TECHNICKÁ UNIVERZITA V LIBERCI
Ekonomická fakulta

VYUŽITÍ NÁSTROJŮ BUSINESS INTELLIGENCE PRO HODNOCENÍ KVALITY PŘÍRODNÍCH VOD

Diplomová práce

Studijní program: N6209 – Systémové inženýrství a informatika
Studijní obor: 6209T021 – Manažerská informatika

Autor práce: **Bc. David Krejbich**
Vedoucí práce: Ing. Vladimíra Zádová, Ph.D.

Liberec 2015

TECHNICKÁ UNIVERZITA V LIBERCI
Ekonomická fakulta
Akademický rok: 2014/2015

ZADÁNÍ DIPLOMOVÉ PRÁCE

(PROJEKTU, UMĚLECKÉHO DÍLA, UMĚLECKÉHO VÝKONU)

Jméno a příjmení:	**David Krejbich**
Osobní číslo:	**E13000292**
Studijní program:	**N6209 Systémové inženýrství a informatika**
Studijní obor:	**Manažerská informatika**
Název tématu:	**Využití nástrojů Business Intelligence pro hodnocení kvality přírodních vod**
Zadávající katedra:	**Katedra informatiky**

Zásady pro vypracování:

1. Nástroje BI
2. Open source nástroje BI a možnosti jejich využití
3. Zpracování dat, tvorba reportů a analýz pro potřeby projektu MARE pomocí zvolených nástrojů
4. Zhodnocení přínosů řešení

Rozsah grafických prací:

Rozsah pracovní zprávy: **65 normostran**

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

Seznam odborné literatury:

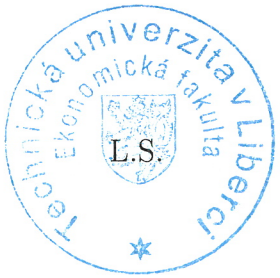
POUR, Jan, Miloš MARYŠKA a Ota NOVOTNÝ. Business Intelligence v podnikové praxi. 1. vyd. Praha: Professional Publishing, 2012. ISBN 978-80-7431-065-2.
BOUMAN, Roland L. a Jos van DONGEN. Pentaho solutions: Business intelligence and data warehousing with Pentaho and MySQL. 1st ed. Indianapolis: Wiley Publishing, Inc., 2009. ISBN 9780470484326.
CASTERS, Matt R., Roland BOUMAN a Jos van DONGEN. Pentaho Kettle solutions: building open source ETL solutions with Pentaho Data Integration. 1st ed. Indianapolis, IN: Wiley Publishing, Inc., 2010. ISBN 9780470635179.
MATTIO, Mariano G. a Dario R. BERNABEU. Pentaho 5.0 Reporting by example: Beginner's guide. B.m.: Packt Publishing Ltd., 2013. ISBN 9781782162254.
LABERGE, Robert. Datové sklady: agilní metody a business intelligence. 1. vyd. Brno: Computer Press, 2012. ISBN 978-80-251-3729-1.
Elektronická databáze článků ProQuest (knihovna.tul.cz).

Vedoucí diplomové práce: **Ing. Vladimíra Zádová, Ph.D.**
Katedra informatiky

Konzultant diplomové práce: **Mgr. Kamil Nešetřil**
Ústav mechatroniky a technické informatiky

Datum zadání diplomové práce: **31. října 2014**

Termín odevzdání diplomové práce: **7. května 2015**



doc. Ing. Miroslav Žižka, Ph.D.
děkan

doc. Ing. Jan Skrbek, Dr.
vedoucí katedry

V Liberci dne 31. října 2014

Prohlášení

Byl jsem seznámen s tím, že na mou diplomovou práci se plně vztahuje zákon č. 121/2000 Sb., o právu autorském, zejména § 60 – školní dílo.

Beru na vědomí, že Technická univerzita v Liberci (TUL) nezasahuje do mých autorských práv užitím mé diplomové práce pro vnitřní potřebu TUL.

Užiji-li diplomovou práci nebo poskytnu-li licenci k jejímu využití, jsem si vědom povinnosti informovat o této skutečnosti TUL; v tomto případě má TUL právo ode mne požadovat úhradu nákladů, které vynaložila na vytvoření díla, až do jejich skutečné výše.

Diplomovou práci jsem vypracoval samostatně s použitím uvedené literatury a na základě konzultací s vedoucím mé diplomové práce a konzultantem.

Současně čestně prohlašuji, že tištěná verze práce se shoduje s elektronickou verzí, vloženou do IS STAG.

Datum:

Podpis:

Poděkování

Na tomto místě bych rád poděkoval Ing. Vladimíře Zádové, Ph.D. za odborné vedení diplomové práce, za její podporu, trpělivost, rady, inspiraci a cenné připomínky.

Mgr. Kamilu Nešetřilovi děkuji za spolupráci na projektu Technologické agentury ČR č. TA02020177 „Informační systém pro podporu rozhodování o využití krajiny po rekultivaci (MARE)“, dále za trpělivost a zejména za cenné rady a zkušenosti ohledně praktické části práce.

Velký dík patří také mé rodině a přítelkyni za podporu a povzbuzování po celou dobu mého studia.

Anotace

Cílem této práce je hodnocení kvality přírodních vod při využití open source nástrojů Business Intelligence. Dílčími úkoly jsou porovnání a zhodnocení vybraných nástrojů. Pro tyto účely je v teoretické části práce nejdříve popsána motivace pro používání nástrojů Business Intelligence v praxi. Dále se teoretická část práce věnuje architektuře Business Intelligence a popisu jednotlivých komponent. Potřebné komponenty procesu Business Intelligence jsou posléze využity v části praktické. Práce je dále zaměřena na vybrané open source nástroje. Ty jsou porovnány a vyhodnoceny podle zvolených kritérií. Pro řešení hodnocení kvality přírodních vod je využito vybraných nástrojů pro reporting a zpracování dat. V poslední části práce jsou formulovány přínosy této práce.

Klíčová slova

business intelligence, datový sklad, etl, pentaho, report, transformace

Annotation

Application of Business Intelligence tools for quality assessment of natural water

The main goal of this thesis is the quality assessment of natural water by using open source Business Intelligence tools. Partial tasks are comparison and evaluation of selected tools. For this purpose, theoretical part describes the motivation for using Business Intelligence tools in practice. Furthermore, the theoretical part deals with the architecture of Business Intelligence and description of each component. Thereafter there are used required components for Business Intelligence process in the practical part. The thesis is also focused on chosen open source tools. These are compared and evaluated according to the own criteria. For solving the quality assessment of natural water are used reporting and data transformation tools. In conclusion, there are formulated the benefits of this thesis.

Key Words

business intelligence, data warehouse, etl, pentaho, report, transformation

Obsah

Seznam obrázků ... **11**

Seznam tabulek ... **12**

Seznam zkratek ... **13**

Úvod ... **15**

1 Zhodnocení současného stavu řešené problematiky ... **17**

1.1 Vysokoškolské kvalifikační práce ... **17**

1.2 Odborná literatura ... **18**

1.3 Internet ... **19**

1.4 Databáze článků ... **20**

2 Business Intelligence ... **21**

2.1 Definice Business Intelligence ... **21**

2.2 Vývoj Business Intelligence ... **22**

2.3 Využití Business Intelligence v praxi ... **22**

2.4 Business Intelligence v praxi ... **23**

2.5 Perspektivy Business Intelligence ... **24**

3 Architektura Business Intelligence ... **25**

3.1 Zdroje dat ... **26**

3.2 Nástroje pro transformace dat ... **27**

3.2.1 ETL vs. ELT ... 28

3.2.2 EAI ... 29

3.3 Metadata ... **29**

3.4 Datový sklad ... **30**

3.5 Datové tržiště ... **31**

3.6 Dočasná úložiště dat ... **32**

3.7 Dolování dat ... **32**

3.8 Multidimenzionalita uložení dat ... **32**

3.8.1 Multidimenzionalita v relačních databázích ... 32

3.8.2 OLAP ... 35

3.9 Reportingové nástroje ... **37**

3.9.1 Dashboard ... 37

3.10 Analytické nástroje ... **38**

3.11 Webové rozhraní ... **39**

4 Open source nástroje Business Intelligence 40
4.1 Pentaho 42
4.1.1 O společnosti Pentaho 42
4.1.2 Pentaho Data Integration 43
4.1.3 Pentaho Reporting 43
4.1.4 Pentaho Metadata Editor 44
4.1.5 Mondrian 44
4.1.6 Pentaho Data Mining 44
4.1.7 Pentaho Business Intelligence Server 45
4.2 SpagoBI 45
4.2.1 O společnosti SpagoWorld 45
4.2.2 Moduly SpagoBI 46
4.2.3 Reporting 47
4.2.4 Multidimenzionální analýza (OLAP) 47
4.2.5 Grafy 47
4.2.6 Data mining a ETL 47
4.3 TIBCO JasperSoft 47
4.3.1 O společnosti TIBCO JasperSoft 48
4.3.2 JasperReports Server 48
4.3.3 JasperReports Library 49
4.3.4 JasperSoft studio 49
4.3.5 JasperSoft ETL 49
4.4 Ostatní projekty 49
4.4.1 BIRT 50
4.4.2 CloverETL 50
5 Porovnání vybraných nástrojů dle zvolených kritérií 51
5.1 ETL nástroje 52
5.1.1 Pentaho Data Integration (Kettle) 52
5.1.2 Talend Open Studio for Data Integration 55
5.1.3 CloverETL 56
5.2 Reportingové nástroje 58
5.2.1 Pentaho Report Designer 58
5.2.2 BIRT 60
5.2.3 JasperReports 62
5.3 Porovnání a zhodnocení vybraných nástrojů 64

6 Zpracování dat ... 68
6.1 Návrh datového skladu ... 68
6.2 Proces zpracování dat v PDI ... 71
6.3 Transformace dat ... 71
6.3.1 Transformace zdrojových dat ... 73
6.3.2 Transformace pro sjednocení dat z různých zdrojů ... 77
6.3.3 Transformace pro import do databáze ... 78
6.3.4 Transformace pro tvorbu dokumentace ... 79
7 Tvorba reportů ... 80
7.1 Časový průběh veličiny ... 80
7.1.1 Technické řešení ... 81
7.1.2 Popis reportu ... 81
7.1.3 Ukázky ... 82
7.2 Report dokumentace (geologických) průzkumných objektů ... 83
7.2.1 Technické řešení ... 83
7.2.2 Popis reportu ... 83
7.2.3 Ukázky ... 84
7.3 Multikriteriální analýza ... 85
7.3.1 Technické řešení ... 85
7.3.2 Popis reportu ... 85
7.3.3 Ukázky ... 86
8 Zhodnocení přínosů řešení ... 87
Závěr ... 91
Citace ... 93
Bibliografie ... 98

Seznam obrázků

Obrázek 1: Vztahy mezi komponenty BI 26
Obrázek 2: Proces ETL 28
Obrázek 3: Schéma hvězdy 34
Obrázek 4: Schéma souhvězdí 34
Obrázek 5: Schéma sněhové vločky 35
Obrázek 6: OLAP kostka 36
Obrázek 7: Dashboard 38
Obrázek 8: Prostředí Pentaho Data Integration 53
Obrázek 9: Prostředí Talend Open Studio 55
Obrázek 10: Prostředí CloverETL 57
Obrázek 11: Prostředí Pentaho Report Designer 59
Obrázek 12: Prostředí BIRT 61
Obrázek 13: Prostředí JasperReports 63
Obrázek 14: Model datového skladu 70
Obrázek 15: Transformace Vrty_Borings 74
Obrázek 16: Transformace PKU Observations 75
Obrázek 17: Funkce PDI pro členění textového dokumentu dle kotevních bodů 76
Obrázek 18: Inklinometrie 77
Obrázek 19: Navazující transformace 78
Obrázek 20: Transformace pro import do databáze 79
Obrázek 21: Časový průběh veličiny 82
Obrázek 22: Dokumentace objektů 84
Obrázek 23: Multikriteriální analýza 86

Seznam tabulek

Tabulka 1: Kritéria hodnocení vybraných nástrojů 52
Tabulka 2: Hodnocení nástroje Pentaho Data Integration 54
Tabulka 3: Hodnocení nástroje Talend Open Studio for Data Integration 56
Tabulka 4: Hodnocení nástroje CloverETL 58
Tabulka 5: Hodnocení nástroje Pentaho Report Designer 60
Tabulka 6: Hodnocení nástroje BIRT Designer 62
Tabulka 7: Hodnocení nástroje JasperReports 64
Tabulka 8: Porovnání hodnocení vybraných ETL nástrojů 64
Tabulka 9: Porovnání hodnocení vybraných reportingových nástrojů 66
Tabulka 10: Popis tabulek datového modelu EI MARE 69
Tabulka 11: Přehled nejčastěji využívaných komponent v PDI 72
Tabulka 12: Kalkulace nákladů 88

Seznam zkratek

BI	Business Intelligence
CRM	Customer relationship management
DMA	Data Mart
DP	Diplomová práce
DSA	Data Staging Area
DWH	Data Warehouse
EAI	Enterprise Application Integration
EDMS	Environmental Data Management Software
ELTA	Extract, Transform, Load, Analyse
ELT	Extract, Transform, Load
ETL	Extract, Transform, Load
EPL	Eclipse Public Licence
ERP	Enterprise resource planning
GPL	General Public License
HOLAP	Hybrid Online Analytical Processing
ICT	Information and Communication Technologies
JDBC	Java Database Connectivity
MOLAP	Multidimensional Online Analytical Processing
MS	Microsoft
OLAP	Online Analytical Processing
OLTP	Online Transaction Processing
PDI	Pentaho Data Integration
PRD	Pentaho Report Designer

ROLAP	Relational Online Analytical Processing
RSS	Really Simple Syndication
SCM	Supply Chain Management
SQL	Structured Query Language
WWW	World Wide Web

Úvod

Business Intelligence (BI) hraje v dnešní době stále větší roli v oblastech informatiky, ekonomie a podnikání. Právě v podnicích je BI nezastupitelným článkem v získávání informací a znalostí pro podporu řízení a rozhodování. Nástroje BI lze využít pro vyhodnocování ukazatelů nejen pro oblast marketingu, financování, logistiky, výroby, řízení lidských zdrojů, ale i v medicíně, přírodních vědách aj. Tématem této diplomové práce (DP) je hodnocení kvality přírodních vod za použití nástrojů BI. Vzhledem k požadavku využít při řešení open source nástroje je stanoven dílčí cíl porovnat a vyhodnotit tyto nástroje. Práce obsahuje část teoretickou a část praktickou a je členěna do osmi kapitol.

Teoretická část práce popisuje principy a metody, jež jsou využity v části praktické. Současným stavem řešené problematiky se zabývá kapitola první. Druhá kapitola je věnována samotnému oboru BI. Kromě definice, stručné historie a využití jsou zde uvedeny moderní trendy. Ve třetí kapitole je popsána architektura BI. Jelikož je celý proces BI velmi komplikovaný a provázaný, jsou komponenty řazeny chronologicky dle průběhu procesu BI. Čtvrtá kapitola se zabývá open source nástroji BI a typy open source licencí. Pro úplnost je zde srovnáno řešení komerční s řešením nekomerčním. Nástroje BI jsou řazeny dle jednotlivých společností, které jsou nejprve stručně představeny a následně jsou popsány samotné nástroje. Důraz v popisu je kladen na odlišnosti, zajímavosti a funkcionalitu jednotlivých nástrojů.

Pátá kapitola je věnována praktické části zabývající se porovnáním vybraných open source nástrojů BI. K těmto účelům jsou definována kritéria, podle kterých jsou nástroje srovnávány. Těmito kritérii jsou: *přehlednost prostředí a ovládání*, *funkcionalita*, *technická a uživatelská podpora*, *instalace a kompatibilita na různých operačních systémech* a *kompatibilita mezi jednotlivými aplikacemi*. Konkrétní řešení problematiky hodnocení kvality přírodních vod je v kapitole šesté a sedmé. Zpracování dat se týká jejich načítání, úprav a ukládání do cílové databáze v jednotném formátu. Reporty jsou pak zaměřeny na analýzy těchto dat. Konkrétně se jedná o *multikriteriální analýzu*, která je

využívána při predikcích projevů eutrofizace[1], dále se jedná o *časový průběh veličiny*, ten je věnován vlastnostem zkoumaných veličin (chemických prvků), a o *dokumentaci (geologických) průzkumných objektů*. Účelem řešení dané problematiky je zlepšení podpory rozhodování o kvalitě přírodních vod.

V závěrečné kapitole jsou formulovány výsledky a přínosy této DP. Dosažené výsledky jsou prezentovány a srovnány s úvodními stanovenými cíli.

Praktická část práce je řešena v rámci projektu MARE, ve kterém je vyvíjen informační systém pro podporu rozhodování o využití krajiny po rekultivaci. Pro projekt MARE řešil autor této DP již bakalářskou práci [1], jejímž cílem bylo porovnání dvou odlišných přístupů (vlastní aplikace a volně dostupné nástroje ETL) při zpracování dat. Stěžejními body práce bylo vytvoření několika transformací sloužících ke zpracování dat o životním prostředí.

[1] *Eutrofizace* (úživnost) je proces zvyšování obsahu živin ve vodách a půdách

1 Zhodnocení současného stavu řešené problematiky

Problematikou systémů pro správu dat o životním prostředí se zabývá několik společností, které vyvíjejí vlastní aplikace a systémy. Tyto systémy jsou známy jako *Environmental data management software* (EDMS). Jejich úkolem je správa dat uložených v databázi. EDMS dále umožňuje provádět úkony specifické pro data o životním prostředí, jako jsou import a export dat, vizualizace a reporting. Jedná se o komerční systémy a neexistuje žádné volně dostupné či alternativní řešení, které by vyhovovalo požadavkům projektu. Z těchto důvodů je kapitola věnována současnému stavu v oblasti open source nástrojů BI. [2]

Tématice BI – od popisu komponent přes srovnávání komerčních i nekomerčních nástrojů až po trendy v oblasti BI – se věnuje nespočet prací, ať už se jedná o knihy, vysokoškolské kvalifikační práce či články v odborných časopisech nebo internetových portálech.

Některé zdroje jsou svým zpracováním a kvalitou přínosné, některé jsou nedostačující. Z tohoto důvodu byla před psaním této DP provedena rešerše v oblasti řešeného tématu. Zdroje jsou rozděleny dle dílčích kategorií.

1.1 Vysokoškolské kvalifikační práce

Rešerše vysokoškolských kvalifikačních prací byla zaměřena na dvě oblasti. První z nich jsou prostředky (komponenty) BI. Druhou oblastí je porovnání open source nástrojů BI.

Zajímavé české vysokoškolské práce zabývající se tématikou BI pocházejí především od studentů Vysoké školy ekonomické v Praze. K odbornosti vybraných kvalifikačních prací přispívá i fakt, že jejich vedoucím byl v mnoha případech J. Pour, odborník, jenž se tématům souvisejícím s BI věnuje řadu let a napsal několik publikací z této oblasti. Některé z knih, jejichž spoluautorem je právě J. Pour, byly použity jako zdroje při psaní této DP.

Diplomová práce Z. Filipčíka: „*Nástroje Business Intelligence jako open source*“ [3] se zabývá popisem a porovnáním vybraných open source nástrojů. K těmto účelům autor

navrhl hodnotící kritérium, kdy jednotlivým kategoriím nástrojů BI přiřazuje body s vahami. Práce se nezaměřuje na projekty jednotlivých společností jako na celky, ale z každé oblasti BI (reporting, ETL atd.) je vybráno několik nástrojů od různých společností, které jsou mezi sebou porovnávány.

Další prací je „*Srovnání komerčních BI reportingových nástrojů s nástroji Open Source*“ [4] od J. Bednáře. Autor stručně, avšak kvalitně zpracoval přehled architektury BI. Srovnání reportingových nástrojů je velice podrobné díky širokému spektru hodnotících kritérií. V závěru práce autor demonstruje získané znalosti na vytvoření vlastního reportu.

Bakalářská práce „*Analýza trhu open source business inteligence*“ [5] V. Formánka detailně popisuje vybrané open source projekty, chybí ovšem výraznější srovnání či doporučení. Tato práce byla spíše inspirací, než zdrojem informací.

1.2 Odborná literatura

Mezi nejznámější české knihy zabývající se touto oblastí patří bezesporu publikace od J. Poura. Zdrojem informací pro teoretickou část práce byly od tohoto autora (a spol.) knihy „*Business Intelligence v podnikové praxi*“ [6] a „*Business Intelligence. Jak využít bohatství ve vašich datech*“. [7] V obou knihách jsou detailně popsány principy, architektura a komponenty BI.

Jedněmi z nejodbornějších zahraničních publikací jsou knihy R. Kimballa. Zejména dílo „*The Kimball Group reader: relentlessly practical tools for data warehousing and business intelligence*“ [8] bylo hodnotným zdrojem při psaní této DP, protože se věnuje konkrétním problémům při operacích s daty nejen v datových skladech, ale i v celém procesu BI. V knize jsou do detailu popsány transformační procesy dat, způsoby manipulace s daty a možnosti jejich uložení. Tato publikace je zajímavá především díky své využitelnosti v praxi, jelikož je pojata jako návod, jak vytvořit funkční proces BI, od plánování procesu přes sběr dat, transformace dat až po analýzu.

Další publikací je „*The performance management revolution: business results through insight and action*“ [9] od H. Dresnera. Kniha je prakticky zaměřena a popisuje cestu,

jakou se má podnik vydat, chce-li úspěšně proměnit strategii v plány, plány v realizaci a realizaci ve výsledky.

Dalším knižním zdrojem, nikoliv teoretických znalostí, ale praktických dovedností, byly knihy věnující se nástrojům BI od společnosti Pentaho. První z nich je *„Pentaho Kettle solutions: building open source ETL solutions with Pentaho Data Integration“*. [10] Tato kniha vysvětluje jak nainstalovat, nastavit a ovládat nástroj ETL od společnosti Pentaho – *Pentaho Data Integration (Kettle)*. Velice podrobně popisuje samotnou aplikaci i s příklady, jak sestavit transformace dat od těch nejjednodušších po transformace komplexní.

Kniha *„Pentaho 5.0 Reporting by example: Beginner's guide“* [11] je příručkou (nejen pro začátečníky) k nástroji *Pentaho Report Designer* a je psána v duchu *user-friendly*[2], přesto odborně. Kapitoly jsou věnovány jednotlivým částem reportingového procesu a obsahují praktické příklady.

Poslední knihou od Pentaho, která posloužila jako přehledný zdroj informací o celém procesu BI, je publikace *„Pentaho solutions: Business intelligence and data warehousing with Pentaho and MySQL“*. [12]

1.3 Internet

Informace o open source nástrojích BI jednotlivých společností byly získány z webových stránek Pentaho [13], SpagoWorld [14], JasperSoft [15], BIRT [16] a CloverETL [17].

[2] Termín zažitý pro uživatelsky přístupný a přívětivý obsah

1.4 Databáze článků

Pro širší přehled o open source nástrojích BI a o rozdílných přístupech jednotlivých společností zabývajících se BI byly využity databáze odborných článků poskytnutých Technickou univerzitou v Liberci.

Článek „*Report Generation using Business Intelligence Tools: A comparative Stu*dy“ [18] se zabývá poměrně podrobným porovnáváním dvou nástrojů *JasperSoft* a *Vanilla*. Porovnány jsou v mnoha kategoriích, od podpory programovacího jazyka přes typy nástrojů až po podporu operačních systémů.

Jiný odborný článek, „*A Five-Layered Business Intelligence Architecture*“ [19], se věnuje popisu architektury podle vrstev. Autor popisuje zdrojovou vrstvu, ETL vrstvu, vrstvu datových skladů a vrstvu koncových uživatelů. Pátou vrstvou jsou metadata, se kterými je možno pracovat ve kterékoliv z předchozích vrstev.

Ve článku „*Knowledge Management through the implementation of Business Intelligence tools“* [20] autor popisuje vztah managementu a BI. Zaměřuje se na obecný popis BI, na nástroje, které BI využívá, ale věnuje se i výhodám využití BI v ekonomické sféře.

Článek „*ELTA New Approach in Designing Business Intelligence Solutions in Era of Big Data*“ [21] je věnován novému trendu v oblasti transformací dat. Tím je proces ELTA (*Extract, Load, Transform, Analyse*). V článku jsou dále porovnány procesy ELT (*Extract, Load, Transform*) a ETL, z nichž proces ELTA vychází.

2 Business Intelligence

První kapitola práce je věnována pojmu BI, jeho definici, historii a trendům. Dále je zde uvedeno, kdo a proč by měl systém BI využívat a jaké jsou problémy při zavádění BI do podniků.

2.1 Definice Business Intelligence

Pojem Business Intelligence se v dnešní době objevuje čím dál častěji. BI se vyskytuje v oborech informatiky a ekonomie, ale také v podnikání. Právě v podnikání Business Intelligence zcela zásadním způsobem ovlivňuje kvalitu, výkonnost i efektivitu ekonomických subjektů.

Česká společnost pro systémovou integraci zabývající se výměnou informací a názorů v oblasti informačních systémů charakterizuje BI jako *„sadu procesů, aplikací a technologií, jejichž cílem je účinně a účelně podporovat rozhodovací procesy ve firmě.“* [22 s. 300]

Jiná definice, dle Gály a kol., říká, že *„business intelligence (BI) je sada procesů, know-how, aplikací a technologií, jejichž cílem je účinně a účelně podporovat řídící aktivity ve firmě. Podporují analytické, plánovací a rozhodovací činnosti organizací na všech úrovních a ve všech oblastech podnikového řízení, tj. prodeje, nákupu, marketingu, finančního řízení, controllingu, majetku, řízení lidských zdrojů, výroby a dalších.“* [23 s. 217]

Z těchto definic vyplývá, že BI pomáhá hlavně manažerům, podnikovým analytikům, ale i běžným uživatelům v rozhodovacích činnostech. V tomto smyslu BI pracuje s daty, která jsou pomocí nástrojů BI upravována, analyzována, interpretována a následně poskytnuta v požadované formě.

2.2 Vývoj Business Intelligence

Jedním z průkopníků v oblasti BI byl pracovník IBM H. P. Luhn. Ve svém článku v roce 1958 charakterizoval BI jako *„the ability to apprehend the interrelationships of presented facts in such a way as to guide action towards a desired goal“* [24 s. 319], tedy *schopnost pochopit vzájemné vztahy prezentovaných faktů takovým způsobem, který umožní provést akci k dosažení požadovaného cíle.* V této době ovšem neexistovaly technické prostředky, kterými by se teoretické základy BI daly přenést do praxe.

S rozvojem využívání počítačových center a terminálů se na konci 70. let minulého století začaly objevovat první pokusy o vytvoření automatického systému pro podporu manažerských a analytických úloh a zpracování dat, ale výrazného pokroku bylo dosaženo až koncem 90. let minulého století. V té době se také začínají prosazovat tzv. *Enterprise Resource Planning* (ERP). Jedná se o informační systémy, které spojují a automatizují řady procesů ve firmách. [25]

O několik let později, v druhé polovině devadesátých let, se začaly masově nasazovat nástroje BI. Způsobeno to bylo zejména díky tehdejšímu nárůstu výpočetního výkonu a velkému množství elektronických dat.

2.3 Využití Business Intelligence v praxi

V dávných dobách ti nejúspěšnější obchodníci, velitelé a vládci věděli, že jedním ze základních pilířů jejich úspěchu je informovanost. Člověk, který se dokázal poučit z chyb, vyvarovat se jich, či je předpovídat, byl vždy o krok napřed před ostatními. Nejinak je tomu dnes.

V dnešní době je konkurence nesmírně obrovská. Aby mohla být firma na trhu úspěšná, potřebuje informace o konkurenci, financích, zaměstnancích, vývoji trhu, zkrátka o všem, co s podnikáním souvisí. Tyto informace by byly takřka bezcenné, kdyby nebyly účinným způsobem zanalyzovány a využity ve prospěch firmy. K těmto účelům slouží právě nástroje BI, které jednoduchým, účinným a přehledným způsobem dokážou data zpracovat a poskytnout v žádané podobě.

Dnes nedokáže větší firma bez oddělení či odborníků zabývajícími se tímto oborem úspěšně fungovat, a proto se stal obor BI nedílnou součástí podnikání. Na dnešním trhu je mnoho společností, které se problematice věnují a poskytují v tomto oboru služby.

Tyto myšlenky shrnul expert na modelování, analýzu a BI J. Wu, který v roce 2005 předpověděl, že během následujících let budou aplikace BI ve všech středních nebo větších organizacích běžnou záležitostí a ti, kteří zapojí technologie BI nejefektivněji, se ve svých oborech stanou lídry a odliší se od svých konkurentů. To je situace, kterou můžeme dnes běžně vidět v praxi. [26]

2.4 Business Intelligence v praxi

Ačkoliv se nasazení systému BI do podniku může zdát jako záležitost, která vždy pomůže podniku ihned získat výhody, v praxi tomu tak není. Ve světě i u nás existuje celá řada problémů, které úspěšnému zavedení systémů BI brání.

Jedním z problémů je vztah mezi manažerskými útvary a útvary ICT (*Information and Communication Technologies*). Funkcionalita systémů BI jde souběžně s funkcionalitou manažerských útvarů a tím vykonává, resp. nahrazuje jejich činnosti. Dalším problémem může být nekvalita zdrojových dat. V případě, že jsou chybná data zpracována a následně vyhodnocena, dochází ke zkreslování či ztrátě informací, a tím dojde ke znehodnocení celého projektu. Proto se dnes více jak 80 % nákladů v systémech BI vynakládá právě na zajištění kvality dat. [26]

Předpokladem pro úspěšné nasazení systémů BI je, aby firma i její útvary akceptovaly změny, které s sebou BI přináší. Všichni, kdo přijdou s těmito systémy do styku, musí být dostatečně a kvalifikovaně připraveni.

2.5 Perspektivy Business Intelligence

Neustále rostoucí počet elektronických dat se zákonitě projevuje i v oblasti BI. Vývoj BI neustále roste a do budoucna se očekává rozvoj jak v rovině aplikační, tak v rovině technologické.

K nejvýznamnějším trendům patří:

- Zpracování nestrukturovaných a semistrukturovaných dat
- Integrace BI a zdrojových aplikací
- Řešení metadat
- Integrace podnikových procesů a BI
- Konvergence technologií a nástrojů
- Více samoobslužnosti, méně IT
- Dotazování přirozeným jazykem
- Informace kdykoliv a kdekoliv
- Zpracování velkých objemů dat

Všechny tyto trendy směřují k rychlejšímu a efektivnějšímu využívání BI systémů nejen pro podnikové účely. Snahou je co nejjednodušeji a nejpřehledněji zpřístupnit potřebné informace skrze BI aplikace pro co nejširší počet uživatelů v rámci podniků a organizací.

V České Republice se BI rozvíjí od počátku devadesátých let minulého století. V současnosti využívá BI aplikací a systémů kolem 50 % podniků a toto číslo se neustále zvyšuje. [6]

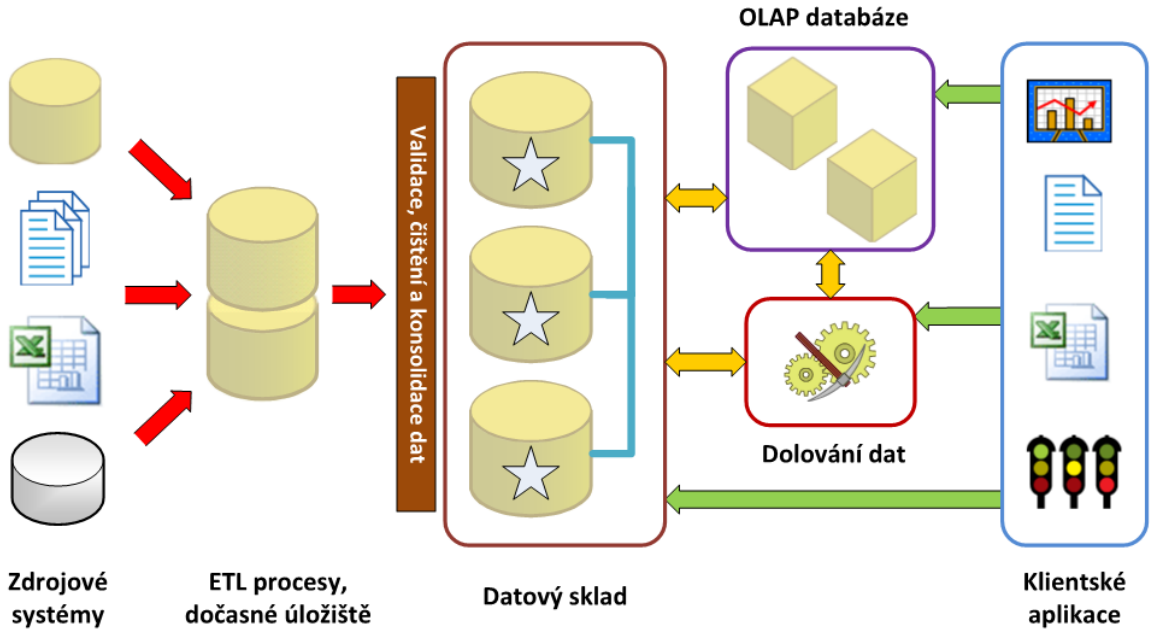
3 Architektura Business Intelligence

Tato kapitola je věnována architektuře BI – tedy prostředkům a nástrojům, které jsou v oboru BI využívány. BI nemá jeden nástroj, který by zajišťoval veškerou funkcionalitu, ale jedná se o souhrn komponent, které dohromady tvoří celek pro řešení úloh v oblasti BI.

Architektura BI se dá obecně rozdělit do pěti kategorií podle toho, v jaké fázi BI a k jakému účelu jsou komponenty, prostředky či nástroje využívány. Jedná se o:

- *Zdrojové systémy* – mezi zdrojové systémy řadíme takové systémy, jež nejsou součástí BI a slouží jako zdroj surových dat pro komponenty BI. Úkolem BI je tato, často heterogenní, data analyzovat a vybrat relevantní data pro potřeby podniku.
- *Komponenty pro transformace dat* – cílem těchto komponent je převést a upravit data do požadované struktury pro potřeby podniku. K těmto účelům jsou využívány nejčastěji dva nástroje, ETL a ELT.
- *Prostředky pro uložení dat* – zpracovaná data je třeba uložit. K tomu slouží datová úložiště v podobě datových skladů, datových tržišť, dočasných úložišť a operativních úložišť.
- *Analytické komponenty* – tyto komponenty pomáhají ke zpracování dat pro potřeby získání požadovaných informací. Tyto informace pak bývají využity pro podporu rozhodování v podniku. Mezi analytické komponenty se řadí dolování dat, OLAP (*Online Analytical Processing*) nástroje a reportingové nástroje.
- *Prezentační vrstva* – jedná se o vrstvu BI, která, jak její název napovídá, slouží k prezentaci informací získaných jejich analýzou. Pro samotnou prezentaci dat slouží analytické nástroje, webové nástroje a mobilní aplikace.

Všechny uvedené komponenty BI mají mezi sebou, jak již bylo v textu naznačeno, určité vztahy. Tyto vztahy popisuje *obrázek 1*. Zdrojem dat pro systémy BI jsou zdrojové (produkční) databáze, ve kterých jsou uložena surová data. Z nich jsou pomocí ETL (ELT) nástrojů data extrahována a zpracovávána dle požadované struktury do datových skladů, jež jsou realizovány v prostředí relačních databázových systémů. Z těchto zpracovaných dat získávají manažeři informace pomocí analytických nástrojů či nástrojů reportingu.



Obrázek 1: Vztahy mezi komponenty BI
Zdroj: [27]

3.1 Zdroje dat

Zdrojem dat nemusí být vždy velká databáze, ale může to být i soubor aplikace databázového typu (Microsoft Access), tabulkový kalkulátor (Microsoft Excel) nebo textové soubory s pevně danou strukturou, tzv. *flat files*.

Další formou uložení dat je zdrojová databáze. Jedná se o nejjednodušší strukturovanou formu uložení dat využitelnou v oblasti BI. Zdrojové (produkční) databáze (OLTP, *Online Transactional Processing*) jsou databáze aplikací, ze kterých BI databáze získávají data. Sami o sobě tedy nejsou součástí BI. [6]

Příkladem zdrojové databáze jsou např. databáze aplikací ERP, CRM (*Customer Relationship Management*) a SCM (*Supply Chain Management*), jež jsou realizovány pomocí databázových systémů, jako jsou ORACLE, MS SQL Server apod. Velmi důležitým faktorem ve zdrojích dat je jejich kvalita, která přímo ovlivňuje úroveň a využitelnost BI systémů. [6]

3.2 Nástroje pro transformace dat

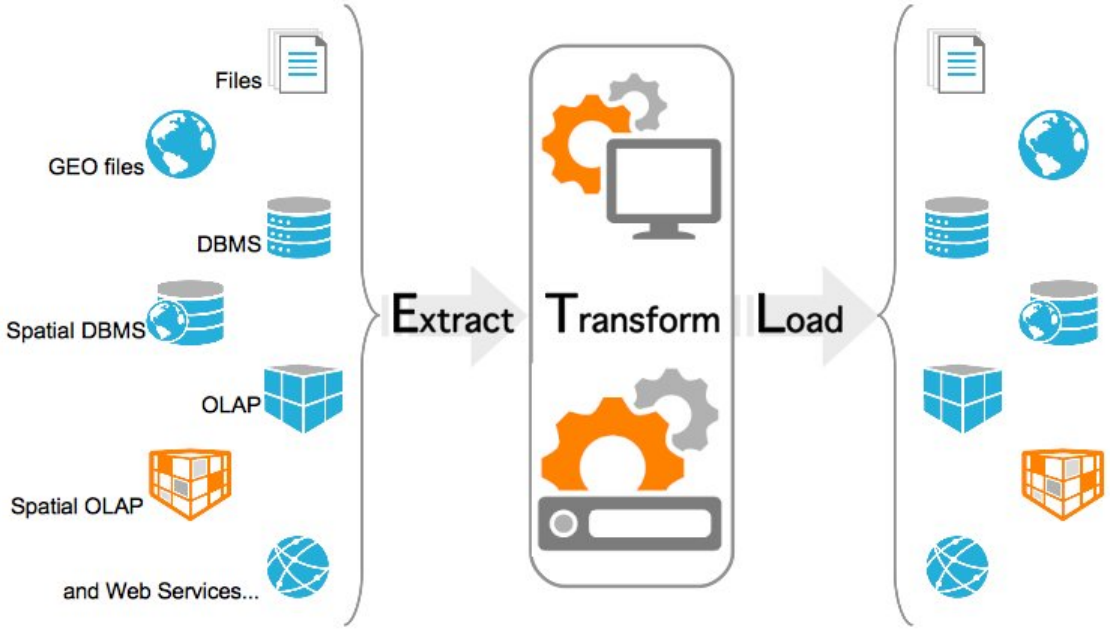
Data ze zdrojových systémů je potřeba před uložením do datových úložišť upravit do struktury, která odpovídá potřebám podniku a umožňuje jejich další zpracování.

Pro tyto účely jsou dnes nejčastěji využívány nástroje ELT, jež postupně nahrazují nástroje ETL. [21] Oba typy nástrojů využívají podobných principů, jež jsou nastíněny v *obrázku 2* a popsány v následujících odstavcích.

První částí obou procesů je *extract*, neboli získání dat, se kterými se dále pracuje. Nejedná se o jednorázovou akci, ale zpravidla je to činnost, která je prováděna po delší časový úsek. *Extract* je nejdůležitější část ETL (ELT) procesu, protože správné extrahování dat je nezbytné pro navazující procesy.

Druhou částí procesu ETL (u ELT se jedná o část třetí), tedy *transform*, je samotná úprava neboli čištění dat. Zde jsou data pomocí vhodných nástrojů převáděna do požadované podoby. Toho je docíleno nástroji jako jsou *filtrování*, *třídění*, *agregace*, *normalizace*, *denormalizace*, práce s hodnotami *null*, *slučování tabulek* apod. Po takovýchto úpravách by měla být zajištěna jednotnost, úplnost a správnost dat. Ve většině případů zabere tato část procesu ETL (ELT) nejvíce času.

Poslední částí procesu ETL (u ELT je to část prostřední) je *load* čili naplnění očištěných dat do cílového úložiště. Nejčastěji jsou data nahrávána do datových skladů, ale obecně stačí data uložit do jakékoliv formátované datové struktury. Tyto formáty musí být navrženy tak, aby co nejlépe odpovídaly potřebám řízení podniku.



Obrázek 2: Proces ETL
Zdroj: [28]

3.2.1 ETL vs. ELT

Rozdíl je na první pohled zřejmý. V ETL nástrojích je aplikován proces transformace po extrakci ještě před načtením. Data jsou tedy nahrávána do úložišť pouze během poslední fáze procesu ETL.

Oproti tomu se v ELT procesu nejdříve data extrahují, poté jsou nahrána do úložiště a až nakonec je aplikována transformace.

Na rozdíl od ETL, kdy jsou data zpracovávána pomocí transformačních nástrojů, je v procesu ELT při transformaci využíváno dotazovacího jazyka v samotných databázích, kde jsou data uložena. I proto je ELT náročnější z hlediska nároků na databázi. Výhodou ELT je oddělení procesu transformace od procesů extrakce a nahrávání. Data mohou být transformována později dle aktuálních potřeb.

3.2.2 EAI

Enterprise Application Integration (EAI) je další způsob, jakým spravovat data. EAI na rozdíl od ETL (ELT) pracuje v reálném čase. Primárně se nezaměřuje na transformaci dat, ale na jejich integraci. Integrací je myšlena snaha o propojení dvou systémů s různou abstrakcí s cílem co nejvyšší nezávislosti mezi nimi. Systém EAI je nejčastěji uplatňován v transformační vrstvě, kde je systému EAI využito pro přenos dat do úložišť v reálném čase. Tento systém je nazýván *Real-Time Data Warehouse* (systém datových skladů v reálném čase). [7]

3.3 Metadata

Pod pojmem metadata se rozumí *data o datech*. Jsou chápána jako popis jakýchkoliv datových prvků. Metadata jsou kromě služby WWW využívána k popisu digitálních fotografií, zvukových nahrávek, ale také je jich využíváno např. ve vyhledávacích katalozích knihoven.

V podniku slouží pro dokumentaci implementací informačních systémů. Z pohledu BI metadata popisují datové modely, funkce, transformační pravidla, reporty apod. Součástí metadat je i vymezení ekonomického (podnikového) obsahu dat. Metadata se dle Poura [6 s. 36] promítají do následujících komponent architektury BI.

Metadata zdrojových systémů popisují zdrojová data, slouží k jejich identifikaci a pochopení vztahů mezi nimi. Výrazně tak pomáhají zlepšit procedury a funkcionalitu zdrojových systémů.

Metadata datových pump[3] nejčastěji popisují původ dat, která jsou pak dále využívána v systému BI. Jelikož jsou data po celou dobu procesu jejich zpracování přenášena a ukládána do navazujících vrstev, je vhodné tato data popsat, aby bylo později možné zjistit jejich původ.

[3] Pojem datová pumpa označuje nástroje sloužící k plnění dat do datových skladů

Metadata uživatelské vrstvy hrají svou úlohu v konečné části systému BI. Uživatelská vrstva, kterou tvoří reporty, dashboardy, analýzy a další rozhraní, umožňuje koncovým uživatelům pohled na požadovaná data. Metadata uživatelské vrstvy pomáhají při pochopení vztahů mezi informacemi (daty), ale také umožňují zpětný pohled na původ dat, která byla do koncové vrstvy přenesena z nižších systémů.

Metadata databázové vrstvy BI slouží nejen k pochopení obsahu samotných dat, ale také zvyšují účinnost databázového systému BI. V databázích slouží metadata k popisu sloupců, řádků, ale i obsahu samotných dat či objektů. Zlepšují vyhledávání v databázích, umožňují lepší pohledy do databáze a mohou dokonce sloužit i jako technická dokumentace.

3.4 Datový sklad

Jak název napovídá, jedná se o místo, kde jsou data shromážděna a uložena na jednom místě, a to ze všech zdrojů, ve kterých se tato data mohou vyskytovat. Data jsou v datovém skladu (*Data Warehouse*, DWH) uchovávána v takovém formátu, aby nad nimi mohly být prováděny analytické operace.

Pour [6 s. 24] uvádí definici W. Inmona[4], která říká, že *„datový sklad je integrovaný, konsolidovaný, subjektově orientovaný, stálý a časové rozlišený souhrn dat, uspořádaný pro podporu potřeb managementu.“*

Pojmy použité v této definici lze vyložit takto:

- subjektově orientovaný – data jsou rozdělena dle jejich typu a ne podle aplikací, kde vznikla
- konsolidovaný – z různých zdrojů a struktur jsou data uspořádána do jedné formy
- integrovaný – data jsou ukládána v rámci celku (podniku)
- stálý – DWH slouží převážně pro čtení, data v nich se (až na výjimky) neaktualizují
- časově rozlišený – datový sklad obsahuje dimenzi času

[4] Zakladatel *datawarehousingu* – oboru zabývajícího se problematikou datových skladů

3.5 Datové tržiště

Datové tržiště (*Data Mart*, DMA) úzce souvisí s datovým skladem. Princip je velice podobný, ale datová tržiště jsou na rozdíl od DWH určena pro specifický okruh uživatelů (např. určité oddělení ve firmě) a z tohoto důvodu obsahuje tržiště data vztahující se k dané oblasti.

DMA jsou tvořeny na základě konkrétních požadavků. Z tohoto pohledu odlišujeme dva přístupy.

Prvním přístupem je postupné vytváření DMA, jehož autorem je R. Kimball[5]. Principem je izolovaný DMA, známý také pod názvem *dvouvrstvá architektura*. Tohoto přístupu je využíváno tehdy, jsou-li požadavky na přístup ke konkrétním datům stanoveny dle jednotlivých skupin uživatelů (oddělení). Pro každou skupinu je pak vytvořen vlastní DMA obsahující všechna potřebná data. Jednotlivá tržiště jsou tvořena postupně – v čase, podle potřeb jednotlivých skupin. V 90. letech minulého století Kimball toto řešení přepracoval a vznikl koncept *sběrnicové architektury*. Myšlenkou tohoto konceptu je nezávislost na technologii a platformě a také budování nezávislých DMA integrovaně. Integračním prvkem jsou myšleny sdílené dimenze, kterých je opakovaně využíváno v jednotlivých DMA. Výhodou tohoto přístupu jsou nižší náklady, zkrácení doby návratnosti investic a menší riziko při jejich zavádění než při zavádění datových skladů. Nevýhodou je redundance dat a nejednotnost údajů v jednotlivých tržištích. [6]

Jiným přístupem je *architektura třívrstvá*, kterou navrhl W. Inmon. Principem je vytváření DMA nad jedním centralizovaným DWH. Redundance dat je tedy minimální, avšak pořizovací náklady vyšší a doba realizace delší než v případě přístupu prvního. [6]

[5] Známý jako „*otec*“ Business Intelligence

3.6 Dočasná úložiště dat

Dočasná úložiště dat (*Data Staging Area*, DSA) slouží pro dočasné uložení dat a zajišťují přepravu a kvalitu extrahovaných dat ze zdrojových databází do datového skladu. V dočasném úložišti jsou data neagregovaná, detailní, nekonzistentní a bez časové dimenze. Po zpracování a uložení dat do datového skladu či tržiště jsou data z úložiště odstraněna. [6]

3.7 Dolování dat

Dolování dat (*Data Mining*) je proces, který je označován jako *hledání skrytých souvislostí*. Principem dolování dat je získání relevantních, předem neznámých či neočekávaných informací z rozsáhlého objemu dat. Jedná se o analýzy tvořené z obsahu dat. Pro tyto účely jsou využívány matematické aplikace, nejčastěji z oboru statistiky. Přínosem tohoto procesu je objevování nových skutečností a skrytých souvislostí a své uplatnění má v medicíně, meteorologii, marketingu aj.

3.8 Multidimenzionalita uložení dat

Pro současné požadavky uložení dat nestačí pouze klasická dvourozměrná tabulka omezena řádky a sloupci umožňující pohled do dvou dimenzí. Proto je dnes stále více využíváno specifického uložení dat, které umožňuje uživateli pohled na data z více dimenzí.

Způsob realizace této technologie se dá rozdělit na dvě základní skupiny, a to vyjádření multidimenzionality v relačních databázích a multidimenzionální datový model založený na kostce. [6]

3.8.1 Multidimenzionalita v relačních databázích

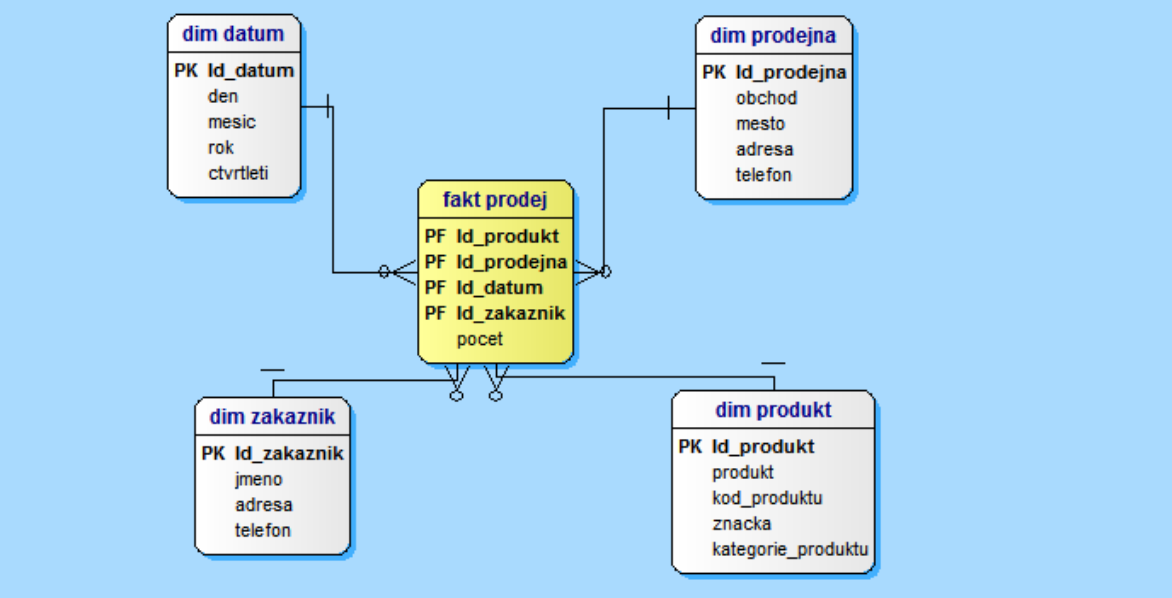
Termín relační databáze definoval v roce 1970 Edgar F. Codd pomocí matematických operací. Dle jeho teorie lze pomocí základních matematických operací *sjednocení*,

kartézský součin, *rozdíl*, *selekce*, *projekce* a *spojení* nebo jejich kombinacemi provést veškeré operace s daty. Dále by měly být dle Codda databáze nezávislé na použitém jazyce a způsobu uložení dat.

Základním prvkem relační databáze je databázová relace. Je to dvourozměrná struktura, jejíž sloupce se nazývají atributy a řádky záznamy. Multidimenzionální modely založené na relačním modelu rozlišují dva typy relací – *tabulky dimenzí* a *tabulky faktů*. Ukazatele, které je třeba analyzovat, jsou ukládány do tabulky faktů. K tabulce faktů se váže pojem *granularita*. Ta představuje určitou úroveň míry podrobností v tabulce faktů. Čím podrobnější data jsou, tím je míra granularity vyšší a naopak. Tabulky faktů jsou propojeny pomocí cizích klíčů s tabulkami dimenzí. Ty obsahují atributy, pomocí nichž můžeme třídit a manipulovat s daty v tabulkách faktů. Tabulka dimenzí je tabulce faktů nadřazená. Vztah mezi tabulkou dimenzí a tabulkou faktů je 1:N.

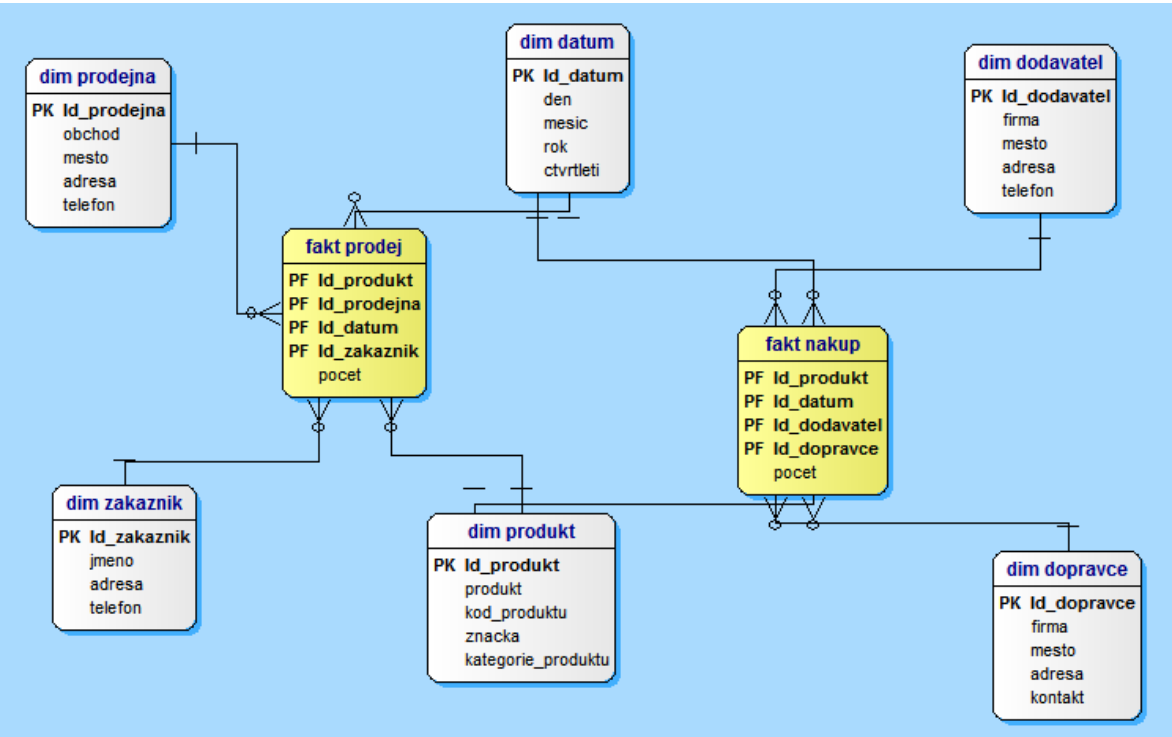
Dvourozměrný model je vhodný pouze pro některá řešení. V případě, že je potřeba využít třetí rozměr (dimenzi), musí se využít rozšířených modelů, jako jsou *schéma hvězdy*, *schéma souhvězdí* či *schéma sněhové vločky*.

Ve *schématu hvězdy* (*star schema*) jsou tabulky dimenzí s tabulkou faktů spojeny pomocí identifikátorů a tvoří tak podobu hvězdy. Kombinací klíčů z tabulek dimenzí jsou určeny sledované hodnoty (ukazatele) v tabulce faktů, mezi dimenzemi neexistují vztahy. Schéma je zobrazeno na *obrázku 3*. Jsou v něm čtyři tabulky dimenzí a jedna tabulka faktů, která obsahuje ukazatel *pocet*, který sleduje počet prodaných kusů.



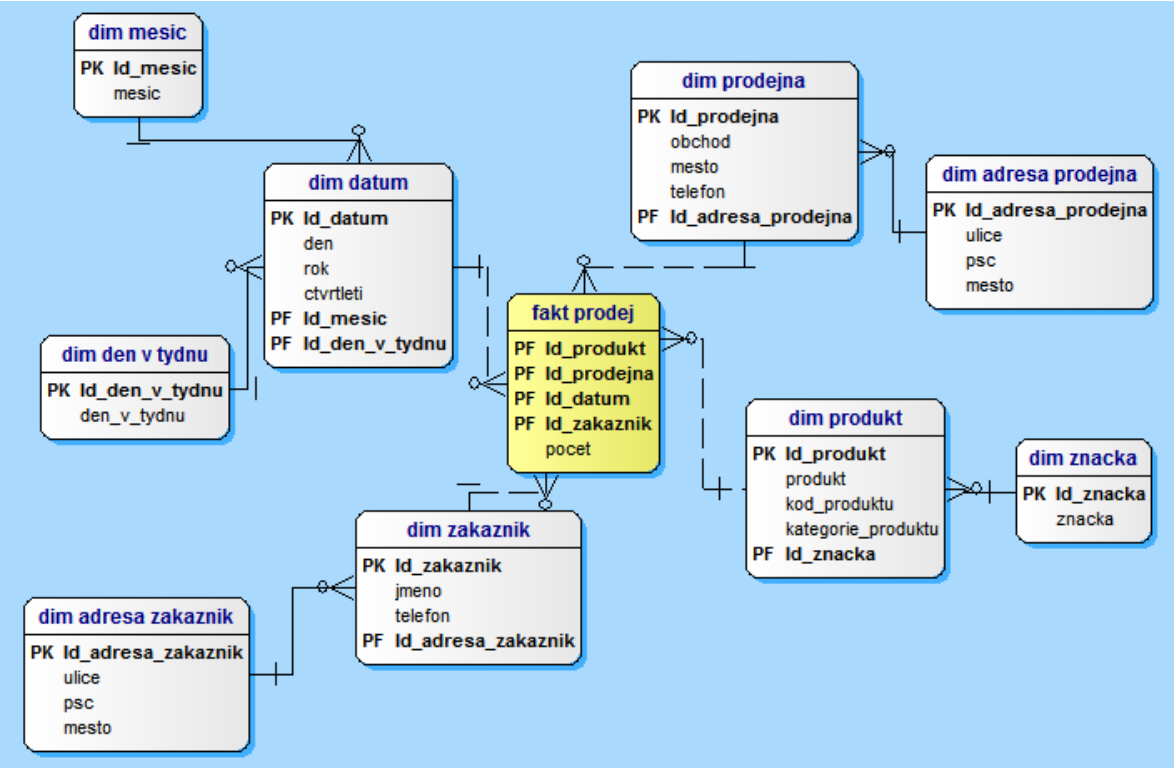
Obrázek 3: Schéma hvězdy
Zdroj: vlastní

Schéma souhvězdí (*constellation schema*) je kombinací několika schémat hvězdy. V modelu je více *tabulek faktů*, z nichž některé sdílejí stejné dimenze. Na *obrázku 4* jsou sledovanými ukazateli počty nakoupeného a prodaného zboží ve žlutě zabarvených tabulkách. V tomto příkladu sdílejí tabulky faktů dimenze *datum* a *produkt*.



Obrázek 4: Schéma souhvězdí
Zdroj: vlastní

Schéma sněhové vločky (*snowflake schema*) je typ hvězdicového schématu. Rozdílem oproti modelu hvězdy jsou normalizované tabulky dimenzí do dílčích tabulek, takže schéma připomíná svým rozložením sněhovou vločku. Na *obrázku 5* jsou tabulky dimenzí *datum*, *produkt*, *zákazník* a *prodejna* normalizovány do dalších tabulek dimenzí. Tabulka faktů *prodej* sleduje počet prodaných kusů.



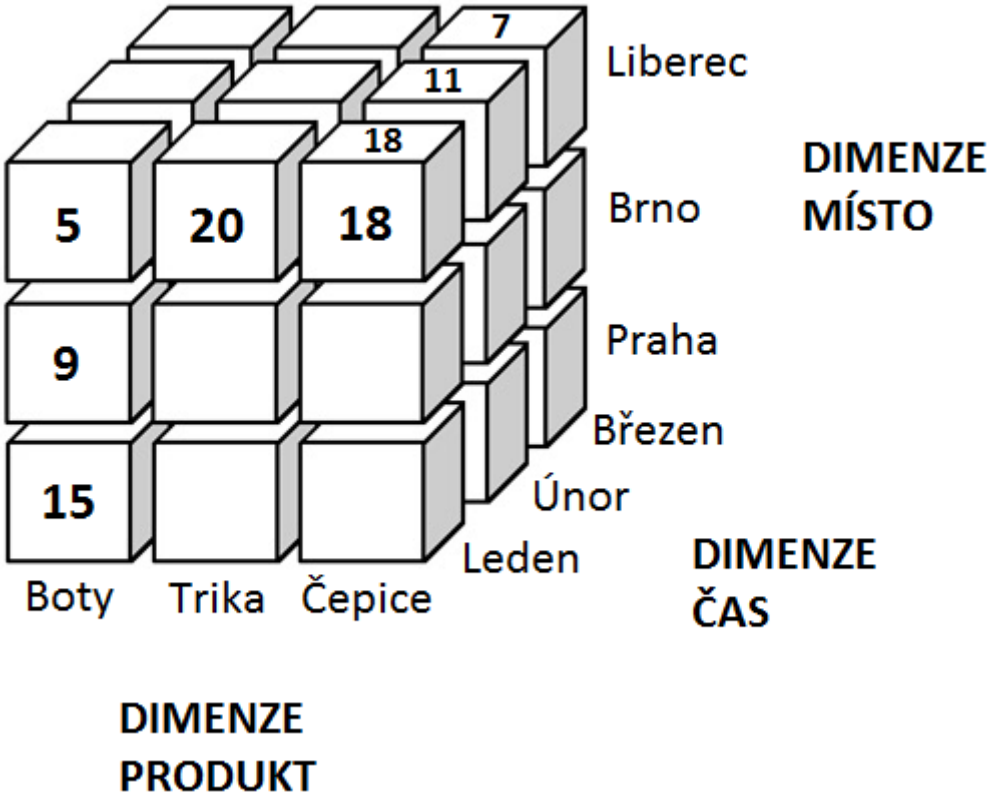
Obrázek 5: Schéma sněhové vločky
Zdroj: vlastní

Důležitou poznámkou je, že databáze datového skladu nemůže být většinou řešena jedním typem z výše uvedených schémat (*schéma hvězdy, souhvězdí a sněhové vločky*). Proto je v praxi využíváno kombinací těchto schémat.

3.8.2 OLAP

OLAP (*Online Analytical Processing*) podporuje uložení velkých objemů dat v databázi, která umožňuje uživatelům náhled na data z různých úhlů pohledu. Tato technologie je důležitou součástí BI, protože velké objemy dat jsou přehledně zpracovány a rychle

přístupny uživatelům z různých pohledů. OLAP databáze jsou uspořádány do tzv. *OLAP* kostek (viz *obrázek 6*). Výhodou takové kostky je to, že každá hrana (osa) kostky slouží jako jedna datová dimenze.



Obrázek 6: OLAP kostka
Zdroj: vlastní

Existují tři základní typy OLAP technologií.

- *MOLAP (Multidimensial OLAP)* je charakteristická uložením dat v multidimenzionálních - binárních OLAP kostkách. Informace jsou ukládány a doplňovány v určitém pravidelném intervalu.
- *ROLAP (Relational OLAP)* využívá technologie relačních databází. Data jsou vybírána z původních datových zdrojů pomocí SQL dotazů.
- *HOLAP (Hybrid OLAP)* je kombinací předchozích dvou typů, kde jsou detailní data uložena v relační databázi a agregované hodnoty jsou uloženy v binárních OLAP kostkách.

Základní operace používané v OLAP kostkách jsou:

- *Drill up* – navigace datovou hierarchií směrem nahoru (k méně obecnějším datům)
- *Drill down* – navigace datovou hierarchií směrem dolu (k detailnějším datům)
- *Slice and dice* – omezení dimenzí na některé hodnoty
- *Pivoting* – operace umožňující *„rotovat“* s kostkou a tím poskytnout náhled na data z různých dimenzí

3.9 Reportingové nástroje

Nástroje pro tvorbu reportů[6] patří mezi nejjednodušší nástroje BI. Jsou využívány k přehlednému zobrazení vybraných dat pro konkrétní potřeby uživatelů. V praxi se s reporty setkáváme velmi často. V podniku se může se jednat o evidenci zaměstnanců s docházkou či o informace o prodeji za určitě období. Z těchto reportů mohou uživatelé jednoduše vyčíst požadované informace a ty dále použít pro jejich potřeby, to ale pouze za předpokladu, že jsou data správně zpracována, upravena a jsou mezi nimi správně popsány vztahy. Pro tyto účely slouží transformační nástroje.

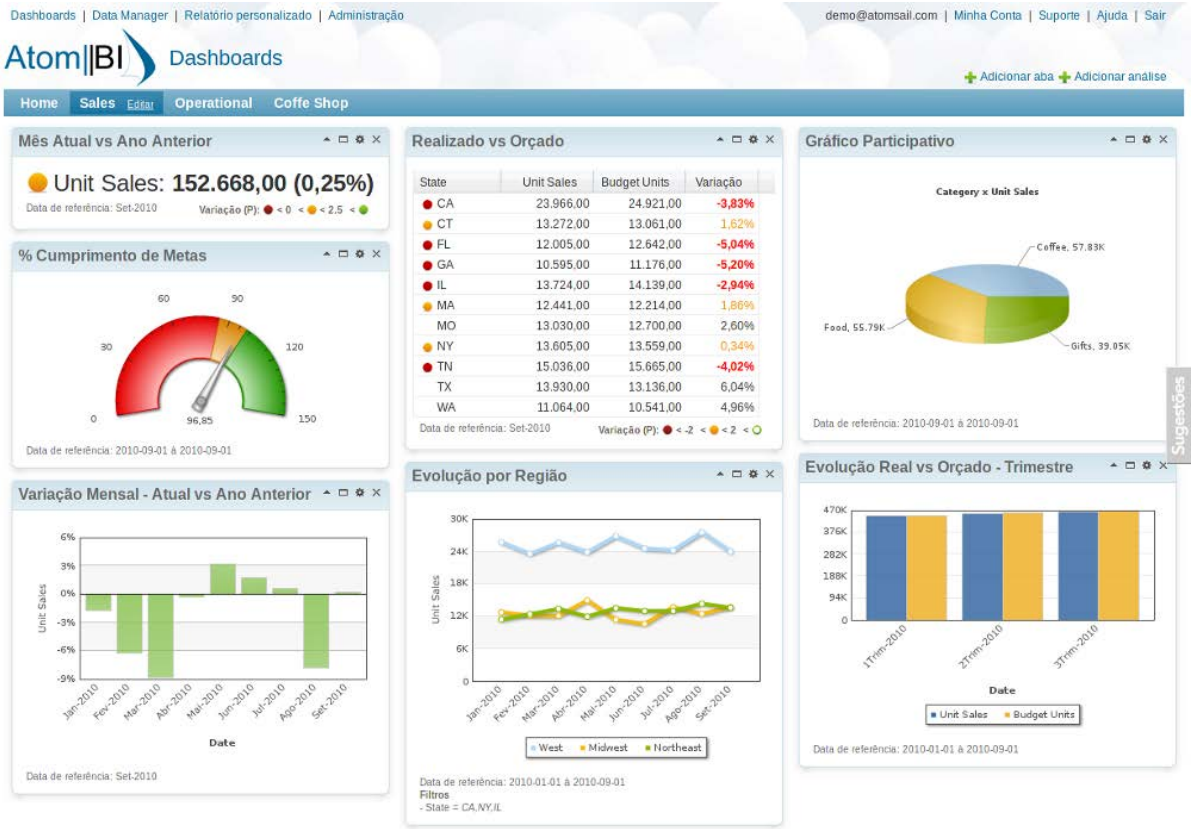
Pro získání požadovaných dat se využívá SQL (*Structured Query Language*) dotazů do databází datových skladů či multidimenzionálních databází. Získaná data jsou načtena do aplikace pro tvorbu reportů a v ní pak pomocí matematického aparátu (suma, průměr, medián aj.) či funkcí aplikace převedena do smysluplných informací a přehledné podoby.

3.9.1 Dashboard

Dashboard (*obrázek 7*) je speciálním typem reportu. Jeho smyslem je, stejně jako v případě reportu, umožnit náhled vybraných dat pro koncové uživatele. Dashboard ovšem umožňuje některou funkcionalitu, jež report neposkytuje. Stejně jako v dopravních prostředcích řídící (palubní) deska (dashboard) zobrazuje informace o stavu v reálném čase, v oblasti BI tomu

[6] Někdy se používá také pojem *tisková sestava*

není jinak. Dashboard je tvořen několika reporty, které uživateli přehledně umožňují pohled na celý systém. Současný stav je zobrazen přehledně na jediné stránce doplněné grafickými prvky. Zdrojem dat pro dashboard je databáze, která umožňuje aktuální (*real-time*) pohled na danou problematiku.



Obrázek 7: Dashboard
Zdroj: [29]

3.10 Analytické nástroje

Jedná se o typ aplikačního software, který využívají především manažeři pro sledování firemních procesů, plnění cílů organizace a získávání informací o aktivitách podniku. Tyto aplikace bývají jednoduše ovladatelné a mívají přehledné grafické rozhraní. Poskytují operace vhodné pro online analýzy, jako jsou např.: *drill up, drill down, slice and dice.*

3.11 Webové rozhraní

S rozvojem Internetu přicházely požadavky na dostupnost komponent BI přes webovou službu. Webové rozhraní umožňuje snadný a rychlý přístup k požadovaným komponentám BI, nejčastěji k reportům, dashboardům a analýzám, tedy koncové vrstvě procesu BI.

Výhodou takového rozhraní je dostupnost potřebných informací na jednom místě, nejčastěji serveru. Koncoví uživatelé či zákazníci mají možnost se rychle a jednoduše podívat na data, která je zajímají, a nemusí řešit jednotlivé komponenty či samotnou strukturu BI. Pro přístup k těmto datům bývají vyžadovány přihlašovací údaje, jako jsou *jméno* a *heslo*. Prostředí bývá jednoduché, účelné a graficky řešené tak, aby byla orientace pro koncové uživatele co nejpohodlnější.

Stejně jako v současnosti přibývá počet mobilních zařízení, přibývá i počet mobilních aplikací, které umožňují přístup právě na takovéto servery. Výhodou takovýchto aplikací je okamžitý přístup k požadovaným informacím (datům) vždy a všude (za předpokladu připojení k síti Internet).

4 Open source nástroje Business Intelligence

Tato kapitola je věnována open source nástrojům BI. Nejprve je nutné definovat, co výraz *open source* znamená a jaké typy licencí[7] vybrané nástroje mají.

Technicky vzato není open source typem licence. Software označovaný jako open source se vyznačuje tím, že má volně dostupný (otevřený) zdrojový kód. Každý má tedy možnost zdrojový kód používat či upravovat, ale pouze v případě dodržení jistých podmínek. Licence spadají pod organizaci *Open Source Initiative*, která je certifikuje.

Typy licencí

Mezi licence, které využívají vybrané nástroje a které spadají pod open source patří:

- *Apache Licence* – licence udává, že uživatel může produkt označený pod touto licencí používat bezplatně pro jakékoliv účely. Licence dále povoluje uživateli upravovat a dále distribuovat produkt bez nutnosti zveřejnit upravené zdrojové kódy. [30]
- *GNU General Public License (GNU GPL)* – jedná se o tzv. *copyleft* licenci. *Copyleft* v tomto případě znamená, že při úpravě originálního díla označeného touto licencí, musí být modifikované dílo označeno licencí jako dílo původní. Licence GNU GPL říká, že díla touto licencí označena mohou být volně užívána, modifikována i distribuována. [31]
- *Eclipse Public Licence (EPL)* – tato *copyleft* licence je používaná společností *The Eclipse Foundation*. Licence byla navržena ve smyslu *business-friendly*[8]. Proto umožňuje volně kopírovat, upravovat a šířit dílo. Zdrojové kódy pod licencí EPL mohou být využívány v programech, které nemají zdrojový kód otevřený. [32]

[7] Softwarová licence je právní nástroj, který umožňuje používat a redistribuovat software chráněný zákonem

[8] *Business-friendly* v tomto případě znamená, že je licence vhodná pro podnikání (volně přeloženo)

Výhody open source

Nespornou výhodou je, že software označovaný jako open source je přístupný zdarma, včetně zdrojového kódu. Komunita pak může opravovat chyby a má možnost se z těchto kódů učit a čerpat z nich. V případě, že aplikace nevyhovuje svou funkcionalitou, má uživatel možnost kód upravovat a přizpůsobit ho tak svým potřebám. Tyto kódy pak může při dodržení licenčních podmínek použít i u jiných projektů.

Nevýhody open source

Otevřenost a přístupnost kódu „všem“ je dvousečná zbraň. Nikde není zaručeno, že je kód napsaný správně a že je program bezpečné používat. To sice není zaručeno ani u komerčního software, ale za komerční produkty vždy zodpovídá společnost či osoba, takže se při potížích je na koho obrátit. Proto je často potřeba před řádným používáním neověřeného open source softwaru ověřit jeho funkčnost. Z tohoto pohledu je zde i nebezpečí napadení programu virem či uživatelem a následného šíření napadené verze, za kterou zpravidla nikdo neručí, což je další nevýhodou open source aplikací. V případě jakýchkoliv problémů není záruka toho, že daný problém někdo opraví nebo že se jím bude vůbec zabývat.

Komerční software vs. open source

Z výše uvedených důvodů vyplývá, že open source nemůže vyhovovat všem. V porovnání se softwarem komerčním není zaručena jeho funkčnost (u softwaru komerčního je vždy zodpovědná osoba a případné nedostatky se řeší většinou *updaty*). U komerčního software je v drtivé většině případů navíc poskytovaný i základní servis či služby. Nevýhodou, či spíše vlastností komerčního software je i to, že stojí peníze.

Proto je nutné se před pořízením jakéhokoliv (nejen BI) software rozhodnout, který typ je pro požadované účely nejvhodnější.

Výběr open source nástrojů

Volba open source nástrojů vychází ze zdrojů [3], [4], [5], [6], [18] a z vlastních zkušeností s touto problematikou. Z ETL nástrojů byly vybrány *Pentaho Data Integration*, *Talend Open Studio for Data Integration* a *CloverETL*. Z reportingových nástrojů to jsou *Pentaho Report Designer*, *BIRT* a *JasperReports*.

Nástroje jsou popsány v podkapitolách dle jednotlivých společností. Důraz v popisu jednotlivých komponent je kladen na odlišnosti či zajímavosti v nástrojích od jednotlivých dodavatelů.

4.1 Pentaho

Prvními popisovanými nástroji jsou open source nástroje od společnosti Pentaho.

4.1.1 O společnosti Pentaho

Firma Pentaho poskytuje komplexní nástroje BI, od nástrojů pro dolování dat, jejich úprav až po nástroje pro tvorbu reportů a analýz. Společnost Pentaho sídlící na Floridě v Orlandu (USA) byla založena v roce 2004 a dnes patří mezi nejvýznamnější společnosti zabývající se free open source BI. Mezi její klienty patří například *Mozilla*, společnost, která je mj. tvůrcem webového prohlížeče Mozilla Firefox, či *Brussels Airport*, společnost provozující mezinárodní letiště v Bruselu. [13]

Společnost Pentaho získala řadu ocenění. Např. *InfoWorld Bossie Award*, ocenění pro nejlepší open source software získala 5 let po sobě (2008 - 2012) či *CRN Big Data 100*, ocenění pro 50 společností, které se nejvýrazněji podílí v oboru zpracování dat. [13]

Pentaho nabízí dvě základní varianty svých produktů, a to *Enterprise* a *Community Edition*. *Enterprise Edition* vyžaduje roční poplatky a oproti *Community Edition* nabízí speciální uživatelskou podporu s přídavnou funkcionalitou. *Enterprise Edition* je volně dostupná všem zájemcům k vyzkoušení po dobu 30 dní.

Většina produktů od firmy Pentaho má možnost rozšíření ve formě *add-onů*, obvykle ve formě *plug-inů*, které poskytuje jak samotné Pentaho, tak komunita uživatelů a nadšenců. Projekt Pentaho má širokou uživatelskou základnu a podporu v podobě internetového fóra, bug trackeru a v neposlední řadě rozsáhlou dokumentaci ke všem dostupným nástrojům.

4.1.2 Pentaho Data Integration

Pentaho Data Integration (PDI, nebo také Kettle) je nástroj ETL s intuitivním, grafickým, *drag & drop* prostředím. Výhodou je tedy úspora času, jelikož uživatel nemusí psát aplikaci ručně, pracuje pouze s ovládacími prvky PDI. Podle samotného Pentaho se jedná o nejvíce využívaný open source ETL nástroj [33]. Přestože jsou ETL nástroje nejvíce využívané ve spojení s datovými sklady, PDI se využívá i pro jiné účely, jako je přesouvání dat mezi aplikacemi či databázemi nebo exportování databází do souborů.

PDI podporuje mnoho vstupních formátů, přes Microsoft (MS) Access, MS Excel, txt, xml, RSS (*Really Simple Syndication*) až po databáze Oracle, MySQL, MS SQL Server, data ve formě OLAP aj. Výstupem pak mohou být stejné formáty jako v případě vstupu.

4.1.3 Pentaho Reporting

Pomocí nástroje *Pentaho Report Designer* (PRD) může uživatel převést všechna svá data do smysluplných informací. Data jsou načítána do uživatelem vytvořeného reportu, který je tvořen pomocí grafických prvků (komponent) z nabídky. Práce s PRD je intuitivní a je možno vytvořit téměř jakoukoliv podobu reportu. Vygenerování reportu obstarává komponenta *Pentaho Report Engine*. Vytvořené reporty mohou být integrovány do *Pentaho BI Platform*.

Zdrojem dat může být libovolná databáze připojená přes JDBC (*Java Database Connectivity*) ovladač, ale také tabulka, OLAP či transformace z PDI. Výstupem pak může být soubor ve formátu HTML, PDF, Excel, rtf aj. Reporty je možno distribuovat pomocí e-mailu.

4.1.4 Pentaho Metadata Editor

Pentaho Metadata Editor (PME) je nástroj, který umožňuje spravovat relační datové modely. Pomocí PME je možné namapovat fyzickou strukturu dat do logického business modelu. Tyto vztahy jsou uloženy v centrálním úložišti metadat a umožňují správci vytvořit popisy komplexních tabulek, nastavit parametry uživatelských oprávnění pro přístup k datům, upravovat formáty dat a jiné.

4.1.5 Mondrian

Mondrian je OLAP server, který se zaměřuje na analýzu velkých a komplexních dat v reálném čase. Nástroj je napsaný v programovacím jazyce Java a Pentaho se chlubí tím, že systém umožňuje odpovídat na dotazy dostatečně rychle, aby mohl uživatel zkoumat data v reálném čase, přestože mohou data o velikosti několika gigabytů obsahovat miliony záznamů.

Prostředí *Mondrianu* je čistě grafické a umožňuje uživateli vytvářet, spravovat a analyzovat datové kostky či na ně nahlížet z různých dimenzí. Vstupem může být jakákoliv databáze připojená přes JDBD ovladač.

4.1.6 Pentaho Data Mining

Software, který má na starosti data mining, se jmenuje *Weka*. Původně se jednalo o univerzitní projekt vyvinutý v roce 1993 v *University of Waikato* na Novém Zélandu. V roce 2005 získala *Weka* ocenění *SIGKDD Data Mining and Knowledge Discovery Service Award*. Jedná se o nejvyšší ocenění v oblasti data miningu. V roce 2006 získalo Pentaho licenci na projekt *Weka* a od té doby je *Weka* součástí rodiny BI nástrojů od této společnosti. [13]

Weka podporuje standardní úkoly pro dolování dat, jako jsou předzpracování dat, transformace, regrese a vizualizace. Vstupem mohou být databáze podporující JDBC ovladač.

4.1.7 Pentaho Business Intelligence Server

Pentaho Business Intelligence Server (*BI Server*) je platforma zprostředkovávající přístup uživatelům přístup k reportům, dashboardům, OLAP analýzám a dalším *business* datům vytvořených v nástrojích Pentaho.

BI Server zajišťuje přístup k obsahu skrze server, na který je software nainstalován. Interakci zajišťuje framework, který poskytuje služby jako přihlašování, nahrávání dat, zobrazování obsahu apod. Komerční (*Enterprise*) variantou je *Pentaho Business Analytics*.

4.2 SpagoBI

V následujících odstavcích jsou popsány nástroje BI od italské společnosti *SpagoWorld*.

4.2.1 O společnosti SpagoWorld

SpagoWorld je italská společnost, založena roku 2001 jako open source iniciativa pod záštitou mezinárodní IT společnosti *Engineering Group*. *SpagoWorld* vede čtyři hlavní projekty:

- *Spagic* – platforma pro řízení *midleware*[9] a vývoj SOA[10] (*Service Oriented Architecture*) aplikací
- *Spago4Q* – platforma pro měření, analýzu a monitoring kvality softwaru, vývojových procesů a managementu softwarových služeb
- *Spago – Java Enterprise Wide Framework* pro vývoj webu v oblasti prostředí SOA
- *Spago BI* – free open source projekt pro služby v oblasti BI.

[9] midleware je typ software, který propojuje softwarové komponenty a aplikace v distribuovaném prostředí

[10] SOA je výpočetní architektura, při jejímž využití softwarové aplikace obsahují pouze logiku specifickou pro jejich vlastní úkol a pro obecnější činnosti využívají služeb dostupných na síti

Zajímavou vlastností produktů společnosti *SpagoWorld* je, že nemají komerční varianty. Jedinou variantou jsou produkty pod open source licencí. Vydávána je vždy pouze jedna stabilní verze aplikace, kterou může používat kdokoliv.

Ke každému projektu je za poplatek nabízena profesionální podpora, konzultace či tréninkový program. To jsou ale jediné služby, za které si firma nechává platit. Přidané funkcionality se tak platící uživatel u produktů *SpagoBI* nedočká.

4.2.2 Moduly SpagoBI

Projekt *SpagoBI* je rozdělen do pěti modulů. Každý modul je jakousi vrstvou, která má na starosti určitou část BI. Jedná se o moduly:

- *SpagoBI Server* – jádro *SpagoBI*, platforma poskytující veškerou základní funkcionalitu (analýza a vizualizace dat) pomocí webového rozhraní, je to jedna z mála komponent vyvíjených přímo společností *SpagoWorld*
- *SpagoBI Studio* – vývojové prostředí, založené na Eclipse[11]
- *SpagoBI Meta* – prostředí pro práci s metadaty
- *SpagoBI SDK* – vrstva umožňující *SpagoBI* interakci s externími nástroji
- *SpagoBI Applications* – kolekce analytických modelů vytvořených ve SpagoBI

SpagoBI je rozsáhlou kolekcí open source aplikací, které tvoří celek pro práci v oblasti BI. Využíváno je nástrojů třetích stran s kombinacemi vlastních nástrojů.

V následujících odstavcích jsou popsány nástroje, které jsou ve *SpagoBI* využívány. Nástroje jsou vázány na modul *SpagoBI Studio*, jehož jsou součástí.

[11] vývojové prostředí určené pro programování v jazyce Java

4.2.3 Reporting

SpagoBI umožňuje pracovat se strukturovanými reporty a exportovat je do formátů HTML, PDF, XLS, XML, TXT, CSV a RTF.

Jako samotné nástroje pro tvorbu reportů využívá *SpagoBI* enginů *JasperReport, BIRT, Accessible report a Business Objects*. Všechno to jsou nástroje třetích stran, jež jsou přístupny ze *SpagoBI Studio*, nejčastěji v podobě doplňků či plug-inů. [14]

4.2.4 Multidimenzionální analýza (OLAP)

Podobně jako v případě reportingu využívá *SpagoBI* dostupných open source nástrojů i pro nástroje OLAP. Pro OLAP analýzy využívá tří enginů, konkrétně *Jpivot/Mondrian, JPalo/Mondrian a JPXMLA*.

4.2.5 Grafy

SpagoBI nabízí specifický analytický engine, založený na knihovně *JFreeChart* pro tvorbu grafů, jako jsou histogram, koláčový graf, sloupcový graf, plošný graf apod. *SpagoBI* umožňuje také tvorbu interaktivních grafů, které mohou být využity v reportech.

4.2.6 Data mining a ETL

Pro data mining využívá (podobně jako Pentaho) nástroj *Weka* a pro transformaci dat využívá nástroj ETL *Talend* od společnosti *Talend Open Studio*.

4.3 TIBCO JasperSoft

Tato kapitola je věnována další ze společností, které poskytují open source nástroje BI – *TIBCO Jaspersoft*.

4.3.1 O společnosti TIBCO JasperSoft

V roce 2001 vznikl projekt jménem *JasperReports*. Ten, jak jeho jméno napovídá, byl zaměřen na reporting. První verze vyšla ve verzi *copyleft JasperReports License*, ale další verze vycházely již pod LGPL licencí. Hlavním produktem firmy je *JasperServer*, který podporuje pokročilé operace. Významným milníkem v historii této firmy byl rok 2014, kdy byla společnost *JasperSoft* odkoupena společností *TIBCO* za 185 milionů dolarů. [15]

Společnost získala několik ocenění. Z těch nedávných je to například ocenění *InfoWorld's 2013 Technology of the Year Award*, která *Jaspesoft* prezentuje jako nejlepší možné BI řešení pro rok 2013. Dalším oceněním je *The Best Overall Use of Technology* udělované na *European Software Testing awards*. Toto ocenění získal konkrétně projekt *JasperReports Server 5.0*. Mezi nejznámější uživatele produktů firmy *JasperSoft* patří např. společnosti *Puma*, *British Telecom* či firma *Ericsson*. [15]

V současnosti se firma zaměřuje zejména na reporting a na analýzu. Nabízí komplexní BI řešení v komerčním i nekomerčním provedení. Komerční edice obsahují na rozdíl od nekomerčních více nástrojů a je zde možnost využít profesionální zákaznickou podporu. *JasperSoft* nabízí služby i ve formě *cloud*, kdy ceny začínají na méně než jednom dolaru za hodinu. Zákazník tímto může využívat služeb serveru podporující služby reportingu a analýz a platí jen za čas, kdy služby skutečně využívá.

V následujících odstavcích jsou popsány jednotlivé komponenty *JasperSoft* verze *Community Edition*.

4.3.2 JasperReports Server

JasperReport Server je webová aplikace, která podporuje ukládání reportů vytvořených v aplikaci *JasperSoft Studio*. Uložené reporty je pak možno spustit, exportovat do požadovaného formátu nebo je nastavit na spuštění v určitý čas. Reporty mohou být ze serveru odeslány na různá média, jako jsou PC, mobilní zařízení, tiskárna či e-mail v různých datových formátech.

4.3.3 JasperReports Library

JasperReports Library je světově nejvíc populární open source engine pro reporting. [34] Je schopný využít data pocházející z libovolných zdrojů a vytvořit dokumenty, které mohou být zobrazeny, tištěny či exportovány do mnoha různých formátů, jako jsou HTML, PDF, Excel, OpenOffice, Word apod. Společnost *TIBCO JasperSoft* na svých webových stránkách [15] dále uvádí, že vytvořené reporty jsou tzv. *pixel-perfect*, což v praxi znamená, že do libovolného formátu exportovaný report bude na pixel odpovídat reportu, který byl vytvořen v návrhovém prostředí.

4.3.4 JasperSoft studio

JasperSoft Studio (dříve pod názvem *iReport*) představuje open source návrhové prostředí pro tvorbu reportů založené na bázi *Eclipse*. Stejně jako PRD (reportingový nástroj od firmy *Pentaho*) umožňuje *JasperSoft* studio tvorbu reportů za pomocí tabulek, grafů a subreportů (report vnořený do reportu). K datům je možno přistupovat skrze JDBC, XML, CSV a tabulkové modely. Výstupem jsou pak formáty PDF, RTF, XML, XLS, CSV, HTML, XHTML, txt, DOCX či OpenOffice.

4.3.5 JasperSoft ETL

Pod názvem *JasperSoft ETL* se skrývá nástroj pro tvorbu transformací dat. Umožňuje navrhovat a spravovat procesy ETL v grafickém prostředí. Zajímavostí je, že *JasperSoft* nevyužívá vlastního enginu, ale používá již stávající platformu pro operace ETL, kterou je *Talend Open Studio* od firmy *Talend*.

4.4 Ostatní projekty

Tato kapitola je věnována menším či méně známým projektům, které jsou ale hojně v oblasti BI využívány nebo jsou něčím zajímavé.

4.4.1 BIRT

Projekt BIRT byl poprvé představen v roce 2004 společností *Actuate Corporation*, která se toho roku připojila k *Eclipse Foundation*. Projekt byl zanedlouho schválen a stal se hlavním projektem komunity *Eclipse*. Kromě *Actuate* je projekt sponzorován i firmou IBM a společností *Innovent Solutions*. Projekt je mohutně podporován komunitou *Eclipse* i samotnými tvůrci. Je licencován pod EPL. [16]

BIRT se může pochlubit počtem stažení, který činí 12,5 milionu. Počet aktivních vývojářů (jak z řad tvůrců, tak z řad komunity) činí úctyhodných 2,5 milionu. BIRT využívá v současnosti celá řada firem. Kromě zmíněné firmy IBM jsou to také *Cisco*, *S1* a *ABS Nautical Systems*. [16]

Projekt se skládá ze dvou hlavních komponent. První z nich je vizuální návrhové prostředí pro tvorbu reportů a druhým je pak komponenta generující vytvořené návrhy. Tyto návrhy jsou generovány do formátu XML a mohou být dále využity v jakémkoliv Java prostředí.

4.4.2 CloverETL

Mezi známějšími zástupci nástrojů ETL se najde i jeden český, konkrétně *CloverETL* od společnosti *Javlin*. Celá platforma je tvořena dvěma hlavními částmi. Jedná se o *CloverETL Designer* a *CloverETL Server*. Jak názvy vypovídají, v případě *Designeru* se jedná o vizuální nástroj umožňující vytvářet, spravovat a spouštět transformace, v případě Serveru se jedná o tu část, která je podnikovým prostředím pro správu uživatelů, umožňuje nastavení automatizace a běh vytvořených transformací.

CloverETL vychází v různých edicích. Od *Community* edice, která je zcela zdarma, přes *Enterprise* edici, určenou pro podnikové nasazení, až po *Cluster* edici umožňující běh na clusteru serverů či v *cloudu*.

5 Porovnání vybraných nástrojů dle zvolených kritérií

Tato kapitola je věnována porovnání vybraných nástrojů podle zvolených kritérií. Kritéria, podle kterých jsou nástroje porovnávány, jsou vypsána v *tabulce 1*. Jelikož mají jednotlivá kritéria různý vliv na hodnocení, jsou jim přiřazeny váhy podle jejich důležitosti. Vybraným nástrojům je pak v každé kategorii přiřazeno jeden až pět bodů dle následujícího hodnocení:

Jeden bod – Zkoumaný software v dané kategorii nesplňuje základní zkoumané vlastnosti a toto hodnocení má výrazný vliv na doporučení, respektive nedoporučení pro používání daného software.

Dva body – Při porovnání byly v dané kategorii zjištěny závažné nedostatky či nevyhovující skutečnosti, které se negativně promítají při práci se samotným software.

Tři body – V dané kategorii trpí software výraznějšími nedostatky, které ovšem nebrání v jeho používání ani nezamezují jeho správné funkcionalitě. Nedostatky se projevují zejména v neefektivnosti práce se software.

Čtyři body – Až na drobné nedostatky splňuje software v dané kategorii kladené požadavky. Tyto nedostatky se výrazněji nepromítají do práce se softwarem a běžný uživatel si jich často ani nemusí všimnout.

Pět bodů – Nejvyšší možné bodové ohodnocení pro zkoumané kategorie může obdržet software, který dokonale splňuje zkoumané parametry a vyhovuje všem kladeným požadavkům.

Celkové skóre je pak vypočteno jako součet všech obdržených bodů vynásobených jejich vahami.

Tabulka 1: Kritéria hodnocení vybraných nástrojů

Kritérium	Popis	Váha (součet 100 bodů)
Přehlednost prostředí a ovládání	Grafické zpracování aplikace je důležitým prvkem při hodnocení. Přehledné a intuitivní prostředí výrazně usnadňuje a urychluje práci s aplikací.	20
Funkcionalita	Základní a přídavná funkcionalita, specifika nastavení a možnost instalace plug-inů jsou kritéria, která rozhodují o možnostech využití aplikace.	30
Technická a uživatelská podpora	Mezi podporu patří aktualizace aplikace, technická dokumentace, bug tracker, diskusní fóra, ale i ostatní možnosti komunikace mezi zákazníkem a uživatelskou podporou.	20
Instalace a kompatibilita na různých OS	Základním požadavkem dnešních aplikací je bezproblémová funkčnost na všech využívaných operačních systémech platformy PC.	15
Kompatibilita mezi jednotlivými aplikacemi	Zkoumána je kompatibilita mezi jednotlivými aplikacemi daného výrobce, ale také možnost integrovat kód do jiných aplikací.	15

Zdroj: vlastní

5.1 ETL nástroje

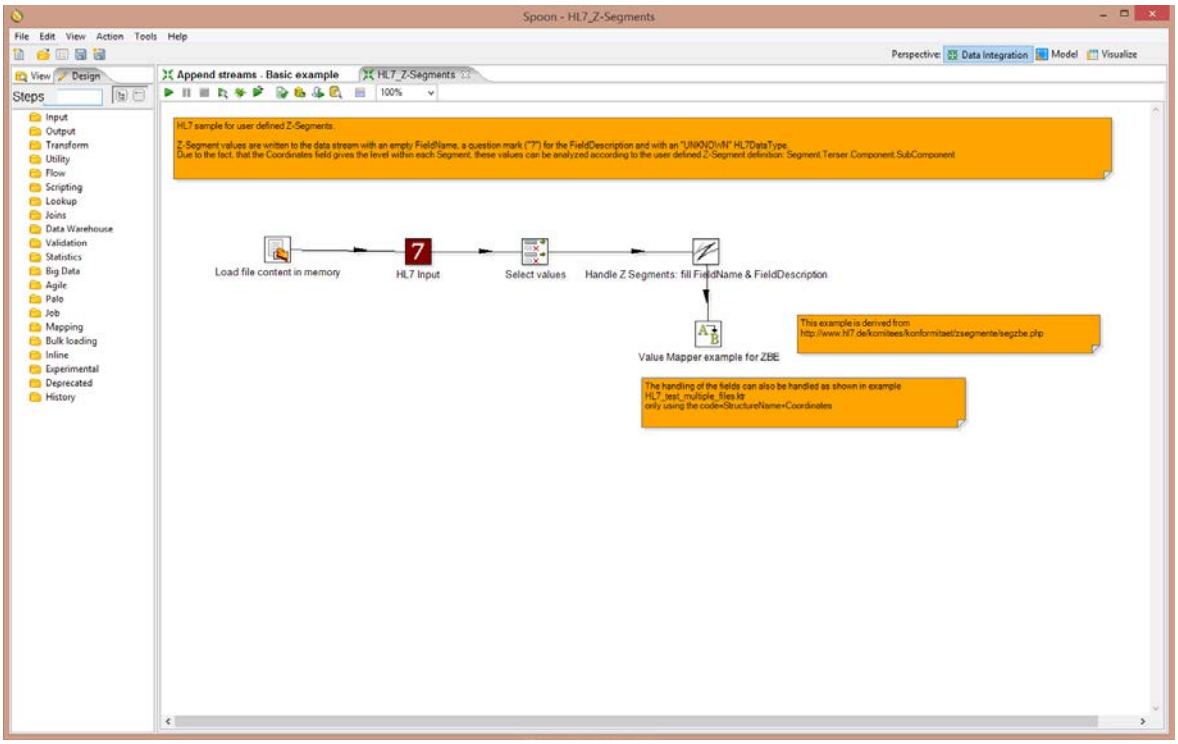
Porovnanými nástroji ETL jsou *Pentaho Data Integration* (Kettle) od společnosti Pentaho, *Talend Open Studio* od společnosti Talend a *CloverETL* od společnosti Javlin.

5.1.1 Pentaho Data Integration (Kettle)

PDI je možno stáhnout ve třech verzích na různé operační systémy, kterými jsou Windows, Linux a MacOSX. U prvních dvou verzí je možno stáhnout jak 32bitovou verzi, tak 64bitovou verzi. Pro MacOSX je dostupná pouze 64bitová verze. Pentaho doporučuje užívat pouze 64bitovou verzi operačních systémů pro optimální využití jejich produktů. PDI je vydáváno pod licencí Apache Licence 2.0. Instalace PDI probíhá pouhým rozbalením zkomprimovaných souborů, hned poté je možno software plně používat. Při

rozbalení se tedy nevytvoří zástupci spouštění na ploše či v nabídkách OS a software se musí spouštět přímo ze složky.

Prostředí PDI (*obrázek 8*) je uživatelsky přívětivé. V levé části jsou v nabídce komponenty pro tvorbu transformací seřazené dle kategorií jako je *Input*, *Output*, *Statistics*, *Transform*, *Flow*, *Lookup*, *Join* apod. Z nich uživatel sestavuje v hlavním okně aplikace transformace. Přestože je prostředí příjemné a intuitivní, má i své drobné nedostatky. Jedním z nich je nemožnost přesunout okno s nabídkou či okno s transformací do jiného prostoru v aplikaci. Další drobnou vadou jsou malé a někdy obtížně rozlišitelné ikony komponent.



Obrázek 8: Prostředí Pentaho Data Integration
Zdroj: vlastní

PDI zajišťuje veškerou potřebnou funkcionalitu při tvorbě transformací zejména díky komponentám, kterých je více jak dvě stě. PDI podporuje všechny nejčastěji používané formáty vstupů, od textových souborů přes tabulkové kalkulátory až po databáze. Zajímavostí je možnost využití některých komponent při zpracování semistrukturovaných dat, která bývají jinak velmi obtížně zpracovatelná. Při běhu transformací má uživatel možnost vidět počet přenesených dat za sekundu a rychlost průchodu dat v jednotlivých

komponentách. Tím může odhalit slabé místo v transformaci a náhradou za jiný blok či změnou nastavení zrychlit celý proces transformace.

Podpora a dokumentace projektu jsou na velmi vysoké úrovni. V samotném PDI je přes padesát ukázkových transformací různých typů, které jsou zdokumentovány a na kterých se může uživatel naučit pracovat s aplikací. Projekt má vlastní internetové fórum, kde bývají dotazy a problémy zodpovězeny v rámci několika hodin či dní. O vynikající úrovni diskusního fóra svědčí i počet založených témat, který se vyšplhal přes 19 000. Projekt má vlastní webové stránky s dokumentací na bázi *wiki*[12]. Na stránkách je celá řada návodů (instalace, spuštění), tipů a triků (zrychlení běhu transformací apod.) jak s nástrojem PDI pracovat. Kromě internetové podpory nabízí Pentaho ke svým produktům knižní publikace.

Produkty Pentaho mají mezi sebou výbornou kompatibilitu. Projekty aplikací jsou mezi sebou provázány a mohou být spouštěny a používány v jiných aplikacích. Například transformace vytvořené v PDI mohou sloužit jako vstup pro reporty v PRD. V *tabulce 2* jsou popsány klady a zápory PDI.

Tabulka 2: Hodnocení nástroje Pentaho Data Integration

Pentaho Data Integration (verze 5.3.0)	
Klady	**Zápory**
Jednoduchá instalace	Některá nastavení (jako využitelnost paměti RAM) nelze nastavit přímo v aplikaci
Přehledné a intuitivní prostředí	Nemožnost manipulace s okny uvnitř aplikace
Silná komunita	Neaktuální nápověda k některým komponentám v aplikaci
Funkcionalita	
Rozsáhlá dokumentace a výborná podpora	
Kompatibilita s ostatními nástroji od Pentaho	

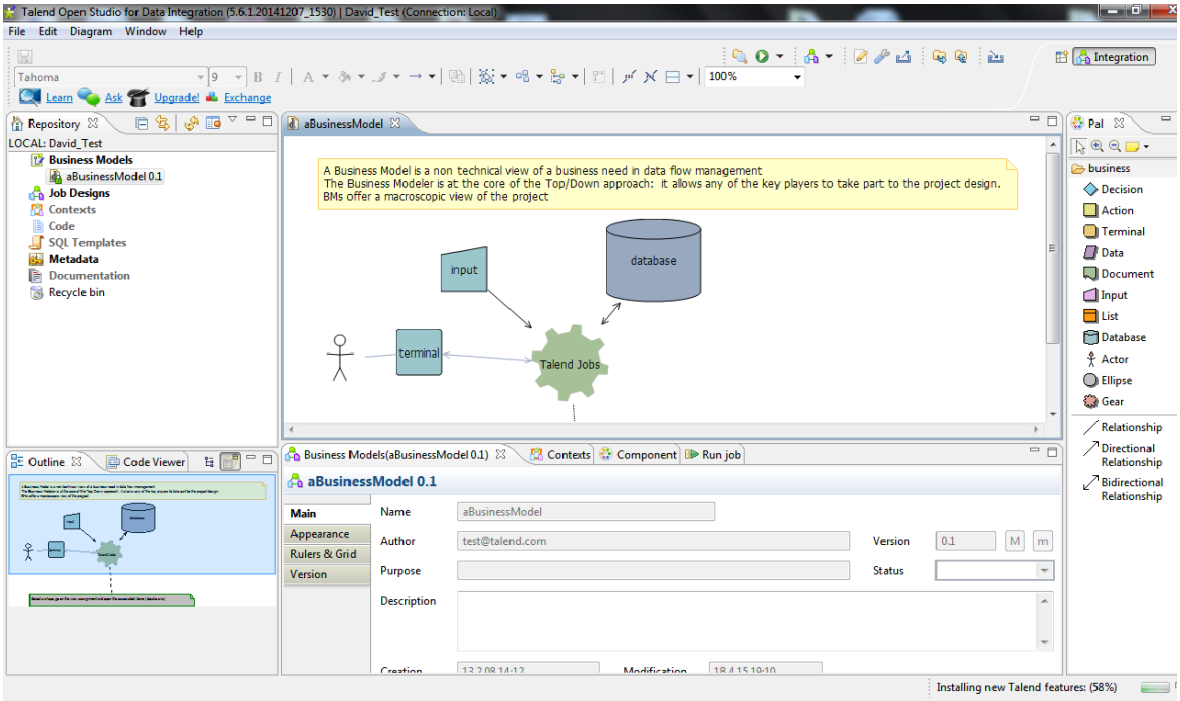
Zdroj: vlastní

[12] Wiki je označení webů, které umožňují uživatelům přidávat a měnit obsah

5.1.2 Talend Open Studio for Data Integration

Talend Open Studio se skládá z několika komponent. Pro operace ETL je to komponenta Data Integration, která je dostupná ve verzích pro Windows, Linux a MacOSX pod licencí Apache Licence 2.0. Instalace probíhá klasickým způsobem, kdy si uživatel stáhne soubor s příponou exe (v případě OS Windows), který spustí a v instalátoru vybere místo na disku, kam se má aplikace nainstalovat. Talend Open Studio plně nepodporuje nejnovější Java verzi 8, je tedy potřeba nainstalovat zpětně starší verzi Java 7, jinak se při spuštění objeví chybové hlášky.

Prostředí aplikace (viz *obrázek 9*) je oproti konkurentovi od firmy Pentaho méně přehledné. Příčinou je množství oken a nastavení, která jsou ihned dostupná při zapnutí aplikace. Okno pro samotnou transformaci zabírá v původním nastavení méně než polovinu plochy v rámci aplikace, což je pro práci poněkud málo. Práce v prostředí je tedy zpočátku pomalejší než u konkurentů a uživateli bude pravděpodobně chvíli trvat, než si na prostředí aplikace zvykne.



Obrázek 9: Prostředí Talend Open Studio
Zdroj: vlastní

Co ztrácí *Talend Open Studio for Data Integration* v přehlednosti a intuitivnosti, to nahrazuje ve funkcionalitě. Funkcionalita je nejsilnější stránkou aplikace. Na výběr je zde přes 800 konektorů a komponent ve více jak 25 kategoriích.

Podobně jako konkurence umožňuje *Talend Open Studio* ladit vytvořené transformace pomocí nástrojů zobrazujících čas průchodnosti dat jednotlivými komponenty či počet zpracovaných řádků za vteřinu. Tímto způsobem může uživatel optimalizovat vytvořené transformace.

Stejně jako PDI má i *Talend Open Studio* rozsáhlou uživatelskou komunitu. K dispozici mají uživatelé internetové fórum, kde mohou řešit problémy a dotazy. Na webových stránkách projektu je možnost stáhnout rozsáhlou dokumentaci popisující všechny komponenty tohoto nástroje.

Tabulka 3 popisuje klady a zápory nástroje *Talend Open Studio*.

Tabulka 3: Hodnocení nástroje Talend Open Studio for Data Integration

Talend Open Studio for Data Integraion (verze 5.6.1)	
Klady	**Zápory**
Jednoduchá instalace	Nekompatibilita s nejnovější Java verzí 8
Rozsáhlá funkcionalita	Nepřehledné prostředí
Silná komunita	
Kvalitní dokumentace a podpora	
Možnost vlastního uspořádání oken v aplikaci	

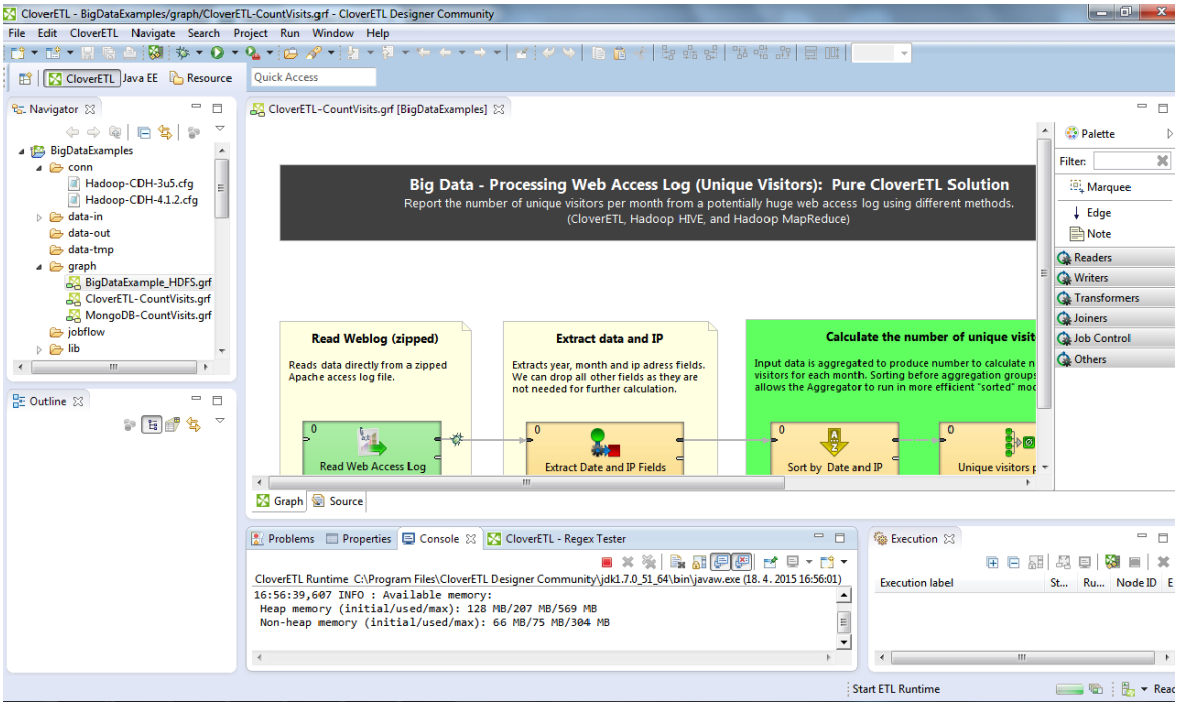
Zdroj: vlastní

5.1.3 CloverETL

Český projekt *CloverETL* (licence GNU GPL*)* je podporován OS Windows, Linux, MacOSX a dokonce je možno ho nainstalovat v podobě *plug-inu* do vývojového prostředí *Eclipse*. Instalace probíhá klasicky pomocí instalátoru, kde si uživatel vybere místo

na disku, kam chce aplikaci nainstalovat. Oproti konkurentům umožňuje při instalaci vytvořit zástupce pro rychlé spouštění.

Jedno z hesel *CloverETL* zní „*Read a thick manual before starting? No, thanks.*“ [35] Pro uživatele, které *nebaví číst tlusté návody před tím, než začnou pracovat* je to jistě povzbuzující motto. To se odráží v celkovém konceptu aplikace. Po spuštění aplikace přivítá uživatele příjemné vizuální prostředí (*obrázek 10*). I bez předchozích znalostí se lze v prostředí poměrně rychle zorientovat a začít v něm rovnou pracovat na transformacích. Ikony jednotlivých komponent pro transformaci jsou graficky velmi povedené a lze se tak v aplikaci snadno orientovat.



Obrázek 10: Prostředí CloverETL
Zdroj: vlastní

Největším problémem *CloverETL community edition* je jeho omezená funkcionalita. Zatímco běžné transformace zvládne *CloverETL* na jedničku, pro komplexnější transformace, kde jsou vyžadovány pokročilejší funkce, je komunitní edice nevhodná. Důvodem je počet komponent v placené a neplacené verzi. V neplacené variantě (*community edition*) je k dispozici pouhých 24 komponent. V placené verzi je jich 123. [35]

Technická dokumentace je na velmi dobré úrovni. Horší je to s diskusním fórem, kde je počet uživatelů i příspěvků daleko nižší než u konkurence, a tak je zde i větší počet nezodpovězených či nevyřešených dotazů.

Přednosti a nedostatky *CloverETL* jsou uvedeny v *tabulce 4*.

Tabulka 4: Hodnocení nástroje CloverETL

CloverETL (verze 4.0.3)	
Klady	**Zápory**
Jednoduchá instalace	Velmi omezená funkcionalita neplacené verze
Přehledné a intuitivní prostředí, ve kterém se lze rychle zorientovat	Nepříliš aktivní diskusní fórum
Výborně graficky zpracované ikony komponent	
Dobrá dokumentace	
Svižné prostředí	

Zdroj: vlastní

5.2 Reportingové nástroje

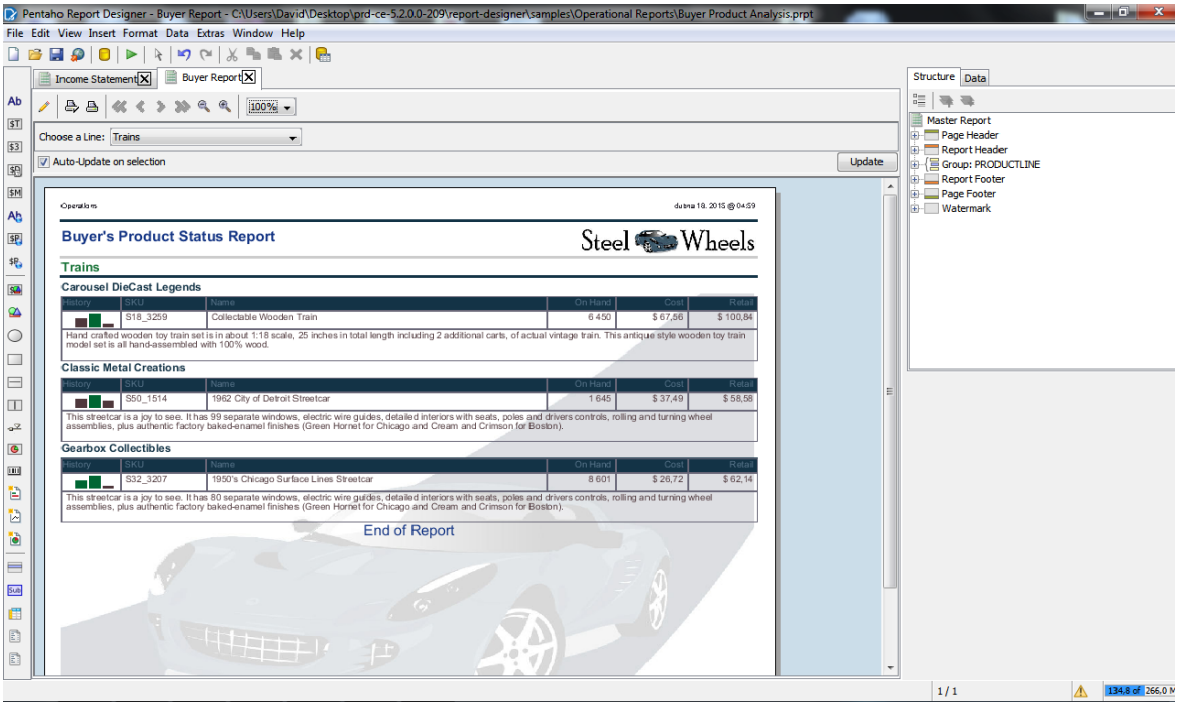
Mezi porovnávané reportingové nástroje patří: *Pentato Report Designer* od společnosti Pentaho, *BIRT* od společnosti Actuate Corporation a *JasperReports* od společnosti TIBCO JasperSoft.

5.2.1 Pentaho Report Designer

PRD (pod licencí GNU GPL) je možno stáhnout ve variantách pro různé operační systémy. Těmi jsou Windows, Linux a MacOSX. Místo instalace stačí stažený soubor pouze rozbalit.

Návrhové prostředí je velice intuitivní (viz *obrázek 11*). Prostřední okno aplikace slouží jako plocha pro návrh reportu. Na ni umisťuje uživatel komponenty, které jsou k dispozici

v levém menu (okně), a tím sestavuje samotný report. Ten je logicky členěn do několika částí, jako jsou hlavička, tělo a zápatí. V pravém okně jsou hierarchicky (dle části reportu, ve kterém se vyskytují) zobrazeny komponenty, které byly při návrhu využity. Dále je zde možnost nastavení vstupních dat, což mohou být databáze, tabulky či transformace vytvořené v PDI.



Obrázek 11: Prostředí Pentaho Report Designer
Zdroj: vlastní

PRD poskytuje veškerou funkcionalitu k vytvoření reportů. Kromě základních funkcí, jako jsou grafy, tabulky či subreporty (report obsahující vnořený report), jsou zde rozšířené možnosti jak obohatit report, jako například funkce *crosstab* (normalizace či denormalizace tabulky). Další zajímavou funkcí je možnost libovolně naformátovat jakýkoliv text a číslo, či napsat vlastní funkci pomocí vestavěného editoru. Výstupem může být pak soubor ve formátu PDF, Excel, HTML a další.

Jedinou „*vadou na kráse*“ (v tomto případě doslova) je pak nepřesnost vykreslování vytvořených reportů do některých formátů. Několikrát se při testování aplikace stalo, že se návrh reportu vytvořeného v PRD neshodoval s vyexportovaným formátem v několika pixelech. Na první pohled nemusí být nepřesnosti zřejmé, ale při přiblížení byl v několika

situacích znatelný posun několika vykreslených objektů o celé pixely. Je proto nutné report pečlivě otestovat pro všechny formáty, jež bude chtít uživatel později využívat. Nestačí se tedy spoléhat pouze na náhled reportu v PRD.

PRD má stejně jako všechny produkty Pentaho výbornou a detailní dokumentaci popisující jednotlivé komponenty a poskytující návody pro vytvoření základních i pokročilých reportů i s ukázkami.

Výhody a nevýhody aplikace PRD jsou popsány v následující tabulce (*tabulka 5*).

Tabulka 5: Hodnocení nástroje Pentaho Report Designer

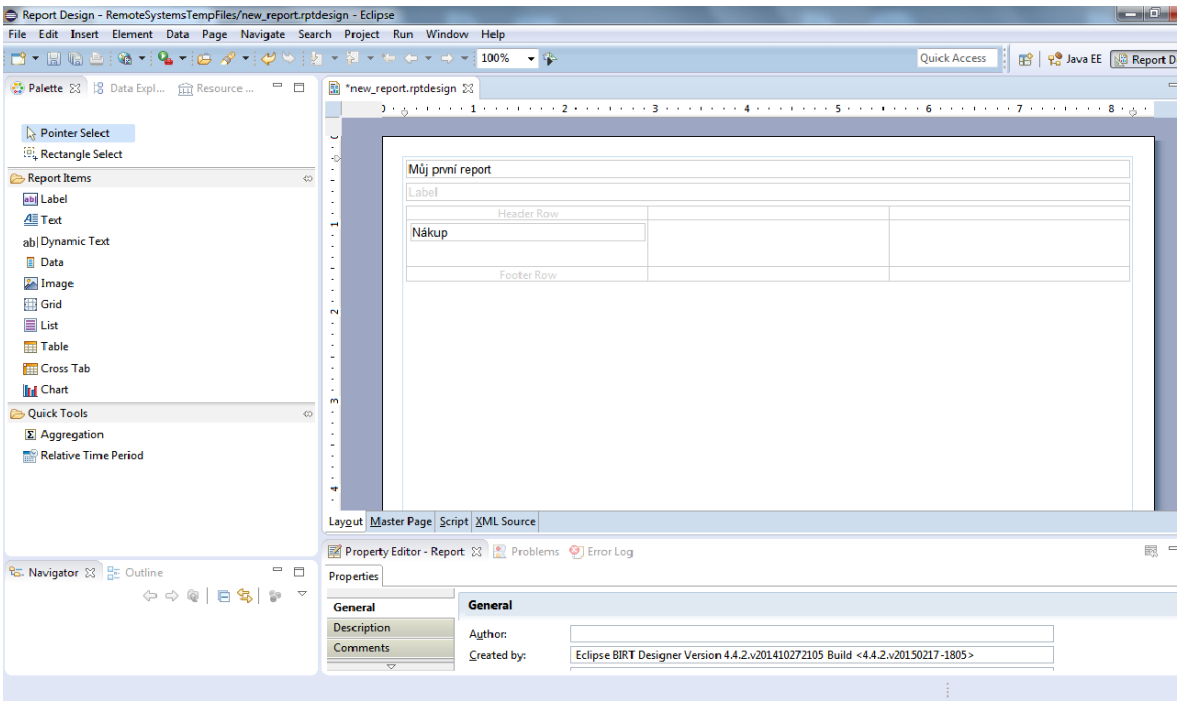
Pentaho Report Designer (verze 5.3.0)	
Klady	**Zápory**
Jednoduchá instalace	Nepřesnosti ve vykreslování u některých reportů v různých datových formátech
Přehledné a intuitivní prostředí	
Silná komunita	
Rychlost prostředí	
Možnost využít transformace z PDI jako zdroj dat	
Velká funkcionalita	

Zdroj: vlastní

5.2.2 BIRT

BIRT Designer (licence EPL) je možno stáhnout ve dvou základních variantách, které se odvíjí od toho, má-li uživatel nainstalované vývojové prostředí *Eclipse*. První variantou je *all-in-one* balíček, který obsahuje všechny potřebné komponenty, včetně knihoven vývojového prostředí *Eclipse*. Druhou variantou je stažení *frameworku*, tedy doplňku do již nainstalovaného prostředí *Eclipse*. BIRT je dostupný standardně na tři základní operační systémy, tedy Windows, MacOSX a Linux. Instalace probíhá pouhým rozbalením stažených souborů.

Prostředí (viz *obrázek 12*) a funkcionalita jsou největšími plusy tohoto projektu. Podobně jako v ostatních projektech je aplikace tvořena oknem s paletou nástrojů, oknem s vlastnostmi vybraného objektu a hlavním oknem, ve kterém probíhá návrh reportu. Struktura i umístění panelů jsou velmi intuitivní a přehledné, což vede ke snadnému používání aplikace. BIRT se může, stejně jako jeho konkurent od Pentaho, pochlubit výbornou prací s grafy. Na výběr je více jak deset typů grafů, u kterých je možno libovolně měnit datové zdroje, osy, písmo, průhlednost aj. Pomocníkem při tvorbě grafů pak může být i implementovaný průvodce.



Obrázek 12: Prostředí BIRT
Zdroj: vlastní

Slabinou tohoto projektu je jeho čistě reportingové zaměření. Součástí projektu nejsou žádné komponenty pro využití již vytvořených reportů v praxi. Pro doručení reportů koncovým uživatelům je třeba využít nástroje třetích stran, což může při rozhodování o výběru reportingového nástroje hrát velkou roli.

Dokumentace a podpora projektu je na skvělé úrovni. Na webových stránkách projektu BIRT lze najít výuková videa, ukázkové reporty či dokumentaci popisující součásti

a komponenty aplikace *BIRT Designer*. Slovní hodnocení nástroje BIRT je uvedeno v *tabulce 6*.

Tabulka 6: Hodnocení nástroje BIRT Designer

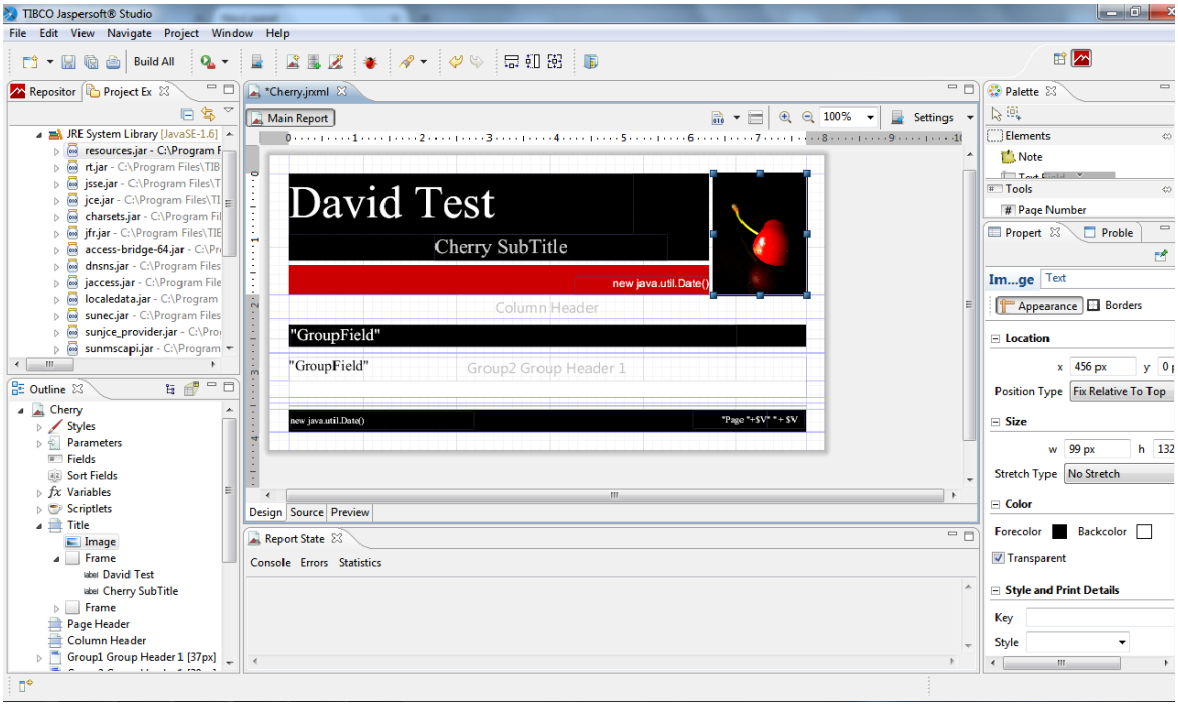
BIRT Designer (verze 4.4.2)	
Klady	**Zápory**
Výborná funkcionalita	Zaměření pouze na reporting
Přehledné a intuitivní prostředí	
Silná komunita	
Dokumentace a podpora	

Zdroj: vlastní

5.2.3 JasperReports

JasperReport je dostupný na operační systémy Windows, Linux a MacOSX pod licencí GNU GPL. Instalace probíhá klasicky pomocí dialogových oken s licenčními podmínkami a výběrem místa pro instalaci.

Prostředí (*obrázek 13*) působí velmi příjemně a lze se v něm rychle zorientovat. V aplikaci se nachází, obdobně jako v případě předchozích testovaných aplikací, okno s paletou komponent, okno s vlastnostmi objektu a plocha pro tvorbu reportů. Jednotlivé komponenty jsou umisťovány na plochu pomocí funkce *drag & drop*. Komponenty lze umístit velice přesně na požadované místo díky funkci, která ukazuje aktuální šířku a výšku vybraného objektu. Oproti konkurentům nabízí *JasperReports* méně funkcionality. Nabídka grafů sice obsahuje více typů než konkurence, ale možnosti jejich nastavení jsou omezeny jen na základní vlastnosti. Často se tak musí uživatel spokojit s tím, co mu aplikace umožňuje, přestože jeho představa a požadavky mohou být odlišné.



Obrázek 13: Prostředí JasperReports
Zdroj: vlastní

S aplikací je to poněkud horší co se týče ovládání. Nabídky, zejména vlastnosti objektů, působí komplikovaně. K základním vlastnostem, jako je velikost písma, se je třeba složitě *proklikávat*. Při práci v *JasperReports* se několikrát stalo, že program několik vteřin nereagoval při kliknutí na některou z nabídek. Výhodou aplikace je její kompatibilita s ostatními komponenty od firmy *TIBCO JasperSoft*. Vytvořené reporty je možno zpřístupnit pro uživatele pomocí komponenty *JasperReports Server*.

Jednotlivé části a komponenty *JasperReports* jsou popsány v detailní dokumentaci na webových stránkách *JasperReports*. *JasperSoft* má ke většině svých aplikací, tedy i k *JasperReports*, dostupná diskusní fóra. Není zde ale velký počet aktivních uživatelů, a tak zůstává mnoho dotazů a připomínek nezodpovězených.

Klady a zápory jsou uvedeny v *tabulce 7*.

Tabulka 7: Hodnocení nástroje JasperReports

JasperReports (verze 6.0.3)	
Klady	**Zápory**
Příjemné prostředí	Oproti konkurenci méně funkcionality
Kompatibilita s dalšími nástroji od TIBCO Jaspersoft	Komplikované nabídky s nastaveními
	Slabší komunita

Zdroj: vlastní

5.3 Porovnání a zhodnocení vybraných nástrojů

V této podkapitole jsou vybrané nástroje bodově ohodnoceny dle zvolených kritérií. V *tabulce 8* jsou ohodnoceny nástroje ETL a v *tabulce 9* jsou přiřazeny body vybraným reportingovým nástrojům.

Vyhodnocení vybraných nástrojů ETL

Tabulka 8: Porovnání hodnocení vybraných ETL nástrojů

	Pentaho Data Integration	Talend Open Studio	CloverETL
Přehlednost prostředí a ovládání (20)	4	3	5
Funkcionalita (30)	5	5	2
Technická a uživatelská podpora (20)	5	5	4
Instalace a kompatibilita na různých OS (15)	4	3	5
Kompatibilita mezi jednotlivými aplikacemi (15)	5	4	1
Celkem bodů	465	415	330
Celkem procent	**93 %**	**83 %**	**66 %**

Zdroj: vlastní

Ze zkoumaných nástrojů ETL se nejlépe umístil *Pentaho Data Integration* (Kettle) s hodnocením 93 %. Silná a aktivní komunita kolem projektu Pentaho je jedním z důvodů, proč se PDI umístilo na prvním místě. Dalším důvodem je výborná kompatibilita mezi jednotlivými aplikacemi od společnosti Pentaho. PDI nemá větších nedostatků, a tak se zaslouženě stává volbou číslo jedna v oblasti open source nástrojů ETL.

Talend Open Studio, který získal 83 %, je rovněž vynikající volbou pro uživatele, kteří hledají nástroj pro transformace dat. Oproti PDI má *Talend* méně intuitivní a uživatelsky přívětivé prostředí. To může být problém pro méně zkušené uživatele, kteří s aplikacemi tohoto typu pracují prvně.

Posledním zkoumaným a zároveň umístěným nástrojem je český *CloverETL*. Přestože se může zdát, že je aplikace schopna konkurovat větším a známějším firmám v oblasti open source BI nástrojů, není tomu tak. Hlavním důvodem je omezení testované *community* verze, které se projevuje v nízkém počtu použitelných komponent pro tvorbu transformací. Přesto si *CloverETL* zaslouží pozornost díky svému zpracování, které je moderní a velmi vstřícné vůči uživatelům. Aplikace dosáhla hodnocení 66 %.

Pro praktické řešení byl vybrán nástroj PDI, který dopadl v hodnocení nejlépe. Nástroj nemá výraznější slabiny a vývojáři společně s velkou uživatelskou základnou vytvořili téměř dokonalý produkt, který konkuruje i projektům komerčním.

Vyhodnocení vybraných reportingových nástrojů

Tabulka 9: Porovnání hodnocení vybraných reportingových nástrojů

	Pentaho Report Designer	BIRT	JasperReports
Přehlednost prostředí a ovládání (20)	5	5	3
Funkcionalita (30)	4	5	3
Technická a uživatelská podpora (20)	5	5	3
Instalace a kompatibilita na různých OS (15)	4	4	5
Kompatibilita mezi jednotlivými aplikacemi (15)	5	1	4
Celkem bodů	445	425	345
Celkem procent	**89 %**	**85 %**	**69 %**

Zdroj: vlastní

Na prvním místě se umístil nástroj *Pentaho Report Designer* od společnosti Pentaho. Důvodem umístění je nepřítomnost výraznějších nedostatků či limitací. PRD získal ve všech kategoriích první či druhé nejvyšší možné bodové ohodnocení, což v součtu znamenalo 89 %. Nástroj disponuje intuitivním prostředím a poskytuje veškerou funkcionalitu pro tvorbu reportů. Navíc je kompatibilní s ostatními nástroji od Pentaho. Jako zdroj dat lze např. využít transformaci vytvořenou v PDI, což se při využívání obou produktů od firmy Pentaho jeví jako velká výhoda oproti konkurenci.

Těsně druhý, s ohodnocením 85 %, skončil *BIRT Designer*. Kromě přehledného uživatelského prostředí zaujme BIRT zejména svou funkcionalitou, kde předčí i konkurenci od Pentaho. Jedinou, zato výraznou slabinou je úzké zaměření projektu. Zatímco konkurence nabízí podpůrné projekty zaměřené na serverové aplikace pro správu a distribuci vytvořených reportů, BIRT se zaměřuje pouze na reporting. To může být rozhodujícím faktorem při výběru vhodného reportingového nástroje, jelikož pro další využití vytvořených reportů je při používání aplikace BIRT nutno nalézt kompatibilní aplikaci pro distribuci reportů, zatímco konkurenti mají tyto aplikace ve svém portfoliu.

S 69 % se umístil na třetím místě *TIBCO Jaspersoft* s nástrojem *JasperReports*. Oproti konkurenci je ve všech důležitých ohledech spíše průměrným produktem. To ovšem neznamená, že by byl nástroj nevyhovující. Běžnému uživateli bude *JasperReports* pro tvorbu reportů naprosto dostačovat. Výhodou je pak možnost využívat vytvořené reporty v nástroji *JasperReports Server*.

Rozhodnutí o výběru reportingové nástroji pro praktickou část práce padlo na aplikaci PRD. Přestože BIRT je velmi silným konkurentem a v celkovém hodnocení zaostal o pouhá 4 %, pro praktickou část práce byl nakonec zvolen reportingový nástroj od Pentaho. Hlavním důvodem volby právě tohoto nástroje je jeho kompatibilita s ostatními nástroji od Pentaho a silná uživatelská komunita.

Shrnutí

Výsledky porovnání open source nástrojů BI se v některých případech více či méně liší od výsledků prací zabývajících se podobnou tématikou. Např. práce Z. Filipčíka [3] hodnotí PRD nejhůře ze tří uvedených nástrojů. Jako celek umístil autor nástroje Pentaho na druhé místo díky poměru náročnosti a funkcionality. V práci V. Formánka [5] vychází Pentaho jako jediný celek, který obsahuje všechny zkoumané komponenty BI. Je tedy zřejmé, že kompletní řešení Pentaho má ve světě open source nástrojů velikou roli.

Je nutno dodat, že v hodnocení hrály roli i zkušenosti s touto problematikou a specifické požadavky pro dané řešení. Proto není vždy vhodné se při výběru nástrojů BI řídit pouze dostupnými recenzemi a hodnoceními. Vždy je důležité brát v potaz všechny požadavky pro konkrétní řešení.

6 Zpracování dat

Tato kapitola je věnována zpracování dat v projektu MARE za využití zvoleného nástroje, kterým se stal *Pentaho Data Integration*. Data jsou v PDI zpracovávána dle jednotlivých částí procesu ETL. Nejprve jsou data extrahována ze zdrojových systémů, poté jsou transformována do podoby cílového formátu a na závěr jsou data načtena do cílového systému, kterým je v tomto případě databáze. Pro pochopení procesu zpracování dat je nejprve nutné popsat model datového skladu.

6.1 Návrh datového skladu

DWA je vybudován podle datového modelu EI (*EnviroInsite*) MARE. Ten je navržen tak, aby bylo možné data načítat do aplikace EI, která slouží k zobrazování geologických a geochemických dat, a zároveň bylo možné data nahrát do DWA sloužící mimo jiné jako zdroj dat pro tvorbu reportů. Model je oproti standardnímu formátu EI rozšířen o tabulky a atributy sloužící k uchování dat pro další potřeby projektu. Model datového skladu je neustále vyvíjen a upravován dle aktuálních potřeb. Tabulky datového skladu, se kterými je nejčastěji pracováno, jsou popsány v *tabulce 10* a jejich vztahy jsou vyznačeny v *obrázku 14*.

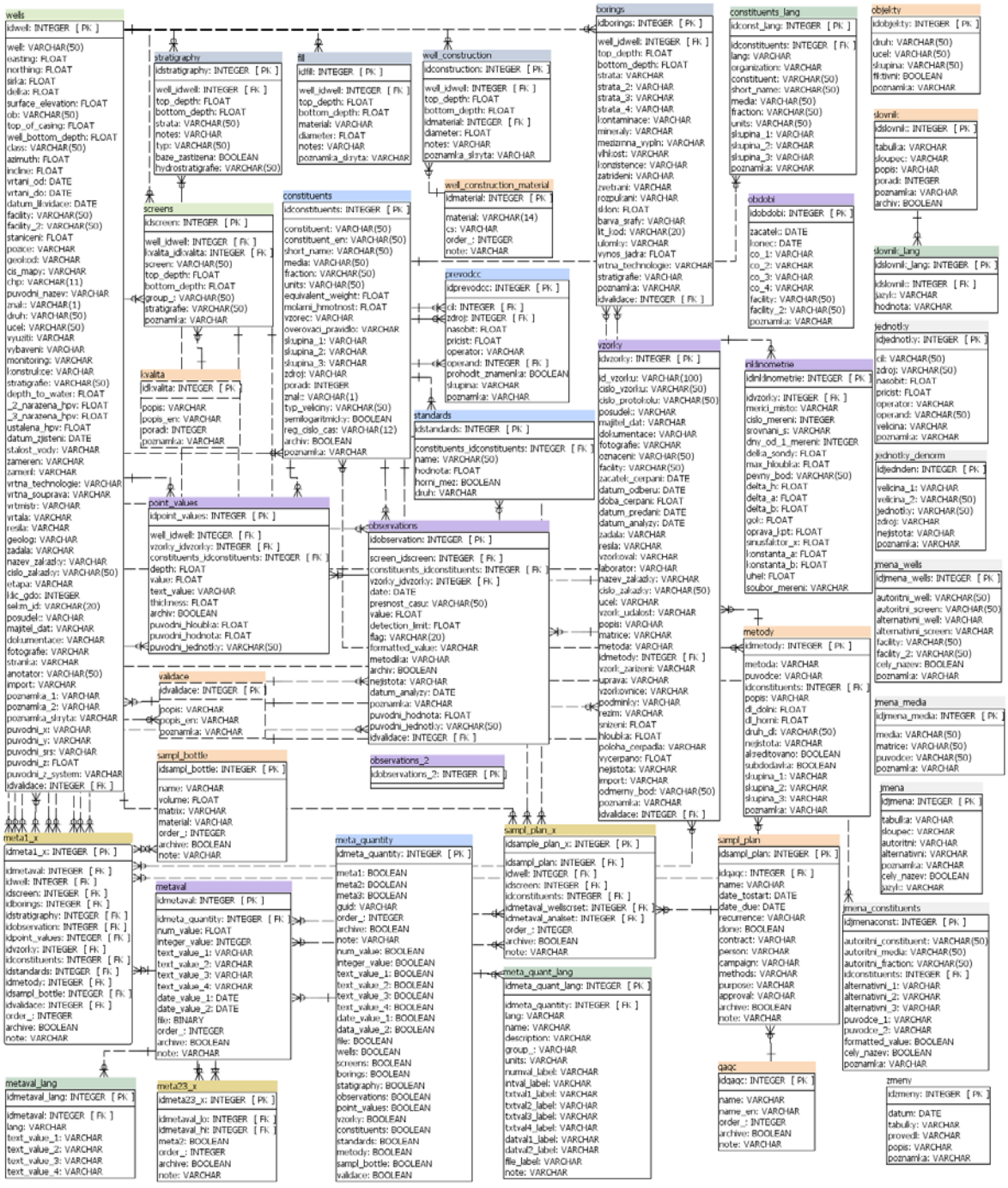
Datový sklad odpovídá částečně normalizovanému *schématu souhvězdí*. Sledovanými ukazateli jsou naměřené hodnoty konkrétních veličin přírodních vod (atribut *Value*), jenž jsou vyhodnocovány v rámci reportingu. Ukazatele jsou uloženy ve dvou *tabulkách faktů*, kterými jsou *Observations* a *Point Values*. Tabulky sloužící jako datový sklad jsou *tabulkami dimenzí*. Zvláštností je atribut *Date* v tabulce *Observations*. Je typem tzv. *degenerované dimenze*. Jedná se o typ dimenze, jenž nemá vlastní tabulku. Tabulka je obsažena přímo v *tabulce faktů* (konkrétně tab. *Observations*). Důvodem je velký počet dat, která by tabulka dimenzí obsahovala. V případě tohoto řešení by *tabulka dimenzí* typu *Date* měla tak vysoký počet záznamů, že by tento způsob realizace vedl k výraznému zhoršení výkonu databáze.

V modelu datového skladu je možno vidět vrstvu *metadat,* jejímž úkolem je zajistit popis vybraných atributů či tabulek.

Tabulka 10: Popis tabulek datového modelu EI MARE

Tabulka	Popis
Wells	Popis a identifikace objektů, zejména vrtů
Screens	Hloubkový interval, ve kterém byly odebrány vzorky
Constituents	Soupis analytů a měřených veličin
Observations	Jednotlivá měření vázána ke vzorkovanému hloubkovému intervalu
Point Values	Jednotlivá měření vázána ke konkrétní hloubce ve vrtu
Borings	Popis geologických vrstev
Stratigraphy	Interpretované vrstvy pro geologický řez
Fill	Obsyp a těsnění vrtu
Well Construction	Výstroj vrtu
Vzorky	Data o měřených vzorcích
Obdobi	Popis časových událostí, vhodných pro tvorbu časových grafů
Inklinometrie	Metadata k inklinometrickým měřením
Jednotky	Přejmenování jednotek, ve kterých byly naměřeny vzorky
Jmena Wells	Přejmenování jmen objektů
Jmena Constituents	Přejmenování názvů veličin
Jmena Media	Přejmenování hodnoty Media, jež charakterizuje typ vzorku
Jmena	Přejmenování textových hodnot

Zdroj: vlastní



Obrázek 14: Model datového skladu
Zdroj: [2]

6.2 Proces zpracování dat v PDI

Vstupní data, detailněji popsaná níže, jsou načítána do PDI v různých datových formátech. Data je nutno transformovat a vyfiltrovat tak, aby odpovídala datovému modelu EI MARE a následně je bylo možno převést do požadovaného formátu, kterým je *MS Excel* a *MS Access*. Data jsou zpracovávána do tabulek popsaných v *tabulce 10* jako *EnviroInsite tabulky* a *MARE datové tabulky*.

Jelikož data pocházejí z různých zdrojů, mezi záznamy v souborech dochází k nejednoznačnosti názvů, kdy bývá užito různých termínů pro jednu a tu samou položku v tabulce (zkoumaná veličina se může v jedné tabulce jmenovat *fosfor*, v jiné *P* apod.). Proto je nutné data (nyní uložené ve formátu *Excel* či *Access*) upravit tak, aby byly všechny názvy jednoznačné a aby nedocházelo k redundanci dat. K těmto účelům je využito číselníků a tabulek *MARE import* popsaných výše. Upravená data odpovídající datovému modelu jsou poté nahrávána do databáze.

V rámci práce byl vytvořen typ transformace, který generuje jednoduchou dokumentaci k již vytvořeným transformacím.

6.3 Transformace dat

Hydrogeologická data pocházejí z různých oblastí a zdrojů. Mají rozdílné formáty a strukturu, kterou je nutno před samotným zpracováním pochopit. Poté je možno data načíst do PDI a následně je zpracovat do požadované podoby. Pro lepší interpretaci a přehlednost vytvořených transformací jsou v *tabulce 11* popsány nejčastěji využité komponenty v PDI. Komponenty jsou spojovány šipkami, které určují směr průběhu transformace. Níže je popsán obecný postup při vytváření transformace. Popis konkrétních řešení je zaměřen na zajímavosti a zvláštnosti v postupech.

Tabulka 11: Přehled nejčastěji využívaných komponent v PDI

Název	Ikona	Popis
Microsoft Access Input		Načítání dat ze souborů ve formátu Microsoft Access
Text File Input		Načítání dat z textových souborů
Microsoft Access Output		Ukládání dat do tabulky databáze Microsoft Access
Sort rows		Seřazení dat ve vybraném sloupci (sloupcích) vzestupně nebo sestupně
Merge Join		Spojení řádků dvou načítaných vstupů pomocí vybraného klíče do jednoho výstupu. Vstupy musí být před spojením seřazeny podle vybraného klíče
Select values		Výběr, úprava nebo vymazání polí. Možnost změnit typ, formát nebo délku dat
User Defined Java Expression		Vlastní výraz napsaný v jazyce Java
Add constant		Přidání jedné nebo více hodnot (konstant) do polí
Filter rows		Filtrování dat řádků pomocí vybraných podmínek
Split Fields		Rozdělení jednoho pole do více polí pomocí vybraných podmínek
Unique rows		Kontrola unikátnosti řádků. Vymazání duplicitních záznamů
Calculator		Vytvoření nového pole pomocí matematických zápisů
Row flattener		Převod sloupce na řádek
Add sequence		Inkrementace hodnot (indexace řádků apod.)
Replace in string		Nahrazení vybraných znaků v řetězcích
Copy rows to result		Zkopírování vybraných polí, která je možno dále načíst a použít v jiných transformacích
Get file names		Získání cesty/jména ke všem souborům ve vybraných oblastech (adresáře)
Get rows from previous result		Získání pole z předchozích komponent, jako je *Copy rows to result*
Concat fields		Slučování více polí do jednoho
Regex evaluation		Vlastní regulární výraz
Stream lookup		Porovnání dvou či více stejných textových polí
Formula		Výraz (matematický, logický, spojování)
Set Variables		Nastavení proměnných

Zdroj: vlastní

Obecný postup při vytváření vlastních transformací v PDI

V PDI je možno vytvořit *transformation* a *job*. Základem je *transformation* sloužící k tvorbě transformací, kde jsou pomocí komponent, které mají specifickou funkci, upravována data. Při založení *job* se pracuje také s komponenty, ale jejich úloha je poněkud jiná. Zatímco v *transformation* jsou komponenty využívány pro práci s daty, v *job* jsou funkce komponent zaměřeny na správu transformací. *Job* je tedy jakousi nadřazenou formou transformace sloužící ke spouštění již vytvořených transformací, správě souborů či zasílání e-mailů s daty.

Po zjištění struktury a formátu uložení zdrojových dat je třeba v nabídce nalézt příslušnou komponentu pro načtení daného typu souboru. Komponenta se pomocí funkce *drag & drop* umístí na plochu pro transformaci. Dle potřeb jsou využity komponenty pro úpravu dat. Nejvíce využívané jsou komponenty pro normalizaci/denormalizaci, spojování a filtrování dat. Pro práci s konkrétními hodnotami záznamů je využíváno komponent pro nahrazování znaků či přejmenování záznamu. Každá z transformací končí komponentou pro výstup, která ukládá data do zvoleného formátu.

Zdrojová data pocházejí od různých subjektů a každý balík dat obsahuje desítky až tisíce souborů. Vstupní soubory jednotlivých subjektů obsahují různorodá data, která je potřeba upravit a poté uložit do různých tabulek datového modelu. Ke každému balíku zdrojových dat byly proto vytvořeny transformace odpovídající jednotlivým tabulkám datového modelu EI MARE. Zároveň byl každé skupině vytvořen jeden *job*, který vytvořené transformace spouští za sebou vždy tak, aby se nejprve do cílového souboru ukládaly tabulky s primárními klíči a až po nich tabulky na nich závislé.

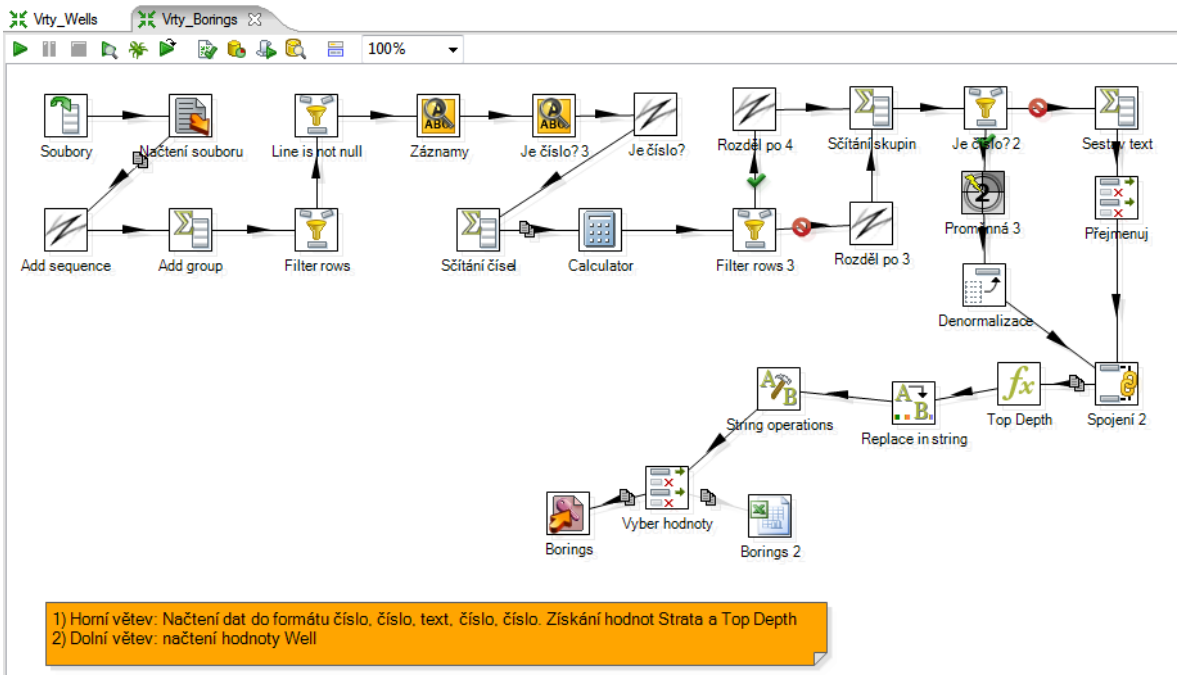
6.3.1 Transformace zdrojových dat

V následujících odstavcích jsou popsány konkrétní transformace dat, které jsou rozděleny podle toho, od jakých subjektů data pocházejí.

Vrty Chabařovice a Ležáky

Vytvořeno: 1x job, 2x transformation (Wells, Borings)

Vstupní data jsou uložena jako textové soubory, kterých je přes 2500. Data jsou semistrukturovaná. Nejsou tedy uspořádaná podle jednoho pevně daného schématu nebo jsou neúplná. V uspořádání lze ale nalézt určité podobnosti či opakující se souvislosti, podle kterých lze data načítat a upravovat. Těchto podobností je využito v transformaci *Borings*. Vstupní soubory mají různý počet řádků a tím se liší i umístění jednotlivých záznamů. Podle počtu řádků a typu záznamu (číslo nebo text) na konkrétních řádcích je v transformaci pomocí větvení vyhodnoceno, jak budou dále data upravována. Na *obrázku 15* je ukázka transformace *Vrty_Borings*, která upravená data nahrává do tabulky *Borings*.

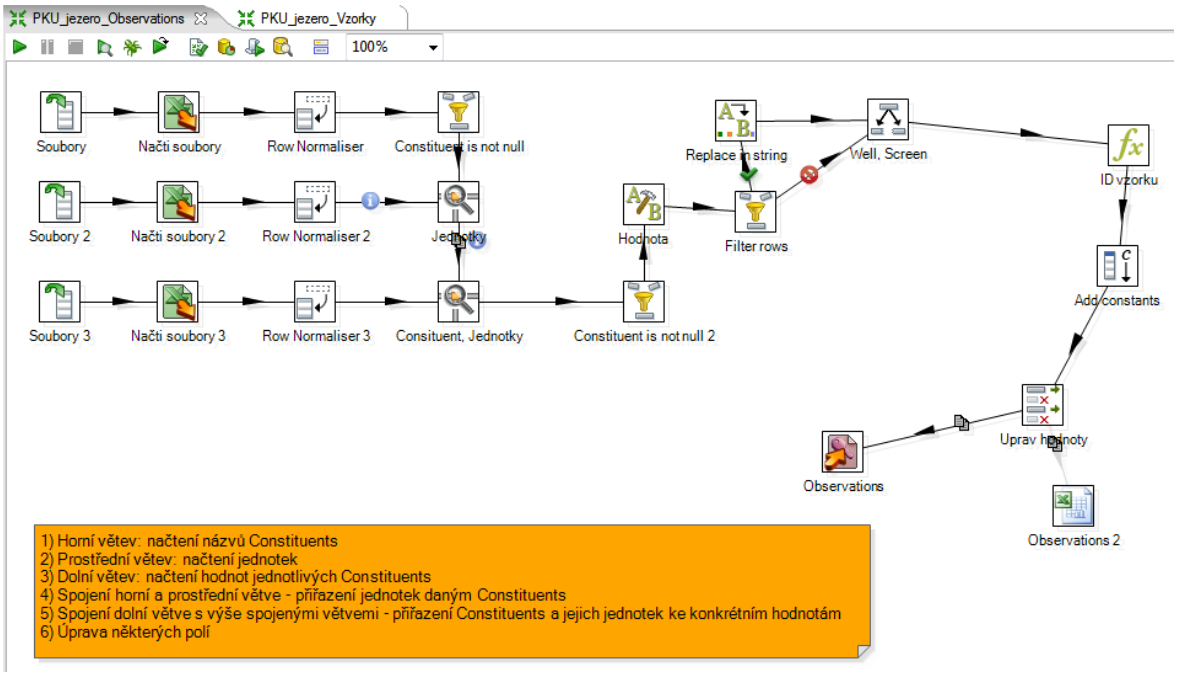


Obrázek 15: Transformace Vrty_Borings
Zdroj: vlastní

Palivový kombinát Ústí, s. p.

Vytvořeno: 1x job, 2x transformation (Observations, Vzorky)

Vstupem je několik desítek souborů ve formátu *Microsoft Excel*. Data mají pevně danou strukturu a nesou informace o měřeních prováděných na jezeře Chabařovice. Velmi využívanou komponentou je *Filter rows*, která propouští data v případě, že řádek tabulky obsahuje pole s konkrétní naměřenou hodnotou. V případě, že by komponenta v transformaci nebyla, ve výstupu by se objevily údaje s nulovou informační hodnotou, což je v dalších procesech nežádoucí. Transformaci lze vidět na *obrázku 16*.



Obrázek 16: Transformace PKU Observations
Zdroj: vlastní

Jezero Ležáky (Chemismus)

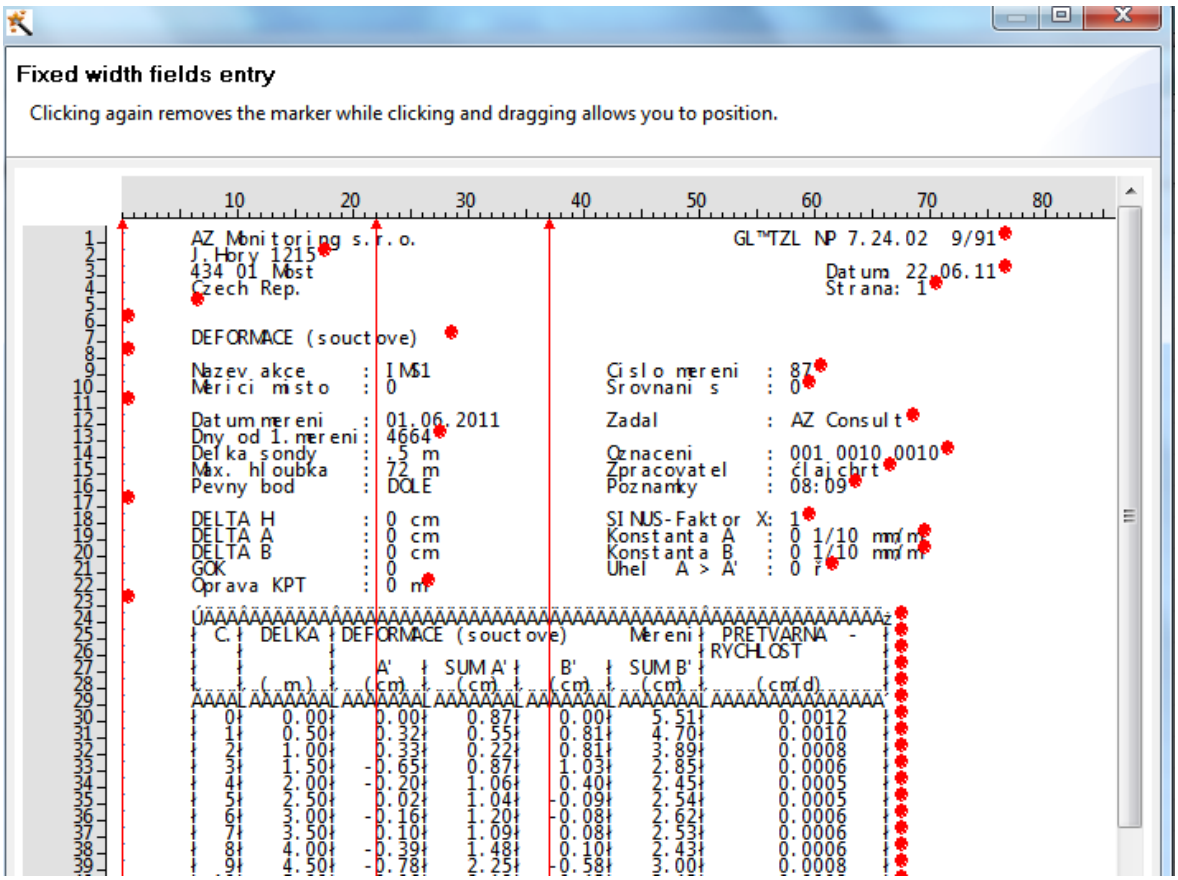
Vytvořeno: 1x job, 2x transformation (Observations, Vzorky)

Zdrojem dat je jediný soubor ve formátu *Microsoft Excel*. Obě transformace spočívají pouze v přejmenování atributů a zabránění redundance dat pomocí komponenty *Unique rows*.

Inklinometrie

Vytvořeno: 1x job, 3x transformation (Inklinometrie, Point Values, Vzorky)

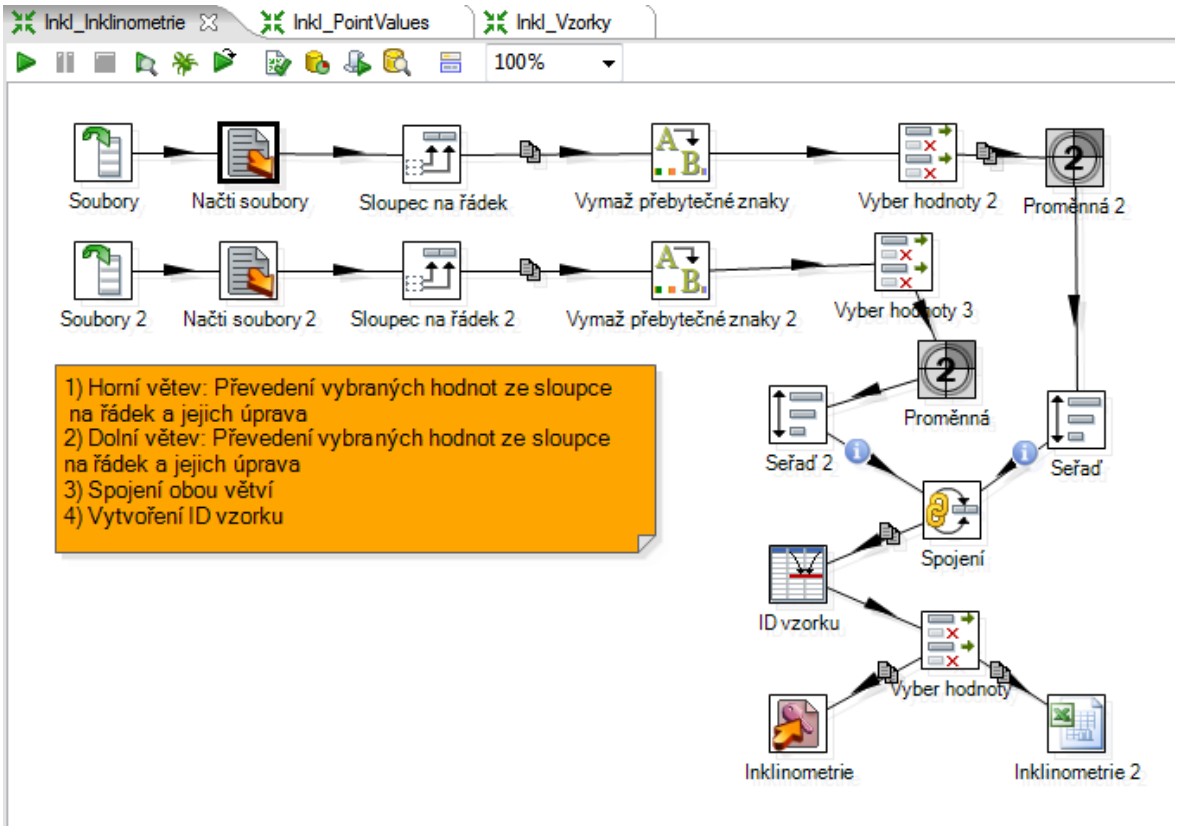
V případě dat Inklinometrie se jedná o desítky souborů s příponou DRT. Po jejich otevření v textovém editoru je zřejmé, že se jedná o soubory se semistrukturovanými daty. V hlavičce souboru jsou uchovány informace o měření, konkrétní hodnoty jsou pak v textové *„tabulce“*, jejíž počet řádků je v jednotlivých souborech odlišný. Těchto tabulek obsahuje každý soubor několik. Jako vstup slouží komponenta *Text File Input*, ve které jsou pomocí vestavěné funkce části souboru rozděleny do předem určených polí, jak je vyznačeno na *obrázku 17*.



Obrázek 17: Funkce PDI pro členění textového dokumentu dle kotevních bodů
Zdroj: vlastní

Zajímavý je způsob řešení různorodé délky textové *„tabulky“*, respektive různý počet naměřených hodnot v souborech. Pomocí komponenty *Filter Rows* jsou díky několika

podmínkám vyfiltrovány pouze řádky, které obsahují číselný údaj o měření. Jedna z transformací zpracovávající inklinometrická data je vyznačena na *obrázku 18*.



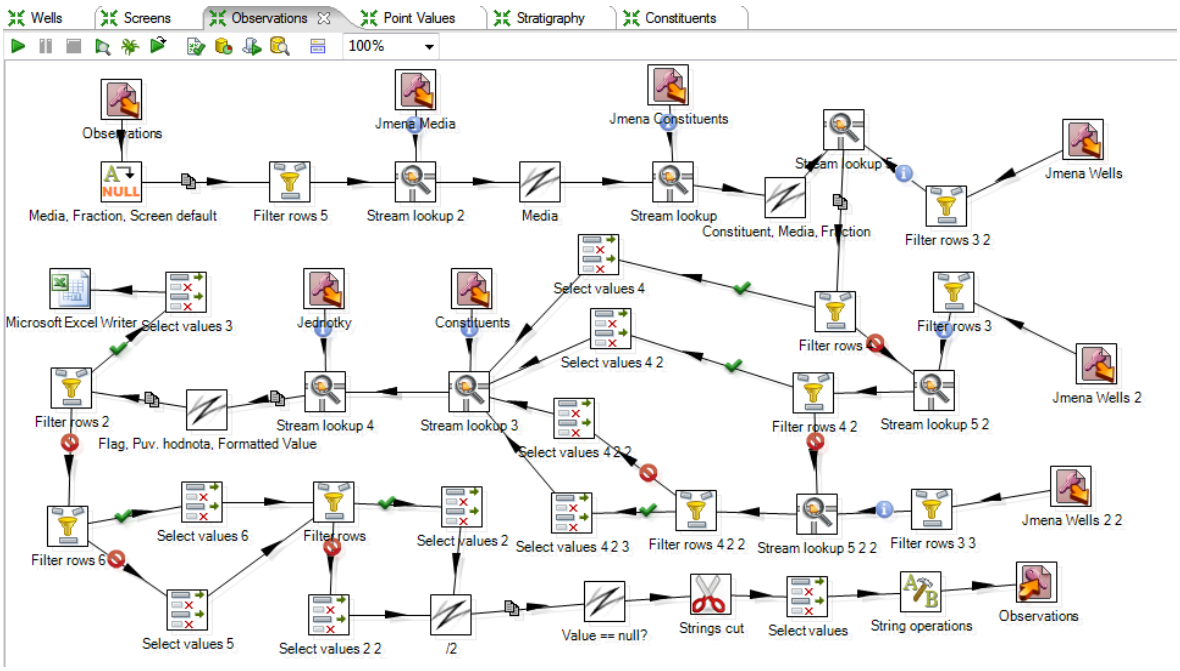
Obrázek 18: Inklinometrie
Zdroj: vlastní

6.3.2 Transformace pro sjednocení dat z různých zdrojů

Vytvořeno: 1x job, 13x transformation

Každá z transformací odpovídá jedné z tabulek datového modelu. Vstupem jsou zde výstupy výše popsaných transformací, tedy soubory ve formátu *Excel* či *Access* s tabulkami odpovídajícími datovému modelu. Úkolem těchto transformací je příslušné tabulky načíst a podle číselníků (tabulky *Jednotky*, *Jmena Wells*, *Jmena Constituents*, *Jmena Media a Jmena*) sjednotit názvy konkrétních záznamů. Výsledkem je pak *jedna verze pravdy*, kde se jeden typ záznamu nevyskytuje pod více názvy a kde mají jednotlivé veličiny stejné jednotky.

Výsledkem jsou často velmi komplikované transformace, které se mnohonásobně větví a ukazují sílu a možnosti PDI. Na *obrázku 19* je transformace, která sjednocuje názvy v tabulce *Observations*.

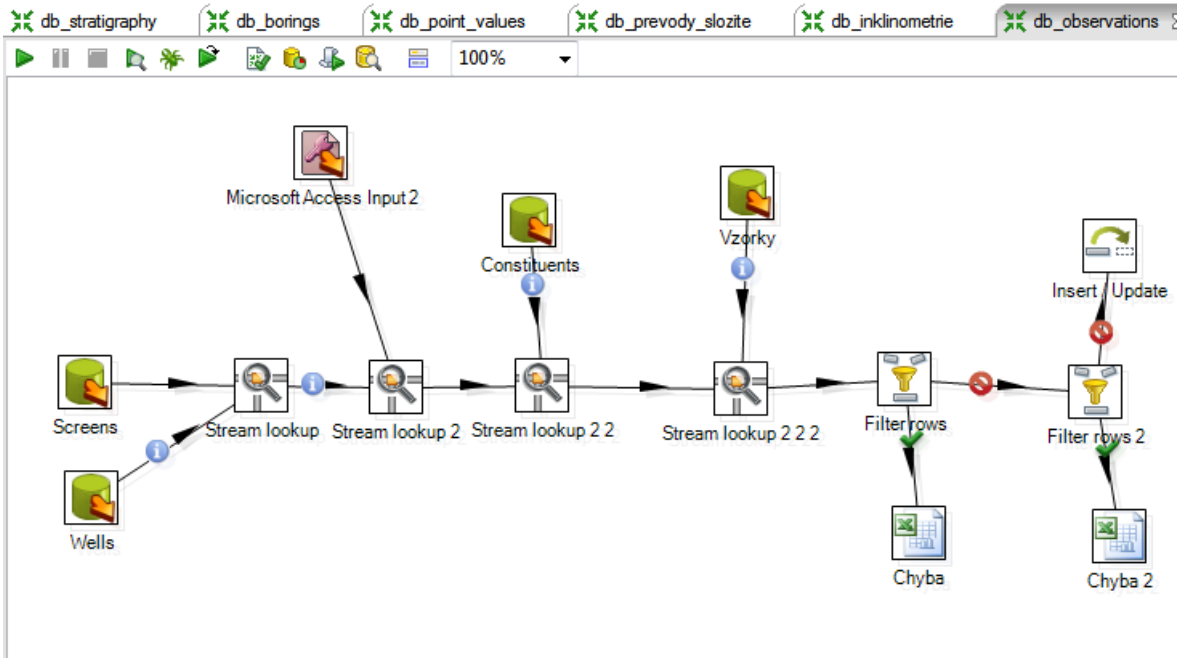


Obrázek 19: Navazující transformace
Zdroj: vlastní

6.3.3 Transformace pro import do databáze

Vytvořeno: 13x transformation

Vytvořené transformace (viz *obrázek 20*) slouží k nahrávání dat odpovídající struktuře datového modelu do databáze. V transformacích dochází k postupnému načítání jednotlivých tabulek datového modelu, které jsou ukládány do relační databáze. Nahrávat je třeba postupně, tedy nejdříve tabulky s primárními klíči a až poté tabulky na nich závislé. U některých tabulek je při nahrávání do databáze vytvořen klíč umělý. Proto je při některých transformacích potřeba tyto tabulky nahrát a spojit tyto klíče s primárními. Použity jsou výhradně databázové komponenty jako *Insert/Update Table*, *Input Table* a *Output Table*.



Obrázek 20: Transformace pro import do databáze
Zdroj: vlastní

6.3.4 Transformace pro tvorbu dokumentace

Vytvořeno: 1x transformation

Tato transformace má za úkol vygenerovat jednoduchou dokumentaci k vytvořeným projektům. Výstupem je HTML soubor obsahující cesty k projektům, screenshoty, datumy posledních úprav apod.

7 Tvorba reportů

Tato kapitola je věnována tvorbě reportů pomocí aplikace PRD. Vstupními daty jsou hydrogeologická data, která byla v rámci praktické části (*kapitola 6*) vyčištěna a převedena do struktury datového modelu EI MARE a poté nahrána do databáze.

V následujících odstavcích jsou popsána konkrétní řešení, kde jsou nejprve stručně představeny smysl a cíle reportů, poté je vysvětleno technické řešení, následováno popisem vytvořeného reportu s jeho ukázkou.

Obecný postup při sestavování reportu

Vstupem pro všechny reporty je databáze s hydrogeologickými daty. Databáze má velmi podobnou strukturu jako datový model popsaný v *tabulce 8*. V každém reportu má uživatel na výběr parametry, kterými určí, jaký objekt (objekty) či veličinu (veličiny) chce v reportu zobrazit. Nabídky parametrů se načítají pomocí SQL dotazů z databáze. Všechny parametry jsou kaskádovitě propojeny, a tak se při zvolení libovolného parametru ostatní nabídky s parametry aktualizují. Díky tomuto řešení zde neexistuje možnost, že by uživatel zvolil kombinaci, která v databázi není uložena.

V PRD jsou komponenty (text, graf, čára apod.) umisťovány pomocí funkce *drag & drop*. Pro načítání dat do reportu je využito SQL dotazu, ve kterém jsou zohledněny parametry, které vybral uživatel. Jednotlivé atributy z databáze jsou pak umisťovány na plochu reportu, kde se při spuštění zobrazí konkrétní hodnoty z databáze.

7.1 Časový průběh veličiny

Cílem je vytvoření reportu pro uživatelem vybrané veličiny. Uživatel má možnost vybrat si z parametrů jako je *Well, Screen, Ucel, Datum* apod. Ke zvolené veličině je pak v reportu vykreslen graf závislosti na čase a zobrazen průměr, počet hodnot a hodnoty minima a maxima. Report slouží k získání informací o veličinách a jejich hodnotách v konkrétních oblastech.

7.1.1 Technické řešení

Zdrojem dat je SQL databáze, kde jsou uloženy záznamy o veličinách a objektech. Uživatel má možnost vybrat si z dvanácti parametrů. Parametry uživatel vybírá ze seznamu.

Trojice *Constituent*, *Media* a *Fraction* určuje veličinu. Zbylé parametry (*Well*, *Screen*, *Class*, *Facility*, *Facility 2*, *Druh*, *Ucel*, *Datum od*, *Datum do*) se vztahují k objektu, ve kterém byly hodnoty dané veličiny naměřeny.

Všechny zvolené hodnoty jsou použity v podobě parametrů v SQL dotazu, který načítá zvolená data z databáze. Z načtených hodnot je pak pomocí funkce *chart* vykreslen graf závislosti veličiny na čase. Funkce *Sum* sčítá naměřené hodnoty a funkce *Count* zjistí počet hodnot. Jejich podílem (*Sum*/*Count*) je získán průměr. Pro zjištění hodnoty minima (resp. maxima) je použita funkce *Minimum* (*Maximum*).

7.1.2 Popis reportu

V horní části reportu je výpis zvolené trojice *Constituent*, *Media* a *Fraction*. Jsou zde i hodnoty minima, maxima, průměru a počtu hodnot. V prostřední části je vykreslen graf závislosti veličiny na čase. Na ose *x* jsou naměřené hodnoty a na ose *y* je vyznačeno datum. V dolní části reportu jsou na řádcích vypsány jednotlivé hodnoty a k nim příslušné informace. V zápatí je pak zobrazeno datum, kdy byl report vytvořen, a číslo strany. Vytvořený report je vidět na *obrázku 21*.

7.1.3 Ukázky

Časový průběh veličiny

Veličina: Teplota vody **Typ:** w **Výběr měření:**-

Jednotky: °C

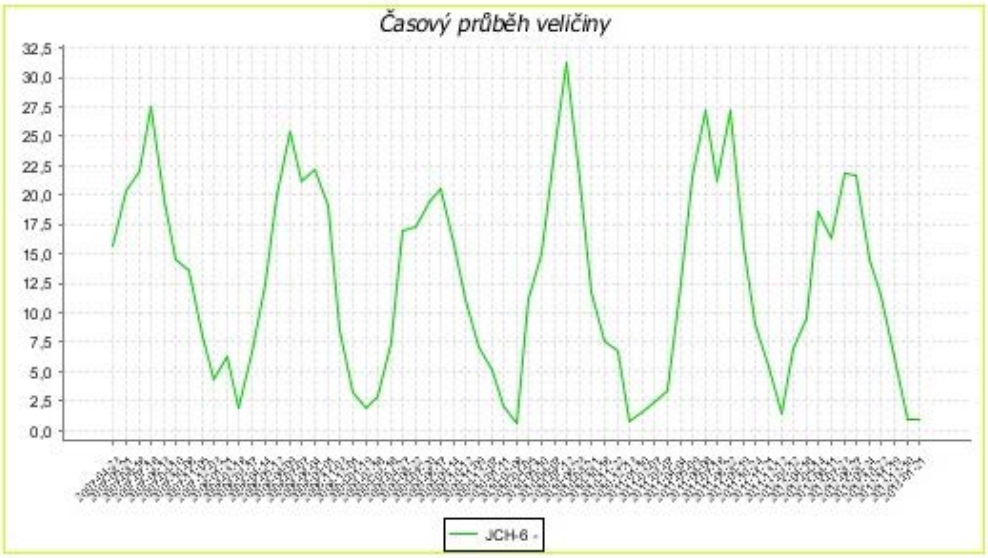
Vybrané objekty: JCH-6

Počet hodnot: 65

Průměr: 12,72 **Variační rozpětí:** 30,6

Minimum: 0,6

Maximum: 31,2



Naměřeno	Stanice	Hloubkový interval	Detail	Lokalita	Dílčí lokalita	Typ	Účel	Datum
15,7	JCH-6	-	-	jCh	-	jezero	-	23.04.2007
20,3	JCH-6	-	-	jCh	-	jezero	-	21.05.2007
22	JCH-6	-	-	jCh	-	jezero	-	18.06.2007
27,5	JCH-6	-	-	jCh	-	jezero	-	16.07.2007
19,6	JCH-6	-	-	jCh	-	jezero	-	13.08.2007
14,6	JCH-6	-	-	jCh	-	jezero	-	10.09.2007
13,6	JCH-6	-	-	jCh	-	jezero	-	08.10.2007
8	JCH-6	-	-	jCh	-	jezero	-	05.11.2007
4,3	JCH-6	-	-	jCh	-	jezero	-	03.12.2007
6,3	JCH-6	-	-	jCh	-	jezero	-	21.01.2008

04.05.2015 str. 1/3

Obrázek 21: Časový průběh veličiny
Zdroj: vlastní

7.2 Report dokumentace (geologických) průzkumných objektů

Cílem je vytvořit report zobrazující informace o vybraných geologických objektech. Uživatel má možnost vybrat si ze seznamu objektů, jež jsou uloženy v databázi. Smyslem reportu je přehledně zobrazit informace o vybraných objektech včetně jejich geologických vrstev.

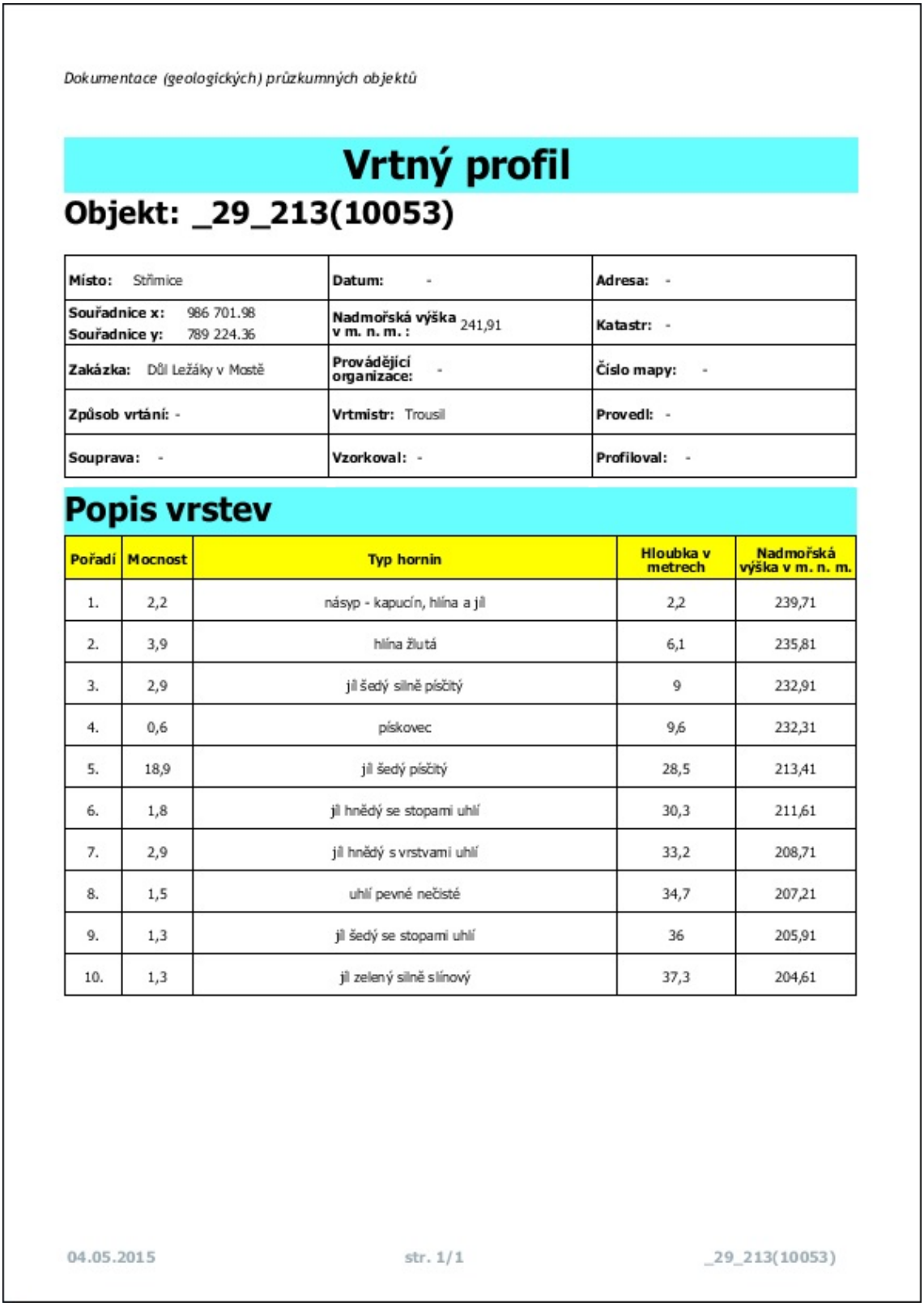
7.2.1 Technické řešení

Data jsou v PRD načítána z SQL databáze. Poté, co si uživatel vybere objekt nebo objekty, ke kterým chce sestavit report, je název objektu předán jako parametr do SQL dotazu, který k daným objektům načte data z databáze.

7.2.2 Popis reportu

Vybraná data z SQL dotazu jsou načítána do předem připravené tabulky, kde jsou základní informace o objektu. Ke každému objektu jsou načítány jednotlivé geologické vrstvy (jíl, uhlí atd.). Každý záznam vrstvy v reportu obsahuje její popis, pořadí, mocnost, celkovou hloubku a nadmořskou výšku. Počet řádků jednotlivých vrstev je dynamický a mění se podle počtu vrstev uložený v databázi. V zápatí stránky je uvedeno datum, kdy byl report vytvořen, název objektu a strana. Vytvořený report je možno vidět na *obrázku 22*.

7.2.3 Ukázky



Dokumentace (geologických) průzkumných objektů

Vrtný profil

Objekt: _29_213(10053)

Místo: Střimice	**Datum:** -	**Adresa:** -
Souřadnice x: 986 701.98 **Souřadnice y:** 789 224.36	**Nadmořská výška v m. n. m. :** 241,91	**Katastr:** -
Zakázka: Důl Ležáky v Mostě	**Provádějící organizace:** -	**Číslo mapy:** -
Způsob vrtání: -	**Vrtmistr:** Trousil	**Provedl:** -
Souprava: -	**Vzorkoval:** -	**Profiloval:** -

Popis vrstev

Pořadí	Mocnost	Typ hornin	Hloubka v metrech	Nadmořská výška v m. n. m.
1.	2,2	násyp - kapucín, hlína a jíl	2,2	239,71
2.	3,9	hlína žlutá	6,1	235,81
3.	2,9	jíl šedý silně písčitý	9	232,91
4.	0,6	pískovec	9,6	232,31
5.	18,9	jíl šedý písčitý	28,5	213,41
6.	1,8	jíl hnědý se stopami uhlí	30,3	211,61
7.	2,9	jíl hnědý s vrstvami uhlí	33,2	208,71
8.	1,5	uhlí pevné nečisté	34,7	207,21
9.	1,3	jíl šedý se stopami uhlí	36	205,91
10.	1,3	jíl zelený silně slínový	37,3	204,61

04.05.2015 str. 1/1 _29_213(10053)

Obrázek 22: Dokumentace objektů
Zdroj: vlastní

7.3 Multikriteriální analýza

Pro hodnocení kvality vody v jezeře byla zpracována tzv. multikriteriální analýza. Zaměřuje se zejména na hodnocení a predikci projevů eutrofizace, které jsou významným rizikovým faktorem. Výsledky jednotlivých kritérií nabývají logických hodnot *„pravda“* či *„nepravda“*. Kritériím jsou přiřazeny váhy podle významu. Vážený průměr kritérií představuje hlavní hodnotící kritérium. Při překonání kritické hodnoty skóre dojde k negativním projevům eutrofie[13].

7.3.1 Technické řešení

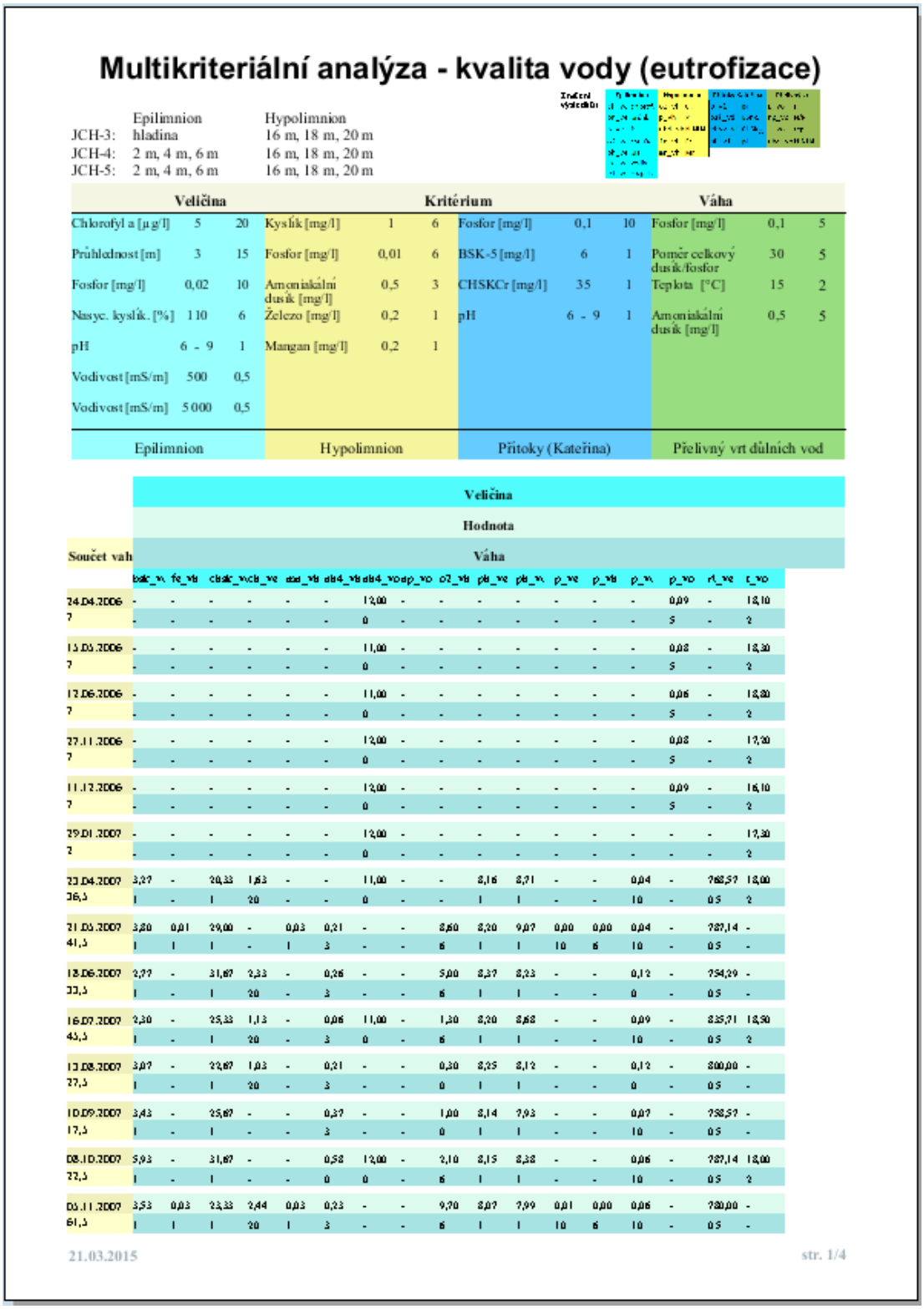
Jelikož se jedná o nejkomplexnější report, je jeho řešení poměrně náročné. Uživatel má na výběr ze šesti parametrů. Pomocí nich volí, v jakém rozmezí (hloubce) chce analýzu provádět. V hlavním SQL dotazu je k veličinám počítán průměr naměřených hodnot za jednotlivá období. Tyto průměry jsou porovnávány s kritérii, které jsou přednastaveny, ale lze je rovněž podle parametrů upravit. V případě, že hodnoty vyhovují kritériím, jsou jim přiřazeny váhy (lze opět ručně nastavit v parametrech). Tyto operace jsou prováděny pomocí funkce PRD *Open formula*. Všechny váhy jsou pak pomocí funkce *Sum* sečteny a slouží jako informace o projevech eutrofie.

7.3.2 Popis reportu

Informace o vybraných hloubkových úrovních jsou zobrazeny v horní části reportu. Parametry kritérií a vah jsou zobrazeny pod nimi. V těle reportu jsou interpretovány samotné výsledky. K jednotlivým datům jsou zde zobrazeny průměry naměřených hodnot různých veličin s jejich vahami. Pod datem je pak jejich součet. V zápatí je informace o datu, kdy byl report vygenerován, a číslo aktuální strany. Popisovaný report je na *obrázku 23*.

[13] stav přiměřené výživy, vývoje a růstu organismu a jeho částí

7.3.3 Ukázky



Multikriteriální analýza - kvalita vody (eutrofizace)

	Epilimnion	Hypolimnion
JCH-3:	hladina	16 m, 18 m, 20 m
JCH-4:	2 m, 4 m, 6 m	16 m, 18 m, 20 m
JCH-5:	2 m, 4 m, 6 m	16 m, 18 m, 20 m

Veličina			Kritérium			Váha		
Chlorofyl a [µg/l]	5	20	Kyslík [mg/l]	1	6	Fosfor [mg/l]	0,1	10
Průhlednost [m]	3	15	Fosfor [mg/l]	0,01	6	BSK-5 [mg/l]	6	1
Fosfor [mg/l]	0,02	10	Amoniakální dusík [mg/l]	0,5	3	CHSKCr [mg/l]	35	1
Nasyc. kyslík. [%]	110	6	Železo [mg/l]	0,2	1	pH	6 - 9	1
pH	6 - 9	1	Mangan [mg/l]	0,2	1			
Vodivost [mS/m]	500	0,5						
Vodivost [mS/m]	5000	0,5						

Epilimnion	Hypolimnion	Přítoky (Kateřina)	Přelivný vrt důlních vod

	Veličina																		
	Hodnota																		
Součet vah	Váha																		
	balc_vr	fe_vs	chsk_vr	ch_ve	dus_vs	dis4_vs	dis4_ve	dsp_vo	o2_vs	ph_ve	ph_vr	p_ve	p_vs	p_vr	p_vo	rt_ve	t_vo		
24.04.2006	-	-	-	-	-	-	12,00	-	-	-	-	-	-	-	0,09	-	18,10		
7	-	-	-	-	-	-	0	-	-	-	-	-	-	-	5	-	2		
15.05.2006	-	-	-	-	-	-	11,00	-	-	-	-	-	-	-	0,08	-	18,30		
7	-	-	-	-	-	-	0	-	-	-	-	-	-	-	5	-	2		
12.06.2006	-	-	-	-	-	-	11,00	-	-	-	-	-	-	-	0,06	-	18,80		
7	-	-	-	-	-	-	0	-	-	-	-	-	-	-	5	-	2		
27.11.2006	-	-	-	-	-	-	12,00	-	-	-	-	-	-	-	0,08	-	17,20		
7	-	-	-	-	-	-	0	-	-	-	-	-	-	-	5	-	2		
11.12.2006	-	-	-	-	-	-	12,00	-	-	-	-	-	-	-	0,09	-	16,10		
7	-	-	-	-	-	-	0	-	-	-	-	-	-	-	5	-	2		
29.01.2007	-	-	-	-	-	-	12,00	-	-	-	-	-	-	-	-	-	17,30		
7	-	-	-	-	-	-	0	-	-	-	-	-	-	-	-	-	2		
23.04.2007	3,27	-	20,33	1,63	-	-	11,00	-	-	8,16	8,71	-	-	0,04	-	768,57	18,00		
36,5	1	-	1	20	-	-	0	-	-	1	1	-	-	10	-	0,5	2		
21.05.2007	3,20	0,01	29,00	-	0,03	0,21	-	-	8,60	8,20	9,07	0,00	0,00	0,04	-	787,14	-		
41,5	1	1	1	-	1	3	-	-	6	1	1	10	6	10	-	0,5	-		
18.06.2007	2,77	-	31,67	2,33	-	0,26	-	-	5,00	8,37	8,23	-	-	0,12	-	754,29	-		
33,5	1	-	1	20	-	3	-	-	6	1	1	-	-	0	-	0,5	-		
16.07.2007	2,30	-	25,33	1,13	-	0,06	11,00	-	1,30	8,20	8,68	-	-	0,09	-	835,71	18,50		
45,5	1	-	1	20	-	3	0	-	6	1	1	-	-	10	-	0,5	2		
13.08.2007	3,07	-	22,67	1,03	-	0,21	-	-	0,30	8,25	8,12	-	-	0,12	-	800,00	-		
27,5	1	-	1	20	-	3	-	-	0	1	1	-	-	0	-	0,5	-		
10.09.2007	3,43	-	25,67	-	-	0,37	-	-	1,00	8,14	7,93	-	-	0,07	-	768,57	-		
17,5	1	-	1	-	-	3	-	-	0	1	1	-	-	10	-	0,5	-		
08.10.2007	5,93	-	31,67	-	-	0,58	12,00	-	2,10	8,15	8,38	-	-	0,06	-	787,14	18,00		
22,5	1	-	1	-	-	0	0	-	6	1	1	-	-	10	-	0,5	2		
05.11.2007	3,53	0,03	23,33	2,44	0,03	0,23	-	-	9,70	8,07	7,99	0,01	0,00	0,06	-	780,00	-		
61,5	1	1	1	20	1	3	-	-	6	1	1	10	6	10	-	0,5	-		

Obrázek 23: Multikriteriální analýza
Zdroj: vlastní

8 Zhodnocení přínosů řešení

Za použití nástroje PDI bylo v rámci DP vytvořeno přes 40 transformací. Pomocí těchto transformací byla zdrojová data načtena, upravena do požadovaného formátu a poté nahrána do databáze. Celkem bylo zpracováno přes 2500 souborů různých formátů a struktur.

Technický přínos řešení

Zpracovaná data a reporty[14] jsou důležitým prvkem pro podporu rozhodování o hodnocení kvality přírodních vod. Vytvořené reporty (*časový průběh veličiny*, *dokumentace (geologických) průzkumných objektů* a *multikriteriální analýza*) umožňují uživatelům pohled na zpracovaná data v podobě přehledných informací.

Přínosem transformací je automatizace zpracování obdržených strukturovaných a semistrukturovaných zdrojových dat. Zatímco ruční zpracování tisíců souborů by bylo neefektivní, časově velmi náročné, resp. takřka neproveditelné, v PDI trvá automatický převod dat maximálně desítky vteřin. Dále je možno vytvořené transformace využít při zpracování nových dat. V případě podobné struktury vstupních dat je ve vytvářené transformaci efektivní použít části již vytvořených transformací. Transformace lze aplikovat nejen na hydrogeologická data, se kterými bylo v rámci této DP pracováno, ale po drobných úpravách je možné využít transformace pro zpracování jakýchkoliv dat stejné či podobné struktury.

V rámci praktické části byly také popsány a porovnány vybrané open source Business Intelligence nástroje. Tento výběr a porovnání může posloužit čtenářům jako pohled na problematiku open source Business Intelligence nástrojů.

---

[14] Vytvořené transformace a reporty jsou přímo vázány na obdržená, veřejně nedostupná data, bez kterých není možné s transformacemi či reporty smysluplně pracovat. Proto je možno v případě zájmu kontaktovat a požádat autora o předvedení výstupů práce, poskytnutí částí kódů či rad ohledně práce ve výše uvedených nástrojích.

Kalkulace nákladů na vývoj, nasazení a provoz IS

Teoretické náklady spojené s vývojem a provozem IS za využití open source nástrojů Pentaho jsou uvedeny v *tabulce 12*. Pro porovnání jsou zde navíc uvedeny *náklady při využití komerčního řešení* a *náklady bez IS*, kde se data zpracovávají ručně (MS Excel apod.).

Tabulka 12: Kalkulace nákladů

Fáze	Náklady na konkrétní položku	Náklady na vlastní řešení (v Kč)		Náklady na komerční řešení (v Kč)		Náklady bez využití IS (v Kč)	
		Tvorba 1 reportu	Tvorba 100 reportů	Tvorba 1 reportu	Tvorba 100 reportů	Tvorba 1 reportu	Tvorba 100 reportů
Vývoj	Zpracování ETL	5000 (25 h)		5000 (25 h)		0	
	Tvorba reportů	1000 (5 h)		1000 (5 h)		0	
	Manuální zpracování dat	0		0		400 (2 h)	40 000 (200 h)
	HW (PC na práci)	20 000		20 000		10 000	
Nasazení	HW (server)	30 000		30 000		0	
	SW (licence)	0		25 000		0	
Provoz	HW (provoz serveru a jeho údržba, hosting atd.)	10 000		10 000		0	
	Optimalizace a aktualizace	2000 (10 h)		2000 (10 h)		200 (1 h)	4000 (20 h)
Celkové náklady (v Kč)		**68 000**		**93 000**		**10 600**	**54 000**

Zdroj: vlastní

Průměrná měsíční mzda IT zaměstnanců v různých pozicích ČR byla převzata z časopisu *CIO Business World*. [36] Z průměrných 174 odpracovaných hodin měsíčně byly rozpočítány náklady na konkrétní položky v tabulce. Zbylé náklady jsou brány jako teoretická částka dle zkušeností z oborů IT a ekonomie.

Počty hodin odvedené práce (v závorkách pod jednotlivými náklady) jsou uvedeny orientačně dle zkušeností v oblasti zpracování dat.

Vliv na výhodnost využití konkrétního řešení mají i další faktory, jako jsou školení pracovníků, prodloužení licence, počet zaměstnanců, různá mzda v jednotlivých typech řešení, distribuce reportů aj. Pro přesnější srovnání uvedených přístupů by bylo třeba nasazení všech řešení do praxe. Jako náhled na problematiku nákladů a optimální řešení je ovšem srovnání dostačující.

Z celkových nákladů vyplývá fakt, že při nízkém výstupu zpracovaných dat za dané období se nevyplatí budovat vlastní IS s nástroji BI. S rostoucím počtem vytvořených reportů (zpracovaných dat) se náklady na jeden report mezi uvedenými přístupy vyrovnávají, protože při využití nástrojů BI jsou náklady na zpracování dat stejné struktury jednorázové (i když vysoké), zatímco při manuální práci náklady stoupají lineárně s počtem zpracovaných dat. Při vytvoření 100 reportů jsou náklady na ruční zpracování dat nižší jen o 20,59 % než při využití vlastního řešení. V případě zpracování 132 reportů by byly částky shodné a při více jak 133 reportech se vyplácí využití vlastního řešení.

Náklady na komerční a vlastní řešení se liší v podstatě jen cenou za licenci. Celkové náklady na komerční řešení jsou vyšší, ale tomu odpovídají služby, podpora a kvalita, které nekomerční nástroje zpravidla nemají na tak vysoké úrovni. Výhodou vlastního řešení je navrhnutí IS podle individuálních potřeb projektu.

V projektu MARE, kde se zpracovává velké množství dat, je vlastní řešení IS z výše zmíněných důvodů výhodnější než využití uvedených alternativ.

Výhody v případě nasazení IS

Z výše uvedeného vyplývá, že přínosem v případě nasazení informačního systému pro hodnocení kvality životního prostředí je snížení provozních nákladů, které je předběžně odhadováno až na 30%. Jedním z důvodů je automatizace zpracování dat, kdy se omezí rutinní zpracování dat člověkem.

Další výhodou je zlepšená podpora rozhodování, která může podnikům přinést konkurenční výhody. Díky automatickému zpracování a vyhodnocování dat se mohou manažerům do rukou dostat důležitá data, která by jinak zůstala nepovšimnuta či opomíjena. Vyvíjený informační systém je v současnosti představován různým subjektům. V budoucnu je plánován prodej licencí a poskytování navazujících služeb (vývoj a provoz systému) subjektům, které by měly o nabízený informační systém zájem.

Další ekonomické ukazatele, jako je např. rentabilita, není možno uvést, jelikož IS není v tuto chvíli optimalizován pro různé skupiny uživatelů.

Závěr

Cílem práce bylo řešení problematiky hodnocení kvality přírodních vod pomocí open source nástrojů BI. Konkrétní řešení spočívá ve zpracování dat z různých zdrojů v různých formátech a jejich následná prezentace v podobě v reportů, které mají podpořit rozhodování o kvalitě přírodních vod. Práce probíhala ve spolupráci na vývoji IS v projektu MARE. K úspěšnému řešení bylo třeba splnit několik dílčích cílů stanovených na začátku práce.

Prvním z cílů byla charakteristika architektury BI. Popsány jsou základní komponenty BI z hlediska jejich využití. Pro vlastní řešení je v práci nejvíce využito komponent ETL, reportingu a DWH.

Pro výběr vhodných nástrojů bylo nutné zjistit jejich aktuální stav na trhu open source a následně vybrané nástroje porovnat, což byl další dílčí cíl práce. K hodnocení bylo využito vlastních kritérií, které jsou důležité z hlediska požadavků na řešení. Vlastní výsledky ve většině případů korespondují s již s dřívějšími publikovanými pracemi a vydanými články o nástrojích BI. Rozdíly ve výsledném hodnocení jsou způsobeny především vlastními zkušenostmi z oboru BI, subjektivním pohledem na některá z kritérií (např. prostředí aplikace) a také novějšími verzemi porovnávaných nástrojů.

Z nastudované problematiky vyplývá fakt, že v pozici lídrů na poli open source nástrojů jsou společnosti poskytující kompletní řešení v oblasti BI. Jedním z takových lídrů je společnost Pentaho, která nabízí nástroje v rámci celého procesu BI. Pentaho poskytuje řešení pro návrh datových skladů, ETL, dolování dat, reporting, OLAP či BI server. Nástroje *Pentaho Data Integration* (ETL) a *Pentaho Report Designer* (reporting) byly využity pro praktické řešení.

Stěžejním bodem práce bylo zpracování dat k hodnocení kvality přírodních vod. V této části jsou využity komponenty pro datový sklad projektu MARE, úpravu dat a reporting. Výstupem je pak několik reportů, které slouží k hodnocení kvality přírodních vod.

Jednou z výhod řešení je automatizace zpracování dat. Očištěná data jsou využívána v reportech, které automaticky generují výstupy podle zadaných parametrů. Odpadá tak potřeba rutinní a časově náročné manuální úpravy dat, která není při větším množství zpracovávaných dat finančně výhodná. Vytvořené transformace mohou být nadále využívány pro zpracování nových dat.

Díky vhodnému využití open source nástrojů BI lze ušetřit na nákladech spojených se zpracováním dat nejen v IS MARE. Podařilo se tedy nalézt vhodné řešení, které může být modifikováno a dále využíváno nejen v oblasti životního prostředí.

Jedním z možných návrhů do budoucna je optimalizace vytvořených transformací. Z postupně nabytých zkušeností vyplývá, že některé transformace by bylo možné zjednodušit a zrychlit za využití jiných komponent v aplikaci PDI. Dalším návrhem je pak optimalizace datového skladu, která by urychlila přístup k datům, jež jsou využívána zejména v reportingu.

Všechny body zadání byly úspěšně splněny a i nadále probíhá spolupráce na projektu MARE. Aktuálně je řešena tvorba dalších transformací pro data v nových formátech a také zpřístupnění vytvořených reportů uživatelům skrz webové rozhraní. To umožní přistupovat k informacím odkudkoliv a kdykoliv.

Citace

[1] KREJBICH, D. *Automatizace konverze datových formátů pro databázový systém*. Liberec, 2013. 50 s., 2 s. příl. Bakalářská práce (Bc.). Technická univerzita v Liberci, Fakulta mechatroniky, informatiky a mezioborových studií.

[2] NEŠETŘIL, K. *Ekvifinalita v modelování podzemní vody a využití nástrojů pro business intelligence*. Teze disertační práce (Ph.D.). Technická univerzita v Liberci, Fakulta mechatroniky, informatiky a mezioborových studií.

[3] FILIPČÍK, Z. *Nástroje Business Intelligence jako Open Source*. Praha, 2012. 114 s., 3 s. příl. Diplomová práce (Ing.). Vysoká škola ekonomická v Praze, Fakulta informatiky a statistiky.

[4] BEDNÁŘ, J. *Srovnání komerčních BI reportovacích nástrojů s nástroji Open Source*. Praha, 2013. 96 s., 1 s. příl. Diplomová práce (Ing.). Vysoká škola ekonomická v Praze, Fakulta informatiky a statistiky.

[5] FORMÁNEK, V. *Analýza trhu open source business inteligence*. Praha, 2010. 70 s., 4 s. příl. Bakalářská práce (Bc.). Vysoká škola ekonomická v Praze, Fakulta informatiky a statistiky.

[6] POUR, J., M. MARYŠKA a O. NOVOTNÝ. *Business intelligence v podnikové praxi*. 1. vyd. Praha: Professional Publishing, 2012, 276 s. ISBN 978-80-7431-065-2

[7] NOVOTNÝ, O., J. POUR a D. SLÁNSKÝ. Business intelligence: Jak využít bohatství ve vašich datech. 1. vyd. Praha: Grada, 2005, 254 s. ISBN 80-247-1094-3.

[8] KIMBALL, R. and M. ROSS. *The Kimball Group reader: Relentlessly practical tools for data warehousing and business intelligence*. Indianpolis, IN: Wiley, 2010, xxiv, 718 p. ISBN 04-706-3053-1.

[9] DRESNER, H. *The performance management revolution: Business results through insight and action*. Hoboken, N.J.: Wiley, 2008, xxii, 321 p. ISBN 978-047-0124-833.

[10] CASTERS, Matt R. *Pentaho kettle solutions: building open source etl solutions with pentaho data integration*. 1st ed. Indianapolis, IN: Wiley Pub., Inc, 2010, p. cm. ISBN 04-706-3517-7. H

[11] GARCIA MATTIO, M. and Dario R. BERNABEU. *Pentaho 4.0 Reporting by Example Beginner's Guide*. Online-Ausg. Birmingham: Packt Publishing, 2013. ISBN 17-821-6225-9.

[12] BOUMAN, Roland L. *Pentaho solutions: business intelligence and data warehousing with pentaho and MySQL*. 1st ed. Indianapolis, IN: Wiley Pub., Inc., 2009. ISBN 04-704-8432-2.

[13] PENTAHO CORPORATION. *Pentaho | Business analytics and business intelligence leaders* [online]. 2005 [vid. 2014-11-29]. Dostupné z: http://www.pentaho.com/

[14] ENGINEERING INGEGNERIA INFORMATICA. *SpagoWorld: The open source initiative supported by Engineering Group* [online]. ROK [vid. 2014-11-29]. Dostupné z: www.spagoworld.org

[15] TIBCO SOFTWARE. *Jaspersoft Business Intelligence* [online]. ROK [vid. 2014-11-29]. Dostupné z: https://www.jaspersoft.com/

[16] BIRT Product Line. OPENTEXT CORP. *Actuate: BIRT Software for Business Analytics & Business Intelligence* [online]. ROK [vid. 2015-01-10]. Dostupné z: http://www.actuate.com/products/

[17] CLOVERETL. *CloverETL: Rapid Data Integration* [online]. ROK [vid. 2015-02-03]. Dostupné z: http://www.cloveretl.com/

[18] KANAKIA, H.T., 2014. Report Generation using Business Intelligence Tools: A Comparative Study. *International Journal of Advanced Research in Computer Science*, 05, vol. 5, no. 5 ProQuest Technology Collection.

[19] ONG, I.L., P.H. SIEW and S.F.WONG, 2011. A Five-Layered Business Intelligence Architecture. *Communications of the IBIMA* ProQuest Central; ProQuest Technology Collection.

[20] ROVCANIN, A., A. MATARADZIJA and A. MATARADZIJA, 2012. *Knowledge Management through the Implementation of Business Intelligence Tools.* Singapore: Global Science and Technology Forum ProQuest Central.

[21] MARÍN-ORTEGA, Pablo M., V. DMITRIYEV, M. ABILOV a Jorge M. GÓMEZ. ELTA: New Approach in Designing Business Intelligence Solutions in Era of Big Data. *Procedia technology* [online]. 2014, č. 16, s. 667-674 [vid. 2015-02-22]. DOI: http://dx.doi.org/10.1016/j.protcy.2014.10.015. Dostupné z: http://www.sciencedirect.com/science/article/pii/S2212017314002424

[22] FOTR, J., E. VACÍK, I. SOUČEK, M. ŠPAČEK a S. HÁJEK. *Tvorba strategie a strategické plánování: teorie a praxe*. 1. vyd. Praha: Grada, 2012, 381 s. Expert (Grada). ISBN 978-80-247-3985-4.

[23] GÁLA, L., J. POUR a Z. ŠEDIVÁ. *Podniková informatika*. 2., přeprac. a aktualiz. vyd. Praha: Grada, 2009, 496 s. Expert (Grada). ISBN 978-80-247-2615-1.

[24] LUHN, H. P. A *Business Intelligence System*. *IBM Journal of Research and Development*. 1958, vol. 2, issue 4, s. 314-319. DOI: 10.1147/rd.24.0314. Dostupné z: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=5392644

[25] HARRIS, CH. The History Of Business Intelligence (Infographic). In: *BCW: IT Leadership* [online]. 2012 [vid. 2015-04-17]. Dostupné z: http://www.businesscomputingworld.co.uk/the-history-of-business-intelligence-infographic/

[26] CHALUPOVÁ, N. a A. MOTYČKA. Situation and trends in trade-supporting information technologies. *Acta Universitatis Agriculturae et Silviculturae Mendelianae Brunensis*. 2008, vol. 56, issue 6, s. 25-36. DOI: 10.11118/actaun200856060025. Dostupné z: http://acta.mendelu.cz/56/6/0025/

[27] PETERKA, M. Seznamte se s BI. In: *DAQUAS* [online]. 2010 [vid. 2015-04-11]. Dostupné z: http://www.daquas.cz/Articles/379-seznamte-se-s-bi.aspx

[28] ANDERSON, D. ETL Input Output. In: *DB Best Technologies* [online]. 2012 [vid. 2015-04-12]. Dostupné z: http://www.dbbest.com/blog/wp-content/uploads/2012/12/ETL_input_output.jpg

[29] Dashboard screenshot. In: *Atom BI: Business Intelligence as a Service* [online]. 2014 [vid. 2015-02-25]. Dostupné z: http://www.atomsail.com/en/images/products/dashboard/atombi-dashboard-screenshot2.png

[30] Apache License, Version 2.0. THE APACHE SOFTWARE FOUNDATION. *The Apache Software Foundation* [online]. 2004 [vid. 2014-12-17]. Dostupné z: https://www.apache.org/licenses/LICENSE-2.0

[31] GNU General Public License. FREE SOFTWARE FOUNDATION. *The GNU Operating System and the free Software Movement* [online]. 2007 [vid. 2014-12-17]. Dostupné z: https://www.gnu.org/copyleft/gpl.html

[32] Eclipse Public License - v 1.0. THE ECLIPSE FOUNDATION. *Eclipse: The Eclipse Foundation open source community website* [online]. ROK [vid. 2014-12-18]. Dostupné z: https://www.eclipse.org/legal/epl-v10.html

[33] ROLDÁN, María C. Pentaho Data Integration (Kettle) Tutorial. In: *Pentaho Community wiki* [online]. 2011 [vid. 2015-01-03]. Dostupné z: http://wiki.pentaho.com/display/EAI/Pentaho+Data+Integration+%28Kettle%29+Tutorial

[34] JasperReports Library: Open Source Java Reporting Library. TIBCO SOFTWARE. *Jaspersoft Community* [online]. ROK [vid. 2015-01-09]. Dostupné z: http://community.jaspersoft.com/project/jasperreports-library

[35] Rapid Data Integration Products. CLOVERETL. *CloverETL: Rapid Data Integration* [online]. ROK [vid. 2015-03-01]. Dostupné z: http://www.cloveretl.com/products

[36] PAVELKOVÁ, K. Platy v českém IT. *CIO BUSINESS WORLD: TOP 100 ICT společností v České republice.* 2014, červen, s. 42. Dostupné z: http://businessworld.cz/top-100-download?

Bibliografie

FOTR, J, L. ŠVECOVÁ. *Manažerské rozhodování: postupy, metody a nástroje*. 2., přeprac. vyd. Praha: Ekopress, 2010, 474 s. ISBN 978-80-86929-59-0.

LABERGE, Robert. *Datové sklady: agilní metody a business intelligence*. 1. vyd. Brno: Computer Press, 2012. ISBN 978-80-251-3729-1.

THE OPEN SOURCE INITIATIVE. *Open Source Initiative* [online]. ROK [vid. 2015-02-22]. Dostupné z: http://opensource.org/

ZÁDOVÁ, Vladimíra. *Pokročilé databázové systémy* (přednáška) Liberec, Technická univerzita v Liberci, 2014/2015