsahovat velké množství chybějících dat. Zároveň ale platí to, že většina nástrojů používaných pro učení a data mining je schopna minimalizovat efekt částečné neúplnosti dat.

## Konzistence

Forma a obsah dat po integračních procedurách musí být konzistentní nehledě na to, z jakého zdroje tato data pochází. Všechna data musí obsahovat stejné jednotky, ať už se jedná o čas, měnu či jiné jednotky.

## Včasnost

Data musí být frekventovaně aktualizovaná. Obvyklá je aktualizace dat pravidelně každý den, popřípadě nejdéle každý týden.

## Zamezení redundancím

Z důvodů prevence plýtvání paměti a nekonzistentnosti dat musí být zamezeno opakování a redundanci dat.

## Relevantnost

Data musí být relevantní vzhledem k požadavkům Business Intelligence systému, aby bylo zaručeno přidání reálné hodnoty v prováděných analýzách těchto dat.

## Interpretovatelnost

Měl by být znám význam a smysl dat. Data by také měla být správně interpretována analytiky.

### Přístupnost

Data musí být jednoduše přístupná pro entity, které je využívají – typicky analytikové a rozhodovací aplikace.

# 5 Business Intelligence v oblastech lidské činnosti

Business Intelligence (dále jen BI) technologie je možno využít v oblastech lidské činnosti, ve kterých je třeba sledování a vyhodnocování určitých hodnot. Gála, Pour a Prokop uvádí následující lidské činnosti a jejich popis:

## *BI v oblasti prodeje*

Aplikace BI se již od počátku orientovaly na tuto konkrétní oblast, protože klíčovou částí této oblasti je výkonnosti a kvalita řízení prodeje. Aplikace BI v této oblasti se postupně rozvíjely a provazovaly s dalšími aplikacemi jako jsou řízení marketingu, nebo analytické CRM.

## *BI v oblasti nákupu*

Umožňují efektivně podpořit všechny činnosti související s plánováním a pořizováním materiálu. Cílem je snížení celkových nákladů za nákup, zvýšení efektivnosti a kontroly nad nákupy v podniku. Problémem této oblasti je nutnost spoléhat na kvalitu dat poskytovaných dodavateli. Tato data se analyzují a poté BI porovnává podmínky nákupů, dodávek, slev, příplatků a podobně.

## *BI v oblasti dopravy*

Škála aplikací v oblasti dopravy je rozsáhlá a můžeme do ní řadit analýzy efektivnosti dopravců, analýzy dopravních nákladů, doby dodávek a podobně.

## *BI v oblasti marketingu*

Cílem aplikací BI v této oblasti je zejména podpora plánování marketingových kampaní a následné vyhodnocení jejich dopadu. Z tohoto důvodu například zahrnují analýzy portfolia produktů a služeb, klasifikace a segmentace zákazníků, správu marketingových kampaní a podobně.

### *BI v oblasti financí*

BI v této oblasti například poskytuje analýzy ukazatelů finanční výkonnosti organizace, jejich závodů, nákladových středisek, projektů, skupin produktů a informace o odchylování se od plánovaných hodnot. Součástí funkcionality BI je obvykle finanční plánování, prognózování, simulace finančního vývoje organizace, finanční výkaznictví a konsolidace, analýza nákladů a ziskovosti, podpora při řízení finančních rizik a různé finanční optimalizace.

### *BI v oblasti řízení lidských zdrojů*

Aplikace BI v této oblasti obvykle využívají velké podniky. Aplikace BI obvykle zahrnují analýzu pracovní síly, analýzu nákladů na pracovní sílu a podobně.

### *BI v oblasti řízení výroby*

Řízení výroby zpravidla probíhá ve spojení s řízením kvality a jedná se o jednu z klíčových domén BI aplikací. Součástí efektivního řízení je přehled o vývoji a stavu výroby a kontrola jakosti.

V aplikaci BI je zahrnuto monitorování klíčových ukazatelů tohoto procesu jako je doba dodávky oproti plánu, rozpracovaná výroba, doba trvání výrobního cyklu, průchodnost výrobní linky, obrat zásob, ziskovost, kvalita výrobků. Tento monitoring lze provádět přes jednotlivé útvary – závody, výrobní linky, dílny, sklady a další útvary společnosti.

### *BI v oblasti webové analýzy*

Webová analýza se zabudovanou funkcionalitou BI slouží pro měření a analýzu ukazatelů získaných z provozu webových aplikací. Uživatelé těchto řešení jsou zpravidla tvůrci webových stránek, dodavatelé obsahu a zákazníci. Cílem těchto aplikací je poskytování statistických údajů pro uživatele internetových serverů (může to být například počet přístupů na jednotlivé stránky, rozdělení v čase, podle typu prohlížeče, operačního systému), analýzu pohybu návštěvního po stránkách a tak dále.

# 6 Trh s produkty Business Intelligence

Trh s BI produkty a službami má pouze krátkou historii, která se ale v současné době ale velmi dynamicky rozvíjí. Pro porozumění trhu je nutné být obeznámen s kategoriemi Business Intelligence produktů a následně s jejich výrobci.

## *6.1 Kategorie BI produktů*

### 6.1.1 Databázové systémy

Databázové systémy v kontextu řešení BI vystupují na straně zdrojových systémů a na druhé straně v roli datových skladů, tržišť, operativních datových skladů i dočasných úložišť dat. Při návrhu a provozu těchto systémů je nutno dbát na následující významné parametry:

- Rozsah datových zdrojů v gigabytech (GB), někdy i terabytech (TB)
- Kapacitní možnosti skladu/tržiště, tím se myslí kapacita v GB/TB při ještě přijatelném výkonu
- Doba odezvy na průměrně složité dotazy při standardním objemu dat
- Kvalita dat a počet chyb přenesených do BI databází
- Úroveň zastarávání dat, časové zpoždění
- Formáty dat – kromě strukturovaných dat i data nestrukturovaná (text, obrázky)
- Existence speciálního jazyka pro multidimenzionální aplikace
- Integrace OLAP nástrojů do základního databázového prostředí

### 6.1.2 ETL Nástroje

Tyto nástroje vykazují silnou dynamiku vývoje. ETL nástroje konvergují s nástroji pro správu metadat a s nástroji pro zajištění datové kvality. Specifické parametry datových pump jsou následující:

- Podporované vstupní a výstupní platformy a jejich konektivita
- Podpora pro správu metadat

- Možnosti vývoje transformačních aplikací
- Možnosti plánování provozu, spouštění datové pumpy dle časového plánu
- Úroveň podpory a dokumentace workflow
- Úroveň dokumentace a protokolování provozu datové pumpy
- Podpora zpracování dat v reálném čase

### 6.1.3 Analytické aplikace a nástroje

Tato kategorie zahrnuje OLAP a reportingové nástroje, a to včetně jejich prezentačních vrstev. Mezi specifické parametry patří:

- Podporované OLAP modely – MOLAP, ROLAP, HOLAP
- Výkon, doba odezvy u srovnatelných požadavků na analýzy
- Integrované dokumentační funkce OLAP řešení
- Možnost změn OLAP databází v průběhu provozu
- Systém přístupů podle ukazatelů a úrovní
- Možnosti zpětného zápisu do OLAP kostky

### 6.1.4 Data Mining technologie

Představuje kategorii technologií, které jsou specifické a jsou charakterizovány těmito parametry:

- Poskytovaná funkcionalita standardními nástroji pro dolování dat
- Úroveň grafických výstupů, prezentační možnosti výsledků složitých analýz
- Výkon, doba odezvy u jednotlivých typů odezvy (brán v potaz standardní vzorek dat)

### 6.1.5 Nástroje řízení kvality dat

V rámci růstu nasazení datových skladů a s nimi souvisejících analytických či manažerských aplikací se zvyšuje i význam nástrojů pro zajištění kvality dat. Je vysoce důležité, aby provoz a užití probíhaly nad korektními daty, které dokumentují reálnou situaci podniku. Specifické parametry jsou zejména:

- Rozsah kontrolních a opravných funkcí
- Rozsah a forma protokolování kontrolních a opravných funkcí

### 6.1.6 Klientské nástroje

Těmito nástroji se rozumí uživatelské rozhraní k relačním databázím datového skladu a dalším databázím. Slouží k vizualizaci dat a snaží se maximálně zjednodušovat přípravu analytických aplikací a reporting. Tyto nástroje vyžadují následující specifické parametry:

- Lokalizace klientských nástrojů – především z pohledu jazykového prostředí
- Jednoduchost ovládání, úroveň uživatelského rozhraní
- Flexibilita v nastavování a generování výstupů
- Podpora grafiky, klikovacích map
- Úroveň integrovaného vývojového prostředí
- Integrace na WWW prostředí

### 6.1.7 Standardní aplikace

Standardní aplikace BI mají v podstatě podobné parametry ostatních standardních programových balíků. Mezi nejvýznamnější patří:

- Funkcionalita standardních aplikací,
- Úroveň standardních analytických a plánovacích funkcí
- Využití matematického a statistického aparátu
- Možnosti customizace (přizpůsobení), standardních aplikací
- Možnosti standardního nastavení výstupů
- Úroveň lokalizace – přizpůsobení produktů po jazykové a legislativní stránce

## *6.2 Výrobci a jejich podíl na trhu*

V roce 2015 celkový trh s Business Intelligence systémy vzrostl o 3 %. Změny devizových kurzů v roce 2015 měly značně negativní dopad na trh, pokud na něj nahlížíme v amerických dolarech. Pokud by se na trh nahlíželo v konstantní měně,

tak by trh vzrostl o 11,7 %. Nepříznivý směnný kurz měl negativní následky především na prodejce, kteří čerpají svůj zisk mimo USA (10).

### 6.2.1 Celosvětový zisk z Business Intelligence a analytického softwaru podle prodejců

Celosvětový zisk za rok 2015 pro výrobce s vysokým ziskem je shrnut v následující tabulce:

| | Zisk (miliony $) | | | Podíl (%) | | | Růst (%) | |
|---|---|---|---|---|---|---|---|---|
| | 2013 | 2014 | 2015 | 2013 | 2014 | 2015 | 2013-2014 | 2014-2015 |
| SAP | 2163,1 | 2305,0 | 2013,5 | 17,0 | 17,0 | 14,4 | 6,6 | -12,6 |
| Microsoft | 1255,3 | 1356,5 | 1488,2 | 9,9 | 10,0 | 10,7 | 8,1 | 9,7 |
| SAS | 1240,4 | 1298,1 | 1376,2 | 9,8 | 9,6 | 9,9 | 4,6 | 6,0 |
| IBM | 1603,3 | 1529,7 | 1374,1 | 12,7 | 11,3 | 9,8 | -4,8 | -10,2 |
| Oracle | 1045,5 | 1059,2 | 1038,4 | 8,2 | 7,8 | 7,4 | 1,3 | -2,0 |
| Tableau Software | 225,2 | 399,3 | 630,6 | 1,8 | 2,9 | 4,5 | 77,3 | 57,9 |
| Qlik | 431,3 | 503,0 | 556,5 | 3,4 | 3,7 | 4,0 | 16,6 | 10,6 |
| MicroStrategy | 437,6 | 444,0 | 428,7 | 3,4 | 3,3 | 3,1 | 1,5 | -3,4 |
| MathWorks | 263,3 | 280,8 | 300,5 | 2,1 | 2,1 | 2,2 | 6,7 | 7,0 |
| Information Builders | 188,1 | 190,0 | 191,0 | 1,5 | 1,4 | 1,4 | 1,0 | 0,5 |
| Palantir | 55,3 | 130,8 | 163,8 | 0,4 | 1,0 | 1,2 | 136,8 | 25,2 |
| TIBCO | 201,5 | 152,6 | 141,3 | 1,6 | 1,1 | 1,0 | -24,3 | -7,4 |
| Panorama Software | 101,2 | 107,6 | 115,5 | 0,8 | 0,8 | 0,8 | 6,4 | 7,3 |
| Mezisoučet | 9214,0 | 9756,8 | 9818,2 | 72,6 | 71,9 | 70,3 | 5,9 | 0,6 |
| Ostatní | 3482,5 | 3804,3 | 4143,6 | 27,4 | 28,1 | 29,7 | 9,2 | 8,9 |
| Celkově | 12696,5 | 13561,1 | 13961,8 | 100,0 | 100,0 | 100,0 | 6,8 | 3,0 |

**Tabulka 2 Celosvětový zisk z Business Intelligence a analytického softwaru podle prodejců. Zdroj: Vesset, Schubmehl, Olofson, Gopal, & Bond, 2016. Upraveno.**

# 7 Praktická část

Praktická část této práce je rozdělena na vysvětlení obsahu dat, se kterými lze pracovat a poté na zkoumání Community edice softwarového balíku Pentaho.

V první části praktické části bylo vysvětleno, co je data set a jaké je jeho využití v Business Intelligence procesu. Následně byla nalezena místa, na kterých jsou k dispozici data sety ke stažení a následnému využití. Byly také specifikovány některé formáty, v kterých se může data set vyskytovat.

Druhá část byla věnována samotnému zkoumání Pentaho Community Edition. V první řadě byly specifikovány kategorie funkcionalit, které vycházely z teoretické části této práce. Po specifikování kategorií byly staženy a zprovozněny všechny komponenty, které jsou součástí Pentaho. Po zprovoznění komponent bylo následně zkoumáno, zda s využitím pouze těchto komponent lze vyhovět specifikovaným očekávaným funkcionalitám. V případě, že byl zjištěn negativní výsledek, bylo zjištěno, zda existuje alternativa, která by danou funkci mohla zastat.

Následně bylo provedeno shrnutí výsledků, kterých bylo dosaženo a bylo provedeno zhodnocení celkového balíku Pentaho. Ze shrnutí výsledků bylo následně proveden závěr a doporučení.

## 7.1 Data set

Vzhledem k tomu, že Business Intelligence se zabývá především prací s daty, je k jeho využívání nutno disponovat určitou bází dat, se kterou může software pracovat.

Existuje několik možností, jak může jednotlivec získat bází dat, kterou může dále využít k analýze pomocí BI softwaru. Jedna z možností je vytvořit si bázi dat ručně. Velkou nevýhodou takového přístupu je fakt, že báze dat není tvořena reálnými daty, popřípadě nedisponuje dostatečném množství reálných dat.

Další z možností je využít již existující data sety od třetích stran. Pro toto využití existují data sety, které jsou šířeny a k dispozici jako *public domain*. V tomto případě je možno dílo volně šířit, využívat a upravovat, bez nároku na další ochranu díla (11). V takovém případě se zpravidla jedná o data sety, které jsou vytvořeny buď vládními subjekty, popřípadě jsou součástí nebo výsledkem výzkumné práce. Komerční společnosti si zpravidla své data sety z pochopitelných důvodů střeží.

Volně přístupným data setům se věnuje množství různých webů, z nichž stojí za pozornost KDnuggets (12), popřípadě velmi rozsáhlý rozcestník vytvořený uživatelem Xiaming Chen vystupující pod přezdívkou caesar0301 (13). Poslední zmíněný nabízí nepřeberné množství data setů rozdělených do třiceti kategorií, v rozsahu od zemědělství, biologii, přes vzdělání až po cestování.

Formáty těchto setů dat se mohou značně lišit. Některé data sety existují například jako soubor pro naplnění databáze, některé data sety jsou reprezentované sbírkou .CSV souborů, některé sety zase například existují v podobně JSON souborů. Ať už je volba uživatele jakákoliv, musí dbát na to, aby jeho zvolený software byl schopen s vybraným formátem souborů pracovat. U některých souborů (jako například .CSV souborů) je mimo jiné také nutno dbát na správné kódování z důvodu využívání některých znaků jako oddělovačů dat.

## 7.2 Pentaho

### 7.2.1 Základní informace

Pentaho není jeden samostatný Business Intelligence program, ale jedná se o kolekci počítačových programů (či komponent), které spolu spolupracují. Některé komponenty zajišťují jednoduché funkcionality jako například autentifikace uživatelů nebo správa přihlašování do databáze. Některé naopak zajišťují složitější funkcionality jako vizualizaci dat (14).

### 7.2.2 Community Edition a Enterprise Edition

Existují dvě různé verze Pentaho. První verze je nazvaná Community Edition, druhá verze se nazývá Enterprise Edition. Rozdíly mezi těmito dvěma verzemi spočívají především v nabízené podpoře. Funkcionálně jsou obě verze téměř totožné, několik rozdílů mezi verzemi ale existuje (14). Bouman a Dongen uvádí následující rozdíly a jejich popis:

- Enterprise Console (podniková konzole) – Enterprise Edition se od Community Edition liší tím, že nabízí funkcionalitu, která je zaměřena na podnikové využití. Mezi tyto funkcionality patří mimo jiné bezpečnostní konfigurace, aplikační diagnostika, monitorování výkonu, logování, zálohování a obnova Pentaho repository. Většina těchto funkcionalit je spustitelná přes podnikovou konzoli. Tyto funkcionality lze využívat i v Community Edition, ale vyžadují značné úsilí ke zprovoznění.
- Pentaho Data Integration rozšíření – přidává možnost monitorování výkonu, vzdálené administrace a nová upozornění. Nabízí také speciální plugin pro data mining, zvaný KnowledgeFlow.
- Tvořič nástěnek – poskytuje uživateli možnost sestavit si vlastní BI nástěnku s různými typy obsahu jako jsou grafy, reporty a mapy.
- Služby a podpora – kromě výše zmíněných výhod poskytuje Enterprise Edition uživateli také podporu, možnost odškodnění v případě selhání, údržbu a dodatečné technické zdroje.

Kromě těchto vyjmenovaných rozdílů si jsou verze rovny. Pro potřeby práce bylo tedy vybráno Pentaho Community Edition 0.7.

### 7.2.3 Zdrojový kód

Veškerý zdrojový kód pro Pentaho a jeho využívané komponenty je k dispozici na webu:

*https://github.com/pentaho/*

Zdrojový kód je rozdělen do několika samostatných komponent a projektů. Každý uživatel, který by si přál načíst zdrojový kód Pentaha nebo jeho komponent, ten poté zkompilovat a vytvořit si vlastní build, musí mít na svém systému zpravidla k dispozici:

- Maven verze 3 nebo vyšší
- Java JDK 1.8
- Specifický soubor *settings.xml* ve svém /.m2 adresáři

### 7.2.4 Komponenty

Všechny Pentaho komponenty lze stáhnout na webu Pentaho - http://community.pentaho.com/. Pro správné porozumění těchto komponent a jejich nabízeným funkcionalitám je třeba popsat a názorně vysvětlit jejich funkčnost.

## Business Analytics Platform

Pod komponentou *Business Analytics Platform* se skrývá vlastní Pentaho Server. Server je velký 1 GB, je zabalený v .zip souboru a je k dispozici pro všechny populární počítačové systémy – Windows, Mac OS a Linux.

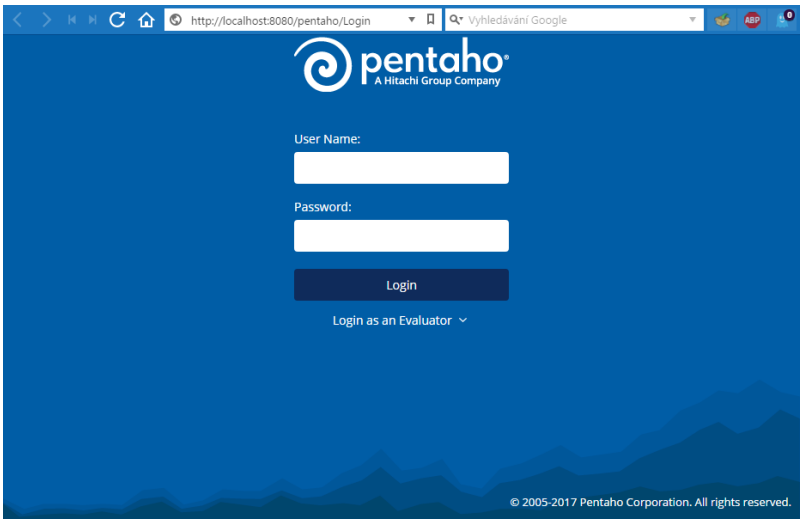
Server se neinstaluje v klasickém slova smyslu, ale rozbaluje se. Uživatel musí pouze otevřít stažený .zip soubor a rozbalit jej do libovolného adresáře. Ke spuštění serveru je také potřeba mít na svém prostředí nainstalovanou 64bitovou verzi Java.

Uživatelé systému Windows poté mohou spustit server pomocí souboru start-pentaho.bat. Uživatelé UNIX systémů (jako Linux nebo Mac OS) mohou server spustit pomocí souboru start-pentaho.sh.

Je třeba brát v úvahu to, že při prvním spuštění dochází k vytváření konfigurací a přípravě prostředí. Z těchto důvodů může první spouštění trvat delší dobu. Po spuštění serveru je možné se k tomuto serveru připojit tím, že do libovolného internetového prohlížeče uživatel napíše následující adresu:

*http://localhost:8080/*

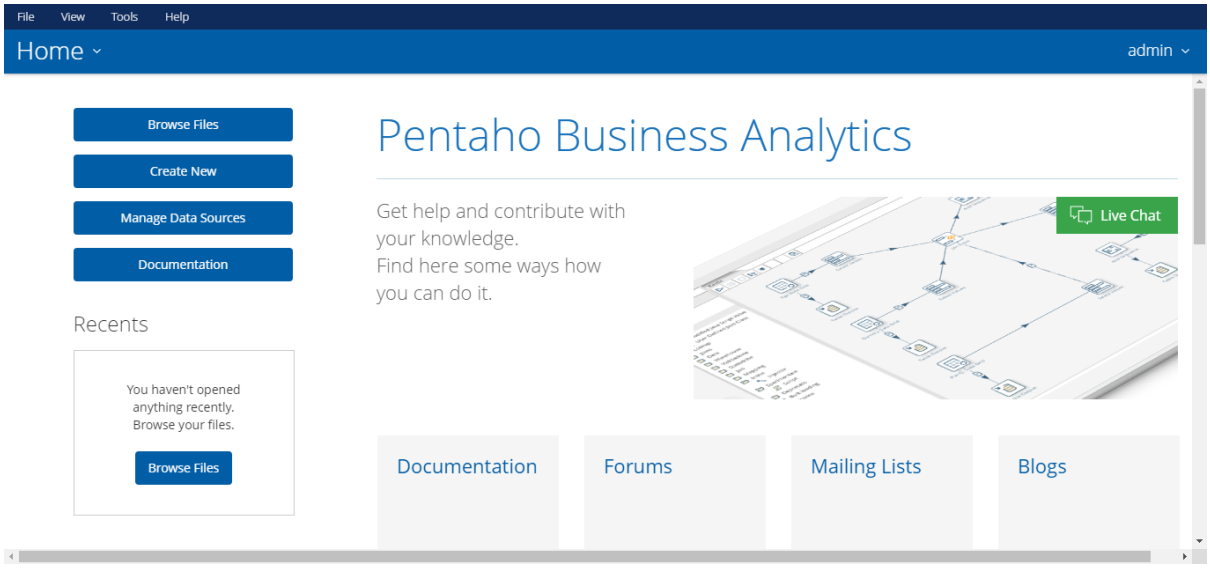
Po zadání této adresy do prohlížeče je uživatel automaticky přesměrován na adresu *http://localhost:8080/pentaho/Login*. Tato adresa slouží jako přihlašovací brána do Pentaho serveru.




**Obrázek 4 Přihlašovací obrazovka do Pentaho serveru. Autor: Jan Poisl.**

## *Prostředí*

Po přihlášení se uživateli zobrazí uživatelská konzole.




**Obrázek 5 Uživatelská konzole Pentaho Serveru. Autor: Jan Poisl.**

Uživatelská konzole je ukázána na snímku obrazovky výše. Ačkoliv slovo konzole může některým lidem evokovat představu textové konzole, tato uživatelská konzole je celá grafická. Vlevo nahoře jsou uživateli k dispozici tlačítka *File*, *View*, *Tools* a *Help*. Pod těmito tlačítky je velké tlačítko Home. Při kliknutí na toto tlačítko se uživateli zobrazí nabídka, ve které lze přepnout uživatelskou konzoli na prohlížení souborů, prohlížení rozplánovaných úloh a v případě, že je uživatel administrátorem, je mu umožněno přepnout do administrace.

Pod tímto tlačítkem se nachází tlačítka Browse Files, Create New, Manage Data Sources a Documentation. Stiskem prvního tlačítka může uživatel prohlížet své soubory. V případě, že je uživatel administrátor, má přístup k souborům všech uživatelů. Tlačítkem Create New lze vytvořit novou nástěnku, JPivot náhled dat, popřípadě přidat nový zdroj dat. Tlačítkem Manage Data Sources lze provádět správu datových zdrojů. Tlačítko Documentation uživatele přesměruje do dokumentace.

### *Administrace*

Pokud je uživatel přihlášen jako administrátor, je mu umožněno vstoupit do administrační části serveru. Administrace je rozdělena do tří kategorií:

- Uživatelé a role
- Mail server
- Nastavení

V kategorii Uživatelé a role je administrátorovi umožněno vytvářet nové uživatele. Při vytváření uživatele musí administrátor vyplnit jméno, heslo a roli uživatele. Dále je mu umožněno vytvářet a upravovat role. Role se od sebe liší množstvím oprávnění. Mezi tato oprávnění patří správa zabezpečení, plánování úloh, čtení obsahu, publikování obsahu, vytváření obsahu, spouštění a správa datových zdrojů.

Mail server je kategorie, kde je možno nastavit připojení k emailovému serveru. Podobně jako ve všech emailových klientech je nutno nastavit host, využívaný port, uživatelské jméno, heslo, typ serveru a název, pod kterým bude Pentaho rozesílat reporty.

V kategorii Nastavení může administrátor nastavit, po jaké době se mají automaticky mazat staré generované soubory, popřípadě je může vymazat ihned. Další možností je naplánovat mazání souborů automaticky.

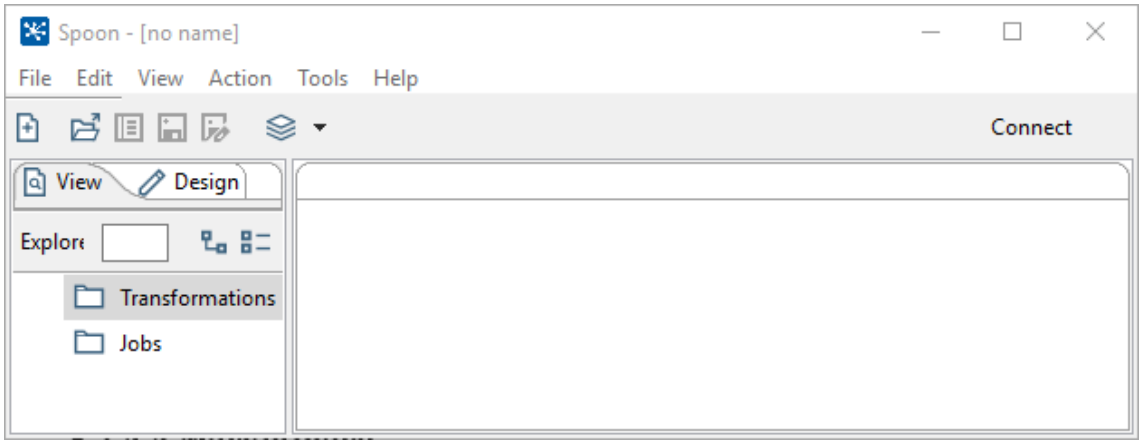
## Data Integration

Data Integration přináší do Pentaha ETL schopnosti. Jak už bylo vysvětleno v teoretické části práce, ETL je zkratka pro *extraction, transformation and loading*. Všechny tyto operace jsou prováděny s daty.

Tento modul je zabalen v 800 MB velkém .zip souboru. Podobně jako Pentaho server jej stačí pouze rozbalit, do libovolně zvoleného adresáře neinstaluje se pomocí instalátoru.

Pentaho Data Integration je přezdívaný Spoon, uživatelé systému Windows jej tudíž mohou spustit pomocí spouštěcího souboru *Spoon.bat*, uživatelé Unixových systému jej spouští pomocí spouštěcího souboru *spoon.sh*.

Po spuštění modulu pomocí spouštěcích souborů se uživateli zobrazí základní rozhraní tohoto modulu.




**Obrázek 6 Rozhraní Spoon modulu. Autor: Jan Poisl.**

Jak je vidět na obrázku výše, prostředí tohoto modulu je specializováno na prohlížení a vytváření Transformací a Prací. Nová transformace nebo práce se dá vytvořit pravým kliknutím na danou položku a vybráním položky New.

Vytváření obou položek probíhá podobně. Při jejich vytváření je využíváno grafické „drag and drop“ prostředí. Uživateli stačí vybrat si žádané bloky chování, přetáhnout je myší do prostředí a poté bloky pospojovat.

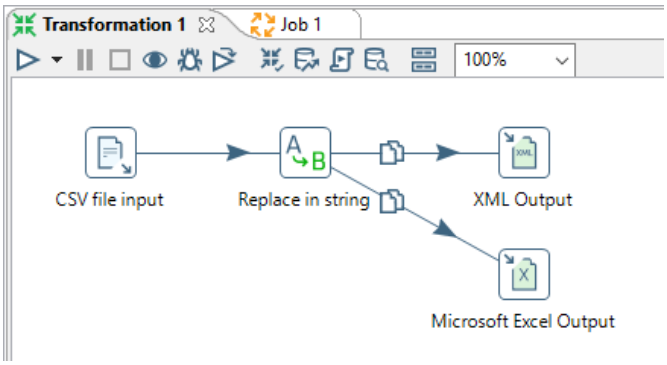
### *Transformace*

Jak už název napovídá, úlohou transformace je přeměna vstupních dat na upravená, výstupní data. Uživatel má k dispozici nepřeberné množství „bloků“, které jsou rozděleny do následující kategorií s vybranými příklady:

- Input – CSV vstup, Datová mřížka, emailový vstup, HL7 vstup, JSON vstup
- Output – Microsoft Access/Excel, JSON výstup, RSS výstup, aktualizace
- Transform – přidání proměnných, přidání XML, výměna znaků
- Ulitity – změna kódování, klonování sloupce, spuštění procesu, email
- Flow – blokování procesu, správce procesů, přepínač procesů
- Scripting – spuštění SQL skriptu, uživatelsky definovaná Java třída
- Pentaho Server – volání koncového bodu, nastavení proměnných
- Lookup – zavolání databázové procedury, HTTP klient, webová služba
- Joins – kartézský součin, spojení sloupců, XML spojení
- Data Warehouse – kombinační vyhledání, upravení

- Validation – validátor kreditních karet, data validátor, email validátor
- Statistics – analytický dotaz, seskupování
- Big Data – Hadoop vstup, MongoDB vstup, MongoDB výstup
- Agile – MonetDB Agile Mart
- Cryptography – PGP dešifrování, PGP šifrování, generátor klíčů
- Palo – Palo Cell vstup, Palo Cell výstup
- Open ERP – OpenERP smazání objektu, OpenERP vstup objektu
- Job – zkopírování řádků do výsledku, získání proměnných
- Mapping – sub-transformace, specifikace vstupu
- Bulk loading – MonetDB Bulk Loader, MySQL Bulk Loader
- Inline – injektor, čtečka soketů
- Experimental – SFTP, skript
- Deprecated – ukázkový krok, starý textový vstup
- History – vstup pro textový soubor, CSV vstup

Z výčtu těchto kategorií je zřejmé, že tento modul nabízí uživateli nepřeberné množství možností transformace dat.




**Obrázek 7 Transformace CSV souboru. Autor: Jan Poisl.**

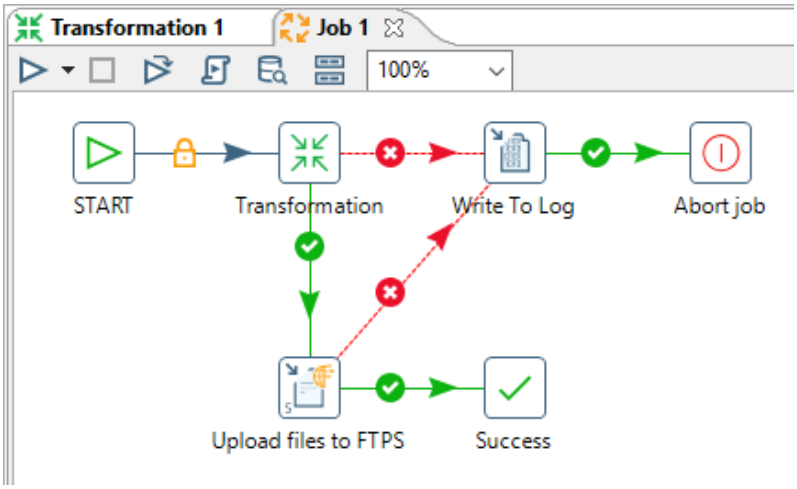
Na příkladu výše je vytvořena velmi jednoduchá transformace, která nejdříve načte soubor ve formátu CSV, nahradí ve vstupu specifické znaky za jiné znaky a takto transformovaná data poté převede do XML souboru a zároveň do druhého Excel souboru.

### *Práce*

Bloky pro práce jsou rozděleny do následujících kategorií s vybranými příklady:

- General – Start, práce, transformace, nastavení proměnných
- Mail – validátor emailu, email, přijetí emailů
- File management – srovnání souborů, vytvoření složky, smazání souboru
- Conditions – kontrola webové servisy, čekat, zjistit, zda soubor existuje
- Scripting – SQL, JavaScript
- Bulk loading – BulkLoad z MySQL do souboru
- Big Data – Sqoop export, Sqoop import
- Modeling – vytvoření modelu, publikování modelu
- XML – XSD validátor, XSL transformace
- Ulitity – čekání na SQL, zrušení práce, zapsání do logu
- Repository – export repository do XML souboru
- File transfer – získání souboru s FTPS, nahrání souboru do FTPS
- File encryption – ověření signatury s PGP, dešifrování souboru s PGP
- Palo – smazání Palo Cube, vytvoření Palo Cube
- Deprecated – MS Access Bulk nahrání

Práce disponují menším seznamem kategorií než transformace, to je ale dáno odlišným zaměřením.




**Obrázek 8 Ukázka Práce. Autor: Jan Poisl.**

Na příkladu výše je vytvořena jednoduchá ukázková Práce. Při startu této práce je provedena transformace. Jestliže transformace není úspěšná, Práce tuto skutečnost zapíše do logu a zruší se. Jestliže je transformace úspěšně provedena, práce nahraje transponované soubory do FTPS. Jestliže nahrání není úspěšné, je o této skutečnosti zapsáno do logu a práce je zrušena. Jestliže je nahrání úspěšné, práce se úspěšně ukončí.

## Report Designer

Report Designer potřebuje ke své funkci modul *Data Integration*. Data z tohoto modulu jsou streamovaná do Report Designeru, který z těchto dat poté generuje reporty. Stejně jako v předchozích modulech je tento modul zabalen v .zip souboru, který je velký 547 MB. Na rozdíl od předchozích je ale rozdělen do dvou souborů podle operačního systému, který uživatel využívá. Jeden .zip soubor je určen pro uživatele systémů Windows a Linux, druhý pouze pro uživatele systému Mac OS X. Stejně jako v předchozích případech stačí soubor rozbalit do libovolné složky. Uživatelé systému Windows poté Report Designer spustí spouštěcím souborem report-designer.bat. Uživatelé Unix systémů využijí soubor report-designer.sh.

Po spuštění se uživateli spustí rozhraní, přes které může vytvářet nebo upravovat reporty. Nový report lze vytvořit dvěma způsoby. Buď lze vytvořit prázdný report kliknutím na File a poté na tlačítko New. Druhý způsob je využití Design Wizardu.

### *Prázdný report*

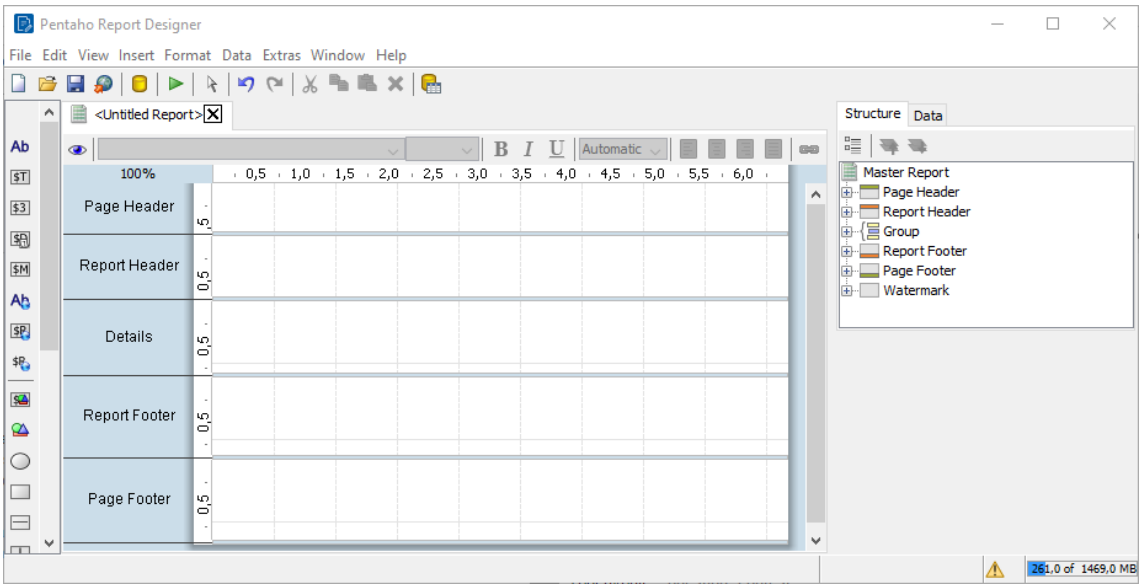
Vytvoření prázdného reportu je přímočaré. Uživatel provede výše uvedený postup, po kterém se mu zobrazí prázdný report. Report je rozdělen do pěti skupin:

- Záhlaví stránky
- Záhlaví reportu
- Detaily

- Zápatí reportu
- Zápatí stránky

Prázdný report už poté lze přizpůsobit dle potřeb konkrétního uživatele. Do reportu lze vložit objekty rozdělené do třech skupin:

- Textové objekty – štítek, textové pole, pole čísel, pole dat, pole zpráv, štítek pro zdroje, pole pro zdroje, zprávy zdrojů
- Grafické prvky – pole pro obrázek, obrázek, elipsa, obdélník, horizontální čára, vertikální čára, rozsah, graf, pruhovaný graf, čárový graf, koláčový graf
- Objekty – skupina, sub-report, kontingenční tabulka, tabulka s obsahem, index




**Obrázek 9 Report Designer s prázdným reportem.**

Prázdný report je k vidění na přiloženém snímku obrazovky. Na levé straně se nacházejí objekty, které může uživatel do reportu vložit. Objekty se do reportu vkládají jednoduše přetažením kurzorem myši.

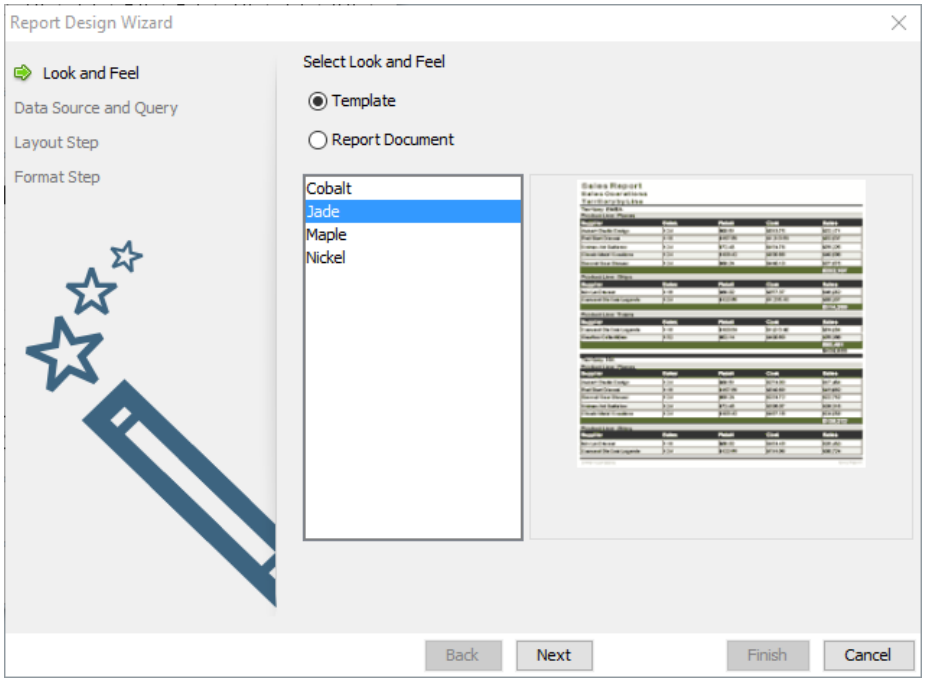
Prostřední část Report Designeru je věnována samotnému reportu. Uživatel může libovolně měnit výšku částí reportu. Do každé části může uživatel vložit libovolný objekt, přičemž zpravidla lze měnit jeho velikost, pozici, viditelnost a nepřeberné množství dalších relevantních parametrů

Na pravé straně Report Designeru lze vidět strukturu reportu. V případě zaplnění reportu prvky lze v této struktuře snadno najít požadovaný prvek. Prvky v reportu a ve struktuře jsou propojeny. Když uživatel označí prvek ve struktuře, označí se i v reportu a naopak. Napravo od struktury je dále Data list, v němž jsou pro uživatele uvedena všechna využitá data v jeho reportu a také seznam zdrojů, odkud tato data pocházejí.

### *Report Design Wizard*

Využití této komponenty výrazně ulehčuje vytváření reportu. Report Design Wizard je rozdělen do čtyř kroků:

- Vybrání šablony – uživatel si může vybrat ze čtyř předdefinovaných šablon, popřípadě může vybrat svou vlastní šablonu
- Vybrání zdrojů dat – jak už název napovídá, uživatel musí zadat zdroje dat pro vytvářený report. Lze vybírat od databází až po XML soubory
- Layout – uživatel si může vybrat z předdefinovaných layoutů
- Formát – vybrání výstupního formátu reportu




**Obrázek 10 První krok v Report Design Wizardu. Autor: Jan Poisl.**

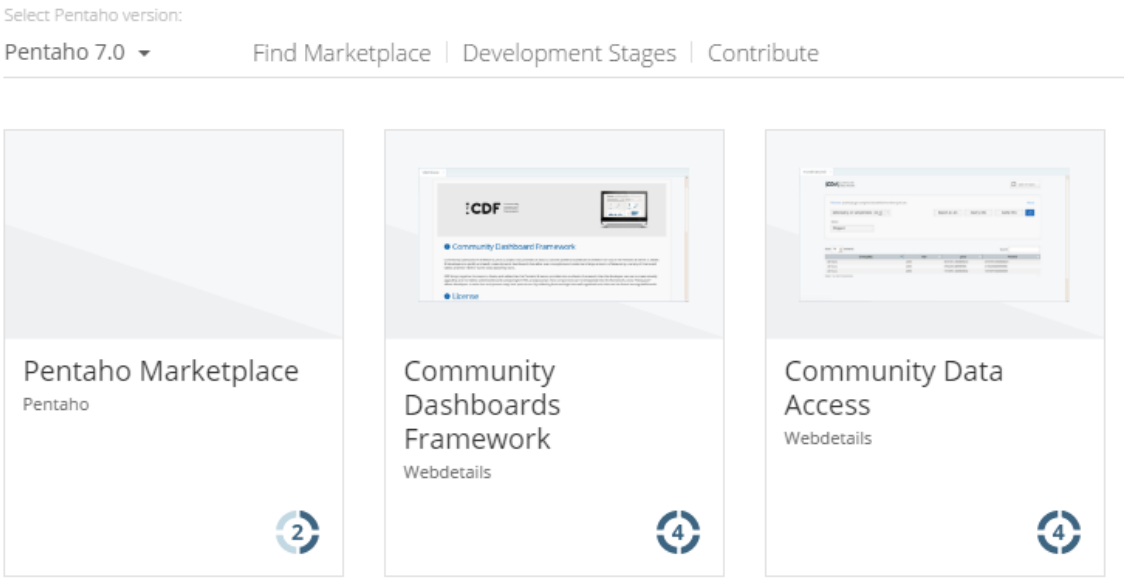
Na obrázku výše je zachyceno vybírání šablony pro report a náhled této šablony. Při správném využití Report Design Wizard si může uživatel značně zjednodušit svou práci, jelikož report s požadovanými daty a informacemi je pro něj pomocí Wizardu vygenerován.

## Marketplace

Modul Markeplace umožňuje uživateli rozšiřovat Pentaho o externí pluginy. Pokud ale uživatel využívá Pentaho verze 6 nebo novější, tento modul stahovat nemusí, jelikož je v něm tento plugin defaultně zabudován. Pokud jej uživatel chci využívat, stačí jej aktivovat úpravou souboru *org.apache.karaf.features.cfg*.

Samotná existence Marketplace znamená to, že Pentaho lze rozšiřovat dle potřeby a možností o extra funkcionality, které defaultně nepodporuje.

Pluginy jsou k dispozici ke stažení na následující webové stránce – *http://www.pentaho.com/marketplace/*, na které lze pluginy vyhledávat dle kategorií. Každý plugin je také označen číslem, které uživateli dává na vědomí to, v jaké fázi vývoje daný plugin je, zda je stabilní a zda je možné jej využívat v produkci.




**Obrázek 11 Seznam pluginů na Marketplace. Autor: Jan Poisl.**

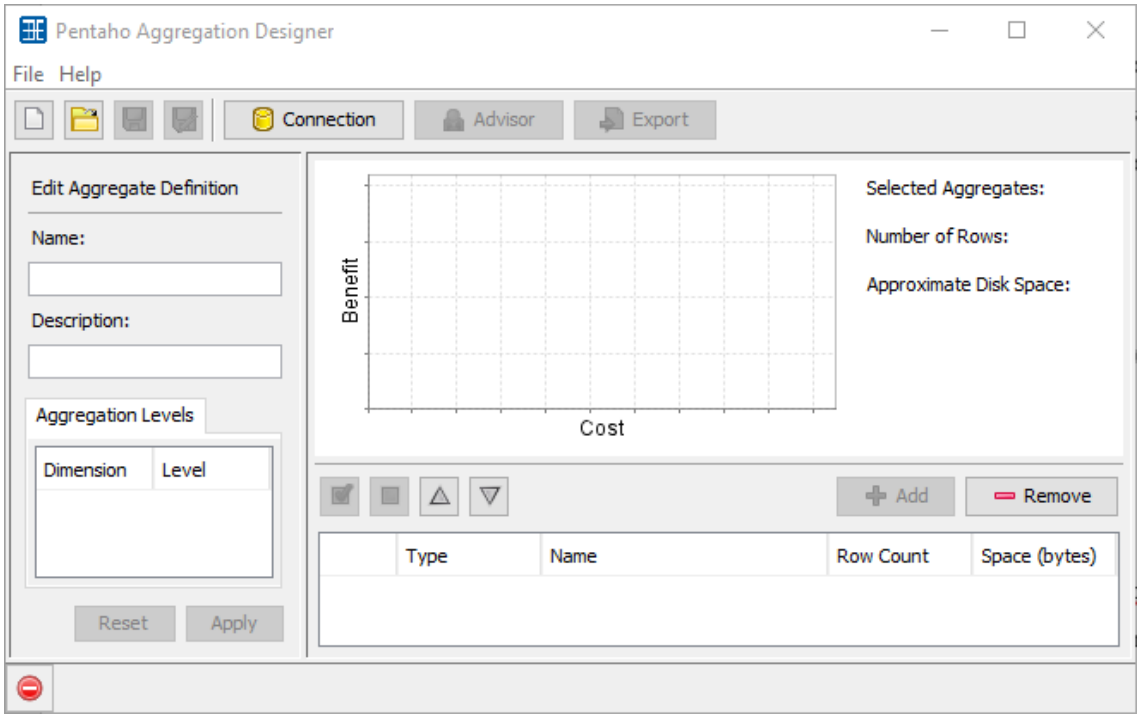
Na přiloženém obrázku lze vidět nabídku několika pluginů z Marketplace. Před tím, že uživatel začne hledat pluginy, musí v levé horní části nastavit svou

verzi Pentaho Serveru, kterou využívá. Plugin vlevo je starý plugin, který je označen jako nestabilní, proto má číslo 2. Pluginy vpravo od něj mají číslo 4, jelikož jsou stabilní a oficiálně podporované. Jsou tedy vhodné k nasazení na produkčním prostředí.

## Aggregation Designer

Tento modul spadá do kategorie designových nástrojů. Přináší jednoduché rozhraní, které umožňuje uživateli tvorbu souhrnných tabulek. Samotný modul je oproti těm ostatním velikostně malý – pouze 24 MB. Stejně jako u ostatních modulů jej akorát stačí rozbalit do libovolné složky. Uživatelé systémů Windows jej mohou spustit spouštěním souborem *startaggregationdesigner.bat*. Uživatelé Unixových systému jej mohou spustit souborem *startaggregationdesigner.sh*.

Ke správné funkci tohoto modulu je třeba mít k dispozici databázi (jako například MySQL, ale k dispozici je nepřeberné množství dalších databází) nebo soubor typu Mondrian Schema.



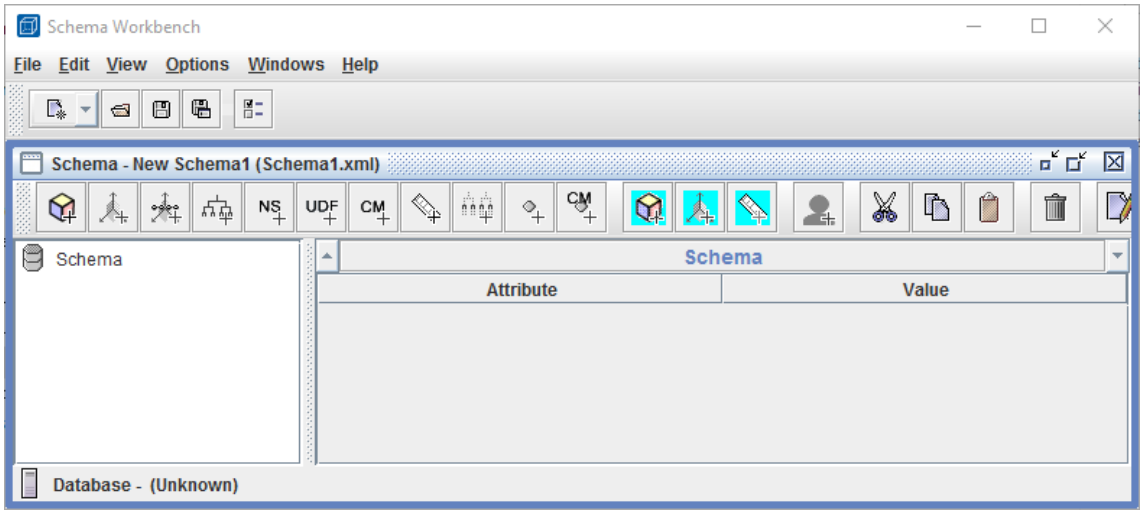

**Obrázek 12 Prázdný Aggregation Designer. Autor: Jan Poisl.**

Aggregation Designer ukazuje uživateli úroveň agregace, rozměry agregace a umožňuje uživateli sledovat statistiky jako počet řádku nebo také dokonce velikost zabíraného místa na disku.

## Schema Workbench

Jak už název napovídá, prací tohoto modulu je vytváření a úprava databázových schémat. Schema Workbench umožňuje uživateli tvořit Mondrian OLAP schémata. Data je v tomto modulu možno reprezentovat vícerozměrně.

Tento modul je stejně jako Aggregation Designer poměrně malý - .zip soubor se zabaleným modulem je 29 MB velký. Po jeho rozbalení do libovolného adresáře ho lze na systému Windows spustit spouštěcím souborem *workbench.bat*. Uživatelé Unix systémů tento modul mohou spustit pomocí souboru *workbench.sh*.




**Obrázek 13 Ukázka Schema Workbench. Autor: Jan Poisl**

Na přiloženém snímku obrazovky je ukázáno rozhraní Schema Workbench. Největší část rozhraní zabírá samotná úprava schématu. Schema Workbench se připojí k nastavené databázi a načte její schémata. Uživatel poté může schémata upravovat pomocí ikonek. Možnosti úprav jsou rozděleny do několika skupin:

- Přidání kostky, přidání rozměru a užívání rozměru, přidání hierarchie,

- Přidání virtuální kostky, přidání rozměru virtuální kostky, přidání měřítka virtuální kostky, pojmenovaného setu, uživatelsky definovaných funkcí, kalkulovaných členů, měřítek, úrovní a vlastností
- Přidání role
- Vyjmutí, kopírování, vložení, smazání
- Upravení rolí

## Metadata Editor

Metadata Editor zjednodušuje vytváření reportů a umožňuje tvorbu metadata domén a relačních data modelů.

Soubor .zip se zabaleným Metadata Editorem je 707 MB velký a stejně jako u předchozích komponent jej stačí rozbalit do libovolné složky. Editor se poté spouští souborem *metadata-editor.bat* v případě užívání operačního systému Windows, popřípadě souborem *metadata-editor.sh* v případě užívání Unix systému.

### 7.2.5 Užití Pentaho pro Business Intelligence

V teoretické části práce jsou kategorie BI produktů rozděleny na Databázové systémy, ETL nástroje, Analytické aplikace a nástroje, Data Mining technologie, Nástroje řízení kvality dat, Klientské nástroje a Standardní aplikace. Cílem je zjistit, kolik těchto kategorií je Pentaho schopno zastat. Určování, zda je daná kategorie splněna, vychází ze získaných vědomostí o všech Pentaho modulech.

## Databázové systémy

Pentaho je schopné pracovat s databázovými systémy, které slouží jako vstupy a zároveň umí pracovat s databázovými systémy, které fungují jako datové sklady, tržiště či dočasná uložiště. Také není vázáno na jeden konkrétní typ databáze, ale podporuje jich celé množství. Tuto kategorii Pentaho splňuje.

## ETL nástroje

Pentaho je schopno využívat ETL nástrojů. Konkrétním nástrojem je v tomto případě modul zvaný Data Integration. Tento modul je specificky určen na ETL operace s daty. Tuto kategorii Pentaho tedy splňuje.

## Analytické aplikace a nástroje

Pod touto kategorií si lze představit OLAP a reportingové nástroje. Na práci s OLAP schématy je zaměřen modul Schema Workbench. Jako reportingový nástroj v tomto případě funguje modul Report Designer. Této kategorii Pentaho vyhovuje.

## Data Mining technologie

Standardní Pentaho moduly nenabízejí možnost Data Miningu. Pokud by ale uživatel měl o Data Mining zájem a chtěl k němu využívat nástroj od výrobce Pentaha, může využít projekt zvaný Data Mining – Weka, který je k dispozici na webové adrese http://community.pentaho.com/projects/data-mining/. Projekt Weka je stejně jako Pentaho open source. Ve striktním hodnocení ovšem čisté Pentaho nepodporuje Data Mining.

## Nástroje řízení kvality dat

Jako nástroj řízení kvality dat může být využit modul Data Integration. Ke splnění těchto úkonů by bylo možné vytvořit Práci, která by v určitém časovém horizontu pravidelně spouštěla Transformaci, ve které by byly definované validace samotných dat a jejich kvality. Lze tedy říci, že Pentaho těmito nástroji disponuje.

## Klientské nástroje

Jestliže klientské nástroje jsou definované jako rozhraní k relačním databázím datového skladu a dalším databázím, pak Pentaho těmito nástroji disponuje. Do těchto nástrojů mohou více či méně patřit moduly Aggregation Designer a Schema Workbench.

## Standardní aplikace

Pentaho disponuje základními standardními aplikacemi. Pentaho disponuje analytickými a plánovacími funkcemi, využívá matematického a statistického aparátu, umožňuje základní úroveň přizpůsobení a je k dispozici ve více jazykových verzích. Tento požadavek je tudíž splněn.

# 8 Shrnutí výsledků

Z výsledků praktické části práce vyplývá, že Pentaho je schopno s vyjímkou Data Miningu splnit všechny úkoly, které uživatelé očekávají od Business Intelligence softwaru. Uživatel tohoto softwarového balíku má výhodu v tom, že Pentaho dokáže zpracovávat takřka jakýkoliv typ nebo formát vstupních dat, ať už se jedná o různé typy databází, soubory ve specifických formátech, či textové soubory.

Pentaho nabízí jednoduchou práci s ETL, protože v data Data Integration modulu se ETL operace designují v přehledném grafickém prostředí, kde uživatel pouze přetahuje operace a spojuje je šipkami. Vzhledem k velké flexibilitě je možné tento modul také využít jako nástroj pro řízení kvality dat.

Dále nabízí dostatečné schopnosti práce s OLAP, a především velkou flexibilitu Reportů. Uživatel není při tvorbě Reportů ničím limitován, každý Report si může přizpůsobit dle svého přání. Pokud uživatel nechce tvořit celý Report ručně, je mu umožněno využít Report Design Wizard, který mu tvorbu reportu značně zjednoduší.

Pokud by si ale uživatel přál nad svými daty provádět Data Mining, Pentaho samotné mu stačit nebude. Pozitivní zprávou v tomto ohledu ovšem je to, že společnost, která vyrobila Pentaho, také vyrobila samostatný nástroj pro provádění Data Miningu a analýz dat, který se jmenuje Weka a je také open source.

Pentaho také disponuje klientskými nástroji pro správu, popřípadě vizualizaci dat z databází a je schopno základních standardních aplikací, které od takého systému uživatel očekává.

V čem Pentaho stojí za zmínku, je také to, že obsahuje Marketplace a je možno jej rozšířit o dodatečnou funkcionalitu. Marketplace je přehledný a obzvláště praktické je hodnocení pluginů podle čísel, díky kterým uživatel ví, v jakém stavu plugin je, zda je dostatečně stabilní a vhodný k nasazení do ostré produkční verze.

Ve striktním hodnocení byla autorova původní hypotéza vyvrácena – samotné Pentaho není zcela kompletním balíkem Business Intelligence a z důvodu chybějícího integrovaného Data Mining řešení. Tento problém lze ale vyřešit užitím projektu Weka.

# 9 Závěry a doporučení

Dle shrnutí výsledků bylo dospěno k závěru, že Pentaho Community Edition 7.0 je schopné s výjimkou Data Miningu zcela kompletně zastat roli Business Intelligence systému, který disponuje požadovanými BI funkcionalitami. K tomu, aby se jednalo o kompletní řešení, mu chybí schopnost Data Miningu, kterou je ovšem schopen zastat další open source projekt od autorů Pentaha zvaný Weka. V případě využití tohoto projektu uživatel disponuje kompletním BI softwarovým balíkem.

Další velkou výhodou je samotný fakt, že Pentaho je open source software. Tento samotný fakt přináší uživatelům značné výhody. V případě jakýchkoliv problémů se uživatel s dostatečnými znalostmi může podívat do zdrojového kódu, identifikovat místo problému a popřípadě problém i vyřešit.

Uživatelé také těží z toho, že Pentaho je vytvořeno v programovacím jazyce Java s využitím Apache Maven, díky čemuž je zcela multiplatformní. Uživatel tedy není limitován konkrétním operačním systémem – může využívat Pentaho ať už má operační systém Windows, Linux nebo Mac OS X. Na Pentaho Server se navíc lze přihlásit přes webové rozhraní, díky čemuž lze využívat i na mobilních zařízeních či tabletech.

Community Edition s sebou ovšem nese několik problémů, především pro velké společnosti. Mezi tyto problémy patří nulová podpora v případě nalezení problému s Pentahem samotným. Dalším značným problémem je to, že v případě poškození či zničení dat nemůže společnost po nikom vymáhat způsobené škody, protože v případě využití Community Edition Pentaho nenese za tyto incidenty žádnou zodpovědnost. Z tohoto důvodu se větším společnostem a subjektům vyplatí Enterprise edice, která je sice placená, ale v případě incidentů je za tyto incidenty Pentaho zodpovědné. Další výhodou Enterprise edice je oficiální uživatelská podpora. Poslední výhodou jsou některé extra funkcionality, které jsou zaměřeny především na zabezpečení.

Autorovo doporučení je, že v případě hledání Business Intelligence systému by uživatel měl brát v úvahu Pentaho, jelikož splňuje naprostou většinu BI funkcí. Pro Pentaho také mluví to, že je open source a Community Edition je zcela zdarma. Větším společnostem je ovšem doporučeno využití placené Enterprise edice.

# 10 Seznam použitých zdrojů


1. **Meeker, J, Heather.** *The Open Source Alternative: Understanding Risks and Leveraging Opportunities.* New Jersey : John Wiley & Sons, Inc., 2008.

2. **What is free software?** *GNU Project - Free Software Foundation.* [Online] 27. Prosinec 2016. [Citace: 20. Leden 2017.] https://www.gnu.org/philosophy/free-sw.en.html.

**3. Vercellis, Carlo.** *Business Intelligence: Data Mining and Optimization for Decision Making.* Chichester : John Wiley & Sons, Ltd., 2009.

**4. Štědroň, Bohumír a Mls, Karel.** *Manažerská informatika I.* Hradec Králové : Gaudeamus, 2007.

**5. Gála, Libor, Pour, Jan a Prokop, Toman.** ***Podniková informatika.*** Praha : Grada Publishing, a.s., 2006.

**6. Janert, K., Philipp.** *Data Analysis with Open Source Tools.* Sebastopol : O'Reilly Media, Inc., 2011.

**7. Olivares, Ramos, Jonathan a Skalská, Hana**. *Review of business intelligence principles in large, medium and small companies.* Hradec Králové : Gaudeamus, 2015. Hradec Economic Days 2015. stránky 70-77.

**8. Turban, Efraim, Sharda, Ramesh a Delen, Dursun.** *Decision Support and Business Intelligence Systems 9th Edition.* New Jersey : Pearson Education, Inc., 2010.

**9. Han, Jiawei a Kamber, Micheline.** *Data Mining: Concepts and Techniques.* San Francisco : Elsevier Inc., 2006.

**10. Vesset, Dan, a další.** *Worldwide Business Analytics Software Market Shares, 2015: Healthy Demand Despite Currency Exchange Rate Headwinds.* Framingham : International Data Corporation (IDC), 2016.

**11. Sýkora, Martin.** Volný software a autorské právo. *pravoit.cz.* [Online] 3. Březen 2010. http://www.pravoit.cz/novinka/volny-software-a-autorske-pravo.

**12. Datasets for Data Mining and Data Science.** *Analytics, Data Mining, and Data Science.* [Online] 2017. http://www.kdnuggets.com/datasets/index.html.

**13. Chen, Xiaming.** GitHub - caesar0301/awesome-public-datasets: An awesome list of high-quality open datasets in public domains (on-going). By everyone, for

everyone! *The world's leading software development platform · GitHub.* [Online] 8. Duben 2017. https://github.com/caesar0301/awesome-public-datasets.

**14. Bouman, Roland a Dongen, van, Jos.** *Pentaho Solutions: Business Intelligence and Data Warehousing with Pentaho and MySQL.* Indianapolis : Wiley Publishing, Inc., 2009.

*Zadání práce*

Univerzita Hradec Králové
Faculty of Informatics and Management
Akademický rok: 2016/2017

Studijní program: Systems Engineering and Informatics
Forma: Full-time
Obor/komb.: Informační management (im3-p)

**Podklad pro zadání BAKALÁŘSKÉ práce studenta**

| PŘEDKLÁDÁ: | ADRESA | OSOBNÍ ČÍSLO |
|---|---|---|
| Poisl Jan | Orlické Záhoří 27, Orlické Záhoří | I1500212 |

**TÉMA ČESKY:**

Open source v úlohách Business Intelligence.

**TÉMA ANGLICKY:**

Open source software in Business Intelligence.

**VEDOUCÍ PRÁCE:**

prof. RNDr. Hana Skalská, CSc. - KIKM

**ZÁSADY PRO VYPRACOVÁNÍ:**

Charakterizace BI, charakterizace úloh BI. Popsání typických úloh a zpracování případové studie. Vyhledání a srovnání dostupného open source BI software.

Struktura práce:

Úvod,
Cíl: obeznámit čtenáře s open source BI systémy,
Metodika: práce s literaturou,
Vlastní obsah práce: vymezení pojmu BI, jejich definice, rozdíly, zpracování přehledu SW
Aplikace: popis dostupného SW,
Výsledky: popis výsledků, ke kterým práce došla,
Shrnutí a závěr.

**SEZNAM DOPORUČENÉ LITERATURY:**

Carlo Vercellis: Business Intelligence,
Jiawei Han, Micheline Kamber: Data Mining,
Jonathan Ramos Olivares, Hana Skalská: Review of business intelligence principles in large, medium and small companies,

Internetový zdroj: KDD - Knowledge Discovery KDnuggets,
Internetový zdroj: Dresner.