FAKULTA MECHATRONIKY,
INFORMATIKY A MEZIOBOROVÝCH
STUDIÍ TUL

Bakalářská práce

Zpracování velkých datových souborů

Studijní program: B0714A270001 Mechatronika
Autor práce: **Patrik Drbohlav**
Vedoucí práce: Ing. Věra Pelantová, Ph.D.
Ústav mechatroniky a technické informatiky

Liberec 2024

FAKULTA MECHATRONIKY,
INFORMATIKY A MEZIOBOROVÝCH
STUDIÍ TUL

Zadání bakalářské práce

Zpracování velkých datových souborů

Jméno a příjmení:	**Patrik Drbohlav**
Osobní číslo:	M21000019
Studijní program:	B0714A270001 Mechatronika
Zadávající katedra:	Ústav mechatroniky a technické informatiky
Akademický rok:	2023/2024

Zásady pro vypracování:

1. Proveďte úvod do problematiky velkých datových souborů v organizaci.
2. Proveďte průzkum stavu problematiky velkých datových souborů v organizaci.
3. Navrhněte softwarové řešení filtrovacího systému pro velké datové soubory v organizaci.
4. Navrhněte efektivní řešení problémů kolem velkých datových souborů v organizaci.

Rozsah grafických prací:	dle potřeby dokumentace
Rozsah pracovní zprávy:	30 až 40 stran
Forma zpracování práce:	tištěná/elektronická
Jazyk práce:	čeština

Seznam odborné literatury:

[1] FAJT, J. *Rozšíření nástroje pro analýzu zakázek malého výrobního podniku.* [Diplomová práce.] Praha: ČVUT, FS, UREP, 2023.

[2] KALABIS, D. *Digitální komunikace napříč dodavatelsko-odběratelským řetězcem.* [Bakalářská práce.] Praha: ČVUT, MÚVS, 2023.

[3] BHARADIYA, J. P. *A Comparative Study of Business Intelligence and Artificial Intelligence with Big Data Analytics.* In: *American Journal of Artificial Intelligence.* Vol. 7, No. 1, 2023, pp. 24-30. DOI: 10.11648/j.ajai.20230701.14

Vedoucí práce:	Ing. Věra Pelantová, Ph.D. Ústav mechatroniky a technické informatiky
Datum zadání práce:	12. října 2023
Předpokládaný termín odevzdání:	14. května 2024

L.S.

prof. Ing. Zdeněk Plíva, Ph.D.
děkan

doc. Ing. Josef Černohorský, Ph.D.
garant studijního programu

V Liberci dne 12. října 2023

Prohlášení

Prohlašuji, že svou bakalářskou práci jsem vypracoval samostatně jako původní dílo s použitím uvedené literatury a na základě konzultací s vedoucím mé bakalářské práce a konzultantem.

Jsem si vědom toho, že na mou bakalářskou práci se plně vztahuje zákon č. 121/2000 Sb., o právu autorském, zejména § 60 – školní dílo.

Beru na vědomí, že Technická univerzita v Liberci nezasahuje do mých autorských práv užitím mé bakalářské práce pro vnitřní potřebu Technické univerzity v Liberci.

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

Současně čestně prohlašuji, že text elektronické podoby práce vložený do IS/STAG se shoduje s textem tištěné podoby práce.

Beru na vědomí, že má bakalářská práce bude zveřejněna Technickou univerzitou v Liberci v souladu s § 47b zákona č. 111/1998 Sb., o vysokých školách a o změně a doplnění dalších zákonů (zákon o vysokých školách), ve znění pozdějších předpisů.

Jsem si vědom následků, které podle zákona o vysokých školách mohou vyplývat z porušení tohoto prohlášení.

14. května 2024 Patrik Drbohlav

Zpracování velkých datových souborů

Abstrakt

Tato bakalářská práce se zaměřuje na efektivní zpracování rozsáhlých datových souborů. Zabývá se obtížemi, spojenými s manipulací rozsáhlých dat a potřebou rychlé a efektivní analýzy. Cílem je prozkoumat současný stav zpracování dat v daném prostředí a navrhnout chytré softwarové řešení filtrovacího systému.

Teoretická část práce zkoumá problematiku zpracování velkých datových souborů v organizaci a získání podkladů pro tvorbu praktické části této práce. Praktická část řeší potíže s různými formáty datasetů, poukazuje na potřebu snadno použitelného nástroje pro částečnou automatizaci procesu vyhodnocování dat.

Na základě teoretických a praktických poznatků byl navržen filtrovací systém v programu MATLAB. Toto řešení nejen eliminuje manuální práci, ale podporuje matematické operace a vytváří vizualizace výsledků testů.

Výsledný nástroj je uživatelsky přívětivý a výrazně zvyšuje efektivitu zpracování rozsáhlých datových souborů. Navrhované řešení přináší technologický pokrok a měřitelné výhody pro rychlejší a přesnější analýzu dat.

Klíčová slova: Velké datové soubory, Filtrovací systém, Automatizace, MATLAB, Analýza dat.

The Big Data Processing

Abstract

This bachelor's thesis focuses on the efficient processing of extensive data files, addressing challenges associated with handling large datasets and the need for quick and effective analysis. The goal is to explore the current state of data processing in a given environment and propose a smart software solution for a filtering system.

The theoretical part of the thesis investigates the issues related to processing large data files in an organisation, providing insights for the practical section. The practical part addresses difficulties with various dataset formats, emphasizing the need for an easily usable tool to partially automate the data evaluation process.

Based on theoretical and practical knowledge, a filtering system in MATLAB has been designed. This solution not only eliminates manual work but also supports mathematical operations and generates visualizations of test results.

The resulting tool is user-friendly, significantly enhancing the efficiency of processing extensive data files. The proposed solution brings technological advancement and measurable benefits for faster and more precise data analysis.

Keywords: Big Data, Filtering System, Automation, MATLAB, Data Analysis.

Poděkování

Rád bych využil této příležitosti k upřímnému poděkování všem, kteří významně přispěli k úspěšnému dokončení této bakalářské práce. Děkuji vedoucí bakalářské práce, Ing. Věře Pelantové, Ph.D., a vedoucímu odborné praxe, Ing. Jiřímu Borovičkovi, za jejich cenné rady a odborné vedení. Velké díky také patří učitelům bakalářského semináře, doc. Ing. Josefovi Černohorskému, Ph.D., a Ing. Janu Koprnickému, Ph.D, za jejich přínos k mému akademickému růstu. Tato práce by nemohla vzniknout bez Vaší podpory, a za to jsem nesmírně vděčný. Rád bych také poděkoval své rodině a přítelkyni za neustálou podporu a trpělivost.

Obsah

Seznam zkratek

VDS	Velké datové soubory
BI	Business Intelligence
AI	Artificial Intelligence
IoT	Internet of Things (Internet věcí)
SW	Software
UI	User interface (Uživatelské rozhraní)
SQL	Structured Query Language
NoSQL	Not only Structured Query Language
NLP	Natural Language Processing
TB	Terabyte
EB	Exabyte
CSV	comma-separated values
MF4/MFD4	Measurement Data Format version 4
BLF	Binary Logging Format
XLSX	Microsoft Excel Spreadsheet

Seznam obrázků

Seznam tabulek

Úvod

V současné době je nezbytně nutné držet krok s technickým pokrokem, neboť každý pracovní proces generuje rozsáhlé datasety, s nimiž je třeba efektivně pracovat. Je zde kladen důraz na důkladnou analýzu a efektivní manipulaci s daty, aby bylo možné držet krok s konkurencí a zůstat inovativní.

Tato bakalářská práce se zaměří na průzkum hlavních výzev, spojených se zpracováním velkých datových souborů. Klíčová zde bude efektivita ve zpracování VDS. V dnešní době platí více, než kdy jindy, přísloví „čas jsou peníze“, a proto se usiluje o analýzu všech pracovních úkonů s cílem optimalizovat časovou náročnost. Podhled práce se zaměří na konkrétní organizaci, ve které se s velkými datovými soubory pracuje velmi často, a usnadnění a zefektivnění práce velmi pomůže celému pracovnímu procesu.

Dále se tato práce bude věnovat fragmentům dat získaných z konkrétní organizace, které poslouží jako demonstrativní model efektivních postupů zpracování a interpretace dat. Pro zpracování těchto dat se využije program MATLAB, který umožní nejen efektivní manipulaci s daty, ale také jejich následné grafické prezentování.

Hlavním cílem této práce je detailně prozkoumat, analyzovat a navrhnout inovativní a efektivní přístupy k zpracování rozsáhlých datových souborů v dané organizaci. Výsledkem bude softwarový program, který zefektivní rychlost zpracování dat získaných z měření a usnadní celkový proces vyhodnocení a uchovávání dat. Tento softwarový program bude schopen zpracovávat rozsáhlé datové soubory na základě specifikací uživatele a poskytne výstup v podobě grafických vizualizací i tabulkových dat.

1 Data, informace a znalosti

Následujících kapitoly se budou věnovat vysvětlení pojmů, které budou provázet celou práci.

1.1 Data

Co jsou to data? Česká terminologická databáze knihovnictví a informační vědy (TDKIV) definuje data jako: „reprezentaci informací vhodně formalizovanou pro komunikaci, interpretaci a zpracování lidmi a automaty.“ [15]. Tato definice zdůrazňuje, že data představují řetězce znaků, čísel nebo příkazů, uložených na informačním nosiči. Samotná data obvykle nemají význam, dokud nejsou pochopena, interpretována, komunikována a využita člověkem nebo počítačem, až poté se stávají smysluplnými informacemi. To naznačuje, že proces interpretace a využití dat je klíčový pro jejich transformaci z jednoduchých řetězců znaků na užitečné informace, které mohou poskytnout hodnotu a podporovat rozhodovací procesy [15].

Podle Marka Macury lze chápat data jako klíčový prvek v dnešním prostředí, kde se rozmanitost zdrojů dat neustále rozšiřuje. Data jsou vnímána jako fakta, události, zprávy a měřitelné charakteristiky, které mají význam nezávisle na pozorovateli, ale stávají se skutečnými „daty“ až tehdy, když jsou sbírána účelově. Skutečná hodnota dat spočívá v jejich správné interpretaci a schopnosti extrahovat užitečné informace pomocí statistické analýzy a vyvozování závěrů pro podporu rozhodování. Tyto výzvy jsou umocněny tím, že data pocházejí z různorodých zdrojů, což přináší další složitosti při integraci a analýze [20].

Cichy a Rass ve svém článku [3, An Overview of Data Quality Frameworks] navrhují, že data lze chápat jako formalizovanou reprezentaci faktů, konceptů nebo instrukcí, která má široké využití v komunikaci a zpracování nejen lidmi, ale i automatickými prostředky. Správný přístup k datům může poskytnout důležité informace o různých aspektech, včetně ekonomických, sociálních nebo technologických trendů a vzorců [3].

Obecně lze říci, že data představují objektivní reprezentaci fakt, událostí a měřitelných charakteristik, které mohou být získány experimentem, měřením, pozorováním nebo šetřením. Jejich skutečná hodnota spočívá v interpretaci a schopnosti extrahovat užitečné informace. Zahrnují různé typy informací, jako jsou numerické, textové, obrazové a zvukové údaje. I když samotná data nemají zpravidla význam sama o sobě, stávají se smysluplnými informacemi, až když jsou pochopena, interpretována, komunikována a využita člověkem nebo počítačem.

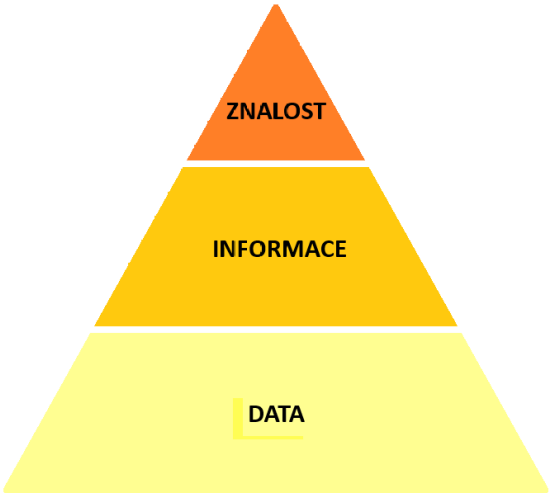
1.2 Informace

Informace jsou klíčovým prvkem v lidské činnosti, neboť snižují neurčitost systému a poskytují poznání o reálném prostředí a procesech, v něm probíhajících. Etymologicky pochází slovo informace z latinského *„informatio“*, což naznačuje proces přenosu myšlenek do komunikovatelné podoby. Jiný pohled na informace je, že jsou to data, kterým příjemce přisuzuje význam na základě svých znalostí a zkušeností, čímž snižuje neurčitost dle svých potřeb. Obecně informace vzniká z dat, kterým se přidává kontext, čímž získávají pro člověka určitý význam a užitečnost [21].

1.3 Znalost

Znalost je výsledkem porozumění zákonitostem a uspořádáním informací.Je také významově rozsáhlejší a hlubší než jen data a informace, ale bez dat, která jsou základem pro tvorbu informací, a bez samotných informací není možné vytvářet a rozvíjet znalosti [21].

Všechny tyto tři aspekty jsou nepostradatelnými stavebními kameny pro porozumění a vytváření smysluplných souvislostí. Jejich fungování mezi sebou je naznačeno na Obrázku 1.1.



Obrázek 1.1: Hierarchie dat (inspirace viz. [23])

1.4 Strukturovaná vs nestrukturovaná data

V této části bude vysvětleno rozdělení dat, která poté ve větším množství tvoří VDS.

1.4.1 Strukturovaná data

Strukturovaná data představují způsob ukládání a organizování dat, který je často reprezentován v tabulkách, jako jsou soubory Excel nebo SQL databáze. Tato data mají definované vztahy mezi jednotlivými datovými body a často jsou uložena ve formě řádků a sloupců, kde je možné identifikovat vztahy mezi jednotlivými položkami dat. Strukturovaná data jsou často snáze zpracovatelná a interpretovatelná, což je výhodné při využití v analýzách a strojovém učení [27].

1.4.2 Nestrukturovaná data

Nestrukturovaná data představují data, která nejsou organizována podle předem definovaného datového modelu nebo struktury. Tyto data se často označují jako kvantitativní a nelze je snadno analyzovat běžnými metodami, určenými pro strukturovaná data. Vzhledem k absenci definovaných vztahů mezi datovými body nelze nestrukturovaná data efektivně organizovat v relačních databázích. Pro jejich ukládání se často používají databáze typu NoSQL nebo jiné nerelační databáze. Zpracování nestrukturovaných dat je náročné a často vyžaduje prozkoumání jednotlivých částí dat k identifikaci potenciálních funkcí [27].

2 Uvedení do problematiky velkých datových souborů (Big Data)

Velké datové soubory (dále jen VDS), často označované jako Big Data, představují obrovské objemy informací, které jsou tak rozsáhlé a složité, že jejich zpracování a analýza tradičními metodami, kterými se rozumí například standardní tabulkové procesory, SQL dotazy, statické metody a SW programy, je velmi obtížná.

2.1 Historie velkých datových souborů

Počátky termínu Big Data lze datovat do konce 20. a počátku 21. století, konkrétně do období od 90. let do první dekády nového tisíciletí. Za jeden z prvních, ne-li úplně první impuls, k jehož užívání lze považovat pozoruhodný počin dvou výzkumníků NASA, Michaela Coxe a Davida Ellswortha. V jejich práci „Application-Controlled Demand Paging for Out-of-Core Visualization“ [4] se potýkali s enormním množstvím dat, že samotný problém, který tato data představovala, označili jako Big Data. Toto významově blízké použití termínu je pravděpodobně jedním z prvních v historii a o této rané fázi se zmiňuje i Steve Lohr ve svém článku „The Origins of „Big Data“: An Etymological Detective Story“ [19]. Tak začíná stopa klíčového pojmu, který se v následujících letech stal ústředním bodem diskusí o zpracování a analýze masivních souborů dat.

S postupem času se pojem Big Data vyvinul a transformoval v reakci na dynamiku technologického pokroku a narůstajících nároků na zpracování informací. Přesné chápání historického vývoje umožňuje lépe porozumět nejen minulým výzvám, spojeným s VDS, ale také poskytuje vodítko pro současný vývoj a budoucí výzvy v oblasti správy a analýzy dat.

2.2 Velké datové soubory v současnosti

Je patrný rychle se měnící svět, kde data hrají klíčovou roli a jsou nepřetržitě dostupná. Oproti historickým trendům, kdy byly změny v oblasti správy dat pomalé a struktury uchovávání dat zůstávaly stabilní po mnoho let, nyní se čelí novým výzvám v oblasti zpracování informací.

Dnešní dynamika datového prostředí nenutí nová data do předem daných forem a struktur. Informace do databází proudí ve vizuální, zvukové nebo textové podobě. Převod těchto formátů do jednotné podoby může být výzvou kvůli komplexnosti a specifičnosti jednotlivých aplikací.

Automatizace rutinních úloh v procesu shromažďování dat výrazně zrychluje a usnadňuje sběr detailních informací o zkoumaných jevech. Tímto způsobem lze získat komplexní soubory dat, které odrážejí podrobný pohled na sledované procesy. Avšak samotné shromažďování dat není konečným cílem, klíčovým prvkem je také následné zpracování, analýza a vizualizace těchto dat. Tyto kroky jsou nezbytné pro dosažení optimálních výsledků a poskytují interpretaci, která umožňuje hlubší porozumění sledovaným jevům a efektivní využití získaných informací.

2.2.1 Sběr dat

Při zpracování VDS je prvním krokem jejich sběr. Organizace obvykle získávají data ze zdrojů, jako jsou senzory, webové stránky, transakce, sociální média, interní databáze a internet věcí (dále jen IoT). Senzory a IoT sledují fyzické parametry, zatímco webové stránky a aplikace poskytují data o uživatelském chování a finanční údaje.

Následuje fáze filtrace, kde jsou z dat vyřazovány duplicitní nebo nepotřebné informace, aby data byla připravena pro konkrétní analýzu nebo zpracování. Tento krok je klíčový pro optimalizaci dat a zajistí, že jsou zachovány pouze relevantní informace pro další fáze. Po filtračním procesu jsou data připravena k okamžité analýze, nebo jsou uložena pro budoucí potřebu a podrobnější zpracování.

2.2.2 Ukládání a archivování dat

Data je nezbytné ukládat, poskytují odlišné pohledy a perspektivy skrze různé způsoby zpracování. Již zpracovaná data by měla být archivována pro možnost následného porovnání. Po sběru dat by bylo neefektivní se jich rychle zbavovat, v případě chyby ve zpracování by bylo obtížné je znovu získat.

Samotný proces ukládání dat není pouze o systematickém archivování, ale také o pečlivém zvážení povahy dat, zejména pokud obsahují citlivé informace. Organizace tak volí různé technologie a strategie pro zabezpečené a efektivní ukládání velkých datových sad.

V poslední době se ukazuje trend ukládání dat v cloudových uložištích. Ta poskytují online úložný prostor, kde organizace i jednotlivci mohou ukládat, spravovat a získávat přístup k datům bez potřeby vlastnit a spravovat vlastní fyzickou infrastrukturu. Většinou se jedná o veřejná cloudová uložiště, nabízející bezplatné nebo zpoplatněné úložné prostředky podle potřeby. Tyto služby jsou obvykle poskytovány třetími stranami. Pro střední a velké podniky, které mají rozsáhlé množství dat a některá jsou citlivá, může být vhodnější vytvořit vlastní soukromé cloudové uložiště. To může být provozováno interně nebo outsourcováno externím poskytovatelem, což organizacím umožní uchovávat data v cloudu s větší kontrolou a zabezpečením.

2.2.3 Analýza a zpracování

Po sběru a ukládání dat přichází na řadu jejich zpracování. Tento proces se liší v závislosti na typu dat, se kterými se pracuje.

Strukturovaná data, která jsou již organizována do tabulek a relací, se zpracovávají pomocí SQL dotazů nebo analytických nástrojů, což umožňuje provádět komplexní analýzy a generovat reporty.

Naopak nestrukturovaná data, vyskytující se v různých formátech a typech, jsou zpracovávána pomocí nástrojů jako Apache Hadoop nebo Apache Spark. Tyto frameworky umožňují distribuované zpracování velkých objemů nestrukturovaných dat pomocí technik mapování a redukce.

Běžně používané nástroje pro analýzu dat zahrnují Python s knihovnami Pandas a Matplotlib, nástroje Business Intelligence (BI) a Matlab pro specifické požadavky na analýzu dat.

Cloudové platformy, jako jsou Amazon Redshift nebo Google BigQuery, nabízí přístup k pokročilým analytickým nástrojům a umožňují rychlou a škálovatelnou analýzu dat v cloudu.

Analýza a zpracování dat jsou klíčové kroky v procesu extrakce hodnoty z velkých datových sad. Poskytují organizacím relevantní poznatky pro strategické rozhodování a zlepšují jejich efektivitu a konkurenceschopnost.

2.2.4 Vizualizace

Vizualizace dat představuje klíčový prvek v převádění komplexních informací na srozumitelné a názorné formy. Tato fáze umožňuje organizacím a jednotlivcům efektivně komunikovat a sdílet poznatky, získané z velkých datových sad.

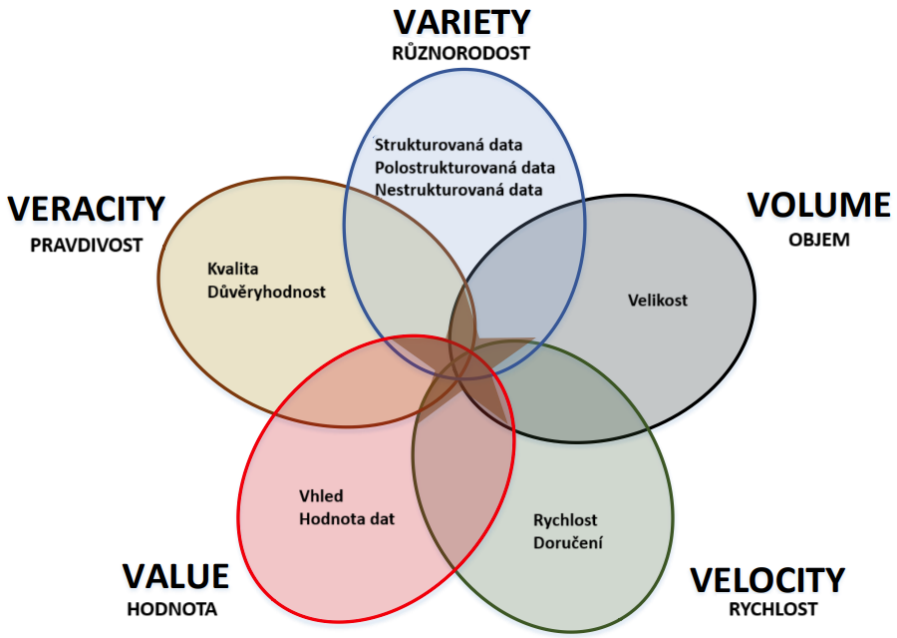
V rámci vizualizace se často využívají různé nástroje. Jedním z nich jsou grafy a grafické reprezentace, jako jsou sloupcové grafy, koláčové grafy nebo čárové grafy, které pomáhají ilustrovat numerické trendy. Dále se využívají heatmapy k zobrazení intenzity nebo vzorů v datech a prostorová rozložení pro zobrazení geografických informací.

V oblasti Business intelligence (BI) se často vytvářejí interaktivní dashboardy, které umožňují uživatelům pohodlně analyzovat data a vytvářet vlastní vizualizace. Kromě toho jsou populární i Word clouds, které vizualizují frekvenci výskytu slov nebo frází, a Treemaps pro hierarchické zobrazení dat.

Experimentace s 3D vizualizacemi a vytváření animací jsou dalšími způsoby, jak představit data v poutavější podobě. Vizualizace dat hraje klíčovou roli v komunikaci a sdílení poznatků s různými zúčastněnými stranami. Správně navržené vizualizace nejen usnadňují pochopení dat, ale také umožňují rychlé rozhodování na základě vizuálních náhledů.

2.3 Vlastnosti velkých datových souborů

V rámci dynamicky se rozvíjejícího digitálního světa se význam termínu VSD neustále prohlubuje, přičemž se stále více zaměřuje na specifické vlastnosti, které charakterizují tato masivní data. Jedním z klíčových konceptů při zkoumání VDS jsou takzvané 3V's – volume, variety a velocity, které tvoří základní pilíře pro porozumění jejich povaze [31] [14].



Obrázek 2.1: Představení „5V's“ velkých dat (inspirace viz. [11])

2.3.1 Objem (Volume)

Objem dat je klíčovým rysem VDS a odkazuje na obrovské množství informací, které tvoří tyto masivní datové soubory. V současném digitálním prostředí se objem dat pohybuje od terabytů (TB) až po exabyty (EB). Zvýšený objem dat vyžaduje efektivní technologie pro jejich ukládání, zpracování a analýzu. Síla dat však nemusí spočívat pouze ve velkém objemu.

Je důležité si uvědomit, že v kontextu VDS není klíčový pouze objem dat. I malé množství dat může být významné v závislosti na jejich povaze a obsahu. Význam dat spočívá i v jejich relevanci a schopnosti poskytovat cenné informace, díky kterým se lze lépe rozhodovat a porozumět sledovaným jevům [31] [12] [14].

2.3.2 Různorodost (Variety)

Různorodost dat v kontextu VDS popisuje rozmanitost typů a struktur informací, které jsou zahrnuty v datových souborech. Zahrnuje širokou škálu formátů, od tradičně strukturovaných dat, přes polostrukturovaná až po nestrukturovaná data, jako jsou textové soubory, obrázky, zvukové záznamy nebo videa.

Tato vlastnost reflektuje skutečnost, že Big Data mohou obsahovat informace, pocházející z různých zdrojů a v různých formátech. Zpracování a analýza takové rozmanitosti představují výzvu, protože vyžadují flexibilitu nástrojů, které jsou schopny efektivně pracovat s různorodými daty. Standardizace a sjednocení těchto informací jsou klíčové pro zajištění konzistence a přístupnosti dat pro všechny zainteresované strany [31] [12] [14].

2.3.3 Rychlost (Velocity)

Rychlost v kontextu Big Data zahrnuje tempo generování a časový rámec pro zpracování a analýzu informací z různých zdrojů, jako jsou například senzory, sociální média a zařízení IoT. Je důležité si uvědomit, že ne všechny aplikace vyžadují okamžité zpracování dat v reálném čase. Některé aplikace generují rozsáhlá data, která uživatel musí následně zpracovat, analyzovat a vyhodnotit. Tato situace zdůrazňuje, že pohled na data jako celek může poskytnout ucelenější perspektivu a umožnit odvozování relevantních závěrů.

Nicméně, s narůstajícím počtem aplikací závislých na datech z jiných zdrojů, roste potřeba zpracovávat a připravovat data v reálném čase. V těchto případech je klíčové zajistit, aby data byla dostupná pro další aplikace bez zbytečných zpoždění. To je především důležité v oblastech, kde synchronizovaná a aktuální data jsou nezbytná pro korektní fungování aplikací a rychlá rozhodnutí [31] [12] [14].

V kontextu zpracování VDS v reálném čase se klíčovou otázkou stává automatizace procesu. Tam, kde je rychlá a okamžitá analýza dat nezbytná, se procesy musí plně automatizovat, aby bylo možné reagovat včas. Naopak, některá data, zejména ta, která nevyžadují okamžitý zásah a mohou být podrobněji analyzována, zůstávají místem pro uživatelskou interakci. I když se lidé snaží automatizovat co nejvíce, role uživatele zůstává klíčová v oblastech, kde je potřeba lidského porozumění, kreativity a rozhodování. Avšak rozvoj umělé inteligence (dále jen AI) otevírá nové možnosti v automatizaci i v těchto komplexnějších sférách zpracování dat.

2.3.4 Další aspekty

S přibývajícím časem a neustále se rozvíjejícím konceptem VDS dochází k identifikaci dalších klíčových vlastností, které jsou nezbytné pro pochopení a efektivní správu těchto objemných datových souborů. Tyto aspekty odrážejí evoluci v přístupu k práci s velkými daty. Zdůrazňují potřebu nejen efektivního shromažďování a zpracování dat, ale také důkladného porozumění jejich významu a kvality.

Pravdivost (Veracity)

Pravdivost, jakožto jedna z klíčových vlastností VDS, se zaměřuje na konzistenci a spolehlivost dat. V kontextu Big Data mohou data obsahovat nejistoty, což ztěžuje kontrolu kvality a přesnost. Často se lze setkat s velkým množstvím nejasností a neověřených informací, které je nutné upravit a zorganizovat, aby sloužily svému účelu [9].

Čištění a organizace dat je klíčovým krokem před jejich využitím, zejména u polostrukturovaných dat, která často obsahují chybějící nebo nepřesné údaje. Problémy mohou vzniknout při shromažďování neplatných nebo duplicitních dat, což může vést k nepřesným výsledkům a zkresleným závěrům [5].

Důvěryhodnost a spolehlivost dat jsou zásadní pro dosažení výsledných hodnot v analýze Big Data. Kvalita dat může být velmi proměnlivá a má vliv na přesnost analýzy. Je důležité zaměřit se pouze na ty proměnné, které jsou považovány za nejspolehlivější, aby analýza vedla k relevantním a důvěryhodným výsledkům [31].

Hodnota (Value)

Hodnota je klíčovým aspektem práce s velkými daty. Bez hodnoty jsou data bezcenná, ať už jde o jejich sběr, analýzu nebo využití. Je nezbytné, aby organizace měly jasné cíle a strategie pro efektivní využívání dat a zajistily, že získané poznatky přinášejí měřitelné a užitečné výsledky [14].

2.4 Výzvy spojené s velkými datovými soubory

V oblasti VDS se lze často setkat s řadou skrytých výzev a překážek, které mohou významně ovlivnit úspěch datových projektů. Jedním z předních problémů je přístup k datům a jejich sdílení. S rostoucím objemem dat je stále náročnější zajistit jejich dostupnost a správné využití. Současně je zásadní chránit soukromí a bezpečnost dat, jakmile se data dostanou do nesprávných rukou, může to mít fatální následky. Dalším důležitým faktorem jsou analytické výzvy, jelikož pracovníci musí být schopni efektivně zpracovávat obrovské množství dat a získávat z nich užitečné informace. Technické problémy, jako je kvalita dat a rozšiřování infrastruktury, představují další zásadní výzvu. Infrastruktura v kontextu velkých dat zahrnuje hardware, software a technologické prostředky, které umožňují sběr, ukládání, správu a analýzu dat. Je nezbytné hledat inovativní řešení a strategie, které pomohou překonat tyto překážky a využít plný potenciál velkých dat.

2.4.1 Přístup k datům a sdílení

Data jsou cennou komoditou v organizacích a často tak dochází k jejich sdílení interně nebo externě. Přístup k datům z veřejných repozitářů může být rovněž obtížný. Je proto nezbytné, aby data byla k dispozici v přesném a úplném provedení, neboť pouze v tomto formátu mohou být efektivně využita v rámci informačního systému organizace [7].

2.4.2 Ochrana soukromí a bezpečnost dat

Ochrana soukromí a bezpečnost jsou klíčovými aspekty při práci s velkými daty. Vzhledem k neustále rostoucí hrozbě kybernetických útoků a zneužití citlivých informací je důležité dodržovat přísné standardy ochrany osobních údajů a respektovat platné právní předpisy. Pravidelné bezpečnostní kontroly a sledování dat v reálném čase jsou nezbytné pro minimalizaci rizik a zajištění ochrany citlivých informací. Pokud by se informace dostaly do nesprávných rukou, mohlo by dojít k vážným následkům pro všechny zúčastněné strany [30].

Je doporučeno, aby organizace pravidelně prováděly bezpečnostní kontroly a zabezpečovaly data, aby se minimalizovala rizika. Některé organizace sbírají informace o jednotlivcích za účelem zlepšení svého podnikání, což může vyvolávat etické otázky, pokud nejsou respektovány zásady ochrany soukromí a práva jednotlivců [7].

2.4.3 Kvalita dat a datová integrita

Ve světě VDS se lze setkat s významnými výzvami, spojenými s kvalitou dat a jejich integritou. Při analýze a zpracování těchto souborů se pracovníci často potýkají s problémy, jako jsou chyby v datech, nekonzistence formátů, duplicity a chybějící hodnoty. Tyto nedostatky mohou výrazně ovlivnit spolehlivost a využití dat pro rozhodování a analýzu. Je klíčové zajišťovat integritní standardy a postupy pro čištění a normalizaci dat, aby byla udržována konzistence a spolehlivost datových souborů. Důkladná péče o kvalitu dat a jejich integritu je nezbytná pro efektivní využití potenciálu velkých dat a minimalizaci rizik [7] [30].

2.4.4 Technické výzvy a infrastruktura

Organizace se často potýkají s výzvami v infrastruktuře při práci s VDS. Klíčové problémy zahrnují potřebu dostatečné kapacity a výkonnosti pro ukládání a zpracování obrovských objemů dat. Zdůrazňuje se nutnost vybudování infrastruktury, která dokáže rychle a efektivně zpracovávat data, aby organizace mohly využít jejich plný potenciál [30].

Dalšími technickými výzvami jsou analýza, vizualizace a interpretace VDS. Je zásadní mít k dispozici efektivní nástroje pro zpracování a interpretaci rozsáhlých dat. Technologický rozvoj a inovace jsou klíčové pro úspěšné využití potenciálu velkých dat v organizacích [7].

2.4.5 Analytické výzvy a zpracování dat

V kontextu VDS se lze setkat s analytickými výzvami, které se objevují zejména při zpracování VDS. Nejčastějšími otázkami jsou, jak efektivně pracovat s daty, pokud jejich objem přesahuje určitou mez, jak identifikovat důležité datové body a jak nejlépe využít dostupná data pro dosažení požadovaných cílů. Existují dva hlavní přístupy k rozhodování: buď zahrnout masivní objemy dat přímo do analýzy, nebo předem určit, která data jsou pro daný účel relevantní [7].

2.5 Trendy velkých datových souborů

Tato část se zaměří na aktuální směry a novinky v práci s VDS. Autor analyzuje trendy a inovace v oblasti technologií, strategií a analytických přístupů, které ovlivňují způsob, jakým organizace zpracovávají a využívají VDS.

2.5.1 Artificial intelligence

Podle IBM lze definovat AI jako: „technologii, která umožňuje počítačům a strojům simulovat lidskou inteligenci a schopnosti řešení problémů.“ [13]. AI je buď samostatně nebo ve spojení s dalšími technologiemi, například senzory, geolokace a robotika, schopna vykonávat úkoly, které by jinak vyžadovaly lidskou inteligenci nebo lidský zásah [13].

Využití umělé inteligence ve velkých datových souborech

V oblasti zpracování VDS hraje AI stále důležitější roli. Díky pokroku v oblasti učících algoritmů a technologií pro analýzu velkých dat se otevírají nové možnosti pro efektivní zpracování obrovských objemů dat. AI umožňuje identifikovat skryté vzory, provádět predikce a extrahovat cenné informace z rozsáhlých datových sad. Například, vývoj speciálních učících algoritmů umožňuje dosahovat vysoké přesnosti i při analýze nestrukturovaných datových sad. Tato kombinace AI a Big Data má potenciál přinést významné inovace a zlepšení v různých odvětvích, od obchodu a průmyslu po vědu a technologii. Díky schopnosti AI, provádět analýzu v reálném čase, je umožněno systémům okamžitě zpracovávat data a reagovat na aktuální události, což je zásadní zejména v situacích, kde výstup z analýzy dat ovlivňuje následné operace nebo rozhodnutí. AI tak přináší do zpracování VDS dynamiku a flexibilitu, potřebnou pro efektivní využití informací, obsažených v datech [32].

Strojové učení

Strojové učení je důležitou součástí AI, která se zabývá vytvářením algoritmů a modelů, které umožňují počítačům „učit se“ ze zkušeností a dat. Tato technika umožňuje počítačům zlepšovat své výkony v konkrétních úlohách nebo oblastech bez explicitního programování. Strojové učení je klíčovým nástrojem pro vytváření inteligentních systémů schopných adaptace a zlepšování výkonu v průběhu času [32].

Neuronové sítě

Neuronové sítě jsou matematickým modelem inspirovaným fungováním lidského mozku, který se skládá z propojených umělých neuronů. Tyto sítě jsou schopny učit se a adaptovat své váhy a strukturu na základě prezentovaných dat. Neuronové sítě jsou základní stavební jednotkou pro hluboké učení a umožňují počítačům zpracovávat složité informace a řešit náročné úkoly, které by jinak byly obtížné nebo nemožné pro tradiční algoritmické přístupy [32].

Hluboké učení

Hluboké učení je specifickým přístupem k strojovému učení, který využívá složité neuronové sítě inspirované strukturou lidského mozku. Díky hlubokému učení mohou počítače rozpoznávat vzory, provádět predikce a řešit složité úkoly, jako je rozpoznávání obrazů nebo překlad jazyka [32].

2.5.2 Cloudové prostředí

S nárůstem datových objemů organizace přecházejí k cloudovým a hybridním cloudovým řešením pro efektivní ukládání a zpracování dat. Tato změna umožňuje organizacím flexibilně reagovat na rostoucí požadavky na úložiště a výpočetní kapacity, přičemž odpadá nutnost správy a provozu vlastních rozsáhlých datových center. Některá odvětví čelí regulačním a technickým omezením ve využívání veřejných cloudových služeb. Proto poskytovatelé cloudových platforem nabízejí hybridní přístupy a infrastrukturu přizpůsobenou specifickým potřebám, čímž podporují další rozvoj cloudových technologií.

Souběžně s pokrokem v oblasti cloudového ukládání a zpracování dat se organizace zaměřují na nové přístupy k architektuře dat, jako je koncept „Data Lake“ (Datová jezera). Datová jezera umožňují ukládání různorodých datových sad v jejich původním formátu a přenášejí odpovědnost za transformaci dat na uživatele. Tento přístup umožňuje organizacím lépe využít jejich datové zdroje a snadněji provádět analýzy, aniž by musely předem strukturovat data do konkrétní podoby [33].

Mezi klíčové analytické nástroje v cloudu patří Amazon Athena, Google BigQuery, Microsoft Azure Analysis Services a Snowflake. Tyto platformy poskytují škálovatelné a rychlé dotazování dat, interaktivní analýzy a podporu pro vývoj analytických aplikací. Využití těchto služeb umožňuje organizacím optimalizovat své analytické procesy a získat komplexní pohled na svá data [35].

Nad rámec komerčních řešení nabízí cloudové prostředí i možnost využití open-source frameworků, jako jsou Apache Hadoop a Apache Spark. Tyto nástroje umožňují organizacím pracovat s velkými objemy dat jak na vlastních serverech, tak i na externím cloudovém uložišti. Hadoop a Spark poskytují distribuované zpracování dat a podporu pro různé programovací jazyky, což umožňuje flexibilní a výkonné analýzy dat bez nutnosti drahých licencí. Takto široká nabídka analytických nástrojů v cloudu a open-source frameworků umožňuje organizacím efektivně využívat svá data a získávat z nich cenné poznatky pro svůj obchodní rozvoj [6] [2] [18].

2.6 Metody analýz velkých datových souborů

Analýza dat představuje klíčový proces v moderním světě a vědeckém prostředí, kde hraje stěžejní roli při získávání užitečných informací a podpoře rozhodovacích procesů. S rozsáhlým spektrem přístupů a technik, které se v různých oblastech uplatňují, nabízí analýza dat možnost detailního zkoumání datových souborů. V kontextu současného podnikání se analýza dat stává nedílnou součástí vědeckých postupů a pomáhá organizacím operovat s větší efektivitou. Tato kapitola se zaměřuje na metody analýzy VDS, které jsou klíčové pro porozumění a využití informací skrytých v rozsáhlých datových sadách.

Tabulka 2.1: Tabulka typů analýz VDS a příslušných nástrojů (Zdroj: vlastní autora)

Typ analýzy	Popis	Příslušné nástroje
Deskriptivní analýza	Shrnutí a vizualizace datových trendů	Excel, SQL, Python
Diagnostická analýza	Identifikace příčin a problémů v datech	R, MATLAB, Python
Prediktivní analýza	Predikce budoucích událostí na základě historických dat	Python, R
Preskriptivní analýza	Stanovení optimálních akcí na základě analýzy dat	Python, R
Explorativní analýza dat	Odhalení vzorů a vztahů v datech	Power BI, Python, SQL
Textová analýza	Zpracování a porozumění textovým datům	Python, SQL
Grafová analýza	Analýza vztahů mezi entitami pomocí grafů	SQL, MATLAB, Python, R

2.6.1 Deskriptivní analýza

Deskriptivní analýza zahrnuje shrnutí a popis dat s cílem porozumět jejich klíčovým charakteristikám. Zaměřuje se na to, co se stalo v minulosti, poskytuje náhled na historické vzorce a trendy dat. Tato analýza pomáhá odpovědět na otázky jako „Co se stalo?“ a „Jaké jsou hlavní rysy dat?“. Zahrnuje techniky jako sumarizace dat a vizualizaci, aby prezentovala data způsobem, který je smysluplný [17].

2.6.2 Diagnostická analýza

Diagnostická analýza si klade za cíl identifikovat kořenové příčiny konkrétních výsledků nebo událostí prozkoumáním dat do hloubky. Zaměřuje se na porozumění tomu, proč existují určité vzorce nebo trendy analýzou vztahů mezi proměnnými. Diagnostická analýza pomáhá odpovědět na otázky jako „Proč se to stalo?“ a „Jaké faktory přispívají k pozorovaným výsledkům?“. Tento typ analýzy je klíčový pro odhalení poznatků, které mohou vést k akčním doporučením pro zlepšení [17].

2.6.3 Prediktivní analýza

Prediktivní analýza se zabývá předpovídáním budoucích trendů nebo výsledků na základě historických vzorců dat. Využívá statistické algoritmy a techniky strojového učení k predikci budoucích událostí. Prediktivní analýza pomáhá odpovědět na otázky jako „Co je pravděpodobné, že se stane?“ a „Jaké jsou potenciální budoucí scénáře?“. Využitím prediktivních modelů mohou organizace předvídat trendy, učinit informovaná rozhodnutí a optimalizovat strategie pro lepší výsledky [17].

2.6.4 Preskriptivní analýza

Preskriptivní analýza jde nad rámec předpovědi výsledků a poskytuje doporučení, jaké kroky podniknout k dosažení požadovaných výsledků. Zahrnuje využití optimalizačních a simulačních technik k navržení nejlepšího postupu na základě prediktivních modelů. Preskriptivní analýza pomáhá odpovědět na otázky jako „Co bychom měli udělat?“ a „Jak můžeme dosáhnout požadovaných výsledků?“. Poskytováním konkrétních doporučení preskriptivní analýza usměrňuje rozhodování a pomáhá optimalizovat procesy pro maximalizaci výsledků [17].

2.6.5 Explorativní analýza

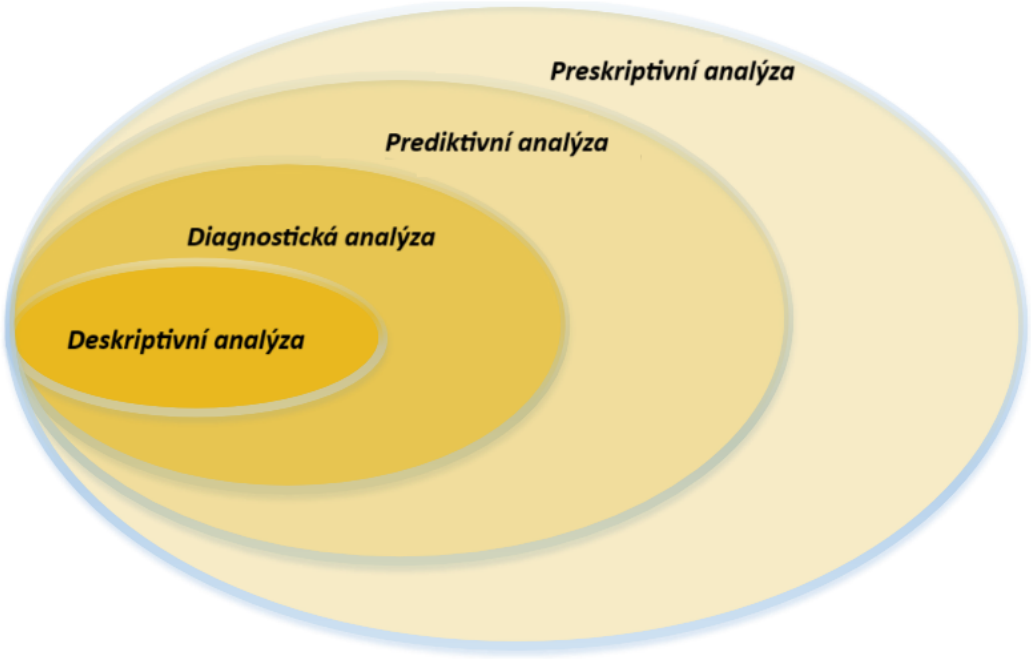
Explorativní analýza se zaměřuje na prozkoumávání dat k objevení vzorců, vztahů a trendů bez předem stanovených hypotéz. Zahrnuje vizualizaci dat a identifikaci skrytých vzorců, které nemusí být okamžitě zřejmé. Explorativní analýza pomáhá porozumět struktuře dat a přivádí člověka na nové hypotézy, které vedou k dalšímu zkoumání. Jedná se o krok v procesu analýzy dat k získání počátečních poznatků a nasměrování dalších analýz [17].

2.6.6 Textová analýza

Textová analýza zahrnuje extrakci strukturovaných dat z nestrukturovaných textových zdrojů, jako jsou e-maily, příspěvky na sociálních médiích a zpětná vazba od zákazníků. Využívá technik zpracování přirozeného jazyka (NLP) k analýze a interpretaci textových dat, umožňující organizacím odhalovat cenné poznatky z textových zdrojů. Textová analýza pomáhá porozumět sentimentu zákazníků, identifikovat trendy a extrahovat smysluplné informace z psaného textu pro potřeby rozhodování [17].

2.6.7 Grafová analýza

Grafová analýza zahrnuje využití vizuálních prezentací, jako jsou grafy, tabulky a diagramy, k prezentaci dat v vizuálním formátu. Pomáhá v komunikaci komplexních informací, identifikaci trendů a zdůraznění vzorců v datech. Grafová analýza zlepšuje interpretaci dat poskytováním jasné a intuitivní reprezentace vztahů mezi daty. Vizuální prezentace hraje klíčovou roli v analýze dat tím, že informace dělá přístupnějšími, usnadňuje rozhodování a zlepšuje data-driven poznatky [17].



Obrázek 2.2: Integrace metod analýzy velkých objemů dat (inspirace viz. [11])

2.7 Nástroje pro analýzu dat

Široká paleta nástrojů a technologií umožňuje uživatelům s různou úrovní technické zdatnosti analyzovat data a extrahovat z nich hodnotné informace. V této kapitole se autor zaměří na základní nástroje pro analýzu dat, jako jsou SQL a NoSQL databáze, a programovací jazyky. Ukáže, jak volba vhodného nástroje závisí na charakteru dat a požadovaném typu analýzy.

2.7.1 SQL a NoSQL

Structured Query Language (dále jen SQL) databáze jsou navrženy pro práci se strukturovanými daty, která jsou uspořádána do tabulek s řádky a sloupci. Mají robustní dotazovací jazyk SQL pro snadné vyhledávání a manipulaci s daty a zajišťují vysokou integritu dat díky pevně definované struktuře. Jsou ideální pro transakční systémy, kde je důležitá přesnost a konzistence dat.

Naproti tomu Not Only SQL(dále jen NoSQL) databáze jsou navrženy pro práci s nestrukturovanými a semi-strukturovanými daty, která nemají pevnou definovanou strukturu. Data mohou být uložena v různých formátech, jako jsou JSON, grafy, obrázky atd. Nabízí vysokou flexibilitu pro ukládání a analýzu nestrukturovaných dat, snadné škálování pro velké objemy dat a jsou vhodné pro aplikace, které vyžadují rychlé zpracování a nízkou latenci [1].

Volba mezi SQL a NoSQL databází závisí na typu dat, s nimiž zájemce pracuje, a na vašich požadavcích na analýzu.

2.7.2 Programovací nástroje

Kromě databází SQL a NoSQL existují další výkonné nástroje pro analýzu dat, a to programovací jazyky a vývojová prostředí. Tyto nástroje nabízí větší flexibilitu a kontrolu nad analýzou dat a umožňují pracovat s různými typy dat a provádět komplexní analýzy a vizualizace.

Python je univerzální programovací jazyk s širokým spektrem aplikací, zejména v oblasti zpracování VDS. Je hojně využíván v datové vědě, strojovém učení, vývoji webových aplikací a automatizaci. Jedná se o jazyk, který je relativně snadný k naučení a disponuje rozsáhlou a aktivní komunitou. Díky tomu má k dispozici rozsáhlé knihovny a frameworky pro různé úkoly, což usnadňuje práci v oblasti zpracování VDS. Python je navíc open-source a volně dostupný, což jej činí atraktivní [26].

MATLAB je vysoce specializované programovací a vývojové prostředí, určené pro numerické výpočty, analýzu dat, vizualizaci a vývoj algoritmů. Je známé svými schopnostmi v oblasti maticových výpočtů, lineární algebry, zpracování signálů a řízení. Tento software disponuje rozsáhlými knihovnami a funkcemi zaměřenými, na vědecké a technické aplikace a poskytuje uživatelům grafické uživatelské rozhraní (dále jen UI), které usnadňuje interaktivní práci s daty. Je však důležité zdůraznit, že MATLAB je komerční software a vyžaduje platbu licenčního poplatku za používání [22].

R je programovací jazyk a vývojové prostředí, zaměřené zejména na statistické výpočty a tvorbu grafů. Jeho hlavní síla spočívá v rozsáhlých knihovnách určených pro statistickou analýzu, vizualizaci dat a strojové učení. Díky těmto nástrojům je R oblíbeným nástrojem mezi statistiky a datovými vědci. Jeho velká a aktivní komunita poskytuje uživatelům nejen podporu, ale také přístup k mnoha užitečným balíčkům a doplňkům. R je rovněž známý pro své výkonné grafické možnosti, které umožňují vytvářet vizualizace dat vysoké kvality [8].

2.7.3 Microsoft Excel & PowerBI

Excel a Power BI od Microsoftu patří mezi nejpoužívanější nástroje pro základní zpracování dat. Tyto platformy poskytují širokou škálu funkcí pro analýzu, vizualizaci a reporting, což výrazně usnadňuje porozumění datům.

Excel představuje univerzální tabulkový procesor, který se využívá pro rozmanité úkoly spojené se zpracováním dat. Kromě základních funkcí pro import, čištění a transformaci dat nabízí také možnosti pro statistickou analýzu, vizualizaci a práci s daty. Díky své dostupnosti a intuitivnímu prostředí je Excel vhodný zejména pro jednotlivce a organizace, které teprve začínají s analýzou dat, nebo pro ty, kteří provádějí jednoduché analýzy dat a potřebují rychlý a snadno použitelný nástroj [24].

Power BI představuje sofistikovanou analytickou platformu, umožňující interaktivní vizualizaci dat prostřednictvím reportů a dashboardů. Tento nástroj disponuje širokým spektrem funkcí pro propojení dat z různých zdrojů, modelování dat a využití strojového učení. Power BI je vynikající pro sdílení informací a umožňuje tvorbu interaktivních prezentací s vysokou mírou vizuální atraktivnosti a užitečnosti [25].

3 Průzkum stavu problematiky v organizaci

V této kapitole se autor práce zaměří na průzkum problematiky VDS v organizaci. Provede charakteristiku organizace, analýzu současných nástrojů, identifikuje problémy se zpracováním dat a zvolí nejvhodnější nástroj, ve kterém vytvoří SW řešení za pomocí Quality Function Deploymen (dále jen QFD) korelační matice.

3.1 Charakteristika organizace

Organizace, vybraná pro praktickou část této bakalářské práce, je jedním z předních hráčů v automobilovém průmyslu s dlouholetou historií, která se promítá v kvalitě a spolehlivosti jejích produktů. Stěžejním úsekem pro tuto práci je oddělení Technického vývoje, zaměřené na nejrůznější zkoušky převodových systémů a pohonů.

Oddělení provádí různé zkoušky převodových systémů a pohonů, jako jsou například zátěžové (cyklické), vibrační, akustické a teplotní. Tyto testy jsou klíčové pro zajištění kvality, spolehlivosti a výkonu převodových systémů a pohonů. Ve finální části těchto zkoušek jsou generována data, která se často liší formátem v důsledku využití různých řídících programů. Tento proces vede k shromažďování velkých objemů různorodých dat, která jsou následně zpracovávána různými způsoby tak, aby výsledné informace byly co nejužitečnější a napomohly k řešení zkoumané problematiky.

Oddělení disponuje moderními testovacími laboratořemi a zařízeními, které umožňují provádět komplexní testování. Tým kvalifikovaných odborníků se stará o přípravu a realizaci zkoušek, monitoruje jejich průběh a provádí kontroly. Toto oddělení úzce spolupracuje s ostatními částmi společnosti, včetně výzkumu a vývoje, výrobních závodů, oddělení kvality a externích partnerů.

Oddělení hraje klíčovou roli v technologickém růstu a konkurenceschopnosti společnosti díky neustálému vývoji a inovacím v oblasti převodových systémů a pohonů. Jeho práce přispívá k udržení vedoucí pozice společnosti na trhu a k jejímu neustálému technologickému růstu.

3.2 Analýza existujících nástrojů ke zpracování velkých datových souborů

V organizaci se převážně pracuje s daty ve formátech XLSX, MF4, BLF a CSV. Data ve formátech MF4 a BLF jsou exportována do XLSX nebo CSV, z důvodu kompatibility využívaných nástrojů ke zpracování. Již zpracovaná data jsou nadále uchovávána ve formátu XLSX, pro jeho snadnou obsluhu a následnou interpretaci dat.

Každá sada dat je zpracovávána individuálně, přičemž metody zpracování se liší v závislosti na zkoumaných aspektech. Například data z akustických zkoušek budou zpracována a vyhodnocena jinak než data z teplotních nebo zátěžových zkoušek.

Pro jednodušší zpracování menších objemů dat se často využívá ruční zpracování pomocí tabulkového procesoru Excel. V tomto nástroji lze nalézt vytvořená makra, která ale nejsou vždy aktualizována, a proto je předpřipravení dat do podoby, se kterou makra pracují, náročné jak časově, tak technicky.

V případech, kdy je zapotřebí sofistikovanější zpracování, se využívají nástroje vytvořené v jazyce C nebo C#. Tyto programy dokáží zpracovat data různými způsoby a pracují s formátem CSV. Před použitím těchto programů je nutné data připravit do požadované podoby, což může být v případě velkých objemů dat časově náročné.

Lze říci, že ve firmě jsou k dispozici různé nástroje pro zpracování dat, a volba mezi nimi závisí na konkrétních potřebách a požadavcích uživatele.

3.3 Identifikace specifických problémů a potřeb organizace v oblasti velkých datových souborů

Jedním z výrazných problémů, se kterými se lze setkat, je rozmanitost formátů a struktury dat, což komplikuje efektivní zpracování a analýzu. Soubory ve formátu XLSX často obsahují názvy sloupců v různých jazycích, nekonzistentní uspořádání sloupců. Jsou zde použity také odlišné formáty, jako například MF4, BLF nebo CSV, které pocházejí ze zkoušek s odlišným řídícím systémem, a kladou tak vysoké nároky na nástroje.

Velký objem dat, generovaný během zkoušek, a vysoká frekvence vzorkování (až 1000 Hz) přináší další výzvy. Po zpracování jsou výsledky interpretovány tabulkami a grafy, které je třeba archivovat pro pozdější využití. Archivovat je nezbytné i nezpracovaná data, kvůli možnosti budoucí potřeby přezkoumání. Je velmi důležité efektivně pracovat s úložištěm, aby byla ukládána pouze nezbytně nutná data. Při nevhodně zvolené metodě zpracování, dochází k navyšování objemu dat z důvodu mezikroků, které generují další nepotřebná data.

Je nezbytné optimalizovat procesy manipulace s daty od sběru po vizualizaci a archivaci. Tato věc bude klíčovým prvkem pro navržení konkrétních řešení a doporučení s cílem zlepšit efektivitu práce s VDS v rámci vybrané organizace.

3.4 Porovnání nástrojů pro zpracování dat pomocí metody Quality Function Deployment

QFD je metoda, která se používá k transformaci potřeb zákazníka do konkrétních technických požadavků a následně do designu, procesů a výsledné kvality produktu. Je to efektivní nástroj, který zajistí, že výsledný produkt plně odpovídá potřebám zákazníka.

Centrálním prvkem QFD metody je korelační matice, která pomáhá analyzovat potřeby zákazníka a převést je do technických specifikací. Tato matice umožňuje identifikovat klíčové vztahy mezi potřebami zákazníka a technickými charakteristikami produktu.

Tabulka 3.1: Porovnání programovacích prostředí a jazyků (Zdroj: vlastní autora)

Požadavek	**Ideal**	**Python**	**R**	**MATLAB**	**Excel & Power BI**	**C++**	**C#**
Snadné použití	10	8	7	6	5	7	6
Rychlost zpracování	10	7	6	9	4	9	7
Spolehlivost výsledků	10	8	8	9	6	9	7
Možnost vizualizace	10	8	8	8	6	5	7
Zpracování různých formátů	10	9	8	7	6	5	7
Export výsledků do různých formátu	10	8	8	7	7	5	6
Podpora vývojových prostředí	10	9	8	9	7	8	8
Kompatibilita s dalšími SW nástroji	10	9	9	8	8	9	8
Váhy	1-10	74	62	63	49	57	56
Výsledek	100 %	92 %	77 %	79 %	61 %	71 %	70 %

Autor provedl komplexní porovnání programovacích prostředí a jazyků pro zpracování dat, aby identifikoval ideální nástroj pro specifické potřeby a požadavky organizace na filtrovací SW. Hodnocení jednotlivých kritérií bylo v rozsahu 1-10, přičemž autor při přiřazování vah čerpal z více zdrojů ([29] [10] [28] [34]) a snažil se zachovat objektivní pohled.

Díky zkušenostem s programovacím prostředím MATLAB se autor rozhodl pro tento nástroj. Jeho znalost umožňuje plynulou a efektivní práci a dosahování požadovaných výsledků. Ovládání známého prostředí snižuje čas potřebný na učení a umožňuje soustředit se na samotný vývoj nástroje pro zpracování dat.

Důležitým faktorem volby byla také dostupnost licence. Organizace již disponovala licencí pro MATLAB, což eliminuje potřebu investovat do dalšího SW a šetří finanční prostředky, což je v komerčním prostředí klíčové.

Je třeba zdůraznit, že rozhodnutí bylo částečně ovlivněno subjektivními preferencemi autora. Nicméně tabulka 3.1 a její hodnocení sloužily jako objektivní podklad pro toto rozhodnutí, zohledňující různé aspekty výběru nástroje pro zpracování dat.

4 Návrh softwarového řešení filtrovacího softwaru pro velké datové soubory v organizaci

Vzhledem k již zmíněným problémům a potřebám (3.3), které autor identifikoval, a k omezenému času je složité navrhnout softwarové řešení, které by všechny zmíněné problémy vyřešilo. Přesto je zde představen program, který částečně automatizuje a zefektivňuje zpracování VDS. Program pracuje s fragmenty dat, aby výstup byl nic neříkající, a sloužil pouze k ukázce funkce programu.

4.1 Charakteristika poskytnutých dat

Tabulka 4.1 reprezentuje část dat, která jsou získávána ze zátěžové zkoušky převodového systému. Je důležité zmínit, že jde o fragmenty dat celé zkoušky a slouží pouze pro demonstraci SW navrženého autorem. Na první pohled si lze všimnout vysoké vzorkovací frekvence, ve které jsou data zaznamenávána (100 Hz). Vysoká vzorkovací frekvence je zde z důvodu potřeby detailního zmapování průběhu zkoušky. Poskytnutý dataset celkově obsahuje dvacet pět sloupců a přibližně sto tisíc řádků. Data jsou převážně numerická, a proto je složité je od sebe odlišit, a tak plně automatizovat jejich zpracování. Každý sloupec má svůj název a znázorňuje jiný aspekt zkoumaného převodového systému. Organizace zpracovává data v různých formátech, přičemž volba formátu a metody závisí na jejich objemu a náročnosti zpracování.

Tabulka 4.1: Příklad poskytnutých dat z organizace [16]

Čas během zkoušky	Čas v zátěžném bloku	# blok	# řádek blok	otáčky_LZ	otáčky_PZ	Moment_LZ	Moment_PZ
[s]	[s]	-	-	[1/min]	[1/min]	[Nm]	[Nm]
375869,61	0	49	4660	0,55776	-0,5488	-0,2172	-0,75
375869,62	0,01	0	0	0	-0,53648	-0,1248	-0,93
375869,63	0,02	0	1	0,47824	0	-0,0552	-0,8964
375869,64	0,03	0	2	-0,38416	0,2856	0,2688	-0,966
375869,65	0,04	0	3	-0,38416	0,2856	0,2688	-0,966
375869,66	0,05	0	4	0	0	0,594	-0,9828
375869,67	0,06	0	5	-0,47264	0,53648	0,9528	-1,2
375869,68	0,07	0	6	0	0,56224	1,5648	-1,2528
375869,69	0,08	0	7	-0,47712	0	0,1968	-1,3104
375869,7	0,09	0	8	0	0	-0,0348	-1,146
375869,71	0,1	0	9	0,31248	-0,26096	-0,0552	-1,092
375869,72	0,11	0	10	0	0	-0,2148	-0,948
375869,73	0,12	0	11	0,4648	-0,55664	-0,1104	-0,84
375869,74	0,13	0	12	0,43792	0	0,036	-0,768
375869,75	0,14	0	13	0,43792	0	0,036	-0,768
375869,76	0,15	0	14	0	0	0,1428	-0,696

Pro zajištění úplné anonymity byla provedena úprava dat autorem tak, aby bylo zcela znemožněno získání jakýchkoli citlivých informací.

4.2 Funkční a nefunkční požadavky na filtrovací software

Autor vytvořil dva typy softwarových řešení. První typ je založen na statickém modelu, kde jsou veškeré parametry a procesy pevně definovány. Druhý typ je rozšířen o UI, které umožňuje interakci s uživatelem prostřednictvím příkazového okna. Toto rozšíření bylo vytvořeno v reakci na proměnlivost dat, kde lze od uživatele získat specifické informace nezbytné pro správné zpracování dat. Tímto přístupem autor zajišťuje robustnost a flexibilitu filtrovacího SW.

4.2.1 Klíčové operace

Je nezbytné stanovit klíčové požadavky, aby filtrovací SW dosahoval optimální účinnosti. Mezi tyto klíčové požadavky patří schopnost načítání dat z Excelu a jejich efektivní zpracování. Autor původně usiloval o maximální automatizaci zpracování s cílem minimalizovat uživatelskou zátěž. Avšak, kvůli odlišnostem v pojmenování sloupců v datech z různých systémů, se ukázalo jako nemožné plně automatické zpracování. Pro zachování funkčnosti a opakovatelnosti využití navrženého nástroje je nezbytné, aby uživatel zadával sloupce v daném pořadí. Detailnější zdůvodnění a instrukce k tomuto postupu jsou k dispozici v popisu kódu (4.3.2).

Základní funkce, které systém provádí jsou výpočty momentů a otočení hřídele na základě definovaných kritérií. To umožňuje uživateli detailní analýzu dat a získání hlubšího přehledu o průběhu zkoušky. Po provedení všech filtrovacích struktur a následných výpočtů se lze dostat k finální vizualizaci formou grafů, ze kterých uživatel snadno rozpozná správnost zpracování. Výsledné tabulky, které již neobsahují takový objem dat, systém exportuje opět do souboru XLSX, který stačí uložit na správné místo, nebo jej rovnou použít pro interpretaci výsledků. Tím se nejen zvýší efektivita zpracování VDS, ale také uspoří náklady na zálohování dosažených výsledků pro pozdější porovnání.

4.3 Architektura a design systému

V této části budou postupně představeny oba filtrovací SW. První, kde jsou všechna nastavení pevně dána. Tento systém poskytuje pevný rámec, který autor následně rozšířil ve druhém filtrovacím SW o komunikaci s uživatelem skrze příkazové okno. Avšak bylo nutné zabezpečit systém proti zadání nesprávných informací, které by mohly vést k nežádoucím výsledkům.

Je důležité poznamenat, že kód, který je zde prezentován, není kompletní. Úplné kódy obou filtrovacích SW se nacházejí v příloze. V rámci bakalářské práce budou uvedeny důležité části kódu, zejména klíčové funkce a výstupy kódu, jako jsou grafy a tabulky.

4.3.1 1. filtrovací Software

Nahrání dat a příprava pro zpracování

```
% Nahrání excelu => MATLABU
Data = 'C:\SKOLA\Bakalarska_prace\
   Data_pro_BP_vyvoj_filtr_SW_komplet - kopie.xlsx';
% Třízení
[data, text, row] = xlsread(Data);
% Filtrování potřebných sloupců
% Číselné
select_col_num = data(:, [2 5:9]); % podle čísla sloupce (
   př.: [1 2 3...] nebo [1:3])
% Podle názvů
% Načtění prvního řádek, který obsahuje názvy sloupců
[~, col_name, ~] = xlsread(Data, 'A1:Z1');
% Názvy vybraných sloupců
col_names = {'Čas v zátěžném bloku', 'otáčky_LZ', '
   otáčky_PZ', 'Moment_LZ', 'Moment_PZ', 'Převodový stupeň'
   };
% Získání indexů sloupců podle názvů
index_col = ismember(col_name, col_names);
% Vybrání sloupců podle indexů
select_col_name = data(:, index_col);
% Unikátní hodnoty rychlostních stupňů
istgang(:,1) = unique(select_col_name(:, end));
```

Kód 4.1: Nahrání dat ze souboru XLSX a výběr potřebných sloupců (Zdroj: vlastní autora)

Tato část kódu se zabývá nahráním dat ze souboru XLSX do MATLABu a přípravou dat pro další zpracování. Nejprve jsou načtena data a poté vybrány potřebné sloupce pro další analýzu. Jak zde autor naznačil, sloupce lze vybrat buďto pomocí názvů, které se v datech (4.1) nachází v prvním řádku, nebo podle čísel sloupců, jak jsou uloženy v tabulce.

Funkce „ismember“ slouží k porovnání dvou množin a vrací logický vektor, kde „1“ označuje prvky, které jsou společné pro obě množiny. V tomto kódu je použita k porovnání názvů sloupců s názvy definovanými v proměnné „col_names“. Tímto způsobem autor vytvořil mechanismus, který může být v budoucnu rozšířen o dynamický výběr sloupců na základě jejich názvů po provedení úprav ve formátování dat v organizaci.

Následující část kódu využívá předchozí výběr sloupců, obsahující rychlostní stupně pro získání unikátních hodnot, což zajišťuje funkce „unique“. Tyto hodnoty jsou následně přiřazeny do prvního sloupce matice „istgang“.

Zpracování dat pro jednotlivé rychlostní stupně a pro kompletní převodový systém

```
% Načtení databáze převodových systémů
[Ratios, Type] = xlsread('Gearboxes.xlsx');
% Získání převodových poměrů pro jednotlivé rychlostní
   stupně
for i = 1:length(istgang)
    istgang(i,2) = Ratios(i+1,2);
end
% Výpočet a zpracování dat pro jednotlivé rychlostní stupně
for i = 1:length(istgang)
    % ... (výpočty a operace pro jednotlivé rychlostní
       stupně)
end
% Výpočet a zpracování dat pro kompletní převodový systém
% ... (výpočty a operace pro kompletní převodový systém)
for j = 1:length(Vysledek)
    %... vynásobení cykly a sčítání hodnot
end
```

Kód 4.2: Načtení dat pro získání převodových poměrů a provedení matematických operací (Zdroj: vlastní autora)

V následujícím úseku autor rozšiřuje SW o databázi „Gearboxes.xlsx“. Tato databáze obsahuje různé typy převodových systémů, společně s převodovými poměry, které jsou následně nahrány do proměnné „istgang“ k odpovídajícím rychlostem.

Při přepočtu otáček na otočení hřídele se vychází z otáček na výstupu, které jsou jako všechna data vzorkované 100 Hz. Při zpracování dat, zahrnujícím všechny rychlosti převodového systému, se s hodnotami počtu otočení i momentu hřídele, pracuje na výstupu ze systému. Naopak u počtu otočení a momentu hřídele u jednotlivých rychlostí, se data přepočítávají za pomocí převodových poměrů na hodnoty, které vstupují do převodového systému. Další operace, která je součástí sekce výpočtů a zpracování, je zjištění maximálních a minimálních hodnot momentu a vytvoření rozsahu.

Volba většího kroku v rozsahu momentu zpracovaných dat, které obsahují všechny rychlosti převodového systému, je zde z důvodu práce s momentem na výstupu z převodového systému. Na druhé straně je zpracování dat pro konkrétní rychlost převodového systému, kde se pracuje s momentem na vstupu do převodového systému. To je hezky vidět na grafech 4.1 a 4.2 nebo v tabulkách zpracovaných dat 4.2 a 4.3. Poté se vypočte a přiřadí suma otočení k rozsahu momentu, do kterého spadá.

V souboru „Vysledek“ se dále k dosavadnímu zpracování přidávají nové sloupce. V prvním sloupci se nachází suma otočení z předchozího sloupce, která je ale, vzhledem k povaze zkoušky vynásobena počtem cyklů, kterými je převodový systém zatížen.

Posledním sloupcem, který se do souboru přidává, je sloupec, kde probíhá kumulativní součet počtu otočení hřídele, vynásobeného cykly z předchozího kroku. Tento sloupec reprezentuje postupné kumulativní sčítání hodnot, kde každá hodnota je spočítána z hodnoty v aktuálním řádku a hodnoty v řádku předchozím. Pro kladné hodnoty rozsahu momentu (>=0), se začíná od nejvyšší hodnoty k nule. V případě prvního řádku, který nemá hodnotu nad sebou, se k aktuální hodnotě přičítá nula.

Pro záporné hodnoty rozsahu momentu (<0) je postup opačný. Zde se prochází soubor od nejnižší hodnoty a postupuje se směrem k nule.

Je důležité zdůraznit, že z důvodu zachování důvěrnosti dat poskytnutých organizací, jsou zde uvedeny pouze obecné informace o funkci kódu, aniž by byly zmiňovány detaily zpracování, které by mohly odhalit citlivé informace.

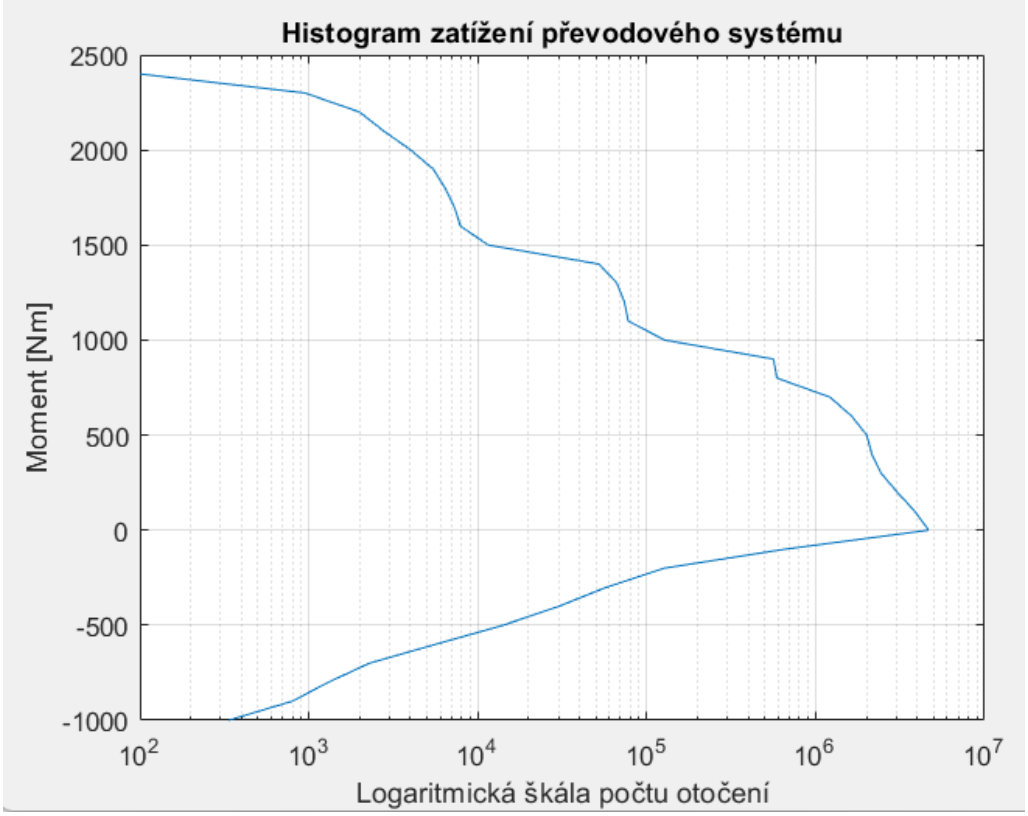
Vizualizace dat

```
x = Vysledek{1}(:, 4);
y = Vysledek{1}(:, 1);
figure;
semilogx(x, y);
title('Graf závislosti');
xlabel('Logaritmická škála posledního sloupce');
ylabel('První sloupec výsledku');
grid on;
figure;
for j = 4:length(istgang(istgang(:,1) > 0))+3
    subplot(2, 3, j-3);
    x = Vysledek{j}(:, 4);
    y = Vysledek{j}(:, 1);
    semilogx(x, y);
    title(['Graf závislosti rychlosti ' num2str(j)]);
    xlabel('Logaritmická škála posledního sloupce');
    ylabel('První sloupec výsledku');
    grid on;
end
```

Kód 4.3: Vizualizace zpracovaných dat (Zdroj: vlastní autora)

Tato část kódu zajišťuje vizualizaci dat pomocí grafů. První graf 4.1 zobrazuje závislost mezi hodnotami momentu a kumulativními součty otočení na logaritmické škále. Další graf 4.2 poskytuje detailní pohled na tuto závislost pro vybraný rychlostní stupeň s nižším krokem rozsahu momentu.

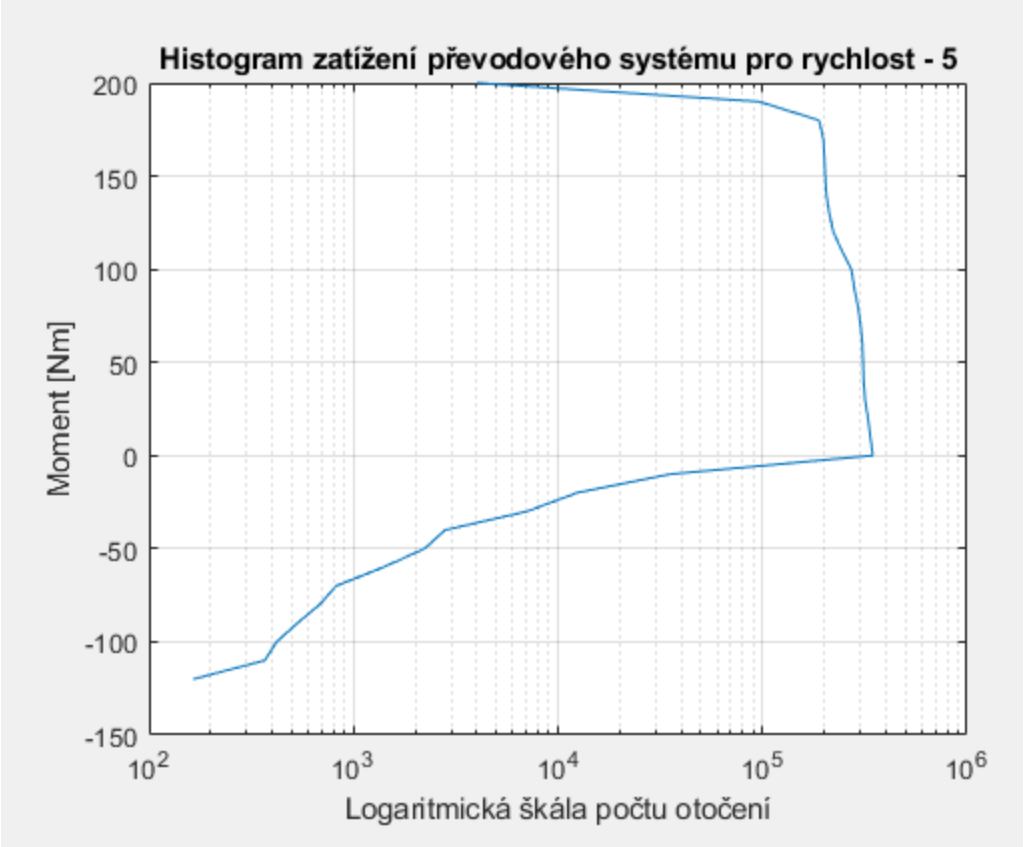
Tyto vizualizace slouží především k orientaci a porozumění datům. Je důležité, aby uživatel byl schopen na první pohled identifikovat případné anomálie nebo neobvyklé vzory, které by mohly indikovat problémy se zpracováním dat.



Obrázek 4.1: Histogram zatížení převodového systému přes všechny rychlosti (Zdroj: vlastní autora)

Graf 4.1 zobrazuje histogram zatížení převodového systému, který byl získán z cyklické zátěžové zkoušky. Tato zkouška slouží k posouzení odolnosti převodového systému a měří, jakou zátěž je schopen převodový systém vydržet v průběhu cyklického provozu. Může být využit k identifikaci oblastí s vysokou zátěží nebo neobvyklých vzorů, což může naznačovat problémy nebo nedostatky v převodovém systému.

Na ose Y je zobrazen rozsah momentu [Nm] a na ose X je otočení hřídele. Všechny tyto hodnoty jsou na výstupu z převodového systému. Osa X je vykreslena v logaritmické škále, což zlepšuje čitelnost grafu.



Obrázek 4.2: Histogram zatížení převodového systému konkrétní rychlosti (Zdroj: vlastní autora)

Na následujícím obrázku 4.2 je zobrazen další histogram zatížení převodového systému. Tento grafů reprezentuje zatížení při konkrétní rychlosti převodového systému. Tento přístup umožňuje detailní analýzu zatížení jednotlivých rychlostí a poskytuje jej uživateli. Detail grafů průběhů zatížení při zařazení různých rychlostí jsou k dispozici v příloze.

Export výsledků do souboru Excel

```
% Vytvoření nového souboru Excel
excelApp = actxserver('Excel.Application');
workbook = excelApp.Workbooks.Add;

% Iterace přes jednotlivé výsledky
for j = 1:length(Vysledek)
    % ... (zapisování výsledků do souboru Excel)
end

% Uložení a uzavření souboru
workbook.SaveAs('C:\SKOLA\Bakalarska_prace\Vysledny_soubor.
   xlsx');
workbook.Close(false);
excelApp.Quit();
```

Kód 4.4: Vytvoření nového souboru Excel (Zdroj: vlastní autora)

Tato část kódu se zabývá exportem výsledků zpracovaných dat do souboru XLSX pro následnou interpretaci a archivaci.

Nejprve je vytvořen nový soubor pomocí aplikace Excel, která je spuštěna pomocí funkce „actxserver“. Následně je vytvořena nová pracovní kniha (workbook), do které budou zapisovány výsledky. Proces exportu je realizován pomocí iterace přes jednotlivé výsledky v matici „Vysledek“. Během všech opakování jsou výsledky zapisovány do vytvořené pracovní knihy. Nakonec je soubor XLSX uložen dle zadané cesty a uzavřen pomocí příkazů „SaveAs“ a „Close“, což vede k ukončení aplikace Excel díky příkazu „Quit“.

Tento postup umožňuje uživateli snadno uchovat a sdílet výsledky zpracování dat ve formě souboru XLSX, což usnadňuje další analýzu a interpretaci dat.

Tabulka 4.2: Tabulka výsledku z filtrovacího SW - kompletní převodový systém (Zdroj: vlastní autora)

Moment [Nm]	**Počet otočení/s**	*** Cykly**	**Kumulativní součet**
2900	0.1508	74.9317	74.9317
2800	0.6048	300.5633	375.4950
2700	2.4837	1234.3952	1609.8902
2600	1.5678	779.1939	2389.0841
2500	1.4417	716.5311	3105.6152
2400	2.5718	1278.2051	4383.8202
2300	2.4499	1217.5900	5601.4102
2200	2.1855	1086.2050	6687.6152
2100	1.5538	772.2270	7459.8423
2000	1.6832	836.5615	8296.4038
1900	1.1041	548.7195	8845.1233
1800	1.8967	942.6843	9787.8076
1700	70.2073	34893.0066	44680.8143
1600	52.0910	25889.2101	70570.0243
1500	12.4585	6191.8725	76761.8968
1400	13.6123	6765.3116	83527.2084
1300	7.1941	3575.4665	87102.6749
1200	61.7070	30668.3694	117771.0443
1100	894.9885	444809.3050	562580.3492
1000	175.1051	87027.2594	649607.6086
900	53.4078	26543.6683	676151.2769
800	2012.5853	1000254.9087	1676406.1856
700	306.9890	152573.5032	1828979.6888
600	813.5284	404323.6109	2233303.2997
500	258.3208	128385.4452	2361688.7449
400	364.1209	180968.0926	2542656.8375
300	670.2526	333115.5567	2875772.3942
200	1336.1575	664070.2877	3539842.6819
100	1559.9384	775289.3933	4315132.0752
0	1559.9384	775289.3933	5090421.4686
-100	1155.1192	574094.2611	737921.7877
-200	150.3323	74715.1396	163827.5266
-300	88.0842	43777.8303	89112.3870
-400	36.1477	17965.4215	45334.5567
-500	27.5308	13682.8321	27369.1351
-600	14.8138	7362.4607	13686.3030
-700	7.3210	3638.5259	6323.8423
-800	1.8340	911.4815	2685.3163
-900	1.4055	698.5465	1773.8348
-1000	0.7361	365.8438	1075.2883
-1100	0.6731	334.5259	709.4446
-1200	0.7544	374.9187	374.9187

Tabulka 4.2 je uložená v proměnné „Vysledek“ společně s ostatními zpracovanými daty. Tato data jsou výsledkem zpracování pomocí nástroje navrženého autorem a jsou připravena k exportu do formátu XLSX, kde mohou být dále analyzována nebo archivována.

Tabulka 4.3: Tabulka výsledků z filtrovacího SW - zaměřeno na 5. rychlostní stupeň (Zdroj: vlastní autora)

Moment [Nm]	**Počet otočení/s**	*** Cykly**	**Kumulativní součet**
200	8.0663	4008.9691	4008.9691
190	185.3039	92096.0578	96105.0269
180	186.5344	92707.6152	188812.6422
170	19.3884	9636.0196	198448.6617
160	4.5847	2278.5858	200727.2476
150	3.7875	1882.4057	202609.6532
140	5.4647	2715.9378	205325.5911
130	13.6770	6797.4637	212123.0547
120	20.2195	10049.1116	222172.1664
110	44.4259	22079.6836	244251.8500
100	56.2267	27944.6681	272196.5181
90	17.6042	8749.3117	280945.8298
80	24.5024	12177.6821	293123.5119
70	17.2588	8577.6328	301701.1447
60	12.5556	6240.1559	307941.3007
50	4.8654	2418.1256	310359.4262
40	4.4681	2220.6577	312580.0839
30	9.8701	4905.4381	317485.5220
20	21.0833	10478.3840	327963.9060
10	18.2065	9048.6212	337012.5272
0	18.2065	9048.6212	346061.1485
-10	46.2169	22969.8055	353631.3359
-20	10.7565	5346.0003	363936.0000
-30	8.5476	4248.1709	368184.1709
-40	1.1548	573.9219	368758.0928
-50	1.6732	831.5980	369589.6908
-60	1.1476	570.3412	370160.0320
-70	0.2831	140.7077	370300.7397
-80	0.3050	151.5632	370452.3029
-90	0.2224	110.5341	370562.8370
-100	0.1054	52.3775	370615.2145
-110	0.4094	203.4925	370818.7070
-120	0.3316	164.8229	370983.5299

Tabulka 4.3 zobrazuje zpracování dat zaměřených na 5. rychlostní stupeň převodového systému. Všechny tabulky zpracovaných dat budou společně s grafy v příloze.

4.3.2 2. filtrovací SW rozšířený o UI

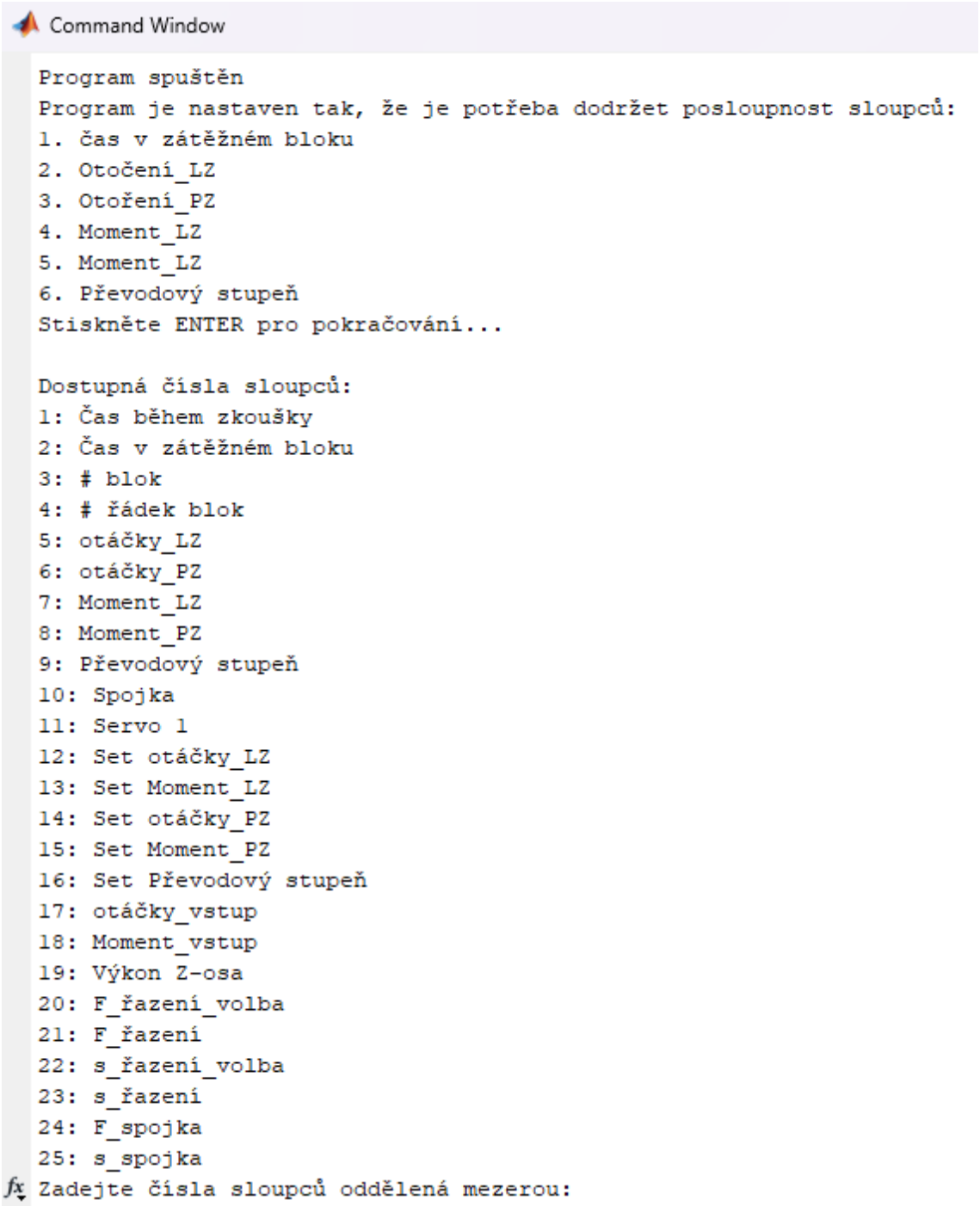
Nahrání a specifikace dat pro 2. filtrující software s uživatelským rozhraním

```
disp('Program spuštěn');

% Nahrání excelu => MATLABU
%... stejné jako u 1. SW

disp(['Program je nastaven tak, že je potřeba dodržet
   posloupnost sloupců:' char(10) '1. čas v zátěžném bloku'
    char(10) '2. Otočení_LZ' char(10) '3. Otoření_PZ' char
   (10) '4. Moment_LZ' char(10) '5. Moment_LZ' char(10) '6.
    Převodový stupeň']);
input('Stiskněte ENTER pro pokračování...');

% Číselné filtrování potřebných sloupce
% Načtění prvního řádku, který obsahuje názvy sloupců
[~, col_name, ~] = xlsread(Data, 'A1:Z1');

% Vypište čísla sloupců, aby si uživatel mohl vybrat
disp([char(10) 'Dostupná čísla sloupců:']);

for i = 1:length(col_name)
    fprintf('%d: %s\n', i, col_name{i});
end

% Zeptejte se uživatele na čísla sloupců, které chce vybrat
selected_col_numbers = input('Zadejte čísla sloupců
   oddělená mezerou: ', 's');
```

Kód 4.5: Načtení dat pro zpracování a následná komunikace (Zdroj: vlastní autora)

Tento úsek kódu se zabývá načtením dat ze souboru XLSX do MATLABu a následnou možností uživatele vybrat potřebné sloupce pro další zpracování. Po spuštění programu se nejprve zobrazí zpráva „Program spuštěn“ a následně se provede načtení dat ze zadaného souboru XLSX.

Poté uživatel dostane instrukce ohledně požadované posloupnosti sloupců, kterou je nutné dodržet. Vstupní dialog umožňuje uživateli stisknutím klávesy ENTER pokračovat dál (viz. 4.3).



Obrázek 4.3: Uživatelská interakce v Command Window: Výběr sloupců z XLSX souboru (Zdroj: vlastní autora)

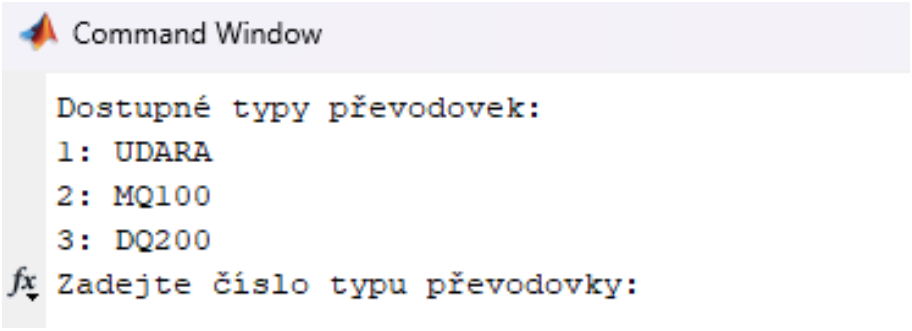
Následuje číselné filtrování potřebných sloupců. Program získá názvy sloupců z prvního řádku souboru XLSX a vypíše je uživateli spolu s jejich čísly. Uživatel je pak vyzván, aby zadal čísla sloupců oddělená mezerou, které chce vybrat pro další zpracování. Program zpracuje vstup uživatele a sloupce, které vybral a uloží je do nové proměnné k dalšímu zpracování.

Specifikace typu převodového systému pomocí uživatelského rozhraní

```
% Unikátní hodnoty rychlostních stupňů
istgang = unique(select_col(:, end));
% Načtení databáze převodových systémů
[Ratios, Type] = xlsread('Gearboxes.xlsx');
% Inicializace proměnné pro výběr typu převodovky
selectedType = [];
while isempty(selectedType) || selectedType == 0
    % Zobrazení dostupných typů uživateli
    disp('Dostupné typy převodovek:');
    for i = 1:length(Type)
        disp([num2str(i) ': ' Type{i}]);
    end
    % Uživatele na výběr typu převodovky
    selectedType = input('Zadejte číslo typu převodovky: ')
       ;
    % Kontrola, zda je vybrané číslo v platném rozmezí
    if selectedType >= 1 && selectedType <= length(Type)
        % Nyní můžeme pokračovat s vybraným typem
           převodovky
        Gearbox = Type{selectedType};
        % Stálý převod
        cons_ratio = Ratios(1, selectedType);
        % Získání převodových poměrů pro vybraný typ
        for i = 1:length(istgang)
            istgang(i, 2) = Ratios(i+1, selectedType);
        end
        % Zobrazit výsledné převodové poměry
        clc;
        disp(['Zadali jste následující převodové poměry pro
            typ ' Gearbox ':' char(10) 'Rychlosti: ' char
           (9) 'Převodové poměry: ']);
        for i = 1:length(istgang)
            disp([char(10) num2str(istgang(i,1)') char(9)
               char(45) char(9) num2str(istgang(i,2)')]);
        end
    else
        disp('Neplatný výběr. Prosím, zadejte platné číslo
           typu převodovky.');
    end
end
```

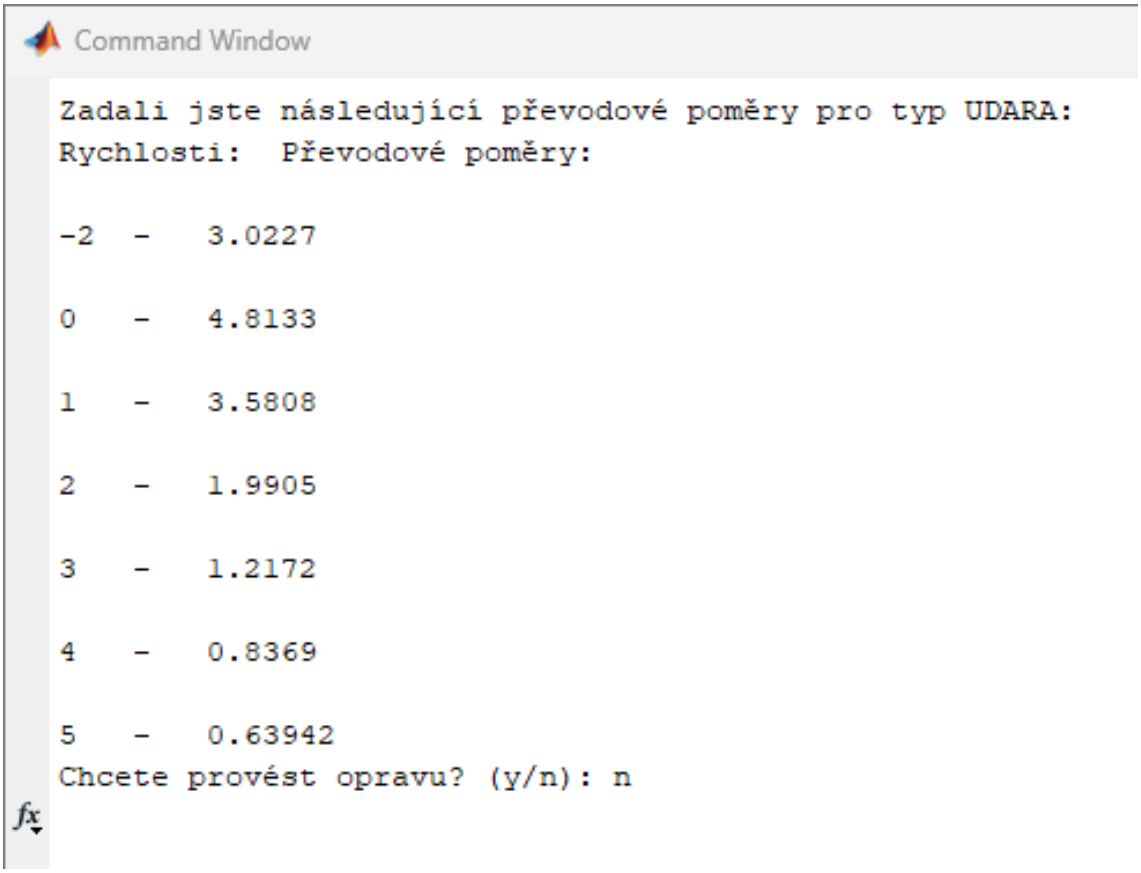
Kód 4.6: Přiřazení převodových poměrů k jednotlivým rychlostem na základě komunikace s uživatelem (Zdroj: vlastní autora)

Zde, podobně jako v případě prvního filtrovacího SW (viz sekce 4.3.1), program připravuje proměnnou obsahující unikátní hodnoty rychlostních stupňů převodového systému. Následně umožňuje uživateli vybrat typ převodového systému z nabídky, která je načtena ze souboru „Gearboxes.xlsx“. Tento soubor může obsahovat různé typy převodových systémů, které organizace využívá, a je navržen tak, aby mohl být snadno rozšiřován o nové typy nebo možnost aktualizovat staré. Program provádí kontrolu uživatelského vstupu, zda se drží v nabízeném rozmezí typů převodových systémů, a případně ho upozorní na chybu, pokud by výběr nebyl platný.



Obrázek 4.4: Uživatelská interakce v Command Window: Výběr typu převodovky (Zdroj: vlastní autora)

Po zadání typu převodového systému (viz. obrázek 4.4) program načte odpovídající převodové poměry pro jednotlivé rychlostní stupně a zobrazí je uživateli (viz. obrázek 4.5). Ten má možnost potvrdit správnost svého výběru, nebo si vybrat možnost opravit svůj výběr. Během této interakce program nadále monitoruje uživatelův vstup a zajišťuje, že se uživatel drží v platných možnostech výběru. Program již nadále pracuje stejně jako 1. filtrovací SW (viz sekce 4.3.1).



Obrázek 4.5: Uživatelská interakce v Command Window: Zobrazení převodových poměrů přiřazených k rychlostem (Zdroj: vlastní autora, data poskytnuta z organizace [16])

Toto rozšíření o UI výrazně zvýšilo komplexnost druhého filtrovacího SW, přičemž zároveň poskytuje robustní nástroj, který odolává změnám v datech. Kontrolované UI umožňuje zpracování dat i uživatelům s minimální znalostí problematiky. To umožňuje, aby i uživatelé, kteří nemají hlubší technické znalosti, snadno zpracovali data a předali je dále k analýze či porovnání s jinými převodovými systémy. Obecně lze říci, že jak první, tak druhý filtrovací SW výrazně usnadňuje vyhodnocení zkoušky a zvyšuje efektivitu práce s daty. Uživatel se tak může zaměřit čistě na analýzu systému a hledání potenciálních chyb a návrhů řešení.

5 Návrh efektivního řešení problémů velkých datových souborů v organizaci

V rámci organizace je práce s rozsáhlými datovými soubory nezbytná a často neoddělitelná součást každodenních činností. Efektivní zpracování dat je klíčové pro celkový úspěch. Aktuální postupy a nástroje, které organizace využívá, jsou důležitým krokem směrem k efektivnímu zpracování dat. Avšak, vzhledem k širokému spektru činností organizace, není možné očekávat dokonalost ve všech oblastech. Dynamika moderního prostředí vyžaduje neustálé optimalizace, aby se zachovala maximální efektivita zpracování dat.

Nástroj autora nejenom usnadňuje zpracování dat, ale může také sloužit jako inspirace pro budoucí inovace v organizaci. Vedle toho je zde prostor pro další obecné prvky, které mohou přispět k efektivnímu zpracování datových souborů, jako je zdokonalení metodik zpracování, vylepšení ukládání a přenosu dat, a využití nových technologií pro analýzu a vizualizaci dat. Efektivní manipulace s daty představuje základní kámen celého procesu. Pouze s tímto pevným základem je možné úspěšně identifikovat problémy, navrhnout vhodná řešení a následně využít zpracovaná data k dosažení stanovených cílů.

Závěrem lze říci, že i přes stávající úroveň efektivity je nezbytné neustále hledat nové přístupy a technologie, které budou odpovídat dynamice organizace a posunou její schopnosti v oblasti zpracování dat ještě dál.

5.1 Analýza dopadu filtrovacího softwaru na řešení problémů

Autorem navržený nástroj pro zpracování VDS byl vyvinut s cílem efektivně řešit identifikované problémy v organizaci v kapitole 3.3 spojené s rozmanitostí formátů a struktur dat, což komplikuje jejich zpracování a analýzu. Díky implementaci tohoto nástroje je dosaženo několika významných vylepšení v pracovním procesu organizace.

Konkrétně nástroj pomohl organizaci zlepšit efektivitu tím, že umožnil rychlejší a přesnější zpracování dat, což vedlo ke zkrácení doby potřebné k analýze a interpretaci dat. Díky tomu došlo k úspoře času a zvýšení produktivity pracovníků.

5.2 Integrace filtrovacího softwaru do celkového řešení

Filtrovací SW pro zpracování VDS je klíčovou součástí celkového navrhovaného řešení pro efektivní správu a analýzu dat. Jeho flexibilita a schopnost zpracovávat data, aniž by data byla předpřipravena do potřebných struktur, umožňuje snadnou obsluhu pro uživatele.

Nástroj přispívá k rychlejšímu a přesnějšímu zpracování dat, pomáhá organizaci dosáhnout strategických cílů v oblasti datové analýzy a rozhodování založeného na datech. Jeho schopnost zpracovat objemná data a poskytovat ihned na výstupu vizualizaci umožňuje uživateli rozpoznat správnost zpracování. Další výhodou je okamžitý export výsledných tabulek do formátu XLSX, ve kterém organizace uchovává data kvůli snadné interpretaci.

5.3 Implementační a provozní aspekty

Pro implementaci a integraci filtrovacího SW do stávající infrastruktury organizace bylo nezbytné provést několik kroků. Prvním krokem byla analýza současného stavu infrastruktury a identifikace potřebných úprav pro zavedení nástroje. Díky tomu, že se licence pro prostředí MATLAB se již v organizace nacházela, odpadla nutnost instalace nového nástroje.

Potenciální výzvy, které se mohou při implementaci vyskytnout, zahrnují odpor pracovníků vůči změnám a nutnost školení pro správné využití nástroje. Pro řešení těchto výzev autor navrhuje komunikaci s pracovníky a poskytnutí dostatečného školení a podpory.

Plán údržby a podpory pro nástroj zahrnuje pravidelné aktualizace softwaru, opravy chyb a poskytování technické podpory pro uživatele. Tímto způsobem bude zajištěna dlouhodobá efektivita a spolehlivost nástroje v organizaci.

5.4 Zhodnocení kvality vlastního přínosu

Autorem navržený SW pro zpracování VDS přináší do organizace přínosy, které významně ovlivňují pracovní proces a využití dat.

5.4.1 Zvýšení efektivity práce s daty

Díky minimálnímu využití uživatele a automatizaci většiny procesů zpracování dat dochází k výraznému zvýšení efektivity. Manuální úkoly jsou nahrazeny automatizovanými postupy, což umožňuje pracovníkům věnovat více času analýze a interpretaci dat namísto rutinním úkonům.

5.4.2 Zlepšení informačního základu pro rozhodování

Díky rychlému a robustnímu zpracování dat poskytuje SW organizaci jistou oporu o informace, na základě, kterých probíhá strategické rozhodování a plánování. To vede k posílení analytických schopností organizace, lepšímu porozumění datům a rychlejší odezvě na případná rizika.

Celkově lze konstatovat, že navržený SW pro zpracování VDS přináší organizaci významné výhody v podobě zvýšené efektivity, úspory času, lepšího využití dat pro strategické rozhodování a s tím spojené posílení konkurenceschopnosti. Jeho implementace má pozitivní dopad na pracovní procesy a pomáhá organizaci dosáhnout svých cílů v oblasti datové analýzy a správy. Dále usnadňuje ukládání výsledků, uživatel jednoduše nahraje nezpracovaná data, která SW zpracuje a rovnou uloží do souboru XLSX. Na konci celého procesu má uživatel původní neporušená data ze zkoušky a současně soubor s vyhodnocenými daty, která již nejsou tak objemná.

5.4.3 Návrh dalších inovativních řešení

Současný filtrovací SW má omezení pro vstupní formáty, a to pouze ve formátu XLSX. Autor by mohl zvážit rozšíření načítání dat i z jiných formátů, které se vyskytují v organizaci, jako například CSV, MF4 nebo BLF, pro zvýšení flexibility. Automatizace všech procesů by ještě více ušetřila čas a snížila riziko chyb, spojených s lidským faktorem. Tato rozšíření by mohla dále zlepšit efektivitu a uživatelskou přívětivost filtrovacího SW.

Závěr

Tato bakalářská práce se zaměřila na efektivní zpracování VDS prostřednictvím filtrovacího softwarového systému vyvinutého v programu MATLAB. Úvodní část práce poskytla čtenáři obecný úvod do problematiky zpracování VDS, zahrnující definici dat, koncept VDS (Big Data), a metody zpracování takovýchto dat. Tento úvodní rámec sloužil k tomu, aby čtenář získal základní povědomí o tématu a porozuměl důležitosti efektivního zpracování dat pro optimální fungování organizace.

Průzkum stavu problematiky VDS v organizaci podtrhl stávající výzvy a nedostatky v procesu zpracování dat. Identifikace těchto problémů poskytla důležitý vstup pro návrh řešení filtrovacího SW pro zpracování dat, který má za cíl automatizovat a zefektivnit práci s daty. Navržený nástroj podporuje matematické operace s daty, vytváří vizualizace výsledků a umožňuje export zpracovaných dat do formátu XLSX.

Přestože současný filtrovací SW pro zpracování dat přináší technologický pokrok a výhody pro rychlejší a přesnější analýzu dat, existují oblasti, které by mohly být vylepšeny. Doporučené inovace zahrnují rozšíření možností načítání dat z různých formátů a úplnou automatizaci procesu zpracování dat. Tyto návrhy by mohly vést k dalšímu zlepšení efektivity a flexibilitě filtrovacího softwaru.

Celkově lze konstatovat, že práce přinesla užitečné poznatky a návrhy pro zpracování VDS v organizaci. Doporučená inovativní rozšíření by mohla dále zlepšit efektivitu a flexibilitu filtrovacího softwaru. Další možná témata pro výzkum a rozvoj by mohla zahrnovat testování a ověření implementace navržených inovací v reálném prostředí organizace. Tím by bylo dosaženo zajištění dalšího technologického pokroku v oblasti zpracování VDS.

Použitá literatura

[1] ANDERSON, Benjamin a Brad NICHOLSON. *SQL vs. NoSQL Databases: What's the Difference?* 2022. Dostupné také z: https://www.ibm.com/blog/sql-vs-nosql/.

[2] *Apache Spark - unified analytics engine for big data.* [B.r.]. Dostupné také z: https://spark.apache.org.

[3] CICHY, Corinna a Stefan RASS. An Overview of Data Quality Frameworks. *IEEE Access.* 2019, roč. 7, s. 24634–24648. Dostupné z DOI: 10.1109/ACCESS.2019.2899751.

[4] COX, Michael a David ELLSWORTH. Application-controlled demand paging for out-of-core visualization. In: 1997, s. 235–244. ISBN 0-8186-8262-0. Dostupné z DOI: 10.1109/VISUAL.1997.663888.

[5] EXCELSIOR, Get. *Big Data, Explained: The 5V s of Data.* 2022. Dostupné také z: https://www.linkedin.com/pulse/big-data-explained-5v-excelsiorites.

[6] FOUNDATION, Apache Software. *Apache Hadoop.* 2019. Dostupné také z: https://hadoop.apache.org/.

[7] GEEKSFORGEEKS. *Big Challenges with Big Data.* 2019. Dostupné také z: https://www.geeksforgeeks.org/big-challenges-with-big-data/?ref=header_search.

[8] GEEKSFORGEEKS. *R Tutorial | Learn R Programming.* 2024. Dostupné také z: https://www.geeksforgeeks.org/r-tutorial/?ref=shm.

[9] GEEKSFORGEEKS. *5 V's of Big Data.* 2023. Dostupné také z: https://www.geeksforgeeks.org/5-vs-of-big-data/.

[10] GOUR, Shivani, Vijay MUTHEKAR a Amol SANER. A Comparative Study of ANN Models developed for predicting Soil Compaction Parameters using MS Excel and MATLAB. 2022, roč. 9, s. 1811–1816.

[11] HUSAMALDIN, Laden a N. SAEED. Big Data Analytics Correlation Taxonomy. *Information.* 2019, roč. 11, s. 17. Dostupné z DOI: 10.3390/info11010017.

[12] CHEN, Min, Shiwen MAO a Yunhao LIU. Big Data: A Survey. *Mobile Networks and Applications.* 2014, roč. 19, č. 2, s. 171–209. Dostupné z DOI: https://doi.org/10.1007/s11036-013-0489-0.

[13] IBM. *What is artificial intelligence (AI)?* 2023. Dostupné také z: https://www.ibm.com/topics/artificial-intelligence.

[14] JAIN, Anil. *The 5 V's of big data - Watson Health Perspectives*. 2016. Dostupné také z: https://www.ibm.com/blogs/watson-health/the-5-vs-of-big-data/.

[15] JONÁK, Zdeněk. Data. *KTD: Česká terminologická databáze knihovnictví a informační vědy (TDKIV)*. 2003. Dostupné také z: https://aleph.nkp.cz/F/?func=direct&doc_number=000000442&local_base=KTD.

[16] KOL. AUTORŮ. *Poskytnutá data pro vývoj filtrovacího softwaru*. 2023. Tech. zpr. Organizace A.

[17] KUDYBA, Stephan. *Big Data, Mining and Analytics: Components of Strategic Decision Making*. 2014. ISBN 1466568704. Dostupné z DOI: 10.1201/b16666.

[18] LLC, Developers Den. *Future of Big Data Trends 2024 and Essential Big Data Technologies You Must Explore*. 2023. Dostupné také z: https://www.linkedin.com/pulse/future-big-data-trends-2024-essential-technologies-you-must-nh85c.

[19] LOHR, Steve. *The Origins of "Big Data": An Etymological Detective Story*. 2013. Dostupné také z: https://archive.nytimes.com/bits.blogs.nytimes.com/2013/02/01/the-origins-of-big-data-an-etymological-detective-story/.

[20] MACURA, Marek. INTEGRATION OF DATA FROM HETEROGENEOUS SOURCES USING ETL TECHNOLOGY. *Computer Science*. 2014, roč. 15, č. 2, s. 109. Dostupné z DOI: 10.7494/csci.2014.15.2.109.

[21] MAREŠ, Milan. *ZDROJE INFORMACE A JEJÍ MĚŘENÍ*. 2018. Dostupné také z: https://docplayer.cz/161264417-Zdroje-informace-a-jeji-mereni.html.

[22] MATHWORKS. *Data Analysis – MATLAB & Simulink*. [B.r.]. Dostupné také z: https://www.mathworks.com/products/matlab/data-analysis.html.

[23] MELANIE. *The Impact of the DIKW Pyramid on Corporate Success*. 2023. Dostupné také z: https://datascientest.com/en/the-impact-of-the-dikw-pyramid-on-corporate-success.

[24] MICROSOFT. *Microsoft Excel, Spreadsheet Software*. 2024. Dostupné také z: https://www.microsoft.com/en-us/microsoft-365/excel.

[25] MICROSOFT. *Power BI - Data Visualization | Microsoft Power Platform*. 2024. Dostupné také z: https://www.microsoft.com/en-us/power-platform/products/power-bi.

[26] MURIITHI, Joe. *Data Processing in Python*. 2024. Dostupné také z: https://www.accel.ai/anthology/2022/3/24/data-processing-in-python.

[27] NELSON, Daniel. *Strukturovaná vs nestrukturovaná data – Unite.AI*. 2020. Dostupné také z: https://www.unite.ai/cs/structured-vs-unstructured-data/.

[28] OWEN-HILL, Alex. *Python vs C++ vs C# vs MATLAB: Which Robot Language is Best?* 2018. Dostupné také z: https://robodk.com/blog/robot-programming-language/.

[29] OZGUR, Ceyhun et al. MatLab vs. Python vs. R. *Journal of data science: JDS*. 2016, roč. 15, s. 355–372. Dostupné z DOI: 10.6339/JDS.201707_15(3).0001.

[30] PARTIDA, Devin. *Big Data Challenges*. 2023. Dostupné také z: https://www.datamation.com/big-data/big-data-challenges/.

[31] ROBINSON, Scott. *5V's of big data*. 2023. Dostupné také z: https://www.techtarget.com/searchdatamanagement/definition/5-Vs-of-big-data?vgnextfmt=print.

[32] RUSSELL, S.J., S. RUSSELL a P. NORVIG. *Artificial Intelligence: A Modern Approach*. Pearson, 2020. Pearson series in artificial intelligence. ISBN 9780134610993. Dostupné také z: https://books.google.cz/books?id=koFptAEACAAJ.

[33] SCHMELZER, Ronald. *Top Trends in Big Data for 2024 and Beyond*. 2024. Dostupné také z: https://www.techtarget.com/searchdatamanagement/feature/Top-trends-in-big-data-for-2021-and-beyond.

[34] SKILLMEA. *11 nejlepších nástrojů pro analýzu dat*. 2021. Dostupné také z: https://skillmea.cz/blog/11-najlepsich-nastrojov-na-analyzu-dat.

[35] UWAJE, BLESSING NKEM. *Choosing the Best Cloud Platform for Data Warehousing*. 2023. Dostupné také z: https://medium.com/@buwaje/over-the-years-data-warehousing-has-proven-to-be-crucial-for-business-intelligence-6b1c0cd86023.

A Přílohy

A.1 BP_Grafy_a_tabulky

Obsahuje výstupy grafů a tabulek z nástroje pro zpracování dat navrženého autorem.

A.2 BP_Filtrovací_SW_verze_1

Obsahuje první verzi filtrovacího softwaru pro zpracování dat navrženého autorem

A.3 BP_Filtrovací_SW_verze_2

Obsahuje druhou verzi filtrovacího softwaru s rozšířeným uživatelským rozhraním, které komunikuje s uživatelem přes okno pro zadávání příkazů.