

Univerzita Jana Evangelisty Purkyně
v Ústí nad Labem
Přírodovědecká fakulta



Využití open-source a komerčních nástrojů pro
vizualizaci a analýzu dat na datové platformě
Portabo

BAKALÁŘSKÁ PRÁCE

Vypracoval: Ladislav Tahal

Vedoucí práce: Ing. Roman Vaibar, Ph.D., MBA

Studijní program: Aplikovaná informatika

Studijní obor:

ÚSTÍ NAD LABEM 2025

Namísto žlutých stránek vložte digitálně podepsané zadání kvalifikační práce poskytnuté vedoucím katedry.

Zadání musí zaujímat právě dvě strany.

Zadání je nutno vložit jako PDF pomocí některého nástroje, který umožňuje editaci dokumentů (se zachováním elektronického podpisu).

V Linuxe lze například použít příkaz pdftk.

Prohlášení

Prohlašuji, že jsem tuto bakalářskou práci vypracoval samostatně a použil jen pramenů, které cituji a uvádím v přiloženém seznamu literatury.

Byl jsem seznámen s tím, že se na moji práci vztahují práva a povinnosti vyplývající ze zákona č. 121/2000 Sb., ve znění zákona č. 81/2005 Sb., autorský zákon, zejména se skutečností, že Univerzita Jana Evangelisty Purkyně v Ústí nad Labem má právo na uzavření licenční smlouvy o užití této práce jako školního díla podle § 60 odst. 1 autorského zákona, a s tím, že pokud dojde k užití této práce mnou nebo bude poskytnuta licence o užití jinému subjektu, je Univerzita Jana Evangelisty Purkyně v Ústí nad Labem oprávněna ode mne požadovat přiměřený příspěvek na úhradu nákladu, které na vytvoření díla vynaložila, a to podle okolností až do jejich skutečné výše.

V Ústí nad Labem dne 21. října 2025

Podpis:

Děkuji vedoucímu bakalářské práce Ing. Romanu Vaibarovi, Ph.D., MBA za jeho odborné vedení a praktické podněty, které výrazně přispěly k analýze dat, hledání efektivních řešení a k celkovému úspěšnému dokončení bakalářské práce.

VYUŽITÍ OPEN-SOURCE A KOMERČNÍCH NÁSTROJŮ PRO VIZUALIZACI A ANALÝZU DAT NA DATOVÉ PLATFORMĚ PORTABO

Abstrakt:

Bakalářská práce se zabývá srovnáním open-source a komerčních nástrojů pro vizualizaci a analýzu dat, datové sklady a OLAP technologie s cílem zhodnotit jejich vhodnost pro využití v datové platformě Portabo. V praktické části byla realizována implementace několika variant datových řešení, zahrnujících kombinace databázových systémů MSSQL, MariaDB, ClickHouse a PostgreSQL s nástroji Power BI, Superset a Cube.js. Bylo provedeno testování výkonu, analýza technických požadavků, nároků na uživatelské dovednosti a srovnání ekonomických aspektů provozu. Výsledkem práce je komplexní přehled výhod a nevýhod jednotlivých přístupů, měření jejich efektivity při zpracování velkých objemů dat a doporučení optimálního řešení pro organizace usilující o efektivní datovou analytiku bez nutnosti vysoké IT expertizy.

Klíčová slova: Business Intelligence, datové sklady, olap, vizualizace dat, Portabo

UTILIZATION OF OPEN-SOURCE AND COMMERCIAL TOOLS FOR DATA VISUALIZATION AND ANALYSIS ON THE PORTABO DATA PLATFORM

Abstract:

This bachelor's thesis focuses on the comparison of open-source and commercial tools for data visualization and analysis, data warehouses, and OLAP technologies, with the aim of evaluating their applicability within the Portabo data platform. In the practical part, several data architecture variants were implemented, combining database systems such as MSSQL, MariaDB, ClickHouse, and PostgreSQL with visualization tools including Power BI, Superset, and Cube.js. The analysis covers performance testing, technical requirements, user skill demands, and economic aspects of operation. The outcome of the thesis is a comprehensive overview of the advantages and limitations of both open-source and commercial solutions, performance evaluation on large data volumes, and recommendations for organizations seeking efficient data analytics without requiring extensive IT expertise.

Keywords: Business Intelligence, data warehouses, OLAP, data visualization, Portabo

Obsah

Úvod	13
1. Přehled současného stavu problematiky	15
1.1. Rešerše v oblasti datových skladů, OLAP technologií a nástrojů pro vizualizaci dat	15
1.2. OLAP technologie	18
1.3. Nástroje pro vizualizaci a analýzu dat	19
1.4. Shrnutí kapitoly	19
1.5. Přehled současných open-source a komerčních řešení	19
1.6. analýza trendů v implementaci	20
2. Teoretická část	21
2.1. Úvod do problematiky datových skladů a OLAP	21
2.2. Nástroje Business Intelligence	21
2.3. Přehled a charakteristika využitých technologií	21
3. Praktická část	23
3.1. Návrh metodiky porovnání	23
3.2. Příprava dat a návrh datového skladu	23
3.3. Provedení srovnávací analýzy	23
3.4. Zhodnocení výsledků	23
4. Sazba ukázek kódu	25
5. Citace	27
5.1. Označování citací	28
5.2. Bibliografický záznam	29
5.3. Často kladené otázky	33
6. Zhodnocení	37
7. Závěr	39
A. Externí přílohy	41
B. Další přílohy	43

Úvod

V současné době organizace generují a shromažďují obrovské množství dat, která se stávají klíčovým zdrojem pro strategické i operativní rozhodování. Společnosti napříč odvětvími se proto zaměřují na to, jak tato data efektivně zpracovávat, ukládat, analyzovat a vizualizovat tak, aby přinášela skutečnou informační hodnotu i uživatelům bez hlubokého technického zázemí. S rozvojem datových technologií vzniká široké spektrum nástrojů, které umožňují tvorbu reportů, analytických modelů a vizualizací nad velkými objemy dat. Tyto nástroje mohou být jak komerční, nabízející komplexní podporu a integrované služby, tak open-source, které poskytují flexibilitu, otevřenosť a nižší náklady na implementaci.

Zároveň však s tímto rozvojem vyvstává otázka, jaké řešení je pro konkrétní organizaci nejvhodnější – z pohledu technického, uživatelského i ekonomického. Volba správného nástroje či architektury datové platformy zásadně ovlivňuje nejen efektivitu zpracování dat, ale také dostupnost a interpretaci výsledků pro koncové uživatele.

Tato bakalářská práce se zabývá využitím open-source a komerčních nástrojů pro vizualizaci a analýzu dat na datové platformě Portabo. Cílem práce je provést srovnávací analýzu vybraných řešení z hlediska technických parametrů, uživatelských požadavků a ekonomických aspektů jejich provozu. Praktická část je zaměřena na implementaci několika variant datové architektury – zahrnujících kombinace databázových systémů MSSQL, MariaDB, ClickHouse a PostgreSQL s nástroji Power BI, Superset a Cube.js. Součástí analýzy je také měření výkonu při práci s rozsáhlým datovým souborem, testování zátěže a vyhodnocení celkových nákladů na provoz (TCO).

První kapitola práce shrnuje současný stav problematiky a základní pojmy z oblasti datových skladů, OLAP a vizualizačních nástrojů. Následující teoretická část rozebírá principy a architekturu zvolených technologií. Praktická část popisuje návrh metodiky, implementaci jednotlivých řešení a provedení testování. Závěrečná kapitola obsahuje shrnutí zjištěných výsledků, jejich interpretaci a doporučení vhodného řešení pro organizace využívající datovou platformu Portabo.

1. Přehled současného stavu problematiky

1.1. Rešerše v oblasti datových skladů, OLAP technologií a nástrojů pro vizualizaci dat

Tato kapitola shrnuje teoretické poznatky o problematice datových skladů, technologií OLAP a moderních nástrojů pro vizualizaci a analýzu dat. Cílem rešeršní části je vytvořit přehled současných trendů, architektur a dostupných open-source i komerčních řešení.

Datové jezero

Datové jezero (*Data Lake*) představuje logický koncept centralizovaného úložiště, které uchovává obrovské množství *surových dat* (*raw data*) v jejich nativním formátu (např. JSON, XML, binární soubory) a to bez nutnosti předchozí definice schématu (princip **schema-on-read**).

Ačkoliv je Datové jezero často spojováno s cloudovými službami (např. AWS S3, Azure Data Lake Storage), tento koncept lze efektivně implementovat i v lokálním (on-premise) prostředí. V našem případě je Datové jezero implementováno právě na relační databázi. Přestože relační databáze vyžaduje definici schématu tabulky, je princip *schema-on-read* zachován. Realizuje se to tak, že veškerá variabilní data jsou ukládána v jednom sloupci (*např. payload*), který je definován jako JSONB nebo jiný typ pro nestrukturovaná data. Tím se fyzicky vytvoří pevné schéma, ale **datový typ jednotlivých vnitřních prvků** se určuje dynamicky **až v rámci ETL procesu** při jejich parsování a vkládání do Datového skladu.

Klíčové cíle a role Datového jezera:

1. **Konsolidace datových toků** (*Landing Zone*): Primární funkcí je převést data z mnoha heterogenních zdrojů na jednu platformu, sloužící jako **vstupní zóna** pro surová data.
2. **Oddělení integrační logiky**: Umožňuje **oddělit logiku připojení** k datovým zdrojům od Datového skladu. Tím se zajišťuje, že náročné transformační procesy (ETL/ELT) v DWH nejsou bezprostředně zatíženy problémy s konektivitou nebo dynamickou strukturou.
3. **Audit (Data Lineage)**: Uchovávání dat v jejich **původním (surovém) formátu** umožňuje kdykoliv ověřit, z jakých zdrojových dat byly odvozeny data v Datovém skladu. To je

nezbytné pro **auditní účely** a pro **reprodukci analytických výsledků** při změnách transformačních pravidel.

Datový sklad

Datový sklad (*Data Warehouse*, DWH) je centrální, časově závislé úložiště historických i aktuálních dat. Jeho primárním účelem je podpora rozhodovacích procesů. Na rozdíl od provozních databází (OLTP) je struktura DWH optimalizována pro rychlé čtení, agregace a dotazování na velkých objemech dat.

Datové tržiště (*Data Mart*) je v korporátním prostředí definováno jako **podsložka datového skladu** zaměřená na data a metriky potřebné pro specifickou obchodní oblast (např. výroba, kvalita, prodej). Jedná se o menší, tématicky specializovanou entitu, která usnadňuje reportování a analýzu pro konkrétní skupinu uživatelů.

Konceptuální návrh DWH se opírá o dvě hlavní, avšak protichůdné, metodiky:

1. **Metodika Billa Inmona (Top-Down Approach):** Inmonova metodika je označována jako **Top-Down (shora dolů)**, protože začíná návrhem centrálního, podnikového datového skladu (*Enterprise Data Warehouse*, EDW).
 - **Schema:** EDW je modelováno ve vysoce **normalizované** formě (typicky 3. normální forma, 3NF), což zajišťuje nízkou datovou redundanci a maximální integritu.
 - **Tok dat a tržiště:** Data jsou nejprve ETL procesem načtena do detailního, normalizovaného EDW (centrální zdroj pravdy). **Datová tržiště** se vytváří **až sekundárně** z dat v EDW a jsou denormalizovaná, aby sloužila pro rychlé reportování.
2. **Metodika Ralha Kimballa (Bottom-Up Approach):** Kimballova metodika je označována jako **Bottom-Up (zdola nahoru)**, protože se zaměřuje na rychlé dodání řešení pro specifické obchodní procesy.
 - **Schema:** Využívá **dimenzionální modelování** (schéma Hvězda nebo Sněhová vločka), které je záměrně **denormalizované**. To zjednoduší dotazování a maximalizuje výkon pro OLAP úlohy.
 - **Tok dat a tržiště:** Data jsou transformována a ukládána **přímo** do dimenzionálních modelů, které **představují Datová tržiště**. Podnikový datový sklad je pak **logickou unií** (sjednocením) těchto jednotlivých tržišť.

V praxi se však často objevují hybridní systémy kombinující prvky obou přístupů. Takovýto hybridní přístup jsem zvolil i já.

Datové modely

Základními modely používanými pro návrh datového skladu v rámci dimenzionálního modelování (Kimball) jsou:

- **Schéma Hvězda (Star Schema):** Jedná se o nejjednodušší a nejčastěji používaný dimenzionální model. Skládá se z centrální tabulky **Faktů (Fact Table)**, která je obklopena několika tabulkami **Dimenzií (Dimension Tables)**. Všechny dimenze jsou přímo napojeny na tabulku faktů, čímž vzniká struktura připomínající hvězdu. Vysoká redundance je vyvážena extrémní rychlostí dotazování.
- **Schéma Sněhová vločka (Snowflake Schema):** Jedná se o rozšíření schématu Hvězda, kde některé dimenze jsou **normalizovány** do několika souvisejících tabulek. Tím se snižuje redundance dat, ale na úkor zvýšení složitosti dotazování (je potřeba více JOIN operací), což může mírně zhoršit výkon.

Datové vrstvy v implementaci DWH

DWH z pravidla obsahuje:

- **Staging Area:** Slouží jako dočasné úložiště, kam jsou data přenesena pomocí ETL (Extract, transform and load). V této vrstvě se provádí **harmonizace a validace dat** před aplikací dimenzionálního modelu. Příkladem je tabulka Stg.CameraCamea ve Vaší implementaci (viz bilina_kamery_lake_to_staging.py).
- **DWH Vrstva (Dimenze a Fakta):** Hlavní vrstva organizovaná dle Kimballova modelu (tedy Datový Mart pro tuto analytickou doménu). Skládá se z:
 1. **Dimenze (Dimensions):** Uchovávají kontext, atributy a popisné detaily (např. DimCity, DimSensor).
 2. **Fakta (Facts):** Uchovávají měřitelné hodnoty a cizí klíče k dimenzím (např. FactCameraDetection viz bilina_kamery_staging_to_fact.py).
- **Sémantická vrstva (Semantic Layer):** Tato vrstva nahrazuje tradiční prezentační vrstvu a slouží jako **jednotná definice obchodní logiky** a metrik nad daty z DWH. Je reprezentována nástrojem **Cube.js** (viz docker-compose3.yml), který zajišťuje předagregaci, cachování a standardizované výpočty. Koncovým BI nástrojům (jako je Apache Superset) pak zpřístupňuje data prostřednictvím **SQL API** nebo **GraphQL**, čímž se zjednodušuje dotazování a zvyšuje výkon.

Datové vrstvy v implementaci DWH

Základní vrstvy DWH, které odrážejí tok dat ve Vašich ETL skriptech, jsou:

- **Vrstva získávání/Staging (Staging Area):** Slouží jako dočasné úložiště, kam jsou data přenesena po jejich transformaci a čištění z Datového jezera. V této vrstvě se provádí **harmonizace a validace dat** před aplikací dimenzionálního modelu. Příkladem je tabulka `Stg.CameraCamea` ve Vaší implementaci (viz `bilina_kamery_lake_to_staging.py`).
- **DWH Vrstva (Dimenze a Fakta):** Hlavní vrstva organizovaná dle Kimballova modelu. Skládá se z:
 1. **Dimenze (Dimensions):** Uchovávají kontext, atributy a popisné detaily (např. `DimCity`, `DimSensor`).
 2. **Fakta (Facts):** Uchovávají měřitelné hodnoty a cizí klíče k dimenzím (např. `FactCameraDetection`).
- **Sémantická vrstva (Semantic Layer):** Tato vrstva nahrazuje tradiční prezentační vrstvu a slouží jako **jednotná definice obchodní logiky** a metrik nad daty z DWH. Je reprezentována nástrojem **Cube.js** (viz `docker-compose3.yml`), který zajišťuje předagregaci, cachování a standardizované výpočty. Koncovým BI nástrojům (jako je Apache Superset) pak zpřístupňuje data prostřednictvím **SQL API** nebo **GraphQL**, čímž se zjednoduší dotazování a zvyšuje výkon.

Použité technologie pro DWH

Mezi nejpoužívanější technologie pro datové sklady patří relační databáze jako **PostgreSQL** s rozšířením **TimescaleDB** (využívané pro ukládání časových řad, jak je patrné z `pg-warehouse`) nebo **Microsoft SQL Server** (využívaný ve Vašich ETL skriptech). Pro extrémně rychlé analytické dotazy se používají sloupcově orientované databáze, jako je **ClickHouse** (viditelné v `docker-compose2.yml`).

1.2. OLAP technologie

OLAP (*Online Analytical Processing*) představuje způsob zpracování dat, který umožňuje rychlé multidimenzionální analýzy. Uživatelé mohou data zkoumat podle různých dimenzi – například času, lokality nebo typu senzoru.

Rozlišují se tři základní typy OLAP řešení:

- **MOLAP (Multidimensional OLAP)** – využívá vlastní vícerozměrnou datovou strukturu, která poskytuje vysoký výkon při agregacích, avšak vyžaduje větší prostor a přípravu dat.
- **ROLAP (Relational OLAP)** – staví přímo na relační databázi, je flexibilní a snadněji škálovatelný.
- **HOLAP (Hybrid OLAP)** – kombinuje výhody obou přístupů, často používaný v moderních datových skladech.

Pro implementaci OLAP vrstev existují nástroje jako Apache Kylin, Mondrian (součást Pentaho), Microsoft Analysis Services nebo ClickHouse s podporou agregací v reálném čase. V kontextu platformy Portabo jsou tyto technologie využívány pro tvorbu tabulek typu *Fact* a *Dimension*, které slouží pro následnou analýzu provozních dat z kamerových systémů.

1.3. Nástroje pro vizualizaci a analýzu dat

Vizualizace dat představuje klíčovou část procesu Business Intelligence (BI). Umožňuje uživatelům rychle pochopit trendy, vztahy a odchylky v datech.

K nejrozšířenějším nástrojům patří:

- **Open-source řešení:** Grafana, Apache Superset, Metabase, Redash – nabízejí integraci s většinou databázových systémů a jsou vhodné pro interaktivní dashboardy.
- **Komerční řešení:** Microsoft Power BI, Tableau, Qlik Sense – poskytují rozšířené funkce pro datovou transformaci, prediktivní analýzy a sdílení výstupů v rámci organizace.

Důležitými kritérii při výběru vizualizačního nástroje jsou:

1. Uživatelská přívětivost a intuitivní rozhraní,
2. Podpora přímého připojení na datový sklad nebo OLAP vrstvu,
3. Možnost automatizace a plánování aktualizací dat,
4. Licenční model a celkové náklady na provoz (TCO).

1.4. Shrnutí kapitoly

Rešeršní část prokázala, že oblast datových skladů a vizualizačních nástrojů je velmi dynamická. V současnosti dochází ke sbližování analytických databází (např. ClickHouse, DuckDB) s vizualizačními platformami, což umožňuje rychlejší zpracování velkých objemů dat bez nutnosti rozsáhlých ETL procesů.

Tato kapitola poskytuje teoretický základ pro následnou praktickou část práce, kde budou jednotlivé přístupy implementovány a porovnány přímo na datech z platformy Portabo.

1.5. Přehled současných open-source a komerčních řešení

\LaTeX podporuje čtyři úrovně vnoření kapitol (kromě kapitol i sekce, podsekce, a podpodsekce), což by mělo být více než dostatečné. Druhou úrovní jsou sekce, jejichž rozsah by měl být alespoň jednu stranu, mohou však být i výrazně delší. Zobrazují se v obsahu a jsou číslované a lze je tudíž snadno odkazovat. Měli by být použity b každé kapitole kromě úvodu a závěru.

1.6. analýza trendů v implementaci

Optimální rozsah bakalářské práce není možné stanovit, neboť závisí na jejím charakteru. Práce jejíž součástí je vytvoření aplikace mohou být menšího rozsahu než práce čistě popisné. I v tomto případě by však měla mít alespoň třicet stran (nepočítaje úvodní stránky až po obsah a stránky od seznamu použitých zdrojů dále a prázdné stránky tzv. vakáty).

Maximální rozsah práce je ještě obtížnější stanovit. Obecně však doporučuji v případě, že pokud rozsah (opět bez povinných částí) přesáhne osmdesát stran je vhodné uvažovat o redukci (ne všechno, co jste napsali musí být nakonec uvedeno ve finální verzi) respektive přesunu do externích příloh umístěných na Githubu (viz příloha A na straně 41). Týká se to především různých schémat, obrazového materiálu apod.

2. Teoretická část

Při tvorbě bakalářské práce byste měli dodržovat základní zásady typografie. Bakalářská práce by měla být přehledná a dobře čitelná. A možná i krásná (uvědomuji si, že krása je subjektivní).

2.1. Úvod do problematiky datových skladů a OLAP

Na naší katedře není předepsáno, jaké písmo by mělo být použito při sazbě práce. Měli byste však dodržovat několik zásad.

- pro sazbu většiny textu by mělo být použito tzv. knižní písmo tj. písmo určené pro sazbu soudobých knih (především odborných). Doporučit lze antikvová písma a resp. písma lineární (bezpatková).
- pro výpisy kódu používejte neproporcionální písmo (pro ostatní text naopak proporcionalní). Pro základní přehled volně šiřitelných písmen se můžete podívat do dokumentu Monotypář.
- pokud používáte více druhů písma, pak by měla být vizuálně kompatibilní (nejlepší je použít písma, která jako kompatibilní vytvořili či alepoň doporučili typografové)

V šabloně je využita rodina písma *Libertinus* (odvozená z písma *Linux Libertine*) s výjimkou neproporcionálního písmo, pro něž bylo zvoleno písmo *Source Code Pro*, které podporuje výrazně větší počet řezů (kurziva tučné).

Standardní velikost písma kvalifikační práce je 12 bodů (přesněji řečeno tzv. pica bodů tj. 12/72 palce). Tato velikost však ve většině případů neodpovídá žádnému viditelnému rozdílu písma (v klasické typografii je to výšky kuželky, na níž je umístěn vystouplý reliéf písma). Z tohoto důvodu se skutečná výška a především šířka písmen může u jednotlivých písmen viditelně lišit. To je problém, pokud se znaky různých písem vyskytují vedle sebe. Často se výrazněji liší například výška antikvy a neproporcionálního písma. V tomto případě je často nutné velikost neproporcionálních písem mírně zmenšit (tato šablona tak v případě využití XeTeXu nebo LuaTeXu činí automaticky).

Optimální řádkování pro bakalářské práce je 1½ (použité i v této šabloně).

2.2. Nástroje Business Intelligence

2.3. Přehled a charakteristika využitých technologií

3. Praktická část

3.1. Návrh metodiky porovnání

3.2. Příprava dat a návrh datového skladu

3.3. Provedení srovnávací analýzy

3.4. Zhodnocení výsledků

4. Sazba ukázek kódu

Sazba ukázek kódu je klíčová pro bakalářské práce věnované implementaci nějaké aplikace či knihovny.

Základní zásady pro sazbu ukázek kódu:

1.

5. Citace

Citace tvoří jeden ze základních pilířů závěrečné práce. Platí zde základní pravidlo: pokud použijete jakoukoliv zdroj informací, pak je nutné tento zdroj citovat, tj. uvést příslušný zdroj.

Zdrojem je ve většině případů text, ale může to být i obrázek, audiovizuální materiál či ve speciálních případech i ústní sdělení. V případě informatických prací je častým zdrojem u zdrojový kód.

Informaci ze zdroje můžete použít dvěma různými způsoby:

- přímo převzít (u textů je to známé Ctrl-C, Ctrl-V)
- použít jako základ vlastního intelektuálního výtvoru (textu, grafiky, programu, apod.), tj. použijete jen informaci, ale její formu změníte.

Prví druh tzv. přímé citace by měli mít v informatických závěrečných prací jen velmi omezený rozsah (méně než stránka), neboť jejich přínos pro hodnocení práce je diskutabilní. Přesto jsou však případy, kdy jsou vhodné:

1. matematické definice a tvrzení (věty, axiomy)
2. definice termínů z neinformatických oborů (např. společenských věd)
3. citace norem resp. standardů

Citace mají tři základní cíle:

1. určují, co je váš vlastní intelektuální přínos a co jste pouze převzali
2. pomáhají určovat primárního autora (resp. autory)
3. definují kontext vaší práce resp. mohou usnadňovat nalezení dalších souvisejících informací

Jakákoliv vědecká práce nevzniká na zelené louce a tak jsou citace její nezbytnou součástí. V rámci práce běžně navazujeme na existující výzkumy, projekty, technologie apod. Stejně tak se můžeme odkazovat na autority či s nimi polemizovat.

První dva cíle také úzce souvisejí s plagiátorstvím. Pokud v práci použijete myšlenku či údaj bez citování je vaše práce plagiátem. I když se to v obecném mínění vztahuje jen na přímé kopírování, není tomu tak. Přímé kopírování se jen snadněji vyhledává a prokazuje. Je také obtížnější jej kvantifikovat.

V případě přímého kopírování, jež není označeno jako přímá citace, postačuje i relativně malý rozsah (například věta, jeden obrázek, jedna procedura), aby byla práce označena jako plagiát.

Plagiát není možno obhájit a v případě většího rozsahu hrozí i vyloučení ze studia. Za přímé kopírování se považují i případy, kde je změna jen formální (změna slovosledu, náhrada synonym, zkrácení, vložení textové výplně, změna barvy či afinní transformace obrázku).

V případě převzetí myšlenky jde o zjevné plagiování, pokud je tato myšlenka důležitou částí práce (podílí se na splnění cílů).

To, že není plagiátorství odhaleno před obhajobou práce, není důkazem, že se nejedná o plagiát. Pokud je plagiátorství zjištěno později, může vám být odebrán titul i zpětně (a jak jste si jistě všimli, plagiátorství je běžně využíváno v politickém boji).

V každém případě si uvědomte, že plagiátorství je druh krádeže a že ani vy nechcete, aby někdo vaše myšlenky nebo dokonce váš text vydával za vlastní.

5.1. Označování citací

Označování citace má dvě části. Za prvé je nutno označit, jaká část práce je citací (rozsah) a jaký je původní zdroj.

Zdroj je vždy určen odkazem na bibliografický záznam, které jsou v případě bakalářské práce uvedeny v kapitole *Použité zdroje* na konci práce. Odkaz může mít různý tvar, ale preferovaný styl je uvedení čísla záznamu v hranatých závorkách. V případě použití našeho latexovského stylu stačí použít příkaz `\cite{id-zaznamu}`. V případě potřeby lze zdroj zpřesnit uvedením např. stránky či kapitoly, jež se uvádí za číslem záznamu (po čárce a ještě před uzavírající hranatou závorkou). V \LaTeX u lze využít nepovinný parametr příkazu `cite`.

Označení rozsahu se poněkud liší u přímých a nepřímých citací.

U přímých citací je označení rozsahu kritické. V případě citací, které jsou kratší než odstavec je nutné text vyznačit kurzívou a zahrnout do uvozovek. Odkaz musí následovat hned za označným textem.

U citací v rozsahu odstavce či více odstavců se využívá zvětšení okrajů na levé i pravé straně odstavců (viditelné na první pohled). V \LaTeX u lze použít prostředí `quote` nebo `quotation`. Text by měl být navíc v uvozovkách. Kurzíva je možná, ale u rozsáhlejších citací není příliš vhodná. Odkaz se umisťuje na konec posledního převzatého odstavce.

Speciální případ je možný v případě matematických definic a vět. Pokud jsou v rámci kapitoly převzaty jen z jediného zdroje, lze na počátku kapitoly uvést hromadný odkaz například v podobě věty: Všechny definice a věty uvedené v této kapitole jsou převzaty z [X].

V případě převzatých obrázků se odkaz umisťuje na konec popisku. Aby však bylo zřejmé, že se jedná o přímé převzetí (kurzívu ani uvozovky nelze použít) je nutné explicitně vyjádřit, že obrázek byl převzat ze zdroje bez podstatných změn například: (převzato z [X]) nebo (překresleno z [X]).

U nepřímých citací je vyznačování rozsahu volnější. Nejjednodušší je uvádění holé citace na konci vět (před tečkou) nebo konci odstavce (za poslední tečkou). V mnoha případech je ale možné citace uvádět explicitněji a stylistiky je provázen s okolním textem.

příklady:

- zajímavá alternativa je popsána v [x]
- údaj je je převzat z [x]
- použití návrhového vzoru poprvé popsal N.N v [x]
- volně přeloženo z [x]
- řešení bylo navrženo uživatelem N v [x] (vhodné např. pro stackoverflow a podobné zdroje)

Explicitnější vyjádření je nutno použít i v případě, že rozsah citace přesahuje odstavec.

- následující příklad je převzat z [x]
- výčet vychází z [x] je však doplněn o ...

Teoreticky lze podobné řešení využít i u celých sekcí či kapitol (kapitola je zpracována na základě [x]). V tomto případě je však nutné předpokládat, že v dané kapitole není žádná autorská myšlenka, a že autor se snažil najít alternativní pohledy či zdroje (a hodnotit tak, lze pouze autorovu schopnost výběru informací či stylistiky).

Výjimečně lze uvádět i několik citací se shodným či překrývajícím se rozsahem např. *následující specifikace je převzata z [x] a [y]*. To je však tolerovatelné jen v případě, v kdy by oddělení oddělení zdrojů bylo obtížné nebo nepřehledné a spojení nepřináší problémy s intelektuálním vlastnictvím (mají stejného autora či copyright). Zcela nepoužitelné jsou v případě většího rozsahu citace (např. na úrovni sekcí či kapitol)!

V případě obrázků je vhodné uvést explicitnější specifikaci, jak byl originální obrázek pozměněn resp. rozšířen.

Příklad:

- (převzato z [x] a doplněno)
- (převzato z [x], přeloženo)
- (upraveno z [x] pro novou verzi technologie ...)
- (inspirováno diagramem [x])
- (viz také [x] pro data X)

5.2. Bibliografický záznam

Bibliografický záznam je datová struktura, jenž má dvě základní funkce:

1. jednoznačné identifikování zdroje
2. určení primární odpovědnosti (typicky je to autor resp. autoři, u webových zdrojů to však často bývá korporace).

Pro každý typ zdrojového dokumentu (zdroje) existuje množina klíčových atributů, které by měly být specifikovány (ne zcela vhodně označované jako povinné) a další, které hrají jen pomocnou roli.

V praxi však může nastat situace, kdy není zřejmé, jaký typ dokumentu pro daný zdroj zvolit resp. nelze zjistit hodnoty klíčových atributů. V tomto případě je nutné improvizovat a snažit se, aby záznam plnil v maximální míře obě funkce.

Struktura bibliografického záznamu je v zásadě dána těmito dimenzemi:

médium – základní dělení je na tištěné dokumenty a online dokumenty (dokumenty na elektro-nických nosičích tvoří jakási přechod mezi oběma typy dokumentů)

samostatnost – zdroj může být samostatný nebo součást rozsáhlejšího zdroje

periodičnost – periodický dokument vychází po jednotlivých částech, přičemž počet částí není předem znám (např. časopis).

Tištěné samostatné dokumenty neperiodické

Typickým příkladem samostatného tištěného dokumentu je kniha či monografie.

Základním zdrojem informací pro bibliografický záznam u knih je tzv. tiráž, tj. soupis vydavatelských údajů uvedený na konci knihy či na stránce za titulem. Využít lze i další zdroje (např. katalogy knihoven či knižní e-shopy, bibliografické záznamy v jiných dokumentech), ale v tomto případě je nutné provádět kontrolu, neboť tyto sekundární zdroje často obsahují chyby.

klíčové atributy:

ISBN : ISBN je celosvětový jedinečný identifikátor neperiodických tištěných dokumentů. Pokud ho kniha má, pak je dokument jednoznačně identifikován (a další identifikace už hraje jen sekundární roli). Pomlčky v ISBN nejsou součástí identifikátoru a lze je vynechávat (i když občas se jedno ISBN přiděluje více svazkům). Navíc existují ve dvou podobách ISBN-10 s deseti číslicemi a ISBN-13 s třinácti. Pokud jsou k dispozici oba je vhodnější uvítat ISBN-13 (i když ISBN-10 lze snadno mapovat na ISBN-13).

název : název knihy je povinný údaj a měl by být vždy vyplněn. Použit by měl být vždy originální název bez úprav. Jedinou přípustnou úpravou je změna velkých písmen (verzálek) na malá, které by mělo odpovídat pravidlům příslušného jazyka.

podnázev : některé knihy mají i podnázev. Někdy je těžké rozeznat, co je název a podnázev. Zde platí pravidlo, že název by neměl obsahovat dvojtečku, tečku, středník apod. Od podnázvu je potřeba odlišit název edice. Podnázev je nepovinný (doporučuji uvádět pokud obsahuje klíčové informace).

autoři : v bibliografickém záznamu by měli být uvedeni všichni primární autoři (tj. není potřeba uvádět překladatele, ilustrátory, apod.)

vydání : označení konkrétního vydání. Je důležité především tehdy, když není známo ISBN a existuje více odlišných vydání (s různým obsahem)

nakladatel : uvádí se jméno nakladatelství, a to především z důvodů odpovědnosti

místo vydání : uvádí se jméno města, popřípadě stát, především tehdy pokud není jednoznačné (např. Cambridge) a to ve stručné podobě (např. stačí *United Kingdom*). Podobně stručný by měl být název nakladatelství (tj. bez označení typy společnosti, apod., rodičovské společnosti, apod.) V dnešní době globalizace je tento údaj v mnoha případech nevýznamný (tj. ho lze vynechat, především tehdy pokud je nakladatelství neznámé).

rok vydání : rok vydání přesněji identifikuje dokument. Pokud ho nelze zjistit, lze jej nahradit rokem copyrightu (v tomto případě je uvozen znakem c např. c2022)

edice : kniha může být vydána v rámci edice. Edici doporučuji neuvádět, výjimkou jsou edice, které jsou všeobecně známé.

URL : uvádí se pouze v případě, že je kniha dostupná onlině a to oficiálně a bez poplatků. Uvedení URL v tomto případě usnadňuje její získání (v tomto případě je ale často lepší citovat ji jako elektronickou knihu).

příklad:

Následující biblografický záznam byl získán z katalogu systému knihovny UJEP (volba Citace Pro v dolní části výpis záznamu).

RASCHKA, Sebastian a Vahid MIRJALILI. *Python machine learning: machine learning and deep learning with Python, scikit-learn, and TensorFlow*. Second edition. Birmingham: Packt, 2017. Expert insight. ISBN 978-1-78712-593-3.

Tento záznam splňuje základní požadavky, neboť obsahuje údaje týkající se odpovědnosti i jednoznačnou identifikaci dokumentu (a to jak ISBN tak přesným určením vydání). Zahrnutí podnázvu je vhodné, neboť obsahuje dodatečné informace (jména frameworků). Nakladatelství je uvedeno ve stručné podobě (tj. *Packt*) je uvedeno ve stručné podobě. Nadbytečné je jen uvedení edice (*Expert insight*).

Online samostatné dokumenty neperiodické

Typickým příkladem je online PDF dokument (včetně elektronické knihy). Dalším příkladem je webové sídlo (*web site*) tj. typicky hierarchický systém více stránek (nikoliv tedy jedna konkrétní web stran).

medium : u online zdrojů se jako médium uvádí slovo **online**.

URL : klíčový údaj pro online zdroje. Některé systémy (např. Wikipedia) poskytují tj. fixní URL, které odkazují na konkrétní verzi dokumentu, resp. stránek. I když jsou tato URL obecně delší, je nutné jim dát přednost, neboť zaručují jedinečnost.

název : název nelze vynechat i když ne vždy je jasné, co je hlavním názvem. V tomto případě je možné využít obsahu elementu title v hlavičce HTML (pokud je zdroj v HTML) nebo jiná metadata (například jak je zdroj pojmenován v odkazu).

autoři : autor nebývá u mnoha online dokumentů dohledatelný (a v tomto případě je nutné ho vynechat). Rozhodně však věnujte čas zjištění autorství (může být uvedeno i mimo dokument).

odpovědná korporace : typicky je to držitel intelektuálních práv (copyrightu). Důležitý je především v případě, že není znám autor, ale uvádějte ho ve všech případech, kdy je dohledatelný. Většina bibliografických stylů tento atribut nepodporuje resp. ho běžně nezobrazuje. Proto je vhodné pro tento účel využívat atribut *nakladatel* (i když to není totéž).

verze/čas poslední aktualizace : nahrazuje rok vydání. V případě, že není použit fixní odkaz, je klíčovým zdrojem informací, jaká z verzí dokumentu byla použita jako zdroj. Online dokumenty se mění často, a tak je vhodné uvádět, co nejpřesnější specifikaci (číslo verze, čas poslední aktualizace). Jen v případě, že dokument není verzován a nelze zjistit přesnější čas poslední modifikace, lze využít vročení (stejně jako u knih může být odhadnuto z copyrightu).

datum použití : je to povinný údaj i když důležitý je jen v případě, kdy nelze určit přesnější verzi. Měl by být v ISO formátu tj. ve tvaru RRRR-MM-DD. Toto datum běžně generuje editor bibliografických citací podle data vytvoření záznamu. V každém případě by mělo ležet v časovém intervalu od poslední modifikace zdroje (je-li uvedeno) do data odevzdání závěrečné práce.

příklad:

Dílčí tištěné dokumenty

U dílčích tištěných dokumentů je typické, že kromě identifikace dílčí části obsahují i identifikaci dokumentu jako celku.

Klasickým příkladem jsou články ve sborníku nebo vědeckém časopise. Kapitoly v knize (monografii) se citují, jen případě, že každou z nich vytvořil jiný autor (či kolektiv autorů)

V zásadě platí tato pravidla:

- uvádí se jen autoři dílčí časti, nikoliv například editoři sborníku nebo časopisu
- uvádí se pozice části v celém dokumentu nejlépe pomocí rozsahu stránek
- pokud má dílčí část vlastní jednoznačný identifikátor (například DOI), není potřeba uvádět identifikátor knihy nebo periodika.

Dílčí online dokumenty

Tento typ citací se používá pro webové stránky, jež jsou součástí webového sídla například pro konkrétní stránky s dokumentací nebo dokumenty uložené na GitHubu. Pro jiné elektronické dokumenty, pokud nejsou výslovně součástí webového sídla (např. elektronického sborníku) je vhodnější použít záznam samostatného dokumentu (viz výše).

Název stránky je doplněn jménem webového sídla (to je typicky uvedeno v záhlaví každé stránky resp. na hlavní stránce webového sídla). Autoři se vztahují ke stránce zatímco korporátní odpovědnost je typicky vztažena k celému sídlu (pokud jsou známy autoři i korporátní odpovědnost je vhodné uvést oba údaje, vždy však musí být uveden alespoň jeden z těchto údajů). Všechny ostatní atributy se vztahují

příklady:

What's New In Python 3.9: Summary – Release highlights. *Python 3.9.0 documentation [online]*. Python Software Foundation, October 14, 2020 [cit. 2020-10-15]. Dostupné z: <https://docs.python.org/3/whatsnew/>

Záznam obsahuje název dílčí části a také název celého webového sídla (v kurzívě). Odpovědná organizace je uvedena na místě nakladatele (autor není uveden a tak je tato informace klíčová). Verze je určena datem poslední modifikace (je uvedeno přímo ve tvaru použitém na stránce). Datum citování je povinné, ale v tomto případě nenesе žádnou přidanou informaci (jen to, že citace byla vytvořena jen den po poslední modifikaci. Poslední součástí je URL.

Python nonlocal statement. *Stack Overflow [online]*. Stack Exchange, 2022-03 [cit. 2022-07-27]. Dostupné z: <https://stackoverflow.com/questions/1261875/python-nonlocal-statement>

Struktura záznamu je stejná. Čas poslední modifikace byl určen z informace, že poslední modifikace proběhla před čtyřmi měsíci (je uvedena v ISO formátu, ale odpovídající by byl i údaj například ve tvaru *březen 2022* nebo *March 2022*).

5.3. Často kladené otázky

Co není potřeba citovat?

Obecně platí, že citovat není potřeba znalosti, které jste získali v průběhu studia a to jak při výuce tak i z učebních materiálů (opor, skript). Citovat není potřeba ani zdroj formálních údajů (např. významu zkratek), pokud je lze snadno získat (například na Wikipedii).

To jest není nutné uvádět citaci při uvedení zkratek HTTP (zkratka je všeobecně známá a běžně využívána v mnoha kurzech). Podobně není nutné odkazovat pojmy jako Internet, počítačová síť, programovací jazyk, procesor, apod.

Běžně se také necitují (původní) myšlenky vedoucího práce, pokud si vedoucí práce nevyžádá jinak. Pokud vám zprostředkuje nepůvodní myšlenku, měl by vám pomoci najít originální zdroj (který uvedete v citaci).

Citování není možné v případě, kdy není znám původní zdroj, resp. je v podobě, kterou není možné citovat (lidová řčení, apod.) Pravděpodobnost výskytu takových textů v informatické bakalářské práci je však velmi nízká.

Jak citovat informace z (podnikových) školení

Pokud se jedná o evidentní výtvar školitele, můžete odkazovat příslušný výukový materiál (i když je neveřejný). Pokud je informace nepůvodní, pak je vhodné citovat primární resp. alespoň dostatečně autoritativní zdroj.

Jak citovat ústní sdělení?

Ústní sdělení je potřeba citovat jen tehdy, když je od autoritativní osoby v oblasti její odbornosti. Pokud například píšete práci o nasazení databáze, pak je autoritativní osobou například správce databázového systému (který vám sdělí například zkušenosti s nasazením).

Jak je uvedeno výše, ve většině případů se necitují ústní sdělení učitelů, školitelů, vedoucího práce a dalších sekundárních zdrojů.

Pokud citujete ústní sdělení je vhodné s tím danou osobu seznámit či získat alespoň neformální souhlas, neboť sdělené informace nemusí být veřejné.

Navzdory důležitosti ústních sdělení v některých typech prakticky zaměřených prací, není citace ústních sdělení standardizována. Jednoduchý návod nabízí například blog na [citace.com](#) [XXX].

Ústí sdělení je však neověřitelné a nelze ho jednoznačně identifikovat. Proto je lepší pokud se sdělení děje například e-mailem. Citace e-mailové komunikace i dalších netradičních zdrojů shrnuje dokument [XXX].

Je možno citovat Wikipedii?

Citování Wikipedie se obecně nedoporučuje, neboť se jedná o terciární zdroj (encyklopédia vytvořená na základě druhotných informací) a její kvalita je značně kolísavá.

Na druhou stranu Wikipedia (především v anglické verzi) často obsahuje i hodnotný a jinak jen obtížně dostupný materiál, a tak nelze citování z Wikipedie striktně zakázat.

Základní doporučení pro citování z Wikipedie:

- citujte jen tehdy, pokud nemáte k dispozici primární zdroje (ty jsou často odkazovány přímo z Wikipedie)
- citujte jen kvalitní články (které nejsou označeny jako problematické), které se v oblasti informatiky a matematiky objevují spíše na anglické Wikipedii
- citace z Wikipedie by měli tvořit jen malou část zdrojů (typicky méně než 10

Z Wikipedie rozhodně necitujte články věnované běžně známým technologiím a poznatkům, které jsou běžnou součástí kurzů.

6. Zhodnocení

7. Závěr

Závěr je klíčovou kapitolou, která může nejvíce ovlivnit vaši obhajobu. Základní částí závěru je přehledné shrnutí výstupů práce tj. co jste udělali pro dosažení cílů práce. Je nutné se vyhnout hodnocení, zda tím byli splněny cíle práce, či nikoliv (to je úkol posudků a především komise).

A. Externí přílohy

Externí přílohy této bakalářské práce jsou umístěny na adrese:

https://github.com/Jiri-Fiser/thesis_ki_ujep.

Na úložišti GitHub mohou být uloženy tyto externí přílohy:

- **zdrojové kódy**
- **doplňkové texty** (například jak instalovat aplikaci, manuály aplikace)
- **schémata** (především, pokud se nevejdou na stranu A4 a jejich vytisknutí je tak problematické)
- **screenshoty** (v textu práce lze použít jen omezený počet snímků obrazovky, které navíc nemusí být při černobílém tisku příliš přehledné)
- **videa** (například ovládání aplikace)

V každém případě by to však měli být pouze materiály, které jste vytvořili sami. Materiály jiných autorů uvádějte v seznamu použité literatury (včetně případných odkazů na jejich originální umístění).

V této kapitole stačí uvést pouze základní strukturu úložiště (co se kde nalézá a jakou má funkci) například v podobě tabulky.

ki-thesis.pdf	text práce v PDF
ki-thesis.tex	zdrojový kód práce v L ^A T _E Xu
kitheses.cls	definice třídy dokumentů (rozšířená třída scrbook)
thesis.bib	bibliografická databáze (exportována z citace.com)
LOGO_PRF_CZ_RGB_standard.jpg	logo fakulty s českým textem
LOGO_PRF_EM_RGB_standard.jpg	logo fakulty s anglickým textem

Všechny tyto soubory jsou potřeba pro překlad dokumentu (logo stačí jedno v příslušné jazykové verzi).

B. Další přílohy

Výjimečně může práce obsahovat i další tištěné přílohy. Obecně však dávejte přednost elektronickým přílohám umístěným na GitHubu (tato kapitola tak bude úplně chybět).