VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
BRNO UNIVERSITY OF TECHNOLOGY

FAKULTA INFORMAČNÍCH TECHNOLOGIÍ
FACULTY OF INFORMATION TECHNOLOGY

ÚSTAV INFORMAČNÍCH SYSTÉMŮ
DEPARTMENT OF INFORMATION SYSTEMS

ANALÝZA A VIZUALIZACE DAT HROMADNÉ DOPRAVY MĚSTA BRNA

ANALYSIS AND VISUALISATION OF BRNO PUBLIC TRANSPORT DATA

DIPLOMOVÁ PRÁCE
MASTER'S THESIS

AUTOR PRÁCE — **Bc. JURAJ LAZÚR**
AUTHOR

VEDOUCÍ PRÁCE — **Ing. JIŘÍ HYNEK, Ph.D.**
SUPERVISOR

BRNO 2023

VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
FAKULTA INFORMAČNÍCH TECHNOLOGIÍ

Zadání diplomové práce

146479

Ústav:	Ústav informačních systémů (UIFS)
Student:	**Lazúr Juraj, Bc.**
Program:	Informační technologie a umělá inteligence
Specializace:	Informační systémy a databáze
Název:	**Analýza a vizualizace dat hromadné dopravy města Brna**
Kategorie:	Informační systémy
Akademický rok:	2022/23

Zadání:

1. Seznamte se s problematikou zpracování a vizualizace dat z hromadné dopravy. Zaměřte se na data reprezentující zpoždění vozidel. Proveďte průzkum existujících systémů určených pro tento účel.
2. Prostudujte oblast zpracování velkých objemů dat (*big data*) a databázových systémů určených pro tento účel.
3. Seznamte se s veřejnými daty reprezentující reálné polohy a časy vozidel hromadné dopravy města Brna poskytnuté Magistrátem města Brna (MMB) – Oddělením dat, analýz a evaluací. Ve spolupráci s MMB analyzujte požadavky na vyhodnocování zpoždění vozidel hromadné dopravy.
4. Navrhněte systém pro vyhodnocování a analýzu zpoždění vozidel hromadné dopravy města Brna.
5. Navržený systém implementujte.
6. Otestujte funkčnost a použitelnost systému ve spolupráci s MMB a vybraným vzorkem uživatelů.

Literatura:

- Kuo, Y. H., Leung, J. M., & Yan, Y. (2022). Public transport for smart cities: Recent innovations and future challenges. *European Journal of Operational Research*.
- Torre-Bastida, A. I., Del Ser, J., Laña, I., Ilardia, M., Bilbao, M. N., & Campos-Cordobés, S. (2018). Big Data for transportation and mobility: recent advances, trends and challenges. *IET Intelligent Transport Systems*, *12*(8), 742-755.
- Johnson, J. (2010). *Designing with the Mind in Mind: Simple Guide to Understanding User Interface Design Guidelines*. Morgan Kaufmann Publishers/Elsevier, ISBN: 9780123750303.
- Tufte, E. (2001). *The visual display of quantitative information*. Cheshire, USA: Graphics Press, ISBN 978-0-9613921-4-7.

Při obhajobě semestrální části projektu je požadováno:
Body 1 až 4.

Podrobné závazné pokyny pro vypracování práce viz https://www.fit.vut.cz/study/theses/

Vedoucí práce:	**Hynek Jiří, Ing., Ph.D.**
Vedoucí ústavu:	Kolář Dušan, doc. Dr. Ing.
Datum zadání:	1.11.2022
Termín pro odevzdání:	17.5.2023
Datum schválení:	21.10.2022

Abstrakt

Plánovanie a správa moderných systémov hromadnej dopravy si vyžaduje znalosť správania systémov v reálnom prostredí. Cieľom tejto práce je zjednodušenie postupu získavania užitočných znalostí z dát o meškaní vozidiel v systéme hromadnej dopravy mesta Brna. Riešenie problému spočívalo v návrhu a implementácií nástroja, ktorý bude tieto dáta autonómne spracovávať a analyzovať. Výstupy generované implementovaným nástrojom poskytujú podporu pri plánovaní, implementácií a vyhodnocovaní zmien v systéme hromadnej dopravy. Zároveň umožňujú detekciu kritických miest a opakujúcich sa vzorov v systéme. Výsledky tejto práce umožňujú zefektívniť fungovanie systému hromadnej dopravy a majú slúžiť k zvyšovaniu spokojnosti cestujúcej verejnosti.

Abstract

The planning and management of modern public transport systems requires knowledge of the behaviour of the systems in a real environment. The aim of this thesis is to simplify the procedure of extracting useful knowledge from vehicle delay data in the Brno public transport system. The solution to the problem consisted in the design and implementation of a tool that will autonomously process and analyze this data. The outputs generated by the implemented tool provide support for planning, implementation and evaluation of changes in the public transport system. They also allow the detection of critical points and recurring patterns in the system. The results of this work allow for a more efficient operation of the public transport system and should serve to increase the satisfaction of the travelling public.

Kľúčové slová

dáta, vizualizácia, hromadná doprava, GTFS, routing, spracovanie dát

Keywords

data, visualisation, public transit, GTFS, routing, data processing

Citácia

LAZÚR, Juraj. *Analýza a vizualizace dat hromadné dopravy města Brna.* Brno, 2023. Diplomová práce. Vysoké učení technické v Brně, Fakulta informačních technologií. Vedoucí práce Ing. Jiří Hynek, Ph.D.

Analýza a vizualizace dat hromadné dopravy města Brna

Prehlásenie

Prehlasujem, že som túto diplomovú prácu vypracoval samostatne pod vedením pána Ing. Jiřího Hynka Ph.D. Uviedol som všetky literárne pramene, publikácie a ďalšie zdroje, z ktorých som čerpal.

.....
Juraj Lazúr
14. mája 2023

Poďakovanie

V prvom rade by som rád poďakoval svojim rodičom za ich podporu a úžasnú výchovu, bez ktorej by som štúdium na vysokej škole a tvorbu tejto práce určite zvládol len veľmi ťažko. Rovnako by som chcel poďakovať celej svojej rodine za úžasné prostredie, v ktorom môžem žiť. Obrovská vďaka patrí vedúcemu mojej diplomovej práce Ing. Jiřímu Hynkovi, Ph.D. za neskutočné množstvo času, ktorý mi venoval počas konzultácií, za poskytnutie cenných rád a študijných materiálov k práci, a za povzdudivé slová keď sa moja práca zasekla na mŕtvom bode. Bez jeho podpory by som túto prácu nikdy nedokončil. Taktiež by som rád poďakoval Mgr. Martinovi Dvořakovi a Mgr. Janovi Zvarovi, Ph.D. z magistrátu mesta Brna za ich čas, cenné rady a kritické postrehy k mojej práci. Moje poďakovanie patrí aj doc. Ing. Radkovi Burgetovi, Ph.D. za možnosť ostestovať moju prácu v reálnych podmienkach.

Obsah

Kapitola 1

Úvod

Nasadzovanie informačných systémov nám umožňuje návrh a riadenie systémov, ktoré by bez použitia informačných technológií neboli s použitím iba ľudskej sily schopné fungovať v reálnom čase. Dokážeme tak spracovávať obrovské množstvá požiadaviek, vstupných dát, distribuovať informácie v reálnom čase. V kontexte verejnej dopravy sa jedná najmä o oblasti plánovania a prevádzky. Používaním informačných systémov však zároveň vzniká cenný zdroj dát, popisujúci správanie skutočného systému v reálnom čase.

Dáta popisujúce správanie systémov hromadnej dopravy sa používajú v dvoch kontextoch, v krátkodobom a dlhodobom časovom horizonte. Pri krátkodobom využití sa s nimi cestujúci stretávajú v podobe vizualizácií na zastávkových optických systémoch, alebo na mapách pohybu vozidiel v reálnom čase. Cestujúci tak môže meniť svoj plán využitia dopravného systému podľa aktuálnej situácie, čo mu v konečnom dôsledku umožňuje efektívnejšie používanie dopravného systému. Podobným spôsobom používajú tieto dáta aj dopravné podniky, koordinátori dopravy. Na základe dát dokážu reagovať na vzniknuté problémy rýchlo a efektívne, či už v podobe odklonu dopravy, zaistenia náhradných spojov, alebo informovania cestujúcich. V dlhodobom horizonte sa tieto dáta používajú pre detekciu vzorov správania, pri plánovaní zmien, a vyhodnocovaní efektivity daného systému.

Zámermi mojej práce sú zlepšenie postupu získavania užitočných znalostí z dát o meškaní, a zjednodušenie prístupu k týmto znalostiam širšej skupine používateľov. Cieľom je navrhnúť a implementovať systém schopný spracovávať a analyzovať tento typ dát. Zamýšľaný systém by mal byť prehľadný, jednoduchý na používanie a čo najviac robustný, aby bol schopný samostatne spracovávať čo najširšiu triedu skutočných dát. Výstupy systému by mali byť ľahko interpretovateľné aj bez hlbších znalostí danej problematiky. V konečnom dôsledku by tak moja práca mala prispieť k zefektívneniu celého systému verejnej dopravy a tým zvýšiť spokojnosť cestujúcej verejnosti.

Kapitola 2 popisuje históriu, vplyv a spôsoby využívania informačných systémov vo verejnej doprave. Rovnako je v tejto kapitole spracovaný prehľad používaných nástrojov pracujúcich s dátami o meškaní. V kapitole 3 je uvedný výber charakteristických vlastností veľkých objemov dát, metódy ich spracovania a vizualizácie, a príklady použitia týchto metód v praxi. Kapitola 4 obsahuje analýzu používateľských požiadaviek, samotného systému hromadnej dopravy mesta Brna, ako aj technické požiadavky na zamýšľané riešenie, a definuje problémy, ktoré bolo nutné vyriešiť. Kapitola 5 podrobne popisuje navrhnutý model dát, návrh architektúry systému, a jeho rozdelenie na jednotlivé moduly. V kapitole 6 je popísaná implementácia samotného systému spolu s niekoľkými významnými modulmi systému. Posledná kapitola 7 sa zaoberá skúšobným nasadením systému, používateľskými testami a ich výsledkami.

Kapitola 2

Systémy verejnej dopravy z pohľadu informačných technológií

Informačné technológie sú v doprave využívané viac ako 50 rokov [17]. Pomáhajú riadiť, plánovať a spravovať všetky druhy dopravy. Ponúkajú riešenia, ako čeliť stále sa zvyšujúcim nárokom na rýchlosť, dostupnosť a odolnosť voči chybám. Vďaka nim je možné ušetriť obrovské množstvo času pri preprave komodít a tovaru, ako aj pri cestovaní nielen verejnou dopravou. Nasadzovanie takýchto systémov do praxe však bolo na začiatku pomalé a opatrné. Dopravné podniky spočiatku neboli ochotné investovať obrovské množstvá peňazí do nových technológií, o ktorých účinkoch nemali žiadne vedomosti. Spolu s obavami, že poskytovanie nepresných informácií by mohlo znížiť počet cestujúcich, prichádzalo k zavádzaniu inovácií len veľmi pozvoľna [31].

Zásluhy za rozšírenie používania informačných a komunikačných technoogií vo verejnej doprave je možné pripísať najmä zníženiu ich zaobstarávacích nákladov v posledných troch dekádach. Cieľom pri ich zavádzaní bolo zvyšovať efektivitu systému, znižovať prevádzkové náklady a prilákať nových cestujúcich. S postupným zavádzaním týchto systémov do prevádzky začali vznikať štúdie [19, 22], ktoré potvrdzovali pozitívny dosah týchto systémov. Práve takéto štúdie sa stali základom zmeny, vďaka ktorej sú dnes informačné technológie neoddeliteľnou súčasťou moderných systémov verejnej dopravy.

Dopravné podniky a autority využívajú informačné technológie v dvoch úzko prepojených oblastiach. Prvou je plánovanie prevádzky a druhou je samotné riadenie prevádzky. Kým prvá oblasť pokrýva úlohy ako plánovanie trás, návrh cestovných poriadkov, grafikonu a plánovanie pracovných turnusov, druhá sa zaoberá distribúciou informácií cestujúcim a riešením neočakávaných udalostí počas prevádzky [31].

Nástroje implementujúce prvú oblasť pracujú hlavne so statickými dátami. Na rozdiel od toho nástroje pokrývajúce druhú oblasť kombinujú dáta statické s dátami získanými v reálnom čase. Nástroje patriace do druhej oblasti odborná literatúra zvykne súhrnne označovať aj ako Systémy automatického monitorovania vozidiel, skrátene AVMS (Automatic Vehicle Monitoring Systems). Opätovným využitím dát z reálnej prevádzky dokážeme docieliť kontinuálne zlepšovanie efektivity a poskytovaných služieb daného dopravného systému. Inými slovami, systém a jeho úpravy plánujeme na základe dát z reálnej prevázky. Keďže sa však táto práca zaoberá spracovaním dát v reálom čase, ďalej sa venujem iba oblasti riadenia prevádzky.

2.1 Systémy automatického monitorovania vozidiel

Systémy kategórie AVMS pokrývajú celú oblasť riadenia prevádzky. Napriek rôznorosti týchto systémov, je možné rozdeliť ich z hľadiska cieľového používateľa na dve kategórie. Prvá kategória týchto systémov slúži na distribuovanie informácií cestujúcim, kým druhá kategória pomáha dispečerom riešiť neočakávané udalosti a zbiera najrôznejšie dáta z prevádzky určené pre ďalšie použitie. Avšak zatiaľ čo prvá dáta sprístupňuje čo najviac používateľom, druhá udržiava dáta striktne vnútri systému. Obe kategórie však primárne pracujú s dátami v reálnom čase a v konečnom dôsledku majú za cieľ zvyšovať spokojnosť cestujúcej verejnosti.

2.1.1 Pokročilé informačné systémy pre cestujúcich

Cestujúcim spôsobuje používanie systémov verejnej dopravy väčšiu mieru neistoty, ako individuálna doprava. To je spôsobé najmä stochastickou povahou poskytovaných služieb a celkovou zložitosťou týchto systémov [13]. Preto sa dopravné podniky a autority snažia tieto faktory eliminovať napríklad aj použitím Pokročilých informačných systémov pre cestujúcich, skrátene ATIS (Advanced Traveller Information Systems). Ich primárnym cieľom je čo najviac sprístupniť a zjednodušiť odovzdávanú informáciu.

Pozitívny vplyv na zvyšovanie spokojnosti cestujúcich popisuje množstvo článkov a výskumov. Napríklad Dziekan a Kottenhoff [8] uvádzajú, že využívanie takýchto nástrojov zmenšuje čakania, zvyšuje pocit istoty a spokojnosť cestujúcich, a celkovo zlepšuje efektivitu využívania systému. Taktiež existujú výskumy, podľa ktorých ak majú cestujúci tieto informácie k dispozícií, sú ochotnejši platiť cestovné lístky [26].

Do tejto kategórie tak patria systémy, ktorých výstup je intuitívny, prehľadný a nesie okamžitú informáciu. Ide napríklad o interaktívne mapy, plánovače ciest, najrôznejšie schémy linkového vedenia, optické informačné systémy vo vozidlách, alebo na zastávkach ako na obrázku 2.1. Práve takéto optické systémy obsahujúce smer a čas do odchodu, častokrát obsahujúci aj údaj o aktuálnom meškaní, cestujúci najviac využívajú ako zdroj informácií [8].

Obr. 2.1: **Optický systém.** Optické informačné systémy sú jedním z najobľúbenejších spôsobov, akým cestujúci získavajú informácie.[2]

[2]Prevzaté z https://commons.wikimedia.org/wiki/File:WMATA_King_Street_PIDS.jpg

2.1.2 Monitorovacie systémy

Do kategórie AVMS patria okrem nástrojov typu ATIS aj nástroje, ktoré slúžia pre potreby riadenia prevádzky dopravných systémov. Na jednej strane sem patria nástroje pre monitorovanie aktuálnej situácie v systéme, stavu vozového parku a komunikačné nástroje. Na strane druhej sa jedná o nástroje, ktoré zaznamenávajú rôzne údaje z reálnej prevádzky. Tieto dáta sú potom využívané ako vstup pre plánovacie nástroje.

Moderné systémy pre správu vozidlového parku sú v dnešnej dobe široko používané. Dopravné podniky ich využívajú predovšetkým pre sledovanie technického stavu vozidiel a plánovanie prehliadok a opráv. Takéto nástroje [10] dokážu zvyšovať efektivitu a šetriť ľudské a finančné zdroje, ktoré môžu byť presmerované na dôležitejšie úlohy. Rovnako tak môžu dispečeri vďaka systémom pracujúcim s aktuálnymi dátami efektívne riešiť neočakávané udalosti ako dopravné nehody, poruchy vozidel alebo infraštruktúry. Rovnako vďaka komunikačným systémom a palubným počítačom je implementácia zmien v systéme, alebo informovanie cestujúcich o akuálnych mimoriadnostiach oveľa jednoduchšie.

Samostatnú skupinu tvoria nástroje zachytávajúce dáta z reálnej prevádzky. Plánovanie dopravnej obslužnosti veľkých dopravných systémov totiž vyžaduje veľké množstvo kvalitných prevádzkových dát [27]. Bez týchto dát nie je možné systém efektívne vylepšovať a sledovať dopad nasadených zmien. Najvhodnejším a najefektívnejším prístupom je tieto dáta získavať a spracovávať automaticky. Keďže sa tieto dáta menia v čase, je nevyhnutné ich zaznamenávať a zbierať periodicky. Zdrojom týchto dát sú predovšetkým vozidlá prevádzkované v danom systéme. Dokážeme tak získať informácie o obsadenosti vozidiel, meniacich sa jazdných časoch medzi zastávkami v závislosti od dennej doby, určovať dopravné prúdy, či sledovať meškanie vozidiel v reálnom čase.

2.2 Spôsoby využívania dát o meškaní

Dáta o meškaní v systémoch verejnej dopravy spolu s aktuálnou polohou vozidiel patria medzi cestujúcimi k najčastejšie požadovaným informáciam. Najvhodnejšou metódou zaznamenávania a poskytovania týchto dát je zameranie sa na meranie rozdielu medzi aktuálnou polohou a polohou interpolovanou na základe grafikonu. Tento prístup zároveň najviac vyhovuje cestujúcim [21].

Okrem cestujúcich využívajú dáta o meškaní aj pracovníci dopravných podnikov. Na základe cieľovej skupiny používateľov je tak možné rozdeliť jednotlivé spôsoby využitia dát na dva prístupy. Prvý prístup pracuje iba s aktuálnymi hodnotami v reálnom čase. Jeho hlavnou úlohou je sprístupniť danú informáciu tak, aby bola okamžite využiteľná a čo najviac dostupná. Tento prístup spravidla neukladá a ani nepracuje s historickými dátami. Dáta o meškaní nikdy nie sú súčasťou obsahu hlavnej informácie, sú iba pridanou informačnou hodnotou. Hlavnú skupinu cieľových používateľov v tomto prístupe tvoria cestujúci. Príkladom tohto prístupu sú nástroje vizualizujúce polohu a aktuálne meškanie vozidiel ako napríklad TRAVIC[3], alebo mapa integrovaného dopravného systému Jihomoravského kraje[4].

Druhý prístup sa špecializuje na analýzu dát o meškaní. Dáta sú spracovávané a ukladané tak, aby bolo možné analyzovať dlhšie časové úseky. Nástroje využívajúce tento prístup teda pracujú najmä s historickými dátami. Cieľom tohto prístupu je hľadanie vzorov, opakujúcich sa problémov a získavanie znalostí. Implementované nástroje pracujú s dátami

[3] https://travic.app/
[4] https://mapa.idsjmk.cz/

o meškaní ako primárnou informáciou, pričom dané informácie spravidla kombinujú s ďalšími dátami najčastejšie geografického charakteru. Cieľovými používateľmi sa tak stávajú hlavne analytici a technický pracovníci dopravných podnikov a autorít. Tento prístup využíva napríklad nástroj Babitron[5], sledujúci štatistiky meškania vlakov na sieti SŽDC.

2.3 Aktuálne používané nástroje

Ako už bolo napísané, vďaka rozšíreniu informačných technológií v tejto oblasti vzniká obrovské množstvo nástrojov, ktoré často presahujú funkcionality popisované kategóriou AVMS. Vzhľadom na túto rôznorodosť obsahuje táto sekcia iba krátky prehľad reprezentatívnej vzorky nástrojov, rozdelených podľa účelu popisovanému v sekcii 2.2. V sekciách 2.3.1 a 2.3.2 je postupne popísaných päť nástrojov, dva využívajúce prvý prístup, jeden nástroj čiastočne kombinujúci oba prístupy, a dva nástroje využívajúce druhý prístup k dátam o meškaní.

2.3.1 Vizualizačné nástroje

Implementujú prvý prístup využívania dát o meškaní. V dnešnej dobe sú široko používané a ľahko dostupné. Ich výstupy predstavujú najčastejší formát dát o meškaní, s ktorým sa stretáva bežný cestujúci. Ako reprezentatívny príklad som sa rozhodol v mojej práci analyzovať dva nástroje tohto druhu. Prvým je švajčiarsko-nemecký nástroj TRAVIC a druhým zástupcom je mapa IDS JMK, ktorá zobrazuje dopravný systém Jihomoravského kraja. Tretí prezentovaný nástroj, Cestovné poriadky, čiastočne kombinuje oba popisované prístupy.

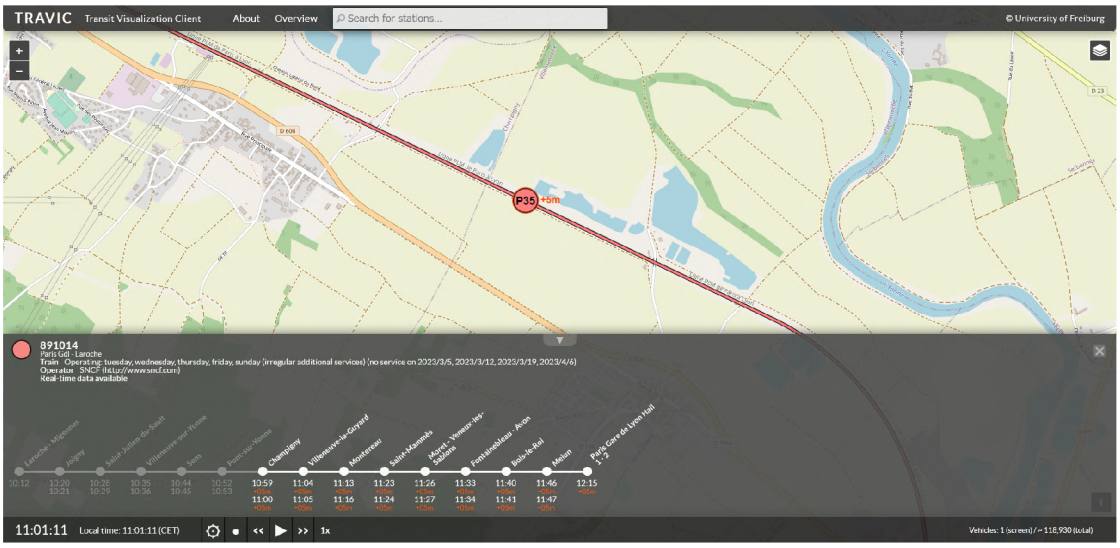
TRAVIC[6]

Online vizualizačný nástroj TRAVIC je projektom firmy geoOps a univerzity Freiburg [2]. Tento nástroj slúži primárne pre vizualizáciu pohybu vozidiel verejnej dopravy vo viac ako 700 dopravných systémoch. Nástroj pracuje so vstupnými dátami v špecializovanom formáte GTFS, pričom každý dopravný systém predstavuje samostatný súbor dát. Podľa dostupnosti dát dokáže nástroj pracovať v dvoch režimoch. Prvý využíva iba statické dáta, druhý využíva dáta v reálnom čase. Práve preto sú dáta o aktuálnom meškaní dostupné a využívané iba v druhom režime. Slúžia ako pridaná informačná hodnota pre cestujúceho, čo je možné vidieť aj na obrázku 2.2, ktorý zobrazuje meškanie vlaku spoločnosti SNCF na trase Laroche – Paris.

Sledování provozu IDS JMK[7]

Nástroj zobrazujúci pohyb vozidiel a linkové vedenie v dopravnom systéme Jihomoravského kraja je druhým zástupcom, ktorého som si zvolil v prehľade vizualizačných nástrojov. Tento nástroj pracuje s dátami v reálnom čase a kombinuje ich so statickými geografickými dátami. Zdroje dát pochádzajú z dopravných dispečingov CEDRIS a RIS, a sú prijímané v špecifickom formáte. Cestujúci prostredníctvom tohto nástroja dokáže rýchlo zistiť aktuálnu polohu a meškanie svojho spoja. Príklad použitia zobrazuje obrázok 2.3, na ktorom je možné vidieť aktuálnu polohu a meškanie autobusu na regionálnej linke 210.

[5] https://kam.mff.cuni.cz/~babilon/zpmapa
[6] https://travic.app/
[7] https://mapa.idsjmk.cz/



Obr. 2.2: **TRAVIC.** Vlak spoločnosti SNCF aktuálne mešká podľa reálnych dát 5 minút. Cestujúci tak okamžite získa dôležitú informáciu a môže podľa nej upraviť svoj plán cesty.

Cestovné poriadky[8]

Tento nástroj existujúci v slovenskej aj českej verzii vyvinutý spoločnosťou CHAPS sa primárne zameriava na vyhľadávanie spojení vo verejnej doprave. Nástroj do výsledkov vyhľadávania zahŕňa aktuálne hodnoty meškania, čím využíva prvý prístup. Zároveň je však schopný do výsledkov vyhľadávania zahrnúť aj štatistiky meškania navrhovaného spoja, čím využíva prístup druhý. Takto kombinované dáta umožňujú cestujúcim jednoducho získať informáciu s vyššou pridanou hodnotou. Reprezentatívny príklad spojenia oboch prístupov zobrazuje obrázok 2.4, kde je zobrazené aktuálne meškanie vlaku IC 520 spolu s informáciou, že náväzný vlak EC 130 pravidelne mešká, kým ďalší náväzný vlak rj74 odchádza štatisticky včas.

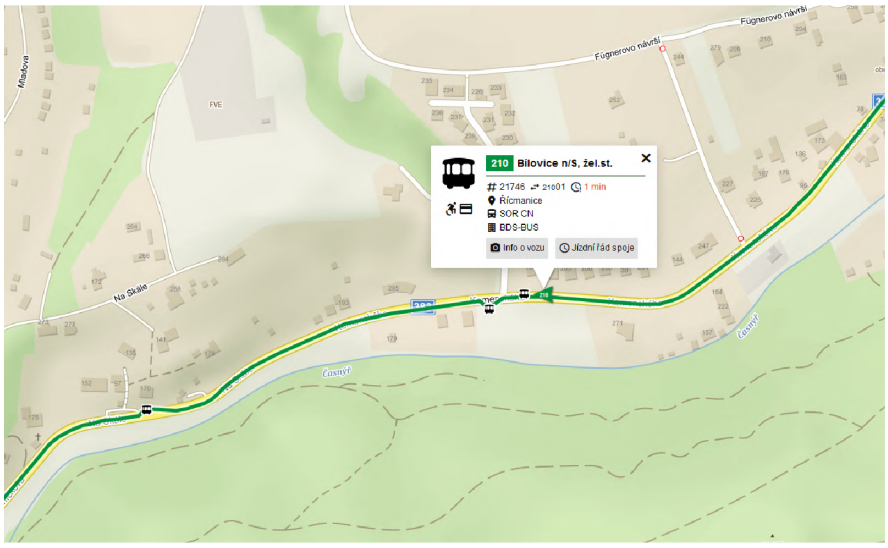
2.3.2 Analytické nástroje

Zastupujú implementáciu druhého prístupu k dátam o meškaní. Pre bežného cestujúceho sa ich výstupy môžu často javiť ako neintuitívne a málo využiteľné. Ich nosnou funkcionalitou je spracovanie, ukladanie a analýza historických dát. V rámci mojej práce som analyzoval dva nástroje tohto druhu. Prvým je český nástroj Babitron, sledujúci polohu a štatistiky meškania vlakov a druhým je americký nástroj Cost of congestion pracujúci s meškaniami autobusov.

Babitron[9]

Nástroj Babitron, ktorý je projektom Informatického ústavu Univerzity Karlovy a využíva dáta Českých drah, pracuje výlučne s vlakovou dopravou prevažne na siete SŽDC. Pozostáva z dvoch hlavných funkcionalít. Prvá je mapa a tabuľky zobrazujúce aktuálne polohy a meškania vlakov ako ukazuje aj obrázok 2.5. Druhou funkcionalitou je zobrazovanie histórie meškania. Pre každý vlakový spoj existuje samostatná tabuľka, ktorá obsahuje štatistiku

[8] https://cp.hnonline.sk/
[9] https://kam.mff.cuni.cz/~babilon/zpmapa



Obr. 2.3: **IDS JMK.** Mapa pohybu vozidiel v dopravnom systéme Jihomoravského kraja vždy zobrazuje polohu a aktuálne meškanie spojov.

meškaní v jednotlivých staniciach rozdelenú do 8 kategórií podľa hodnoty meškania. Používateľ si taktiež môže zvoliť časový rozsah skúmaných dát od 3 do 28 dní do histórie. Pre používateľa môžu byť zaujímavé aj agregované hodnoty spočitané pre každú stanicu. Ilustráciou výstupu je obrázok 2.6 zobrazúci štatistiku časti trasy vlaku EC 103 v časom horizonte 14 dní.

Cost of congestion[10]

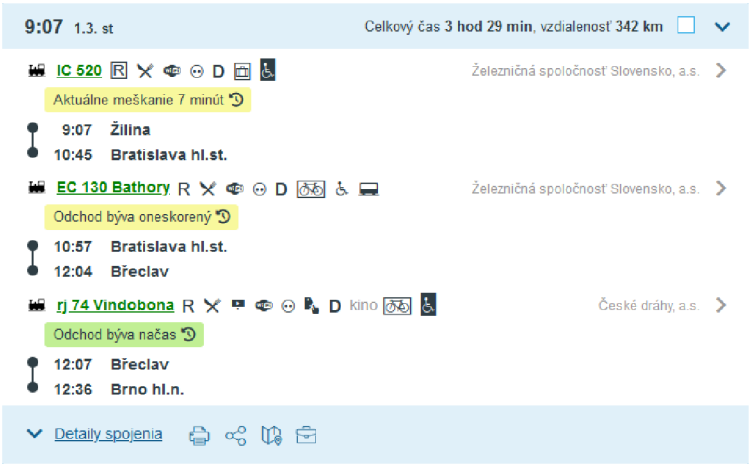
Americký nástroj vyvinutý na North Carolina State University zobrazuje, aký majú dopravné zápchy vplyv na meškania sutobusov verejnej dopravy. V tomto prípade sa jedná o výstup analýzy 3 datasetov pre 2 rôzne okresy. Cieľom bolo dáta spracovať pre ďalšie využitie dopravnými podnikmi a mestskými správami. Samotný výstup pozostáva z mapy zobrazujúcej hodnoty priemerného meškania rozdeleného do niekoľkých kategórií. Používateľ si taktiež môže zvoliť hodinový časový rozsah podľa toho, o ktorú dennú dobu sa zaujíma. Zaujímavou funkcionalitou je možnosť spustenia simulácie, ktorá zachytáva postupný vývoj hodnôt meškania počas dňa. Príklad výstupu daného nástroja uvádzam na obrázku 2.7, ktorý zobrazuje hodnoty priemerného mešakania školských autobusov v okrese Durham v popoludnajšej špičke.

2.4 Zhrnutie

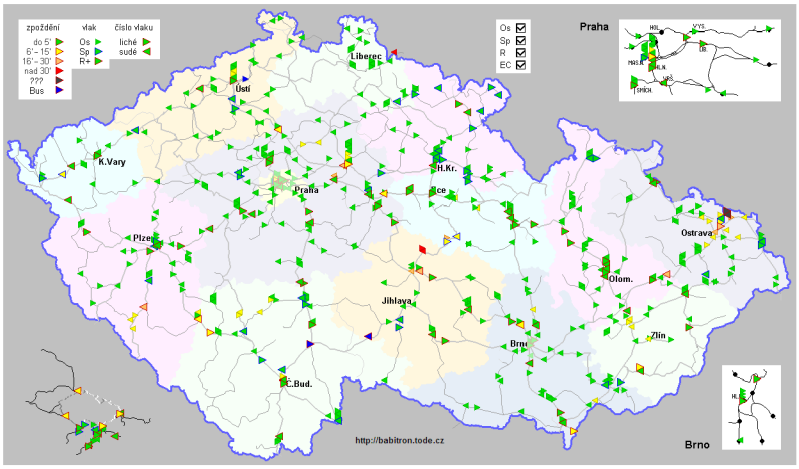
Nasadzovanie informačných technológií pri riešení problematiky verejnej dopravy nebolo vždy samozrejmosťou. K expanzii ich používania v tejto oblasti došlo až v posledných troch dekádach. Ako preukazujú mnohé štúdie, používanie informačných technológií so sebou prinieslo zefektívňovanie systémov a zvyšovanie spokojnosti cestujúcej verejnosti.

Jednotlivé nástroje môžeme rozdeľovať podľa rôznych kritérií, avšak vždy tvoria komplexný a prepojený systém. Pomáhajú nám pri plánovaní, ako aj prevádzke systémov ve-

[10] https://transitportal.org/cost_of_congestion.html



Obr. 2.4: **Cestovné poriadky.** Výsledky vyhľadávania spojenia kombinujú informáciu o aktuálnom meškaní a štatistickú informáciu o meškaní prípojných vlakov.



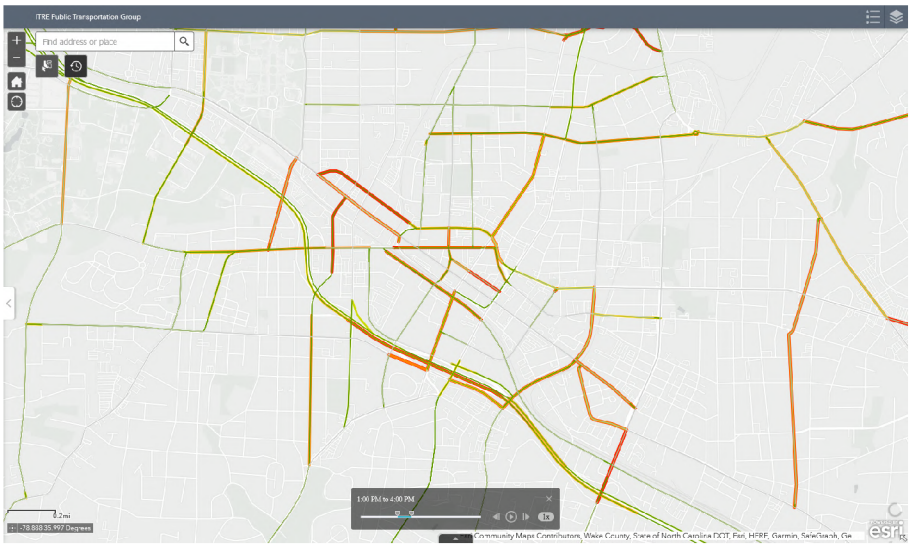
Obr. 2.5: **Babitron mapa.** Zobrazovanie aktuálnej polohy a meškania je jednou zo základných funkcionalít daného nástroja.

rejnej dopravy. Využívajú nielen dáta statické, ale aj dáta dynamické, čím v konečnom dôsledku zlepšujú fungovanie týchto systemov.

Z analýzy aktuálne používaných nástrojov je možné vyvodiť, že existujú dve úplne odlišné skupiny nástrojov s odlišnými skupinami používateľov a účelom použitia. Avšak oba druhy nástrojov využívajú rovnaké vizualizačné techniky a zdroje dát. Spoločné požiadavky kladené na obe tieto skupiny spočivajú predovšetkým v efektívnom spájaní geografických a časových dát do jednoducho interpretovateľných vizualizácií. Celkovo má však čo do počtu nástrojov aj používateľov prevahu práve prvá skupina vizualizačných nástrojov, pretože ju častejšie využívajú cestujúci v dopravných systémoch. Druhá skupina analytických nástrojov ale dokáže poskytovať pokročilé anaýzy, z ktorých získavame cenné znalosti o dopravných systémoch.

zpoždění vlaku 103 ve dnech 30.12.-12.1.												
		min.	průměr	max.	do 5'	6'-10'	11'-15'	16'-20'	21'-30'	31'-60'	nad 60'	???
Petrovice u Karviné	10:18 (příj.)	0'	7'	51'	10	1	0	0	1	1	0	1
Petrovice u Karviné	10:18 (odj.)	1'	8'	53'	11	1	0	0	1	1	0	0
Závada odbočka	10:21 (odj.)	0'	8'	53'	11	1	0	0	1	1	0	0
Dětmarovice	10:24 (odj.)	0'	8'	53'	10	2	0	0	1	1	0	0
Bohumín	10:30 (příj.)	0'	7'	53'	10	2	0	0	1	1	0	0
Bohumín	10:42 (odj.)	1'	11'	59'	7	5	0	0	1	1	0	0
Bohumín předn.	10:44 (odj.)	1'	11'	59'	7	4	1	0	1	1	0	0
Bohumín-Vrbice	10:45 (odj.)	2'	12'	59'	7	4	1	0	1	1	0	0
Ostrava-Hrušov	10:46 (odj.)	3'	12'	59'	5	5	2	0	1	1	0	0
Ostrava hl.n.	10:49 (příj.)	0'	10'	59'	8	4	0	0	1	1	0	0
Ostrava hl.n.	10:51 (odj.)	3'	11'	60'	8	2	2	0	1	1	0	0
Ostrava-Mar.Hory	10:53 (odj.)	3'	12'	61'	6	3	3	0	1	0	1	0
Ostrava-Svinov	10:57 (příj.)	2'	12'	60'	6	4	2	0	1	1	0	0
Ostrava-Svinov	10:59 (odj.)	2'	12'	60'	6	3	3	0	1	1	0	0
Polanka n.Odrou výh.	11:01 (příj.)	2'	12'	61'	6	3	3	0	1	0	1	0
Polanka n.Odrou výh.	11:01 (odj.)	2'	13'	61'	6	3	3	0	1	0	1	0
Polanka nad Odrou	11:02 (odj.)	2'	12'	61'	6	3	3	0	1	0	1	0
Jistebník	11:03 (odj.)	2'	13'	61'	6	3	2	1	1	0	1	0

Obr. 2.6: **Babitron štatistika.** Štatistiky jázd vlaku EC 103 nám môžu slúžiť pri plánovaní prestupov medzi jednotlivými spojmi.



Obr. 2.7: **Cost of congestion.** Zobrazovanie kumulácie meškaní priamo na mape nám pomáha ľahko detekovať problematické miesta.

Kapitola 3

Spracovanie a vizualizácia veľkých objemov dát

Slovné spojenie veľké objemy dát *(Big Data)* označuje explóziu dostupných informácií [9]. Zásluhy na zvýšení miery zaznamenávania, ale najmä ukladania veľkých objemov je možné pripísať zníženiu nákladov, ktoré si tieto úlohy vyžadujú. Dáta dnes produkuje prakticky každá oblasť ľudskej činnosti. Od medicíny, cez ekonomiku až po správanie používateľov sú generované obrovské množstvá dát z najrôznejších zariadení [30]. Veľké objemy dát tak nachádzajú využitie napríklad pri podpore rozhodovania, predikcii, simuláciach, hľadaní vzorov a množstve iných úloh [3]. Ako ukazuje súčasnosť, skutočnosť, že údaje sa dajú masívnejšie a lacnejšie vytvárať a uchovávať, sa v budúcnosti pravdepodobne zachová alebo dokonca ešte zrýchli [7].

Celkovo je však náročné veľké objemy dát zaznamenať, ukladať, spravovať, zdieľať, analyzovať a vizualizovať prostredníctvom štandartných softvérových nástrojov pre databázy [24]. Množstvo a objem týchto dát predstavujú nové výzvy v oblasti metód spracovania a analýzy dát. Jednou z týchto výziev je venovanie pozornosti správnemu spracovaniu a interpretácií dát, aby nedochádzalo k chybným analytickým a vedeckým záverom [9].

Hnacou silou v tejto problematike je snaha tieto dáta efektívne využiť. Pri riešení každej úlohy, ktorá sa z veľkých objemov dát snaží získať užitočné znalosti, predstavuje základ zhrnutie kladných a negatívnych prínosov, ktoré so sebou veľké objemy dát prinášajú. Univerzálnym nositeľom týchto prínosov sa stávajú charakteristické vlastnosti tohto druhu dát. Pomocou týchto vlastností je možné určiť vhodnosť použitia veľkých objemov dát v danej úlohe.

Ďalším spoločným krokom pri návrhu riešenia je vlastné spracovanie dát. Spracovanie veľkých objemov dát si vyžaduje použitie špecifických metód, ktoré zodpovedajú zamýšľanému cieľu úlohy. Princíp fungovania jednotlivých metód pritom reflektuje požiadavky, ktoré stanovujú charakteristiky veľkých objemov dát.

Aby bolo možné z dát v konečnom dôsledku získať užitočné znalosti nasleduje po spracovaní dát analýza. Jedným zo spôsobov analýzy je aj vizualizácia dát. Tento termín vyjadruje myšlienku, že vizualizácia dát obsahuje viac, ako len reprezentáciu dát v grafickej podobe. Informácie ukryté za dátami by mali byť rovnako dobre zobrazené, tak aby grafika pomohla čitateľom vidieť v dátach štruktúru [6]. Samotná vizualizácia veľkých objemov dát so sebou prináša nové problémy, ktoré riešia štandartné, ale aj novo vyvynuté druhy vizualizácií.

3.1 Charakteristiky veľkých objemov dát

Veľké objemy dát je možné charakterizovať rozličnými vlastnosťami v závislosti od zvolenej metodológie, či hľadiska. Kým staršia literatúra [15, 18] pracuje s menším počtom vlastností, novšie výskumy a analýzy [1, 16] pridávajú podrobnejšie charakteristiky, prípadne ich rozdeľujú do niekoľkých úrovní [25]. V nasledujúcich podsekciách je tak uvedený iba výber vlastností, ktoré sa najviac dotýkajú zámeru tejto práce. Každá charakteristika zároveň popisuje pozitívny, alebo negatívny prínos, ktorý veľké objemy dát prinášajú.

3.1.1 Objem

Objem je prvou vlastnosťou, ktorú ľudia pomenujú pri termíne veľké objemy dát [15]. Odkazuje na množstvo dostupných dát, ktoré sú spracovávané, ukladané a analyzované. Súčasné datasety nadobúdajú veľkosti v zettabajtoch (10^{21}), pričom nielen technologické spoločnosti denne spracovávajú desiatky petabajtov (10^{15}) dát. Prínosom spracovania takýchto objemov je možnosť získať skryté vzorce a informácie z analýz, čo poskytuje možnosť predikcie budúcich vzorov správania [11].

3.1.2 Rýchlosť

Charakterizuje trend stúpajúcej rýchlosti, s akou sú dáta generované v reálnom čase [18]. Okrem toho vyjadruje aj rýchlosť, s akou sa veľké objemy dát presúvajú [15]. Práve organizovanie a spracovanie veľkých objemov dát počas ich zbierania tak, aby mohli byť používané pri rozhodovaní v reálom čase, alebo v aplikáciach, je považované za najdôležitejšiu technickú výzvu [29]. Kým objem poskytuje možnosť nájsť skryté vzory, rýchlosť generovania a analýz umožňuje včas robiť cenné rozhodnutia na základe dát potrebných a platných v čase ich prijatia [12].

3.1.3 Vierohodnosť

Vierohodnosť hovorí o presnosti a pravdivosti dát. V prípade kombinácie veľkého objemu, vysokej rýchlosti vzniku a rôznorodosti dát nie je možné aby boli všetky dáta správne [15]. Vstupné dáta tak môžu byť zašumené, neúplne alebo chybné. V oblastiach akými sú financie alebo obchodovanie predstavuje dôležitý aspekt [20]. Preto predstavuje odstraňovanie nepresností z dát nevyhnutnú súčasť spracovania veľkých objemov dát. Vierohodnosť dát reprezentuje daň za výhody, ktoré plynú z objemu a rýchlosti.

3.2 Metódy spracovania veľkých objemov dát

Špecifické vlastnosti veľkých objemov dát popisované v predchádzajúcej sekcii vytvárajú problémy, ktoré nie sú s použitím štandartných metód efektívne riešiteľné. Veľké objemy dát si tak vyžadujú nové metódy, nástroje a techniky pre riešenie týchto problémov [4]. Postupne tak začali vznikať metódy, ktorých cieľom je pokrývať požiadavky vyplývajúce z charakteristík veľkých objemov dát. Najstaršou metódou spracovania je dávkové spracovanie, ktoré neskôr doplnilo spracovanie dát v reálnom čase. Oba prístupy k spracovaniu potom zjednocuje aktuálne používaný hybridný model [4]. Jednotlivé metódy implementujú rôzne systémy pre spracovanie, pričom každý systém ponúka odlišné funkcie, ktoré sú vhodné pre rôzne prípady použitia [23].

3.2.1 Dávkové spracovanie

Technika dávkového spracovania sa sústreďuje najmä na riešenie objemu dát. Hlavnou výhodou tejto metódy je škálovateľnosť. Aby dávkové spracovanie dokázalo zabezpečiť vysokú škálovateľnosť a zároveň zvládlo spracovať veľký objem dát, využíva paralelné distribuované spracovávanie. Tým dokáže efektívne riešiť najrôznejšie úlohy nad veľkými vstupmi. Zároveň obaľuje celý proces spracovania, čím zjednodušuje použitie samotných paralelných algoritmov a dokáže sa samostatne vysporiadať s výpadkami hardvéru aj softvéru. Preto je táto metóda považovaná za veľmi spoľahlivú. Nevýhodou je nárast výpočetnej doby v prípade nárastu objemu dát, hlavne pri iteratívnych aplikáciach [14]. Celkovo je tak dávkové spracovanie pomalšie a nie je vhodné pre aplikácie vyžadujúce nízku latenciu odpovedí na svoje požiadavky. Využíva sa v prípade, že dáta sú zbierané, alebo ukladané do veľkých súborov [23]. Príkladom systému pre spracovanie dát využívajúceho túto metódu je technológia Hadoop[1], ktorej jadro tvorí programovacia paradigma Map and Reduce.

3.2.2 Spracovanie v reálnom čase

Kým dávkové spracovanie rieši najmä problém objemu dát, spracovanie v reálnom čase sa sústreďuje na riešenie problémov spojených s rýchlosťou vzniku veľkých objemov dát. Spracovanie v reálnom čase využíva rovnaké princípy ako dávkové spracovanie, teda paralelizmus a distribuovanie výpočtov. Zjednodušene povedané, jedná sa o nekonečný počet malých dávkových spracovaní za sebou [4]. Na rozdiel od prechádzajúcej metódy spracovanie v reálnom čase využíva namiesto súborov dynamickú pamäť. Táto metóda je teda vhodnejšia pre aplikácie, ktoré vyžadujú rýchle spracovanie heterogénnych dát a nízku latenciu odpovedí. Uplatnenie tak nachádzjú najmä v oblasti odporúčacích systémov a sociálnych sietí, kde je nevyhnutné spracovávať veľké množstvá nestatických, kontinuálne vznikajúcich, dát [14]. Túto metódu využívajú napríklad technológie Storm[2], alebo Samza[3].

3.2.3 Hybridný model spracovania

Mnoho aplikačných oblastí však vyžaduje kombináciu oboch predchádzajúcich prístupov k spracovaniu dát [4]. Toto spojenie implementuje hybridný model spracovania dát, označovaný aj ako Lambda model [28]. Tento model sa skladá z dvoch častí. Prvou je dávková časť, ktorá spravuje dáta v distribuovanom súborovom systéme a výsledky analýz nad týmito dátami. V tejto časti analýzy trvajú dlhšie a nie je možné počas bežiacej úlohy zahrnúť novo získané dáta. Druhou je rýchla časť, ktorá pracuje s novými dátami, ktoré vstupujú do systému počas behu výpočtov v dátovej časti, a ich analýza si vyžaduje nízku latenciu. Pre získanie kompletných výsledkov je potom nevyhnutné výsledky z oboch vrstiev spojiť do jedného výstupu. Vo výsledku tak táto metóda kombinuje výhody oboch predchádzajúcich metód. Avšak architektúra založená na dvoch vrstvách zahŕňa netriviálne úlohy ako synchronizácia a skladanie dát [4]. Technológiami využívajúcimi hybridný model spracovania sú napríklad systémy Spark[4], alebo Flink[5].

[1] https://hadoop.apache.org
[2] https://storm.apache.org
[3] https://samza.apache.org
[4] https://spark.apache.org
[5] https://flink.apache.org

3.3 Vizualizácia veľkých objemov dát

Vizualizácia dát je vo všeobecnosti dôležitá najmä pre zoskupenie veľkého množstva dátových bodov, pochopenie vzťahov v dátach, môže slúžiť ako podklad pre diskusiu o otázkach v reálnom čase, a urýchliť proces rozhodovania, kam zamerať prieskum [5]. Grafická reprezentácia dát je zároveň jednoduchá na pochopenie, čím sa jej výstupy stávajú vhodným prezentačným prostriedkom dát odbornej aj laickej verejnosti. Hlavným cieľom vizualizácie dát je podpora pri identifikácií problematických miest, alebo pri hľadaní zaujímavých vzorov. Avšak nutnosť analyzovať veľké objemy dát priniesla do oblasti vizualizácií nové výzvy a problémy.

3.3.1 Problémy vizualizácie veľkých objemov dát

Zdrojom problémov vizualizácie veľkých objemov dát sú špecifické vlastnosti tohto druhu dát. Rovnako ako charakteristiky veľkých objemov dát, je možné podľa rôznych kritérií rozlišovať rôzne problémy. Pre účely tejto práce sú tak popisované iba dve úzko prepojené skupiny, obsahujúce problémy s podobnou povahou.

Vysoké požiadavky na výkon

Prvá skupina je tvorená problémami, ktoré súvisia s hardwarovými požiadavkami. Vizualizácia veľkých objemov dát kladie zvýšené nároky na výpočetný výkon najmä počas generovania výstupov a aplikovania agregačných funkcií. Tento problém sa ale stáva oveľa viditeľnejším v dynamickej vizualizácií dát, ako napríklad počas analýzy správania v reálnom čase. Je to spôsobené rýchlosťou a najmä vysokým počtom zmien medzi jednotlivými stavmi sledovaného systému. Nedostatočný výkon sa tak prejavuje najmä zvyšovaním latencie generovania výstupu, čím môže dôjsť až k strate potencionálne dôležitých informácií vyplývajúcich zo zmien hodnôt. Riešením pritom nemôže byť spomalenie generovanie vizualizácií, pretože to môže viesť k strate relevantnosti informácií. Riešenie tak spočíva v paralelizácií výpočtov, a transformácií dát s cieľom zmenšenia ich objemu.

Vysoký počet vizualizovaných dát

Druhá skupina problémov súvisí s použiteľnosťou samotných vizualizačných výstupov. Vzhľadom na obrovský objem dát, nie je vizualizácia do jedného grafického výstupu prakticky využiteľná. Jednotlivé záznamy od seba nie je možné odlišiť, keďže sa nachádzajú v malej vzdialenosti a vplyvom rozlíšenia obrazovky sa zlievajú do jedného bodu. Dochádza tak k strate prehľadnosti dát, čim sa grafický výstup stáva nepoužiteľným.

Zvyšovaním rozlíšenia výstupného zariadenia, alebo použitím viacerých výstupných zariadení, je možné vizuálny šum potlačiť. Jednotlivé záznamy sa tak nespájajú, ale zostávajú oddelené. Avšak toto riešenie naráža na obmedzené ľudské vnímanie. Používateľ dokáže vo svojom zornom poli zachytiť a prijať iba ohraničené množstvo vnemov, čím dochádza pri nadmerne veľkých vizualizáciach k strate potenciálne využiteľných informácií. Vhodným riešením vizuálneho šumu a obmedzeného vnímania je redukcia informácií pomocou rôznych agregácií a filtrovania, na základe preferencií koncového používateľa. Avšak tento prístup obsahuje riziko straty skrytých a potenciálne užitočných vzorov a informácií.

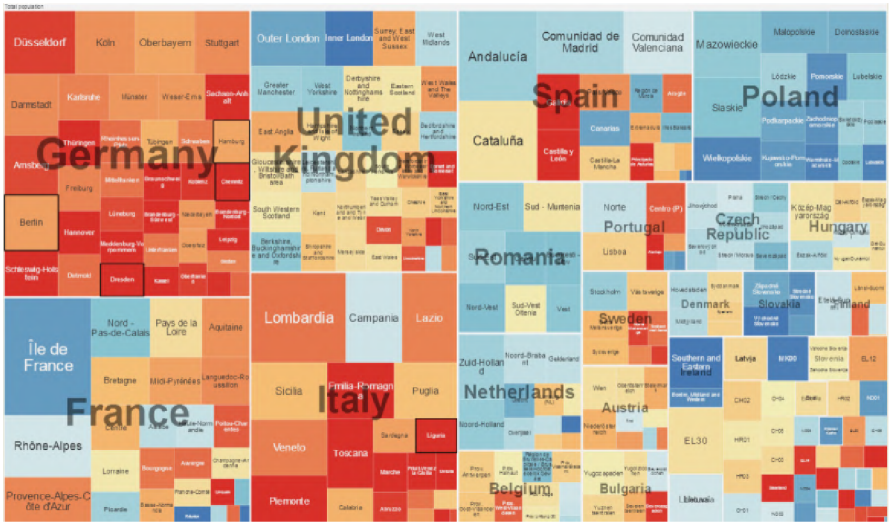
3.3.2 Nové vizualizačné metódy

Riešenia, ktoré vznikli ako odpoveď na problémy spôsobené nutnosťou vizualizácie veľkých objemov dát, spočívajú v modifikácií existujúcich a vývoji nových vizualizačných metód. Oba prístupy riešia spoločnú výzvu, ktorá spočíva v tom, ako s veľkými objemami dát efektívne pracovať, a ako zobraziť výsledky vizualizácií a analýz dát tak, aby boli zmysluplné a užitočné [5]. Hlavnými cieľmi týchto metód sú nízka latencia pri generovaní výstupu a formátovanie výstupu tak, aby bolo možné jednoducho identifikovať zaujímavé vzory. Zároveň je dôležitou úlohou zvoliť vhodný počet dimenzií zobrazovaných dát. Nízky počet dimenzií môže viesť k strate informácií, kým vysoký počet vedie k vizuálnemu šumu.

Vzhľadom na vysoký počet metód, ktoré sa pri vizualizácií veľkých objemov dát používajú, je v tejto práci uvedený iba reprezentatívny výber metód. V nasledujúcich podsekciách sú tak popísané dve vybrané novovyvinuté metódy, pre vizualizáciu veľkých objemov dát.

Stromová mapa

Metóda stromovej mapy vizualizuje usporiadané dáta a ich vzťahy pomocou obdĺžnikovej mapy. Rámec celej vizualizácie je tvorený hlavným obdĺžnikom, ktorý je rozdelený na presne usporiadané menšie obdĺžniky. Veľkosť plochy obdĺžnika je priamo úmerná hodnote zvoleného atribútu záznamu, ktorý obdĺžnik reprezentuje. Jednotlivé obdĺžniky je potom možné zoskupovať do rôznych tried pomocou farebnej palety, ktorá je závislá od zvoleného atribútu. Jej nevýhodami sú problematické zobrazovanie nulových a záporných hodnôt, schopnosť vizualizovať dáta iba v jednom stave v čase, a zložité skladanie jednotlivých odĺžnikov do kompaktného celku. Oproti tomu výhody spočívajú v možnostiach vizualizovať širokú triedu vstupných dát a široké možnosti modifikácie podľa požiadaviek používateľa. Príklad použitia tejto metódy zobrazuje obrázok 3.1, na ktorom je vizualizovaná veková štruktúra obyvateľstva vybraných častí Európy.

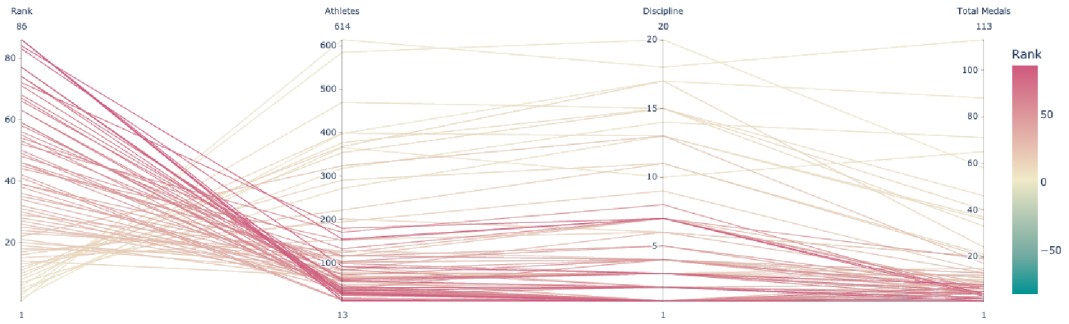


Obr. 3.1: **Metóda stromovej mapy.** Je schopná prehľadne vizualizovať širokú triedu dát, ako napríklad vekovú štruktúru obyvateľstva[7]

[7]Prevzaté z https://ncva.itn.liu.se/education-geovisual-analytics/treemap

Paralelné súradnice

Hlavnou oblasťou použitia tejto metódy je vizualizácia vzťahov v dátach s vysokým počtom dimenzií. Základom vizualizácie je graf, kde os x obsahuje jednotlivé dimenzie dát a na osi y sú vynesené hodnoty atribútov platných pre danú dimenziu. Každý záznam je v konečnom dôsledku reprezentovaný sadou úsečiek prepájajúcich body reprezentujúce jednotlivé hodnoty atribútov daného záznamu vzhľadom na dimenzie. Pomocou farebných škál a hrúbok úsečiek je taktiež možné vytvárať skupiny záznamov, alebo zvýrazňovať dôležité záznamy. Nevýhodou tejto metódy je nemožnosť jej využitia nad kategorickými dátami. Naopak hlavnou výhodou je schopnosť prehľadne zobraziť vysoký počet dimenzií a záznamov. Obrázok 3.2 zobrazuje použitie tejto metódy nad generickými dátami so štyrmi dimenziami.



Obr. 3.2: **Metóda paralelných súradníc.** Použtím tejto metódy je možné vizualizovať viacero samostatných dimenzií a vzťahov medzi nimi.[9]

3.3.3 Aktuálne používané nástroje

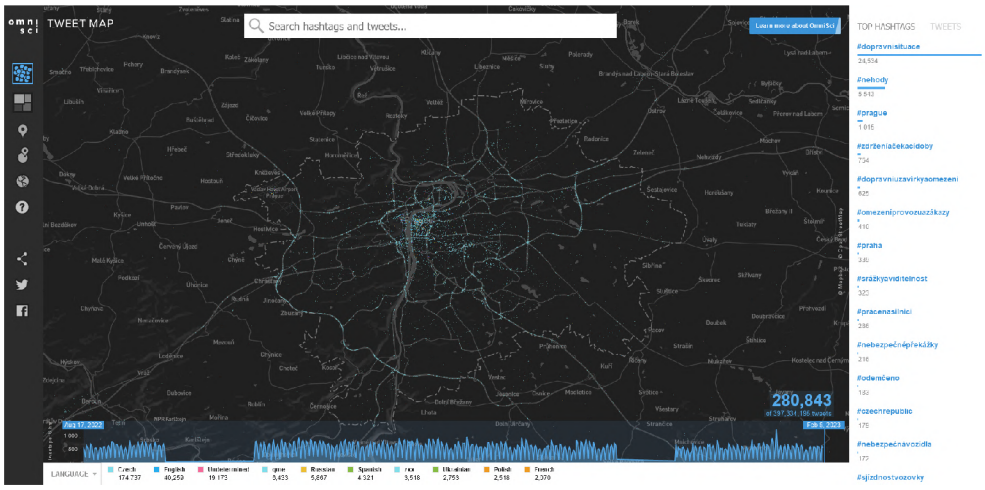
Kým predchádzajúce podsekcie popisujú samotné novo vyvinuté vizualizačné metódy, nasledujúce podsekcie obsahujú reprezentatívny výber aktuálne používaných nástrojov, ktoré ilustrujú riešenia problémov veľkých objemov dát v praxi.

Tweet Map

Jedným zo spôsobov, akými je možné vysporiadať sa s vysokými požiadavkami na výpočetný výkon súvisiacimi s veľkými objemami dát je využitie GPU. Takýmto riešením disponuje aj vizualizačná platforma OmniSci, ktorá využíva pre spracovanie a vizualizáciu masívnu paralelizáciu a grafické procesory. Uvedený príklad použitia tejto platformy, aplikácia Tweet Map[10], vizualizuje aktivitu na sociálnej sieti Twitter. Aplikácia je schopná spracovávať tisíce záznamov v reálnom čase, pričom celkovo vizualizuje niekoľko stoviek miliónov záznamov. Grafickým výstupom je potom mapa vizualizujúca záznamy v zvolenom časovom období. Okrem času môže používateľ záznamy filtrovať na základe jazyka, alebo hashtagu. Výhodou takéhoto riešenia je schopnosť poskytovať analýzy obrovského množstva dát v reálnom čase s nízkou latenciou, bez nutnosti predvypočítavnia výstupov. Príklad použitia platformy ilustuje obrázok 3.3 v podobe výstupu z aplikácie Tweet Map.

[9] Prevzaté z https://www.analyticsvidhya.com/blog/2021/11/visualize-data-using-parallel-coordinates-plot

[10] https://www.heavy.ai/demos/tweetmap



Obr. 3.3: **Tweet Map.** Vizualizácia veľkých objemov dát je náročná na výpočetný výkon. Jedným z používaných riešení je, ako aj v tomto prípade, využitie GPU procesorov.

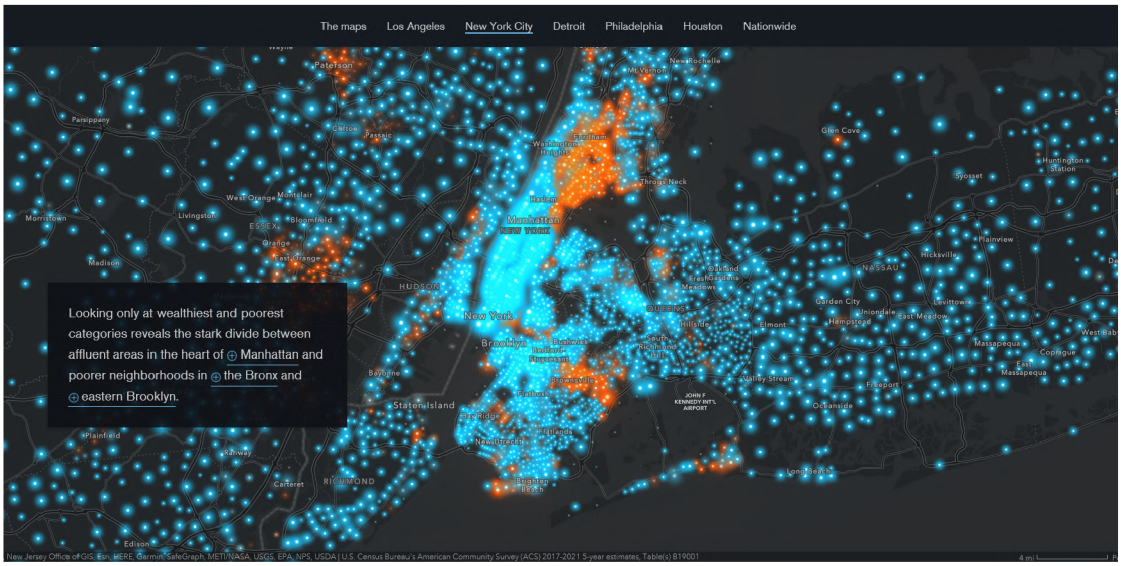
Mapping incomes

Kým niektoré vizualizácie veľkých objemov dát riešia problémy zvyšovaním výpočetného výkonu, iné nástroje využívajú metódu predspracovania výstupov. Takýmto príkladom je aj platforma ArcGIS StoryMaps[11] spoločnosti Esri, konkrétne vybraná mapa príjmov[12] amerických domácností. Nástroje využívajúce tento prístup zväčša pracujú s už predspracovanými datasetmi, pričom ich hlavnou úlohou je generovanie prehľadných výstupov používaných pri ďalších analýzach, alebo prezentácií dát laickej a odbornej verejnosti. Výhodou takéhoto riešenia je zníženie nárokov na výpočetný výkon na úkor pomalšieho generovania výstupov. Uvedený príklad takéhoto prístupu na obrázku 3.4 ilustruje jednoduchosť, s akou je možné veľké objemy dát prezentovať v prehľadnej podobe.

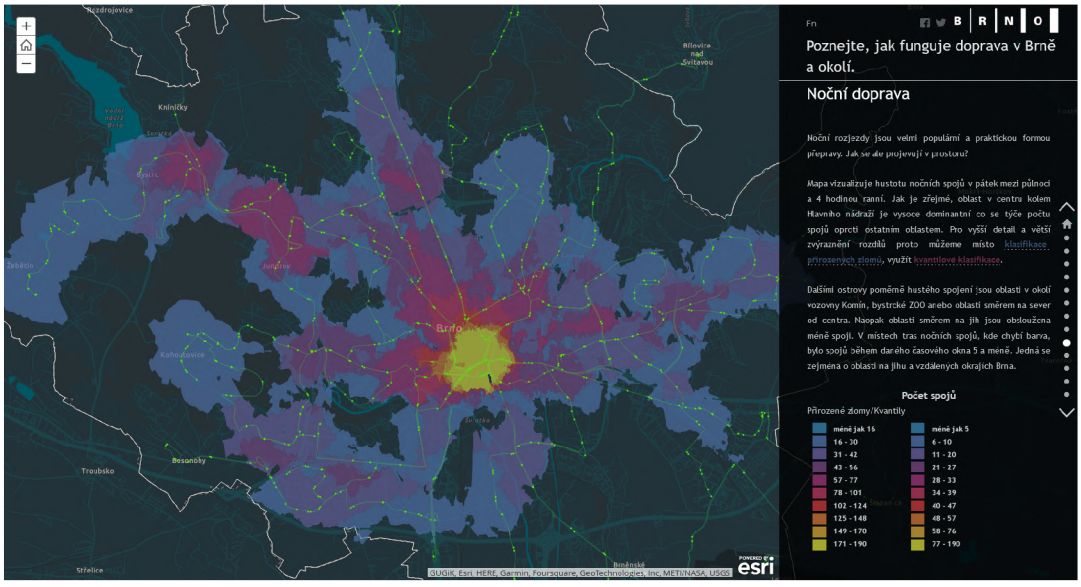
Poznejte, jak funguje doprava v Brně a okolí

Do rovnakej skupiny ako predchádzajúca platforma patrí aj aplikácia Poznejte, jak funguje doprava v Brně a okolí[13], ktorá používateľa postupne prevedie vybranými vlastnosťami brnenskej verejnej dopravy. Základným princípom nástroja je geovizualizácia, teda rozmiestnenie dát nad podkladovou mapou v priestore, pričom aplikácia využíva ako svoj základ platformu ArcGIS[14]. Aplikácia postupne vizualizuje rôzne druhy dopravnej dostupnosti verejnej dopravy v rámci času a priestoru. Používateľ tak získava možnosť sledovať, kde sa dopravná obslužnosť zhoršuje a kde sa zlepšuje. Aplikácia taktiež obsahuje simuláciu zmeny dopravnej obslužnosti počas dňa. Výhodou tohto nástroja je prehľadná vizualizácia veľkého množstva dát do jednoducho pochopiteľného výstupu, z ktoré je možné získať užitočné znalosti. Nevýhodou je neaktuálnosť dát, ktorá je spôsobená statickým zdrojom dát z roku 2018. Fungovanie aplikácie zobrazuje obrázok 3.5, na ktorom je vizualizovaná hustota spojov centra Brna v čase nočných rozjazdov.

[11] https://www.esri.com/en-us/arcgis/products/arcgis-storymaps
[12] https://storymaps.arcgis.com/stories/424cee66a3b144a88b65508443ba88a4
[13] https://mestobrno.maps.arcgis.com/apps/MapJournal/index.html?appid=073a3d9303b54cd08c48e4158c7fe1f4
[14] https://www.esri.com/en-us/arcgis/products/arcgis-pro



Obr. 3.4: **Mapping incomes.** Predgenerovanie výstupov z veľkých objemov dát môže slúžiť ako prehľadný formát pre následné analýzy a prezentáciu ich výsledkov.



Obr. 3.5: **Nočná doprava.** Aplikácia Poznejte, jak funguje doprava v Brně a okolí vizualizuje rôzne vlastnosti dopravného systému Jihomoravského kraje.

Kapitola 4

Analýza problému

Vďaka rozšíreniu nástrojov monitorujúcich aktuálny stav, poskytujúcich podporu pri plánovaní, alebo informujúcich cestujúcich v reálnom čase sú dnešné integrované systémy hromadnej dopravy oveľa efektívnejšie. Vďaka týmto nástrojom sú dopravné autority schopné riešiť rôzne neočakávané udalosti ako dopravné nehody, alebo poruchy vozidiel a infraštruktúry v reálnom čase.

Kľúčovou výhodou týchto systémov je však ich schopnosť pracovať s dátami z reálnej prevádzky. Tieto dáta slúžia ako zdroj vstupných informácií pre samotné plánovanie, návrhy zmien, a sledovanie dopadov na systémy verejnej dopravy. V konečnom dôsledku je tak možné hovoriť o metóde iteratívneho zvyšovania kvality dopravnej obslužnosti, optimalizácií dopravných výkonov, a zlepšovanie dopravných spojení. Vďaka tomu sa systémy hromadnej dopravy stávajú ekonomicky efektívnejšie, dochádza k skracovaniu cestovných časov, a celkovo sa zvyšuje spokojnosť cestujúcej verejnosti.

V prípade väčších systémov hromadnej dopravy tieto benefity spravidla prináša celá sada rozličných, navzájom prepojených nástrojov. Jedným z týchto nástrojov sú aj systémy pre analýzu správania celého dopravného systému v čase. Aby takýto navrhnutý a implementovaný nástroj plnil svoj účel a prinášal tak požadované benefity, je v prvom rade nevyhnutné analyzovať a definovať požiadavky z troch odlišných oblastí.

4.1 Analýza požiadaviek

Prvou oblasťou sú požiadavky, ktoré vyplývajú z analýzy profilu potencionálnych používateľov. Druhou kategóriou je analýza reálneho systému a jeho správania tak, aby zamýšľaný nástroj dokázal toto správanie s určitou mierou abstrakcie simulovať. Posledná skupina požiadaviek vyplýva z predchádzajúcich dvoch kategórií a sú ňou požiadavky na návrh technického riešenia. Dobre definované požiadavky môžu byť následne využité ako základ návrhu potencionálneho nástroja.

4.1.1 Analýza používateľov

Informácie o aktuálnom stave a správaní dopravného systému využíva široká skupina používateľov. Avšak, je možné identifikovať dve podskupiny, ktorých požiadavky na nástroj sledujúci toto správanie sa od seba do istej miery líšia. Prvú skupinu tvoria bežný cestujúci, ktorý systém iba používajú. Druhou podskupinou sú pracovníci dopravných autorít, ktorých úlohou je fungovanie systému vyhodnocovať a zlepšovať jeho účinnosť.

Cestujúca verejnosť

Prevažnú skupinu používateľov, ktorí s informáciami o správaní dopravného systému pracujú, tvoria bežní cestujúci. Ich najčastejšou požiadavkou je získavanie aktuálnych dát o polohe a meškaní vozidiel v dopravnom systéme. Tieto dáta následne využívajú pri prispôsobovaní svojho plánu využitia systému v reálnom čase. Cestujúci si tak napríklad môže zvoliť iný spoj, alebo úplne zmeniť zamýšľanú trasu. Naopak historické dáta využívajú cestujúci pri plánovaní ciest a prestupov v systéme, kedy napríklad na základe pravidelnej hodnoty meškania odhadujú pravdepodobnosť prestupu medzi dvoma spojmi. Získavanie a využívanie týchto informácií tak umožňuje cestujúcim skracovať čas svojich ciest, čo v konečnom dôsledku vedie k zvyšovaniu ich spokojnosti.

Z prípadov použitia, ako aj z povahy sledovanej skupiny používateľov, je tak možné definovať dve skupiny požiadaviek:

- Prvú skupinu tvoria požiadavky na dostupnosť nástroja. Potencionálne riešenie by malo byť čo najdostupnejšie, bez nutnosti inštalácie, a spustiteľné na širokej škále zariadení. Vzhľadom na absenciu odborných znalostí, musí byť zamýšľané riešenie ovládateľné bez špecifických vedomostí. Samotné používateľské rozhranie by malo byť čo najprehľadnejšie, a čo najviac intuitívne na používanie.

- Druhá skupina požiadaviek sa zameriava na spôsob, akým sú informácie poskytované. Samotný výstup by mal byť prehľadný, ľahko pochopiteľný bez špecifických znalostí, a mal by umožňovať filtráciu dát na základe času a spojenia. Rovnako prehľadne by mali byť modelované aj vzťahy medzi dátami a reálnym systémom, ako aj vzťahy medzi dátami samotnými.

Zamestnanci dopravných autorít

Pomerovo menšiu skupinu používateľov tvoria zamestnanci dopravných podnikov, koordinátorov systémov hromadnej dopravy, prípadne rôzni analytici. Hlavný rozdiel oproti predchádzajúcej skupine používateľov spočíva v odbornej znalosti problematiky a cieľom využitia informácií o správaní systému. Z týchto rozdielov vyplývajú aj odlišné prípady použitia. Zamestnanci dopravných autorít využívajú prevažne historické dáta pri plánovaní zmien v systéme, vyhodnocovaní predchádzajúcich zásahov, a pri identifikácií problematických miest. Ide najmä o miesta s permanentne vznikajúcim meškaním, prípadne slabou dopravnou obslužnosťou. Na rozdiel od cestujúcej verejnosti tak nevyužívajú tieto informácie primárne pre svoju potrebu, ale pre zlepšovanie fungovania dopravného systému prostredníctvom zmien v štruktúre systému, prípadne zmenou nasadzovaných vozidiel. Rovnako tak sú tieto dáta využívané pri plánovaní rozširovania systémov hromadnej dopravy.

Napriek rozdielnym prípadom použitia sú požiadavky definované pre predchádzajúcu skupinu používateľov platné aj v tejto skupine. Avšak vzhľadom na odlišnú povahu prípadov použitia je nutné definovať aj nové používateľské požiadavky:

- Prvá skupina požiadaviek je tvorená rozšírenými nárokmi na filtráciu a agregáciu dát. Zamýšľaný nástroj by mal používateľovi umožniť dáta filtrovať a agregovať v rôznych dimenziách. V časovej dimenzii je nevyhnutné umožmiť agregáciu od jednotlivých spojov až po celé dni až týždne. Z hľadiska priestoru ide najmä o požiadavky od agregácie nad spoločnými úsekmi vybraných liniek až po celé geografické a administratívne oblasti. Špecifické požiadavky na filtráciu potom tvoria dimenzie, akými sú napríklad druh dopravného prostriedku, alebo typ vozidla, ktoré spoj obsluhuje.

- Druhá skupina požiadaviek je tvorená potrebou hlbšej analýzy správania systému. Aby pracovníci dopravných autorít mohli napríklad efektívne identifikovať kritické miesta, rozlišovať medzi anomáliami a vzormi, je nevyhnutné, aby zamýšľaný nástroj poskytoval ľahko ovládateľnú možnosť dáta rôznymi spôsobmi analyzovať. Táto požiadavka zároveň reflektuje potrebu pripraviť sa na pridávanie nových spôsobov analýzy dát, ktoré prinesie zvyšovanie komplexnosti dopravných systémov.

4.1.2 Analýza dopravného systému

Kým požiadavky vyplývajúce z analýzy používateľov ovplyvňujú najmä výstup, analýza sledovaného systému hromadnej dopravy prináša požiadavky na vnútornú štruktúru zamýšľaného nástroja. Systém hromadnej dopravy mesta Brna je súčasťou väčšieho integrovaného systému Jihomoravského kraje. Kostru tohto dopravného systému tvoria električkové linky pokrývajúce najväčšie dopravné prúdy. Koľajová sieť je najmä v západnej časti mesta doplnená trolejbusovými linkami, ktoré dopĺňajú základnú štruktúru systému. Zvyšok systému je tvorený autobusovými linkami, ktoré pokrývajú odľahlejšie mestské časti a zabezpečujú ich spojenie s prestupnými uzlami, alebo zvyšujú dopravnú obslužnosť. Špecifickým doplnkom celého systému je lodná doprava na brnenskej priehrade.

Z pestrej skladby rôznych druhov dopravy a štruktúry samotného systému následne vyplývajú bežné, ale aj špecifické vlastnosti brnenského systému hromadnej dopravy. Prvou vlastnosťou je počet výluk a zmien v dopravnom systéme. V rámci roka pripadá na jeden deň 1,5 výluky, čím sa štruktúra systému mení každý deň. Osobitnú kategóriu potom predstavujú dlhodobejšie a rozsiahlejšie výluky najmä v letných mesiacoch. Druhou vlastnosťou je vysoká miera diverzity štruktúry celého systému. V rámci jednotlivých liniek existujú rôzne varianty trás, od čiastočne skrátených trás, cez výjazdy a dojazdy do vozovní, až po prejazdy na iné linky počas jedného spoja. Niektoré linky tak obsahujú variant trasy vykonávaný iba jedným spojom za deň. Tretia vlastnosť súvisí s fyzickou infraštruktúrou samotného systému. V prípade väčších prestupných uzlov existujú zastávky s rovnakým názvom v niekoľkých polohách, pričom sú oddelené číslom nástupiska. V rámci niektorých zastávok taktiež dochádza k zdieľaniu nástupiska viacerými druhmi dopravy.

Z charakteristík tohto systému vyplývajú netriviálne požiadavky, ktoré je nutné splniť, aby bolo možné správanie daného systému sledovať. Tieto požiadavky je potom možné rozdeliť od dvoch samostatných skupín:

- Prvá skupina je tvorená požiadavkami na samotné sledovanie správania dopravného systému. Aby bolo možné toto správanie vyhodnocovať, je nevyhnutné spracovávať dôverhodný zdroj dát, ktorý toto správanie popisuje v danom čase. Zamýšľaný nástroj tak musí byť schopný spracovať a anotovať takéto dáta v čase ich platnosti, aby následne získané znalosti boli skutočne platné pre zvolený čas.

- V druhej skupine sú obsiahnuté požiadavky na modelovanie samotného dopravného systému. Aby bolo možné dáta o správaní anotovať, je nevyhnutná existencia referenčného modelu, ako sa daný dopravný systém v ideálnej podobe správa. Tento model musí byť generovaný automaticky každý deň, aby sa zabezpečila jeho aktuálnosť vzhľadom na časté zmeny, ktorými systém hromadnej dopravy mesta Brna prechádza.

- Tretia skupina požiadaviek smeruje na modelovanie vzťahov medzi jednotlivými stavmi systému. Aby bolo možné sledovať a porovnávať správanie systému v čase je nevyhnutné prepájanie na sebe závislých častí systému, akými sú zastávky, linky, alebo spoje, medzi sebou.

4.1.3 Požiadavky na technické riešenie

Analýza požiadaviek používateľov a dopravného systému vytvára základný rámec, v ktorý by malo potencionálne riešenie napĺňať. Aby však bol tento rámec úplny, je nevyhnutné analyzovať technické požiadavky, ktoré bude musieť zamýšľané riešenie spĺňať. Na rozdiel od predchádzajúcich skupín požiadaviek, požiadavky na technické riešenie budú mať vplyv na návrh potencionálneho nástroja ako celku. Tieto požiadavky je rovnako ako predtým možné rozdeliť do viacerých podskupín:

- V prvej skupine sa nachádzajú požiadavky na spôsob implementácie potenciálneho nástroja. Narvhnuté riešenie by malo byť čo najviac prenositeľné, výpočetne úsporné a v čo najväčšej miere automatizované. Dôležitou požiadavkou je taktiež udržateľnosť tohto riešenia a jednoduchý sposôsob pridávania nových funkcionalít.

- Druhá skupina požiadaviek súvisí s predpokladaným objemom dát. Vzľadom na časový a priestorový rozsah potencionálne ukladaných dát, sa ako nevyhnutná požiadavka javí navrhnúť úspornú metódu ukladania dát. Okrem toho by mali byť ukladané dáta anotované tak, aby aj v prípade väčších objemov nepresahovalo modelovanie vzťahov a generovanie analýz exponencionálnu časovú zložitosť.

- Tretia skupina požiadaviek vyplýva z povahy samotného problému. Modelovanie reálnych systémov v digitálnej podobe pracuje s určitou mierou abstrakcie. Reálne systémy sú však popisované reálnymi dátami, ktoré sú spravidla nekompletné a obsahujú rôzne anomálie. Z tohto dôvodu je nevyhnutné, aby bol zamýšľaný nástroj schopný s takýmito dátami pracovať, a modelovať systém z poskytnutých dát s čo najväčšou presnosťou.

4.2 Definícia problémov

Napriek široko rozšírenému používaniu informačných systémov vo verejnej doprave, v oblasti správy a plánovania stále existujú problematické, alebo potenciálne zlepšiteľné miesta. Tieto problémy znižujú využiteľnosť získaných dát, čo v konečnom dôsledku znamená pomalšie a neefektívnejšie prispôsobovanie systému verejnej dopravy potrebám cestujúcich. Riešenie, ktoré by mohlo zlepšiť súčasný stav, preto musí na jednej strane spĺňať požiadavky získané analýzou a zároveň riešiť problémy v súčastnosti používaných nástrojov.

4.2.1 Nedostupnosť dát

Samotné získavanie dát o stave a fungovaní systémov hromadnej dopravy predstavuje problém. Sprístupňovanie takýchto dát mimo interné systémy jednotlivých dopravných podnikov prebieha najčastejšie vo forme jednorázového zverejnenia archívu s neaktuálnymi, alebo agregovanými dátami. Zároveň je častým javom zverejňovanie týchto záznamov v strojovo nečitateľnom formáte. Rovnakým problémom sa vyznačujú aj dáta popisujúce štruktúru dopravného systému. V tomto prípade dochádza najčastejšie k zverejňovaniu informácií o štruktúre systému iba v rovine časovej, ale nie priestorovej. Aj v tomto prípade sú dáta často zverejňované v strojovo nečitateľnom formáte. Tieto problémy následne vedú k nedostupnosti analýz, ktoré by vznikali s rovnakými časovými odstupmi. Výsledkom je nekompletný obraz vývoja správania systému v čase.

4.2.2 Nedostupnosť nástrojov

Špecifickým problémom je aj nedostupnosť existujúcich nástrojov, ktoré správanie dopravných systémov sledujú. Existujúce nástroje v rámci dopravných autorít nie sú vo väčšine prípadov dostupné cestujúcej verejnosti. Cestujúci tak nemá možnosť založiť plánovanie svojich ciest v systéme na historických dátach, ale iba na jeho aktuálnom stave. Výnimkou sú niektoré plánovače ciest, ktoré poskytujú informácie o agregovaných hodnotách meškania navrhovaných spojení. Nedostupnosť týchto nástrojov znižuje možnosti efektívneho využívania systému hromadnej dopravy cestujúcimi, čo v konečnom dôsledku vedie k nevyužitému potenciálu pre zlepšovanie spokojnosti cestujúcich.

4.2.3 Neprenositeľnosť

Absencia spoločného vnútorného formátu dát a výstupov z analytických nástrojov predstavuje ďalší problém existujúcich riešení. Vzhľadom na rozdielne hardwarové a softwarové riešenia v jednotlivých dopravných systémoch sú možnosti porovnania závislé na definovaní spoločného formátu, ktorý by daný systém a jeho správanie popisoval. Takýmto formátom je napríklad GTFS ktoré môže obsahovať kompletnú štruktúru daného systému verejnej dopravy. Avšak používanie tohto formátu stále nie je najmä v priestore východnej Európy bežným štandartom. Spolu s absenciou dát a nekonzistenciou formátov je tak porovnávanie efektivity jednotlivých systémov medzi sebou zložitou úlohou. Rovnako tak vznikajú duplicitné implementácie rovnakých analytických metód nad rôznymi dátovými modelmi jednotlivých dopravných systémov.

Požiadavky získané analýzou spolu s definovanými problémami predstavujú základný rámec, z ktorého by mal vychádzať dobre definovaný návrh riešenia. Zamýšľané riešenie by tak v konečnom dôsledku malo prispieť k zvýšeniu efektivity systému verejnej dopravy mesta Brna, a tým okrem iného zvýšiť spokojnosť cestujúcej verejnosti.

Kapitola 5

Návrh riešenia

Cieľom mojej práce je navrhnúť nástroj, pomocou ktorého bude možné analyzovať dáta o meškaní vozidiel v systéme hromadnej dopravy mesta Brna. Z analýzy používateľských a technických požiadaviek, ako aj z definície problémov vyplynulo ako najlepšie riešenie navrhnúť komplexný systém pokrývajúci proces spracovania, správy a analýzy týchto dát. Hlavným zámerom mojej práce je doplniť a obohatiť paletu v súčasnosti používaných nástrojov pre správu systému verejnej dopravy mesta Brna vytvorením nového analytického nástroja. Účelom tohto nástroja by malo byť zvýšenie množstva využiteľných znalostí z týchto dát spolu so zjednodušením spôsobu ich získavania. Zároveň tento návrh sleduje zámer prenositeľnosti tohto nástroja pre prípadné použitie nad inými systémami hromadnej dopravy. Samotný návrh zamýšľaného nástroja sa skladá z analýzy dátových zdrojov, ktoré by mali byť v riešení použité, celkovej architektúry systému, podrobného modelu dát, a prehľadu funkcionalít. Všetky ilustrácie v tejto kapitole boli vytvorené pomocou programu Inkscape[1] s využitím mapových podkladov služby OpenStreetMap[2].

5.1 Vstupné zdroje dát

Základom celého návrhu je vyriešenie požiadaviek a problémov, ktoré vyplývajú z potreby pravidelného získavania surových informácií o štruktúre a správaní systému. Bez týchto dát je navrhovaný systém prakticky nepoužiteľný. V rámci prieskumu dostupných zdrojov dát som sa rozhodol pre výber dvoch oddelených dátových zdrojov. Prvý zdroj obsahuje informácie o štruktúre dopravného systému, kým druhý zdroj obsahuje informácie o správaní systému v reálnom čase.

5.1.1 Kolekcia súborov GTFS

Tento zdroj dát skladajúci sa z 9 samostatných textových súborov obsahuje podrobné informácie o štruktúre integrovaného dopravného systému Jihomoravského kraje. Dostupný je verejne v rámci dátového portálu mesta Brna[3]. Podrobná analýza tohto dátového zdroja sa dá zhrnúť do niekoľkých bodov:

- Samotné súbory sú pravidelne aktualizované v týždňovom intervale, pričom dochádza aj k náhodným aktualizáciam medzi pravidelným intervalom v prípade korekcie chyb-

[1] https://inkscape.org/
[2] https://www.openstreetmap.org/
[3] https://data.brno.cz/

ných dát. Napriek týždňovému intervalu aktualizácie sú v rámci súborov zahrnuté aj plánované zmeny v systéme, čo zabezpečuje aktuálnosť dát platných pre daný deň.

- V kolekcii týchto súborov je obsiahnutá časová, aj priestorová štruktúra systému. V rámci oddelených súborov sú obsiahnuté polohy zastávok, trasy liniek a časové polohy spojov. Taktiež sú tu obsiahnuté dáta o garantovaných prestupoch medzi jednotlivými spojmi.

- Problém predstavujú chyby, ktoré tieto dáta obsahujú a neúplnosť, ktorá sa prejavuje najmä pri identifikácií spojov. V rámci dát napríklad dochádza k zámene čísla nástupiska zastávky v rámci trasy, alebo k projekcii neexistujúcich spojov. Nepresnosti vykazujú aj geografické súradnice jednotlivých objektov v priestore, kedy poloha niektorých označníkov je posunutá oproti reálnemu umiestneniu o desiatky metrov.

Celkovo sú však tieto dát spoľahlivé a vhodné pre využitie v rámci návrhu a implementácie zamýšľaného nástroja.

5.1.2 Záznamy o stave vozidiel

Druhým kľúčovým zdrojom dát sú záznamy o stave vozidiel v určitom čase. Podobne ako predchádzajúci zdroj dát je aj tento zdroj verejne dostupný v rámci dátového portálu mesta Brna. Záznamy pochádzajú zo všetkých vozidiel v rámci integrovaného dopravného systému Jihomoravského kraje. Ziskavanie týchto dát prebieha spájaním viacerých dátových zdrojov, pričom napríklad v rámci Brna sa jedná o dáta zo systému RIS II[4]. Analýzu tohto zdroja dát je tak možné zhrnúť do nasledujúcich bodov:

- Vlastná štruktúra záznamov je pevne daná a osahuje rôzne atribúty od polohy vozidla, cez linku a trasu až po aktuálne meškanie oproti cestovnému poriadku, a časovú značku záznamu.

- Samotné záznamy je možné získavať dvoma spôsobmi. Prvým je prúd dát, v ktorom sa nachádzajú záznamy o vozidlách aktuálne vykonávajúcich službu. Nové záznamy prichádzajú každých 10 sekúnd, pričom v dopravnej špičke môže každá aktualizácia obsahovať až 1400 záznamov. Druhým spôsobom je databáza, ktorá uchováva všetky prijaté záznamy za posledných 48 hodín. Samotná databáza funguje na princípe FIFO zásobníka, kedy s každým uložením nových záznamov sa najstaršie vymažú. Kompletné záznamy za predchádzajúci deň je tak možné získať v 24 hodinovom okne.

- Tieto záznamy taktiež obsahujú dáta popisujúce mimoriadne udalosti v reálnom čase. V záznamoch sa tak napríklad nachádajú zmeny trasy, operatívna náhradná doprava, či odklony spôsobené dopravnými nehodami.

- Problematickým je nepresnosť a redundancia dát, ktorá sa prejavuje najmä odosielaním dát aj v pripade, že vozidlo žiaden spoj aktuálne nevykonáva. Opačným problémom je absencia dát, ktorá je spôsobená neodosielaním záznamov vozidlom pre fyzické prekážky, alebo poruchy informačného systému.

[4] https://www.dpmb.cz/ridici-informacni-system-pro-mhd-brno-ris-ii

5.2 Architektúra systému

Návrh architektúry zamýšľaného nástroja vychádza z návrhového vzoru Model–view–adapter. Hlavnou riadiacou jednotkou bude adapter, ktorý bude riadiť všetky interné funkcionality vrátane komunikácie medzi model a view. Celkovo je tak možné rozdeliť systém a jeho funkcionality na tri časti:

- Samotný model systému bude uchovávať spracované informácie o meškaní, indexované podľa jednotlivých dní. Zároveň bude uchovávať aj nevyhnutné dáta popisujúce dopravný systém a jeho zmeny.
- View bude reprezentovaný používateľskou aplikáciou, ktorá bude umožňovať dáta o meškaní vizualizovať, filtrovať a agregovať. Zároveň bude zodpovedná za generovanie podrobnejších analýz nad poskytnutými dátami.
- Celá riadiaca logika systému bude uložená v adaptéry. Ten bude zodpovedný za samotný proces spracovania a ukladania vstupných dát, a za sledovanie a zaznamenávanie zmien v dopravnom systéme. Rovnako bude zodpovedný za odpovede na dátové požiadavky používateľskej aplikácie, ako aj za predprípravu dát na podrobnejšie analýzy realizované samotnou používateľskou aplikáciou.

Z požiadaviek sa ako najlepšie javí navrhnúť systém ako webovú a serverovú aplikácie s dátovým úložiskom. Koncepčne by sa tak malo jednať o databázu implementujúcu model, používateľskú webovú aplikáciu realizujúcu view, a serverovú aplikáciu zastupujúcu adaptér. Tým bude zabezpečená vysoká prenositeľnosť a dostupnosť zamýšľaného systému.

Vzhľadom na povahu dát sa ako najlepšie riešenie používateľskej aplikácie javí využitie geovizualizácie. Základ tak bude tvorený vizualizáciou mapových podkladov so základnými mapovými funkcionalitami, do ktorých budú vizualizované samotné dáta o meškaní. Nástroje pre filtráciu, agregácie a generovanie analýz budú riešené cez integrované grafické rozhranie.

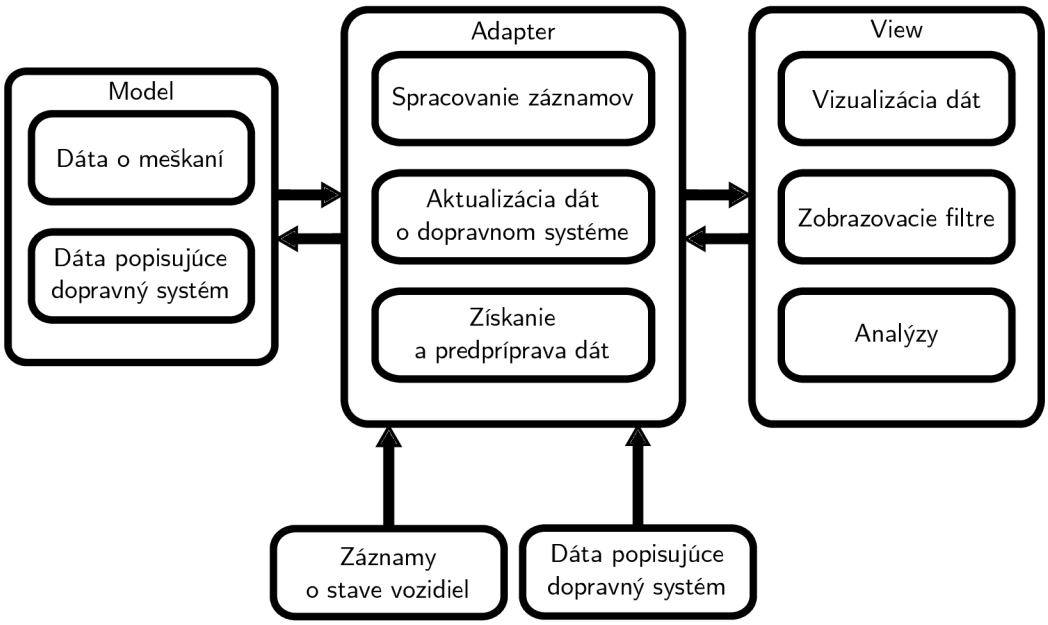
Prehľadná schéma celej aplikácie je zobrazená na obrázku 5.1. Jednotlivé vzťahy medzi procesmi, dátami a požiadavkami ukazujú prepojenie medzi jednotlivými prvkami použitého návrhového vzoru.

5.3 Dátový model

V navrhnutom systéme sa budú vyskytovať dve rozdielne skupiny dát, ktoré však spolu úzko súvisia. Prvá skupina dát zahŕňa dáta o meškaní, ako v podobe vstupných záznamov, tak ich spracovanej podobe. Spracovanie dát zahŕňa komprimáciu objemu a extrakciu užitočných informácií a úzko súvisí so zvoleným spôsobom vizualizácie týchto dát používateľovi. Podrobne je navrhnutý princíp reprezentácie spolu s algoritmom pre spracovanie dát popísaný v sekcii 5.3.2. Druhou skupinou sú dáta popisujúce systém verejnej dopravy mesta Brna a jeho zmeny v čase.

5.3.1 Druhy dát

Rozdelenie dát v návrhu na dve skupiny súvisí s navrhnutým princípom ich reprezentácie a spracovania. Keďže dáta o meškaní neobsahujú dostatok informácií, aby bolo možné vizualizovať ich vzťah k samotnému dopravnému systému, je nevyhnutné samotný systém



Obr. 5.1: **Návrh architektúry.** Zamýšľaný systém vychádza zo vzoru Model-view-adapter.

reprezentovať vlastnou skupinou dát. V konečnom dôsledku je tak prvá skupina bez druhej nekompletná, kým druhá skupina bez prvej nemá v rámci mojej práce využitie.

Dáta o meškaní

Ako vyplýva z analýzy, vstupom navrhnutého nástroja budú záznamy o stave vozidiel v konkrétnom čase. Tieto záznamy sa v systéme budú nachádzať iba počas procesu zoskupovania a extrakcie. Samotné záznamy teda nebudú v systéme ukladané. V systéme budú ukladané extrahované informácie o meškaní v internej reprezentácií. Táto reprezentácia sa bude skladať z časovej značky, identifikátoru spoja a trasy, a sady zaznamenaných hodnôt meškania. Uložené hodnoty meškania budú zodpovedať geografickej reprezentácií trasy daného spoja v deň zodpovedajúci časovej značke. Súčasťou návrhu je aj automatická redukcia a odstraňovanie záznamov po určitom časovom období, kedy stratia svoju relevanciu.

Dáta popisujúce dopravný systém

Samotné záznamy, ktoré tvoria vstup zamýšľaného nástroja, neobsahujú dostatok informácií, aby z nich bolo možné generovať štruktúru dopravného systému. Pričom práve táto šturuktúra v podobe geografickej definície trás liniek je nevyhnutná pre navrhnutý spôsob spracovania. Na druhej strane je štruktúra systému v podobe polohy zastávok, ich názvov a vzťahov medzi nimi základným kameňom generovania výstupov celého systému. Preto existuje druhá skupina dát, ktorá reprezentuje stavy systému v čase. Každý záznam o stave sa bude skladať zo sady vzájomne prepojených objektov, pričom v rámci aktualizácie dát bude dochádzať k automatickému prepájaniu objektov medzi jednotlivými stavmi.

5.3.2 Princíp spracovania a vizualizácie dát

Pre splnenie požiadaviek daných analýzou, ako aj samotným zadaním mojej práce, som považoval za kľúčové navrhnúť princíp reprezentácie dát o meškaní. Z niekoľkých návrhov som vybral ten, ktorý považujem osobne za najvhodnejšie riešenie v podmienkach predpokladaného prostredia nasadenia systému. Tento návrh sa snaží minimalizovať časové, priestorové, a výpočetné požiadavky. Cieľom je taká reprezentácia dát, ktorá poskytne čo najlepšie predpoklady pre získavanie užitočných znalostí zo vstupných dát.

Princíp reprezentácie dát

Základnou myšlienkou tohto návrhu je, že meškanie vzniká najmä vplyvom fyzickej infraštruktúry. Aby bolo možné pochopiť tento vplyv, je nevyhnutné zachytiť vzťah medzi samotným meškaním a priestorom, kde toto meškanie vzniká. Podstatou sa tak stáva reprezentácia v podobe sady hodnôt meškania na jednotlivých prvkoch infraštruktúry v konkrétnom čase. Medzi týmito hodnotami je potom možné modelovať vzťahy jednak na úrovni fyzickej infraštruktúry, ale aj na úrovni systému verejnej dopravy. Inak povedané, je možné napríklad sledovať vývoj meškania na konkrétnej ulici, alebo trase konkrétnej linky. Vďaka univerzálnosti takéhoto princípu reprezentácie bude systém založený na tomto návrhu schopný generovať celú paletu rôznych analýz.

Algoritmus spracovania dát

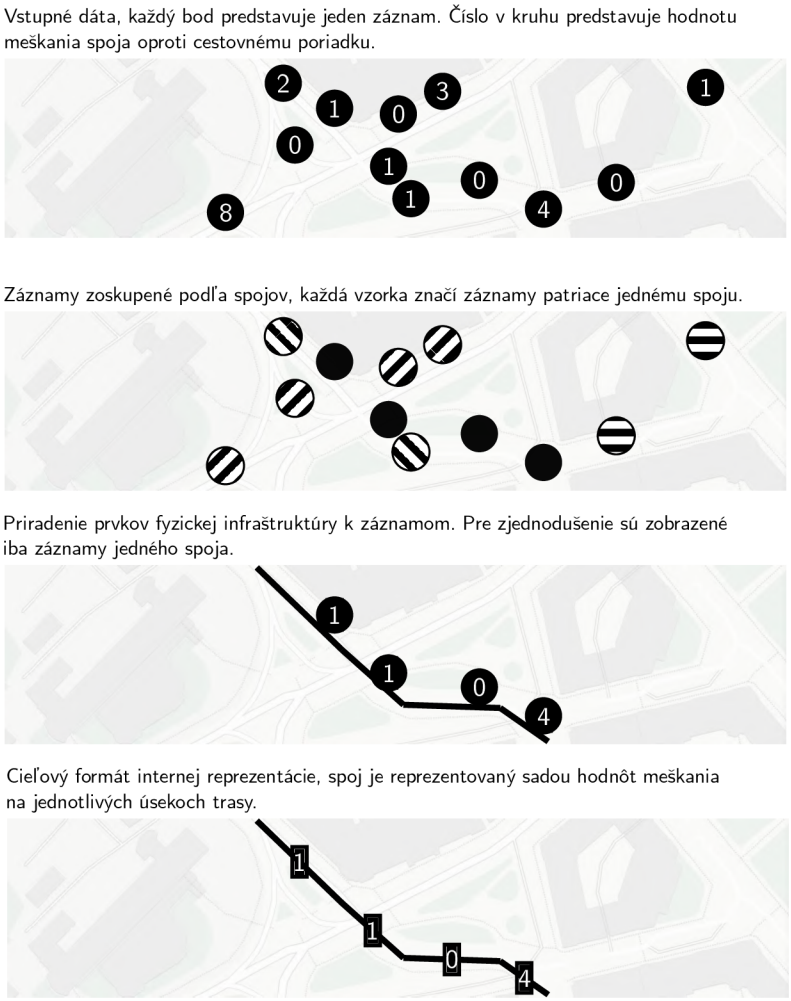
Aby bolo možné navrhnutý princíp realizovať, je nevyhnutné vstupné záznamy o stave vozidiel spracovať do formátu zodpovedajúcemu navrhnutému princípu. Algoritmus realizujúci toto spracovanie bude založený na dávkovom spracovaní záznamov za jeden konkrétny deň. Samotná štruktúra ukladania spracovaných záznamov bude založená na rozdelení na jednotlivé spoje. Tento spôsob je pamäťovo aj výpočetne vhodnejší, ako rozdelenie na základe objektov fyzickej infraštruktúry. Dôležitým prvkom bude princíp ukladania iba jednej hodnoty pre každú časť fyzickej infraštruktúry. Ignorované sú tak záznamy, ktoré obsahujú veľmi podobné dáta. Takáto redukcia záznamov však neznižuje informačnú hodnotu redukovaných dát. Navrhnutý algoritmus spracovania vstupných záznamov je možné zhrnúť do troch krokov:

1. Zoskupenie záznamov podľa jednotlivých spojov. Vlastné záznamy získané z databázy budú roztriedené podľa svojich atribútov na jednotlivé spoje.

2. Obohatenie každého záznamu o informáciu, ktorému prvku fyzickej infraštruktúry patrí na základe svojich geografických atribútov. Následne bude nad každou sadou záznamov reprezentujúcich jednotlivé spoje vykonaná redukcia. Pre každú časť trasy spoja, časť fyzickej infraštruktúry, bude uložený iba jeden najnovší záznam.

3. Uloženie každej sady záznamov ako poľa hodnôt meškania, kde každý prvok reprezentuje časť trasy, časovej značky, a identifikátoru spoja.

Podrobnú grafickú reprezentáciu navrhnutého algoritmu zachytáva obrázok 5.2. Nad takto uloženými dátami je potom možné jednoducho modelovať časové, priestorové, aj štrukturálne vzťahy.



Obr. 5.2: **Algoritmus pre spracovanie dát.** Vizualizácia zodpovedá trom krokom spracovania, pričom prvá vizualizácia zobrazuje formát vstupných záznamov.

5.4 Prehľad navrhnutých funkcionalít

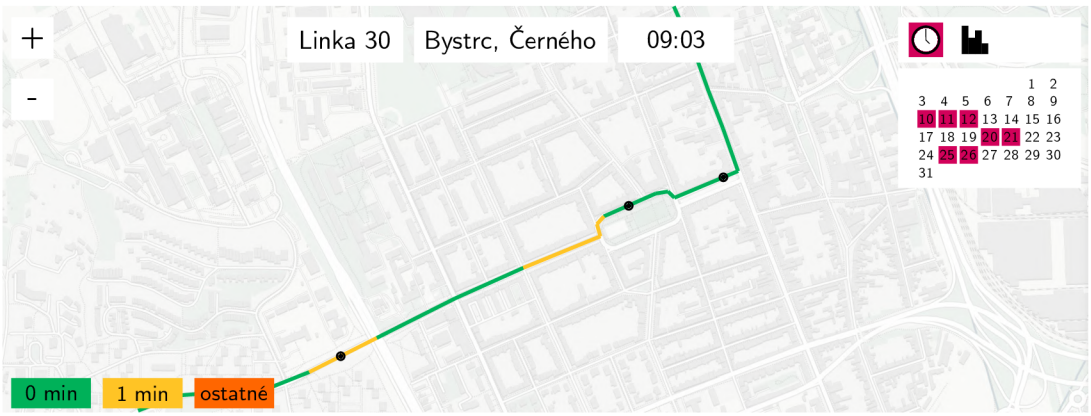
Kým návrh architektúry a dátového modelu popisujú ideový základ zamýšľaného nástroja, navrhnuté funkcionality sú ich konkrétnym zobrazením do budúcich implementačných celkov. Definovaný návrh sa tak stáva základom postupu, ako zamýšľaný nástroj implementovať. Celkovo je možné navrhované funkcionality rozdeliť na tie, ktoré budú vyžadovať používateľský vstup, a tie, ktoré budú pracovať autonómne.

5.4.1 Používateľské funkcionality

Implementované budú v klientskej webovej aplikácií a ich hlavnou úlohou je umožniť používateľovi vizualizovať a analyzovať dáta. Ich vstupom budú používateľské požiadavky. Každá funkcionalita spracuje vstup, požiadaja o zodpovedajúce dáta serverovú aplikáciu, a následne tieto dáta príslušným spôsobom vizualizuje. V prípade analýzy dát budú niektoré údaje interpolované zo získaných dát priamo v klientskej aplikácií.

Základná vizualizácia dát

Základom tejto funkcionality bude geovizualizačná metóda mapy spojení. Používateľ si zvolí linku, trasu a konkrétny spoj. Potom si zvolí časový rozsah, z ktorého chce dáta vizualizovať. Na výstupe sa potom zobrazí celá trasa v podobe prepojených zastávok. Používateľ si tiež bude môcť definovať vlastné kategórie hodnôt meškania a farebne ich označiť. Jednotlivé úseky trasy budú zodpovedať príslušným častiam cestnej, alebo koľajovej siete. Zároveň budú farebne rozlíšené na základe priemernej hodnoty meškania jednotlivých spojov v jednotlivé dni zvoleného časového rozsahu. Ilustratívny príklad tejto funkcionality poskytuje obrázok 5.3. Vizualizovaná je časť štatistiky spoja na linke 30. Používateľské rozhranie bude okrem hlavných ovládacích prvkov obsahovať aj správu vlastných kategórií meškania umiestnenú vľavo dole. Na pravej strane sa bude nachádzať ovládanie časového rozsahu a analytické nástroje. Rozšírením základnej vizualizácie bude možnosť agregovať viacero spojov do jedného výstupu, alebo porovanie jednotlivých spojov medzi sebou.



Obr. 5.3: **Používateľské rozhranie.** Návrh klientskej aplikácie je založený na geovizualizácií s integrovaným ovládaním.

Analýza štatistiky meškaní

Prvou zo základných analýz bude štatistika meškaní pre jednotlivé zastávky. Používateľ si opäť zvolí konkrétny spoj a časový rozsah. Opäť však bude mať možnosť agregácie viacerých spojov do jedného výsupu, alebo porovnania spojov medzi sebou. Používateľská aplikácia následne na základe nastavených kategórií meškania spočíta pre každú zastávku na trase pomer spojov v jednotlivých kategóriach meškania.

Analýza meškaní vzhľadom na zastávky

Predstavuje druhú základnú analýzu nad dátami. Podobne ako predchádzajúce funkcionality aj tu si používateľ môže zvoliť jeden spoj, agregáciu viacerých spojov, alebo ich porovnanie. Podstatou tejto funkcionality je agregácia hodnôt celkového meškania pre jednotlivé zastávky. Výstup tak zobrazuje údaje o tom, do ktorej zastávky prichádzajú spoje s najvyššími hodnotami meškania.

5.4.2 Systémové funkcionality

Táto skupina funkcionalít obsahuje najmä procesy spojené so správou dát. Okrem toho je tu zahrnutá aj správa údajov o dopravnom systéme a ich aktualizácia, ako aj procesy zotavovania sa z chýb. Spoločnými menovateľmi sú autonómnosť a robustnosť týchto procesov.

Správa dát o meškaní

Celé spracovanie vstupných záznamov, spolu s ukladaním a organizáciou spracovaných dát bude plne automatické. Spracovanie bude prebiehať raz za deň, kedy budú spracované všetky záznamy z predchádzajúceho dňa. Spolu so samotnou implementáciou algoritmu pre spracovanie budú fungovať aj diagnostické a štatistické funkcie pre riešenie anomálií a mimoriadností v dátach. Samotný algoritmus bude implementovaný s dôrazom na robustnosť a spoľahlivosť.

Správa dát o dopravnom systéme

Aktualizácia údajov popisujúcich systém verejnej dopravy mesta Brna bude prebiehať každý deň. V systéme budú okrem aktuálneho stavu uložené aj dáta popisujúce predchádzajúce stavy systému. Historické dáta o meškaní tak bude možné použiť v kontexte takého stavu systému, v ktorom boli tieto meškania zaznamenané. Základom celej správy dát v tejto funkcionalite budú dáta popisujúce fyzickú koľajovú a cestnú dopravu, ktoré bude v prípade zmien aktualizovať osobitne.

Správa požiadaviek klientskej aplikácie

Serverová aplikácia tiež bude zodpovedná za odpovede na požiadavky z používateľskej aplikácie. Na základe parametrov budú z dátového úložiska načítané a predspracované požadované dáta. V rámci týchto požiadaviek taktiež budú odosielané dáta popisujúce dopravný systém v príslušnom časovom okamihu.

Riešenie problémov a zotavenie z chýb

V tejto skupine funkcionalít sú funkcie kontrolujúce dostupnosť dátových zdrojov a úložísk, a riešiace ich nedostupnosť. Tieto funkcie budú kontrolovať kompletnosť spracovania dát o meškaní, záznamov o dopravnom systéme a v prípade problémov sa budú snažiť tieto dáta skompletizovať. Taktiež budú ukladať diagnostické informácie a v nevyhnutných prípadoch upozornia správcu systému na nevyhnutnosť jeho zásahu.

Kapitola 6

Implementácia

Popis samotnej implementácie na základe návrhu z predchádzajúcej kapitoly pozostáva z troch častí. V prvej časti sú uvedené jednotlivé technológie a knižnice, ktoré riešenie využíva. V druhej časti je popísaná celková architektúra systému, spôsob modelovania skutočného systému hromadnej dopravy mesta Brna na základe dostupných dát, a cyklus spracovania záznamov o stave vozidiel. Posledná časť je venovaná dvom význačným modulom, ktoré riešenie obsahuje. Prvým je routovací algoritmus, ktorého implementácia vyplýva z potreby doplnenia vstupných dát metódou interpolácie, a analytického modulu, ktorý nad dátami umožňuje vykonávať základné analýzy.

6.1 Použité technológie

V súlade s návrhom architektúry je riešenie rozdelné na 3 samostatné celky. Prvým je databáza, pričom riešenie využíva technológiu MongoDB[1]. Komunikáciu medzi databázou a serverovou časťou zabezpečuje knižnica Mongoose[2]. Samotné riešenie je koncipované tak, aby bolo možné zvolený typ databázy jednoducho zameniť.

Druhú časť tvorí serverová aplikácia, ktorá je hlavným riadicim prvkom celého implementovaného riešenia. Funkčnosť serverovej aplikácie zabezpečuje technológia Node.js[3], konkrétne je základom aplikácie framework Express[4]. Tento framework zabezpečuje komunikáciu s klientskou aplikáciou, získavanie dát z externých zdrojov, ako aj komunikáciu s databázou cez uvedenú knižnicu.

Tretia časť pozostáva z klientskej aplikácie. Opäť je využitá technológia Node.js v kombinácií s knižnicami React[5], Leaflet[6] a PrimeReact[7]. Tieto knižnice poskytujú nástroje pre návrh a implementáciu ovládacích prvkov, ako aj pre implementáciu použitých geovizualizačných metód.

[1] https://www.mongodb.com/
[2] https://mongoosejs.com/
[3] https://nodejs.org/
[4] https://expressjs.com/
[5] https://react.dev/
[6] https://leafletjs.com/
[7] https://primereact.org/

6.2 Architektúra

Samotná architektúra implementovaného nástroja vychádza z návrhového vzoru, ktorý je popisovaný v predchádzajúcej kapitole. Model realizovaný formou databázy zabezpečuje ukladanie a správu dát o systéme a jeho správaní, ako aj dáta popisujúce fyzickú infraštruktúru. Adapter predstavuje hlavný riadiaci prvok celého systému. Jeho úlohou je sledovať zmeny v dopravnom systéme, spracovávať a anotovať záznamy o stave vozidiel, ako aj spracovávať požiadavky klientskej aplikácie a poskytovať jej požadované dáta. View realizovaný ako klientská aplikácia potom slúži pre vizualizáciu uložených záznamov. Zároveň taktiež umožňuje ich filtráciu a zabezpečuje generovanie základných analýz.

6.2.1 Model

V rámci implementácie bol zvolený prístup, kedy sú všetky ukladané entity modelované ako objekty so svojimi atribútmi. V prípade dát o fyzickej infraštruktúre sa jedná o dve samostatné sady pre koľajovú a cestnú sieť. Každá sada je zložená z bodov, ich súradníc a susedov, do ktorých sa dá z daného bodu, rešpektujúc určité pravidlá, prejsť. Vzťahy v rámci týchto sietí sú tak definované pred načítaním do databázy.

Entitity modelujúce dopravný systém rešpektujú reálnu štruktúru systému hromadnej dopravy. Jednotlivé entity, ako aj podrobná štruktúra vzťahov medzi nimi je popísaná v sekcii 6.2.2. V rámci databázy sú tieto vzťahy modelované prostredníctvom odkazovaní sa na unikátne id objektu.

Samotné záznamy o priebehu spojov sú ukladané ako samostatná entita, pričom sú previazané s príslušnými prvkami dopravného systému. Okrem týchto záznamov sú v databáze uložené aj štatistiky pre jednotlivé dni. Tieto napríklad obsahujú počet vykonávaných spojov podľa cestovného poriadku, alebo mieru úspešnosti spracovania vstupných záznamov.

6.2.2 Adapter

Obsahuje riadiacu logiku celého nástroja. Taktiež je adapter zodpovedný za všetky výpočetne náročnejšie úlohy, ktoré systém vykonáva na pravidelnej báze. Celkovo je možné všetky úlohy, ktoré adapter plní, zhrnúť do niekoľkých bodov:

- Prvou úlohou je načítanie a aktualizácia dát o fyzickej infraštruktúre. Tieto dáta sú v počiatočnnej fáze reprezentované dvojicou predspracovaných súborov a sú jediným prvkom, ktorý je nutné ručne aktualizovať. Systém potom zodpovedá za ich načítanie a poskytovanie týchto dát pri routingu, alebo na základe používateľských požiadaviek.
- Druhou úlohou systému je na dennej báze aktualizovať model samotného systému hromadnej dopravy mesta Brna z hľadiska časového aj priestorového. Systém pre plnenie tejto úlohy využíva ako zdroj dát kolekciu súborou formátu GTFS, podrobne popísaných v sekcii 5.1.1.
- Tretia úloha systému je na základe aktuálneho modelu spracovávať záznamy o stave vozidiel, čím vznikajú záznamy o správaní systému v konkrétnom čase. Toto spracovanie prebieha každý deň, pričom je využívaná databáza záznamov podrobne popísaná v sekcii 5.1.2.
- Štvrtou úlohou je správa požiadaviek klientskej aplikácie. Adapter poskytuje informácie o dostupných spojov vo vybraných dňoch, ako aj dáta o priebehu spoja v rámci

zvoleného časového rozsahu. Tieto dáta sú poskytované vždy spolu s geografickými údajmi o danom spoji, akými sú zastávky a trasa spoja v mapovom priestore.

V rámci týchto úloh sú výpočetne, aj komplexne zložitejšie najmä úlohy modelovania systému a spracovania samotných záznamov. Z tohto dôvodu sú im venované nasledujúce podsekcie, ktoré ich podrobnejšie popisujú.

Model dopravného systému

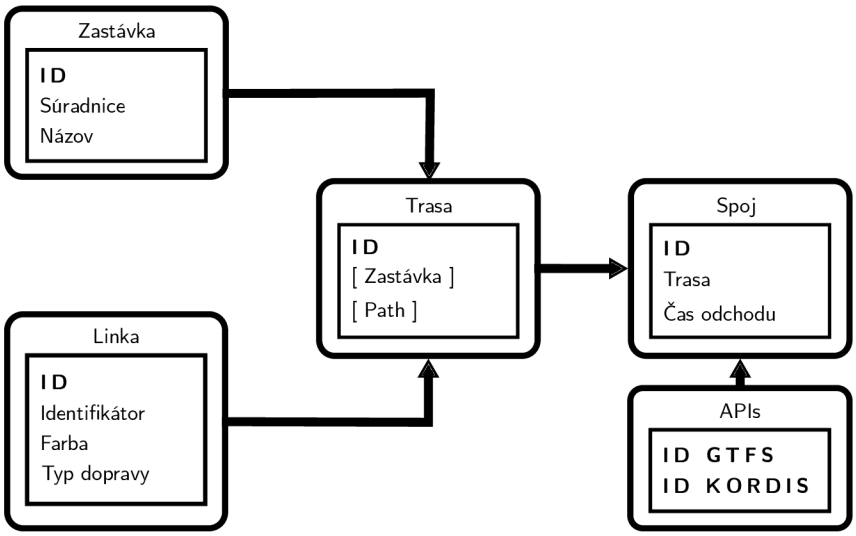
Systémy hromadnej dopravy je, v prevažnej väčšine, možné popísať modelom rozdeleným do niekoľkých úrovní. Toto rozdelenie vychádza zo štruktúry samotného systému, kedy sa niektoré prvky a ich vzťahy menia s vyššou frekvenciou ako iné. Zároveň sú prvky vyššej úrovne vždy naviazané na prvky úrovne nižšej. Typicky je možné rozlišovať nasledujúce tri úrovne:

- Fyzická úroveň – obsahuje zastávky, ich polohy, koľajovú a cestnú sieť. Zmeny v tejto úrovni prebiehajú najmenej často.
- Priestorová úroveň – je tvorená linkami a nim prislúchajúcimi trasami. Na tejto úrovni dochádza k zmenám častejšie, ako na úrovni predchádzajúcej.
- Časová úroveň – je nadynamickejšou úrovňou. Jej obsah je tvorený najmä spojmi, spolu s dátami o platnosti jednotlivých spojov v rámci rôznych režimov dopravného systému.

Samotná implementácia modelovania tohto systému vychádza jednak z popisovanej štruktúry a zároveň reflektuje potrebu neustálej aktualizácie daného modelu. Vzťahy medzi jednotlivými entitami sú tak modelované presne podľa popisovaného modelu. Avšak, aby bola splnená požiadavka reflektovať zmeny v systéme, existuje okrem samotnej reprezentácie prvku aj sada odkazov na aktuálne platné verzie všetkých prvkov danej entity. V rámci každodennej aktualizácie sú tak vzťahy, definované v sade súborov GTFS, modelované medzi najaktuálnejšími verziami entít systému. Aktualizácia jednotlivých entít prebieha na základe sady súborov GTFS platnej pre daný deň. Takýmto spôsobom je možné modelovať jednak vzťahy, a zároveň udržiavať históriu stavov systému. To umožňuje spracované dáta vnímať v rámci kontextu takej podoby systému, aká bola platná v dobe vzniku týchto dát. Z takto uložených dát je potom možné rekonštruovať správanie dopravného systému v čase. Výhodou tohto riešenia je pokrytie dynamických zmien v systéme a zároveň pamäťová nenáročnosť. Okrem toho je tento princíp prenositeľný na akýkoľvek systém hromadnej dopravy popísaný formátom GTFS.

Osobitý problém predstavuje neúplnosť dát v rámci GTFS. Tieto dáta neobsahujú definície trás v priestore, pričom tieto definície sú kľúčovým prvkom v navrhnutom spôsobe spracovania a ukladania dát. Vzhľadom na nedostupnosť týchto dát je v rámci riešenia implementovaný modul routingu, popisovaný v sekcii 6.3.1, ktorý jednotlivé definície trás interpoluje na základe polohy a poradia zastávok, a fyzickej infraštruktúry.

Celkový obraz modelu a vzťahov medzi entitami znázorňuje obrázok 6.1. Špecifickým prvkov je entita APIs, ktorá obsahuje zoznam prepojení medzi indexáciou spojov v rámci formátu GTFS a interného formátu KORDIS. Táto entita má vždy jednodňovú platnosť a existuje iba po dobu spracovania dát o stave vozidiel.



Obr. 6.1: **Model systému hromadnej dopravy.** Jednotlivé entity a ich vzťahy sú modelované tak, aby implementovaný nástroj dokázal reagovať na zmeny a zároveň si uchovávať minulé stavy systému.

Spracovanie dát o stave vozidiel

Správanie systému hromadnej dopravy je možné popísať rôznymi spôsobmi. Na základe prieskumu existujúcich riešení a dostupných dát je táto práca založená na sledovaní priebehu jednotlivých spojov v rámci času a priestoru. Najdôležitejšou informáciou sa tak stáva hodnota meškania daného spoja v určitom čase a mieste. A práve tieto dáta sú získavané a spracovávané v rámci implementovaného nástroja presne podľa algoritmu navrhnutého v sekcii 5.3.2.

Spracovanie záznamov prebieha v jednodňových intervaloch, pričom sú spracované všetky platné záznamy za uplynulý deň. Celý proces od získania, cez spracovanie, až po uloženie do cieľovej podoby je možné zhrnúť do nasledujúcich krokov:

1. Získanie a predpríprava spojov a definícií trás na základe aktuálnych prepojení v entite APIs. Zároveň sú takto pripravené dáta zoskupené na základe koreňovej linky.

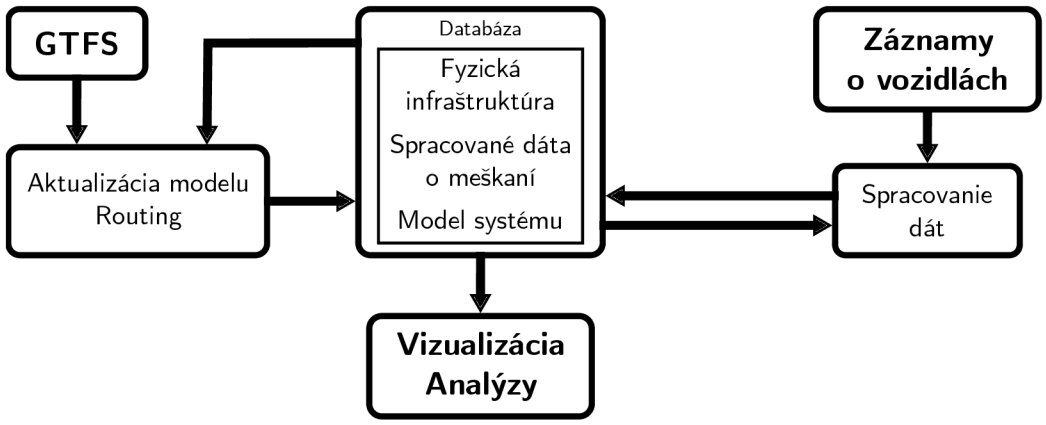
2. Dávkové spracovanie záznamov na základe liniek. Pre každú linku sú stiahnuté všetky záznamy, pričom následne sú záznamy na základe prepojení v APIs priradené konkrétnym spojom. Pre každý spoj tak vzniká sada záznamov, ktoré k nemu patria.

3. Jednotlivé sady záznamov sú následne spracovávané do cieľového formátu. Na základe trasy spoja a geografického atribútu záznamu sú jednotlivé meškania priradené častiam trasy, ktorým prislúchajú.

4. Výsledok v podobe poľa hodnôt je následne uložený do databázy spolu s časovou značkou spracovávaného dňa. Pre každý spoj je tak uložený jeho priebeh v daný deň na danej trase.

Takto navrhnutý spôsob spracovania dokáže úspešne eliminovať duplicitné, alebo veľmi podobné záznamy, celú triedu možných anomálií dát a zároveň zachovať hodnotu informácie

na úrovni pôvodných záznamov. Zároveň tento spôsob ukladania redukuje objem uložených dát. Dáta sú tak anotované a uložené na najnižšej úrovni abstrakcie, čo umožňuje široké možnosti filtrácie a agregácie. Nevýhodou je možná strata zaujímavých anomálií, ako aj strata užitočných dát v prípade chýb v kolekcii súborov GTFS. Celkový pohľad na vzťahy medzi dátovými vstupmi a ich internou reprezentáciou v rámci implementovaného nástroja zobrazuje obrázok 6.2.



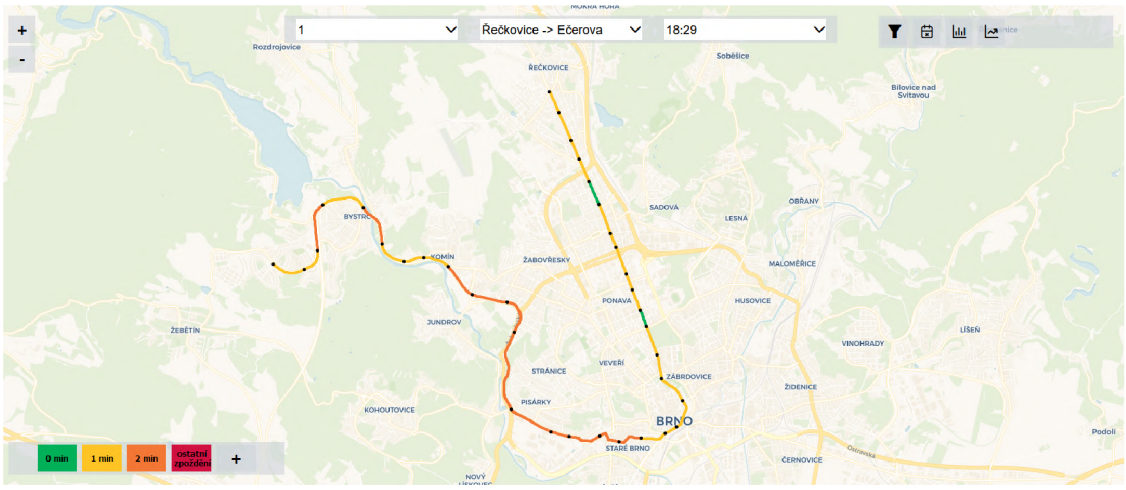
Obr. 6.2: **Vzťahy medzi dátovými entitami** Pri aktualizácií dát a ukladaní záznamov o správaní sú využívané dva vstupné zdroje dát spolu s dátami uloženými v databáze. Vzniká tak cyklická závislosť.

6.2.3 View

View je implementovaný klientskou aplikáciu, a poskytuje vizualizáciu dát a platformu pre ich analýzy. Základom používateľského rozhrania je metóda geovizualizácie. Nad podkladovou mapou sú potom implementované jednotlivé ovládacie prvky. Prehľadne zobrazuje používateľskú aplikáciu v stave po inicializácií obrázok 6.3. Základom ovládania sú tri selektory pre výber linky, trasy a konkrétneho spoja. V rámci týchto selektorov sú jednotlivé prvky systému zoradené vzostupne v lexikografickom poradí a je možné v nich vyhľadávať.

V ľavej časti sa okrem možností zmeny mapovej mierky nachádza aj panel pre definíciu kategórií meškania. V tomto si používateľ môže rozdeliť jednotlivé hodnoty meškania do farebne rozlíšených kategórií podľa svojej potreby. Definovanie týchto kategórií následne ovplyvňuje aj výstupy generované jednotlivými analýzami.

V pravej časti sa nachádza ovládací panel, ktorý obsahuje implementované analytické moduly spolu s možnosťami filtrácie vizualizovaných dát. V prvej záložke sa nachádza modul kalendára, cez ktorý je možné vyberať dni, z ktorých budú použité dáta v rámci vizualizácie a analýz. Po zvolení vybraných dní adapter pripraví pre view zoznam spojov, ktoré v rámci všetkých týchto dní premávali. Používateľ má tak istotu, že získa všetky dostupné dáta. Druhá záložka pre filtrovanie obsahuje možnosti zobrazenia presných hodnôt meškania, ako boli spracované, a možnosť agregácie meškania v rámci trasy spoja.



Obr. 6.3: **Používateľské rozhranie.** Základom klientskej aplikácie je podkladová mapa, nad ktorou sú implementované ovládacie prvky.

6.3 Systémové moduly

Implementované riešenie obsahuje dva význačné moduly, pričom jeden je implementovaný v rámci adapteru a vznikol ako riešenie problému vzniknutého počas implementácie. Druhý modul je realizáciou časti návrhu nástroja a je implementovaný v rámci view.

6.3.1 Routing

Routing, alebo hľadanie trasy, predstavuje v rámci riešenia tejto práce dôležitý prvok kompletizácie vstupných dát. Jeho vstupom je postupnosť zastávok a ich súradníc na trase. Výstupom je pole indexov uzlov tvoriacich trasu. Navrhnutý algoritmus potom nad mapou fyzickej infraštruktúry spočíta trasu v podobe uzlov, cez ktoré trasa prechádza. Samotný algoritmus bol navrhnutý tak, aby bol znovupoužiteľný s rôznymi parametrami pre všetky druhy dopravy v rámci systému hromadnej dopravy mesta Brna. Riešenie s využitím vlastného routovacieho algoritmu vychádzalo z potreby zachovať nezávislosť nástroja od externých služieb, ako aj predpoklad externého využitia tohto algoritmu.

Samotný algoritmus funguje na základe hľadania spojenia medzi dvoma zastávkami. Z týchto spojení je potom poskladaná následná trasa. Základom algoritmu je metóda A*, pričom heuristická funkcia je založená na výpočte vzdialenosti aktuálneho a cieľového uzla. Parametre algoritmu, závislé od druhu dopravy, spočívajú najmä v nastavení počtu aktuálne prehľadávaných možností, maximálnej dĺžky cesty, po ktorej je routing neúspešný, a vzdialeností používaných pri hľadaní počiatočných a koncových uzlov na základe súradníc príslušnej zastávky.

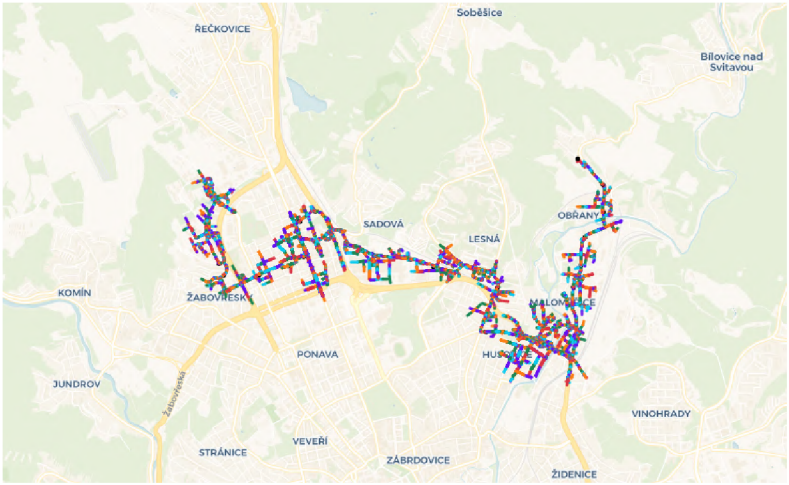
Jednotlivé kroky algoritmu je možné zhrnúť do nasledujúcej postupnosti:

1. Nájdenie počiatočné a koncového uzla – na základe súradníc počiatočnej a koncovej zastávky algoritmus určí v závislosti od druhu dopravy niekoľko vhodných kandidátov, kde by trasa mala začínať a končiť. Algoritmus pritom pracuje vždy s dvojicou uzlov a súradnicami zastávky, pričom vychádza z predpokladu, že súradnice a uzly musia tvoriť ostrouhlý trojuholník. Najvhodnejšími kandidátmi sú potom uzly, kde vzdialenosť súradníc zastávky od priamky medzi týmito uzlami je čo najmenšia.

2. Prehľadávanie stavového priestoru – z počiatočného uzla sú potom v iteráciach prehľadávané susedné uzly. To, ktoré uzly budú prehľadávané závisí od fyzickej štruktúry systému, ako aj uhla odbočenia. Vzniká tak sada možností, ktorých počet je pravidelne redukovaný. Počet, ako aj frekvencia redukcie možností je závislá od druhu dopravy. V každej iterácií je aktualizovaná iba spodná polovica možností na základe aktuálnej dĺžky ich trasy v priestore. Tým dochádza k rovnomernejšiemu prehľadávaniu stavového priestoru, pričom je ignorovaný počet navštívených uzlov. Prehľadávanie končí v prípade, ak nejaká možnosť dosiahne koncový uzol a všetky ostatné možnosti nadobudnú vyššie skóre. Toto skóre je počítané ako súčet do teraz navštívenej trasy, počítanej zo vzdialeností medzi uzlami, a z predpokladanej vzdialenosti do množiny koncových uzlov.

3. Uloženie trasy – výsledná postupnosť uzlov je uložená ako časť trasy, pričom koncový uzol je použitý pri hľadaní nasledujúceho úseku trasy. Zároveň sa ukladajú indexy zastávok, aby pri následných analýzach bolo jednoduché určiť, ku ktorej časti trasy daná zastávka patrí.

Príklad prehľadávania stavového priestoru zachytáva obrázok 6.4, na ktorom je vizualizované prehľadávanie stavového priestoru. Každá farba predstavuje jednu iteráciu, pričom je možné vidieť rovnomerné rozvetvovanie možností.



Obr. 6.4: **Routing.** Prehľadávanie stavového priestoru a pravidelná redukcia možností na základe skóre umožňujú získať trasu linky vo fyzickom priestore.

6.3.2 Analytická platforma

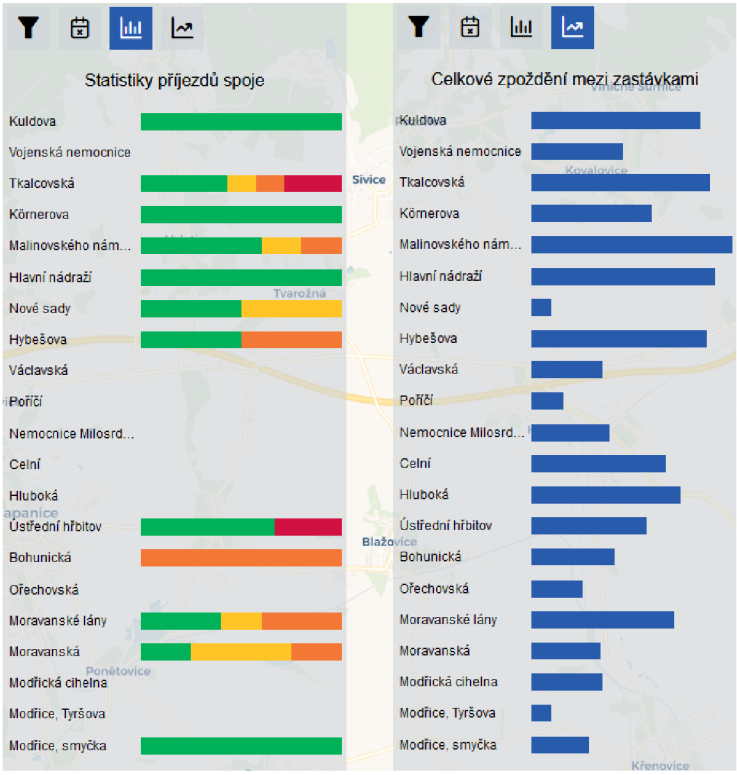
Samotné spracovanie, ukladanie a vizualizácia správania systému predstavujú nevyhnutný základ definovaný návrhom. Aby však bolo možné zo správania systému získavať užitočné znalosti, je nevyhnutné tieto dáta ďalej analyzovať. Pre tento účel je implementovaná analytická platforma, ktorá v rámci klientskej aplikácie z poskytnutých dát generuje na základe používateľských požiadaviek rôzne analýzy. Táto platforma je zároveň navrhnutá tak, aby bolo možné implementovaný nástroj jednoducho rozšíriť o ďalšie analýzy. V rámci tejto práce sú v súlade s návrhom implementované dve základné analýzy.

Analýza štatistiky meškaní

Prvá zo základných analýz vizualizuje štatistiku meškania spoja v jednotlivých dňoch pre jednotlivé zastávky. Používateľ tak získa prehľad o tom, s akou pravdepodobnosťou príde jeho spoj na zastávku včas. Hodnoty meškania sú získavané z časti trasy, ktorá náleží danej zastávke. Keďže tieto dáta sú vzhľadom na nerovnomerné rozloženie v priestore nekompletné, vyskytujú sa v generovanej štatistike prázdne miesta. Jednotlivé štatistické kategórie zodpovedajú kategóriam meškania, ktoré si používateľ nastavuje pri vizualizácií dát. Príklad výstupu takejto analýzy zobrazuje obrázok 6.5.

Agregácia meškaní medzi zastávkami

Druhá zo základných analýz poskytuje informáciu o meškaniach medzi zastávkami. Pre každý úsek medzi jednotlivými zastávkami sú spočítané všetky hodnoty meškania, ktoré sa na danom úseku vyskytli. Na rozdiel od predchádzajúcej analýzy sa v tomto prípade nekompletnosť dát neprejavuje. Príklad výstupu tejto analýzy spolu s porovnaním s analýzou štatistiky meškaní poskytuje obrázok 6.5.



Obr. 6.5: **Základné analýzy.** Porovnanie výstupov analýzy pravdepodobnosti príchodu spoja s určitým meškaním do konkrétnej zastávky, oproti agregovanému meškaniu na celom úseku medzi dvoma zastávkami.

Kapitola 7

Testovanie

Vlastné testovanie implementácie prebiehalo počas celej doby vývoja nástroja. Postupne boli testované jednotlivé moduly adaptera a view. Následne prebiehalo integračné testovanie a testovanie celkovej funkčnosti systému. Počas vývoja vznikli a boli postupne testované tri prototypy nástroja. Záverečná fáza testovania zahŕňala skúšobné nasadenie nástroja v rámci školského servera. Súčasne prebiehalo testovanie používateľského rozhrania v rámci pravidelných stretnutí s Oddělením dat, analýz a evaluací města Brna, ktorého požiadavky stáli jednak na začiatku tejto práce, a zároveň formovali jej smerovanie. V nasledujúcich podsekciách sú postupne popísané tri vybrané oblasti testovania, pričom v každej časti je okrem metodiky a výsledkov aj podsekcia venovaná smerovaniu budúceho vývoja aplikácie v popisovanej sekcii.

7.1 Testovanie routovacieho algoritmu

Výsledná podoba implementovaného routovacieho algoritmu je výsledkom niekoľkých iterácií implementácie a testovania. Cieľom testovania bolo, aby výsledná implementácia v priemere dosahovala úspešnosť aspoň 85%. Úspešnosť bola vypočítavaná ako pomer medzi všetkými vstupnými trasami a trasami, ktorých priestorová definícia sa zhodovala s referenčným riešením na viac ako 98%. Za referenčné riešenie bola stanovená interaktívna mapa IDS JMK. Chybné trasy boli určované samotným algoritmom, ako aj na základe následnej porvnávacej kontroly. V rámci tejto kontroly boli porovnávané súradnice bodov na testovanej a referenčnej trase. Postupne bolo testovaných niekoľko prístupov k algoritmu, pričom výsledná implementácia dosahuje v niektorých prípadoch presnejšie výsledky, ako referenčné riešenie.

7.1.1 Testované postupy

Základnou myšlienkou pri návrhu algoritmu bolo, že cieľom by malo byť inteligentné prehľadávanie stavového priestoru. Na základe tohto predpokladu potom boli navrhované nasledujúce prístupy. Prvým bolo naivné prehľadávanie do šírky, pričom algoritmus bol v rovnakej podobe testovaný pre všetky druhy dopravy. Prvotným pokusom o zlepšenie bola parametrizácia algoritmu na základe druhu dopravy. Tento prístup priniesol skokové zlepšenie úspešnosti, a preto bol použitý vo všetkých neskorších prístupoch.

Druhým prístupom bolo použitie algoritmu Greedy Search, pričom ako heuristická funkcia bolo zvolené počítanie vzdialenosti medzi aktuálne prehľadávaným uzlom a cieľovým

uzlom. Tento prístup vykazoval zlepšenie najmä v oblasti cestnej dopravy, avšak stále nedosahoval požadovanú mieru úspešnosti.

Tretí prístup už využíva cieľový algoritmus A*. Zvolená heuristická funkcia je rovnaká, ako v predchádzajúcom prístupe. Testovaním tohto prístupu bola dosiahnutá najvyššia úspešnosť routingu s použitím čistej implementácie algoritmu. Následne boli testované rôzne vylepšenia v oblasti počítania skóre, ako aj voľbe najvhodnejších uzlov. Ako prínos sa ukázala penalizácia potencionálnych trás s ostrými uhlami odbočenia a rovnomerné prehľadávanie jednotlivých nenavštívených uzlov na základe prejdenej vzdialenosti. Algoritmus tak v každom kroku prechádza a aktualizuje iba tie kandidátne trasy, ktorých doteraz prejdená dĺžka je pod priemerom dĺžky všetkých potencionálnych trás. Takto implementované riešenie už dosiahlo požadovanú presnosť a mohlo byť využité v riešení tejto práce.

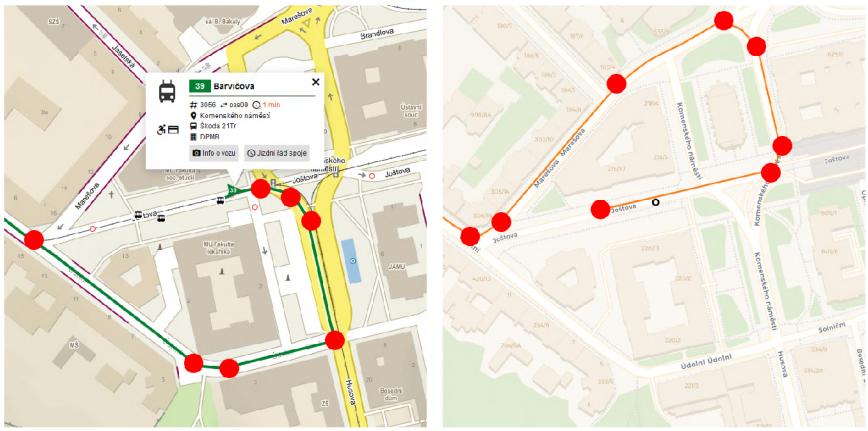
7.1.2 Výsledky testovania

Testovanie každého prístupu prebiehalo nad rovnakou sadou dát. Táto bola tvorená všetkými električkovými, vlakovými, trolejbusovými a vybranými autobusovými linkami. Pre každú linku boli zvolené tri rozdielne časovo rozdielne verzie GTFS súborov. Prehľad výsledkov testovania zobrazuje tabuľka 7.1. Pri každom prístupe je uvedená priemerná hodnota úspešnosti všetkých troch verzií GTFS.

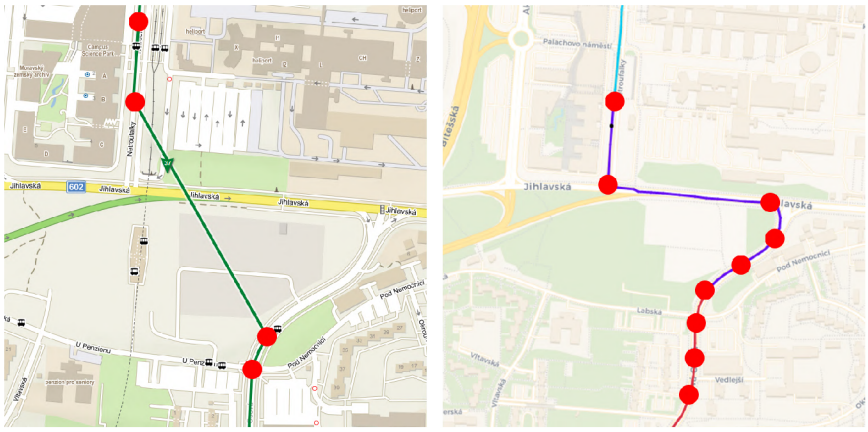
	Električky	**Vlaky**	**Trolejbusy**	**Autobusy**
Naivné prehľadávanie do šírky	45%	60%	38%	36%
Naivné prehľadávanie s parametrizaciou	69%	76%	60%	55%
Greedy Search	78%	85%	70%	68%
A*	88%	92%	83%	73%
A* s penalizáciou zalomení	94%	96%	90%	77%
A* s penalizáciou zalomení a s rovnomerným prehľadávaním	97%	98%	94%	83%

Tabuľka 7.1: **Testovanie prístupov k routingu.** V jednotlivých iteráciach vývoja routovacieho algoritmu boli testované rôzne metódy. Najúspešnejšia metóda bola následne využitá v rámci ďalšej implementácie.

Nepresnosti, s ktorými sa implementovaný algoritmus nedokáže vysporiadať, vychádzajú jednak z nedokonalej implementácie, ako aj z chybných vstupných GTFS dát. Budúci vývoj by tak mal smerovať k ďalšiemu zvyšovaniu úspešnosti routingu, najmä v oblasti schopnosti sa s takýmito anomáliami vyrovnať. Kým obrázok 7.1 zobrazuje lepšiu úspešnosť referenčného riešenia, obrázok 7.2 ukazuje príklad, kedy je naopak úspešnejšie implementované riešenie. Na oboch týchto obrázkoch sú zároveň zvýraznené body, ktoré boli použité v rámci porovnávania zhody. Obrázok 7.3 potom ukazuje príklad nepresností spôsobených chybnými GTFS dátami. V uvedenom príklade je zastávka posunutá do protismeru, čo spôsobuje následnú deformáciu celej trasy.



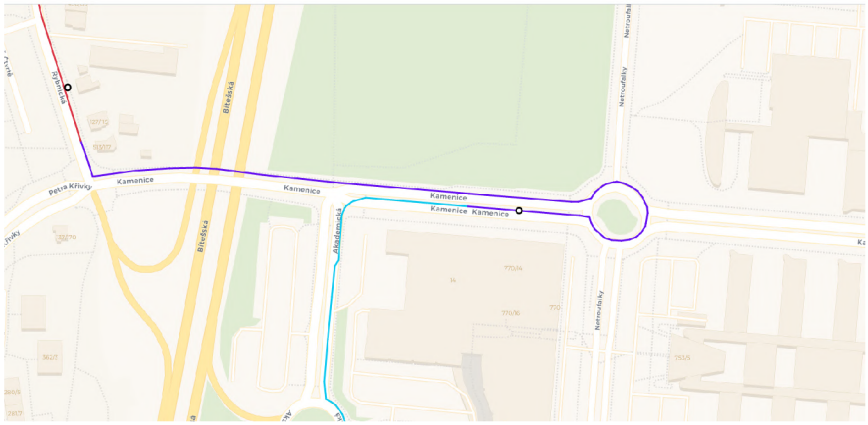
Obr. 7.1: **Presnosť referenčného riešenia.** Referenčné riešenie vľavo vizualizuje správnu trasu linky 39, kým implementované riešenie v pravej polovici vizualizuje trasu mimo trolejové vedenie. Tieto anomálie nie sú v súčasnej implementácií riešiteľné.



Obr. 7.2: **Presnosť implementovaného riešenia.** V niektorých prípadoch implementované riešenie vizualizuje presnejšiu trasu, ako riešenie referenčné. Vizualizácia referenčného riešenia vľavo tak narozdiel od implementovaného riešenia vpravo vizualizuje trasu mimo cestnej siete.

7.2 Testovanie spracovania dát

V rámci analýzy vstupných zdrojov dát boli vybrané dva potencionálne zdroje dát. Prvá fáza testovania sa tak venovala porovnaniu týchto zdrojov. Následne bol na základe výsledkov zvolený jeden zdroj dát. Samotné testovanie spracovania dát prebiehalo v dvoch intervaloch. Na základe výsledkov z prvej časti testovania potom riešenie prešlo sadou úprav a optimalizácií. Druhá časť testovania potom mala za cieľ overiť vhodnosť týchto úprav. V rámci testovania bol sledovaný čas spracovania a počet správne uložených záznamov. Za správne uložené záznamy sú považované tie, ktoré boli priradené prislúchajúcemu spoju a správnej časti trasy. Vyhodnocovanie prebiehalo porovnávaním vstupných a uložených záznamov. Samotné testovanie bolo vykonané lokálne v prostredí WSL 2.



Obr. 7.3: **Nepresnosť GTFS.** Nepresne umiestnené zastávky v externých GTFS dátach spôsobujú deformovanie trás pri routingu.

7.2.1 Testovanie zdrojov dát

Na základe povahy každého z dvoch testovaných zdrojov dát bola navrhnutá metodika spracovania, ktorá bola implementovaná a testovaná. V prípade prvého zdroja v podobe prúdu dát bolo nevyhnutné udržiavanie dočasných spojov a ukazovateľov na trasy. Výhody a nevýhody vyplývajúce z testovania je možné zhrnúť do nasledujúcich bodov:

- Oproti dávkovému spracovaniu je prístup spracovania dát pamäťovo aj výpočetne menej náročný.
- Dáta spracovávané priebežne v reálnom čase sú ihneď k dispozícií.
- Problémom spracovania záznamov v reálnom čase je nutnosť udržiavať zoznam aktuálne vykonávaných spojov, pričom v niektorých prípadoch je zložité určiť, kedy spoj skončil.
- Pri použití tohto zdroja dát je taktiež zložitejšie identifikovať niektoré anomálie bez celkového obrazu v čase.

V prípade druhého zdroja v podobe databázy bolo kľúčové vyriešiť dávkové spracovanie. Prehľad výhod a nevýhod vyplývajúcich z testovania tohto zdroja dát je opäť možné hrnúť do niekoľkých bodov:

- Dávkové spracovanie záznamov umožňuje lepšiu detekciu anomálií.
- Oproti spracovaniu v reálnom čase odpadá réžia spôsobená udržiavaním zoznamu aktuálne vykonávaných spojov. Z hľadiska implementácie sa jedná o riešenie menej náchylné na chyby.
- Toto riešenie je časovo, pamäťovo aj výpočetne náročnejšie.

Porovnaním oboch zdrojov dát, ich výhod a nevýhod, bola pre riešenie tejto práce zvolená metóda využívajúca ako zdroj dát databázu záznamov. Následne bola implementácia tejto metódy testovaná v dvoch iteráciach. Prvé testovanie malo za cieľ odhaliť problémy, ktoré neodhalila analýza ani návrh riešenia. Druhé testovanie refaktorizovanej a optimalizovanej verzie malo za cieľ overiť úspešnosť implementovaných zmien. Výsledky oboch iterácií testovania sú popísané v nasledujúcich podsekciách.

7.2.2 Priebežné testovanie – 1. fáza

Ako referenčná sada v rámci tohto testovania boli zvolené všetky električkové a vlakové linky, ktoré premávali v danom období. Tabuľka zobrazuje prehľadné záznamy o testovaní spracovania. Pre zjednodušenie sú pracovné dni v rámci daného týždňa zjednotené do jedného záznamu, ktorý obsahuje priemerné hodnoty za uvedené obdobie. V prípade odchýlok od normálu sa jedná o deň, kedy spoje v dopravnom systéme premávali v rámci špeciálneho režimu.

Dátum	**Dĺžka spracovania**	**Počet spracovávaných spojov**	**Počet úspešne uložených spojov**
11. 4. 2023 – 14. 4. 2023	00:44:32	4125	2812
15. 4. 2023	00:33:12	2728	2355
16. 4. 2023	00:36:41	2806	2658
17. 4. 2023 – 21. 4. 2023	00:47:41	4109	2765
22. 4. 2023	00:31:15	2705	2308
23. 4. 2023	00:35:40	2815	2607

Tabuľka 7.2: **Testovanie spracovania – Fáza 1.** Prvotné testovanie spracovania v rámci 3 verzie prototypu prinieslo neuspokojivé výsledky z hľadiska anotácie spojov, a odhalilo neočakávané problémy.

Výsledky testovania poukázali na potrebu zavedenia vlastnej indexácie trás, nakoľko indexácia trás vo formáte GTFS podliehala zmenám častejšie, ako bolo predpokladané. Nevyhnutné bolo taktiež upraviť spôsob, akými sú spoje prepájané medzi interným kódovaním KORDIS a formátom GTFS. Testovanie upraveného riešenia popisuje nasledujúca podsekcia.

7.2.3 Priebežné testovanie – 2. fáza

V prípade testovania optimalizovaného riešenia boli ako referenčná sada zvolené všetky električkové, vlakové a trojbusové linky. Výsledky zobrazené v tabuľke sú opäť zjednotené v rámci pracovných dní do jedného záznamu, ktorý obsahuje priemerné hodnoty za uvedené obdobie.

Celkové výsledky testovania ukázali, že optimalizované riešenie dokáže úspešne spracovať a anotovať v priemere 95% zázanamov. Problematické miesta predstavujú linky zachádzajúce mimo oblasť Jihomoravského kraje, ako aj chyby v kolekcii súborov GTFS, ktoré nie je možné identfikovať bez ďalšieho zdroja dát. Budúci vývoj nástroja by sa tak mal sústrediť aj na odstránenie týchto problémov, napríklad použitím ďalšieho zdroja dát popisujúceho systém hromadnej dopravy mesta Brna.

Dátum	Dĺžka spracovania	Počet spracovávaných spojov	Počet úspešne uložených spojov
1. 5. 2023	00:49:59	3769	3671
2. 5. 2023 – 5. 5. 2023	01:33:32	6319	5899
6. 5. 2023	01:02:52	3882	3681
7. 5. 2023	01:00:14	3838	3714
8. 5. 2023	00:55:21	3782	3678
9. 5. 2023 – 12. 5. 2023	01:38:20	6441	6218

Tabuľka 7.3: **Testovanie spracovania – Fáza 2.** Upravená implementácia bola testovaná na väčšej triede vstupných dát.

7.3 Testovanie použiteľnosti

Samostatným testovaním prešiel aj modul view. Samotná funkčnosť vizualizácií, ako aj výstupov analytickej platformy bola potom overovaná pomocou porovnávacích výpočtov nad vstupnými dátami. Toto overovanie zahŕňalo najmä priestorový a časový súlad vizualizovaných dát so vstupom. Testovanie zahŕňalo okrem vizualizácie rôznych generických dát aj používateľské testy. Tie boli vykonávané v rámci konzultácií na magistráte mesta Brna, pričom podľa následných pripomienok bolo používateľské rozhranie rôzne upravované.

7.3.1 Používateľské testy

Cieľom používateľských testov bolo overenie, či implementácia napĺňa ciele stanovené v návrhu riešenia. Išlo najmä o intuitívnosť ovládania a možnosti získavať z výsledných vizualizácií užitočné znalosti. Prehľad vybraných pripomienok, získaných počas testovania, je zhrnutý v nasledujúcich bodoch:

- Zachovanie presných dát bez vyhladzovania v prípade chýbajúcich záznamov.
- Pridanie podrobnejších možností filtrovania pri vizualizácií dát.
- Pridanie úvodných informácií a nápovedy o používaní aplikácie.

V rámci testovania výsledného riešenia boli taktiež diskutované možnosti ďalšieho použitia a vývoja nástroja. Budúci vývoj nástroja by mal z pohľadu použiteľnosti smerovať k vývoju nových analytických funkcií v rámci implementovanej analytickej platformy. Medzi návrhmi sa nachádzali napríklad porovnávanie správania rôznych liniek v rámci rovnakého času a úseku trasy, alebo sledovanie dodžiavania garantovaných prestupov.

Kapitola 8

Záver

Cieľom mojej práce bolo navrhnúť a implementovať nástroj, ktorý by dokázal sledovať a ukladať správanie systému hromadnej dopravy mesta Brna. Samotnej analýze a návrhu riešenia predchádzal podrobný prieskum existujúcich nástrojov a spôsobov riešenia tejto problematiky. Vzhľadom na povahu dát, ktoré sa mali stať vstupom budúceho riešenia, bolo taktiež nevyhnutné spracovať prehľad riešenej problematiky z hľadiska veľkých objemov dát. Z následnej analýzy, ako aj z prieskumu v súčasnosti používaných nástrojov, potom vyplynuli možnosti ako návrh riešenia posunúť na kvalitatívne vyššiu úroveň. Samotná analýza požiadaviek používateľov a dátových zdrojov priniesla problémy, ktoré si vyžadovali implementáciu pôvodne neplánovaných modulov. Záverečná fáza testovania okrem overenia funkčnosti priniesla aj nové podnety, kam by sa budúci vývoj tejto práce mohol uberať.

Výsledkom je implementovaný nástroj, ktorý autonómne sleduje správanie systému hromadnej dopravy mesta Brna. Implementované riešenie sa dokáže úspešne vysporiadať s veľkou množinou anomálií a je schopné reagovať na dynamické zmeny v systéme. Vizualizácie dát a výstupy analýz môžu slúžiť ako prostriedok pre jednoduché získavanie užitočných znalostí, sledovanie dopadov zmien vykonaných v systéme, a pri plánovaní ďalšieho rozvoja. Zároveň je navrhnutý spôsob spracovania, aj samotná implementácia jednoducho prenositeľná. Môže tak byť použitá na ľubovoľný systém hromadnej dopravy popísaný formátom GTFS. Limitujúcimi prvkami sú nedokonalosti v implementácií, ako aj nie úplne spoľahlivé vstupné dáta. Táto kombinácia potom vedie k odchýlkam pri sledovaní a modelovaní samotného systému.

Súčasná podoba nástroja je optimalizovaná pre použitie v rámci dopravného systému mesta Brna. Samotný návrh, ako aj jednotlivé moduly, sú však pripravené na rozšírenie pokrytia celého systému IDS JMK. Budúci vývoj by sa však okrem rozširovania pokrytia mal sústrediť aj na optimalizáciu routovacieho algoritmu, ako aj na pridávanie nových analytických metód, ktoré umožnia získavanie ďalších znalostí o fungovaní systému.

Správa a návrh systémov hromadnej dopravy v súčasnosti nie je bez využitia informačných technológií možná. V rámci jednotlivých systémov existujú celé kolekcie nástrojov, ktoré udržiavajú tieto systémy v pohybe. Výsledky práce, ktoré boli publikované aj v rámci konferencie Excel@FIT 2023, spočívajú najmä v jednoduchosti ovládania implementovaného nástroja a jeho prenositeľnosti. Využitie tejto práce by tak malo spočívať v pridaní ďalšej možnosti, ako jednotlivé dopravné systémy zefektívňovať a tým zvyšovať spokojnosť cestujúcej verejnosti.

Literatúra

[1] Al Mekhlal, M. a Khwaja, A. A. A Synthesis of Big Data Definition and Characteristics. In: IEEE. *2019 IEEE International Conference on Computational Science and Engineering (CSE) and IEEE International Conference on Embedded and Ubiquitous Computing (EUC)*. IEEE, Aug 2019. DOI: 10.1109/cse/euc.2019.00067. ISBN 978-1-7281-1664-8. Dostupné z: https://doi.org/10.1109%2Fcse%2Feuc.2019.00067.

[2] Bast, H., Brosi, P. a Storandt, S. TRAVIC. In: ACM. *Proceedings of the 22nd ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*. ACM, Nov 2014. DOI: 10.1145/2666310.2666369. ISBN 9781450331319. Dostupné z: https://doi.org/10.1145%2F2666310.2666369.

[3] Becker, T. Big Data Usage. In: BIG. *New Horizons for a Data-Driven Economy*. Springer International Publishing, 2016, s. 143–165. DOI: 10.1007/978-3-319-21569-3_8. ISBN 978-3-319-21568-6. Dostupné z: https://doi.org/10.1007%2F978-3-319-21569-3_8.

[4] Casado, R. a Younas, M. Emerging trends and technologies in big data processing. *Concurrency and Computation: Practice and Experience*. Wiley. oct 2014, zv. 27, č. 8, s. 2078–2091. DOI: 10.1002/cpe.3398. Dostupné z: https://doi.org/10.1002%2Fcpe.3398.

[5] Chawla, G., Bamal, S. a Khatana, R. Big Data Analytics for Data Visualization: Review of Techniques. *International Journal of Computer Applications*. Foundation of Computer Science. oct 2018, zv. 182, č. 21, s. 37–40. DOI: 10.5120/ijca2018917977. Dostupné z: https://doi.org/10.5120%2Fijca2018917977.

[6] Chen, C. houh, Härdle, W. a Unwin, A. *Handbook of Data Visualization*. Springer Berlin Heidelberg, 2008. ISBN 978-3-540-33036-3. Dostupné z: https://doi.org/10.1007%2F978-3-540-33037-0.

[7] Donoho, D. L. et al. High-dimensional data analysis: The curses and blessings of dimensionality. *AMS math challenges lecture*. Citeseer. 2000, zv. 1, č. 2000, s. 32.

[8] Dziekan, K. a Kottenhoff, K. Dynamic at-stop real-time information displays for public transport: effects on customers. *Transportation Research Part A: Policy and Practice*. Elsevier BV. jul 2007, zv. 41, č. 6, s. 489–501. DOI: 10.1016/j.tra.2006.11.006. Dostupné z: https://doi.org/10.1016%2Fj.tra.2006.11.006.

[9] Fan, J., Han, F. a Liu, H. Challenges of Big Data analysis. *National Science Review*. Oxford University Press (OUP). feb 2014, zv. 1, č. 2, s. 293–314. DOI: 10.1093/nsr/nwt032. Dostupné z: https://doi.org/10.1093%2Fnsr%2Fnwt032.

[10] G, P., S, S., A, T., N, V., S, C. et al. Real Time Automatic Vehicle Monitoring System Using IoT. In: IEEE. *2022 8th International Conference on Smart Structures and Systems (ICSSS)*. IEEE, Apr 2022. DOI: 10.1109/icsss54381.2022.9782293. ISBN 978-1-6654-9761-9. Dostupné z: https://doi.org/10.1109%2Ficsss54381.2022.9782293.

[11] Hashem, I. A. T., Yaqoob, I., Anuar, N. B., Mokhtar, S., Gani, A. et al. The rise of "big data" on cloud computing: Review and open research issues. *Information Systems*. Elsevier BV. jan 2015, zv. 47, č. 1, s. 98–115. DOI: 10.1016/j.is.2014.07.006. Dostupné z: https://doi.org/10.1016%2Fj.is.2014.07.006.

[12] Hassan, M. K., Desouky, A. I. E., Elghamrawy, S. M. a Sarhan, A. M. Big Data Challenges and Opportunities in Healthcare Informatics and Smart Hospitals. In: Springer. *Security in Smart Cities: Models, Applications, and Challenges*. Springer International Publishing, Nov 2018, s. 3–26. DOI: 10.1007/978-3-030-01560-2_1. ISBN 978-3-030-01559-6. Dostupné z: https://doi.org/10.1007%2F978-3-030-01560-2_1.

[13] Hickman, M. D. a Wilson, N. H. Passenger travel time and path choice implications of real-time transit information. *Transportation Research Part C: Emerging Technologies*. Elsevier BV. aug 1995, zv. 3, č. 4, s. 211–226. DOI: 10.1016/0968-090x(95)00007-6. Dostupné z: https://doi.org/10.1016%2F0968-090x%2895%2900007-6.

[14] Inoubli, W., Aridhi, S., Mezni, H., Maddouri, M. a Nguifo, E. M. An experimental survey on big data frameworks. *Future Generation Computer Systems*. Elsevier BV. sep 2018, zv. 86, č. 1, s. 546–564. DOI: 10.1016/j.future.2018.04.032. Dostupné z: https://doi.org/10.1016%2Fj.future.2018.04.032.

[15] Ishwarappa a Anuradha, J. A Brief Introduction on Big Data 5Vs Characteristics and Hadoop Technology. *Procedia Computer Science*. Elsevier BV. 2015, zv. 48, č. 1, s. 319–324. DOI: 10.1016/j.procs.2015.04.188. Dostupné z: https://doi.org/10.1016%2Fj.procs.2015.04.188.

[16] Kitchin, R. a McArdle, G. What makes Big Data, Big Data? Exploring the ontological characteristics of 26 datasets. *Big Data &amp Society*. SAGE Publications. feb 2016, zv. 3, č. 1, s. 205395171663113. DOI: 10.1177/2053951716631130. Dostupné z: https://doi.org/10.1177%2F2053951716631130.

[17] Lampkin, B. a Wren, A. Computers in Transport Planning and Operation. *Operational Research Quarterly (1970-1977)*. JSTOR. sep 1972, zv. 23, č. 3, s. 404. DOI: 10.2307/3007903. Dostupné z: https://doi.org/10.2307%2F3007903.

[18] Laney, D. et al. 3D data management: Controlling data volume, velocity and variety. *META group research note*. Stanford. 2001, zv. 6, č. 70, s. 1.

[19] Lehtonen, M. a Kulmala, R. Benefits of Pilot Implementation of Public Transport Signal Priorities and Real-Time Passenger Information. *Transportation Research Record: Journal of the Transportation Research Board*. SAGE Publications. jan 2002, zv. 1799, č. 1, s. 18–25. DOI: 10.3141/1799-03. Dostupné z: https://doi.org/10.3141%2F1799-03.

[20] Loshin, D. Big Data Tools and Techniques. In: Elsevier Inc. *Big Data Analytics*. Elsevier, 2013, s. 61–72. DOI: 10.1016/b978-0-12-417319-4.00007-7. ISBN 978-0-12-417319-4. Dostupné z: https://doi.org/10.1016%2Fb978-0-12-417319-4.00007-7.

[21] Mishalani, R. G., McCord, M. R. a Lee, S. The value of real-time bus arrival information under various supply and demand characteristics. In: *ITS America 10th Annual Meeting and Exposition: Revolutionary Thinking, Real ResultsIntelligent Transportation Society of America (ITS America)*. 2000.

[22] Nijkamp, P., Pepping, G. a Banister, D. Public Transport Information Systems: An English Case Study. In: *Telematics and Transport Behaviour*. Springer Berlin Heidelberg, 1996, s. 137–165. DOI: 10.1007/978-3-642-80139-6_7. ISBN 978-3-642-80141-9. Dostupné z: https://doi.org/10.1007%2F978-3-642-80139-6_7.

[23] Saadoon, M., Hamid, S. H. A., Sofian, H., Altarturi, H. H., Azizul, Z. H. et al. Fault tolerance in big data storage and processing systems: A review on challenges and solutions. *Ain Shams Engineering Journal*. Elsevier BV. mar 2022, zv. 13, č. 2, s. 101538. DOI: 10.1016/j.asej.2021.06.024. Dostupné z: https://doi.org/10.1016%2Fj.asej.2021.06.024.

[24] Sagiroglu, S. a Sinanc, D. Big data: A review. In: *2013 International Conference on Collaboration Technologies and Systems (CTS)*. IEEE, May 2013. DOI: 10.1109/cts.2013.6567202. ISBN 978-1-4673-6404-1.

[25] Sun, Z., Strang, K. a Li, R. Big Data with Ten Big Characteristics. In: *Proceedings of the 2nd International Conference on Big Data Research*. ACM, Oct 2018. DOI: 10.1145/3291801.3291822. ISBN 9781450364768.

[26] Swanson, J., Ampt, L. a Jones, P. Measuring bus passenger preferences. *Traffic engineering & control*. 1997, zv. 38, č. 6.

[27] Symes, D. Automatic vehicle monitoring: A tool for vehicle fleet operations. *IEEE Transactions on Vehicular Technology*. Institute of Electrical and Electronics Engineers (IEEE). may 1980, zv. 29, č. 2, s. 235–237. DOI: 10.1109/t-vt.1980.23846. Dostupné z: https://doi.org/10.1109%2Ft-vt.1980.23846.

[28] Warren, J. a Marz, N. *Big Data: Principles and best practices of scalable realtime data systems*. Simon and Schuster, 2015. ISBN 9781638351108.

[29] Yadranjiaghdam, B., Pool, N. a Tabrizi, N. A Survey on Real-Time Big Data Analytics: Applications and Tools. In: *2016 International Conference on Computational Science and Computational Intelligence (CSCI)*. IEEE, Dec 2016. DOI: 10.1109/csci.2016.0083. ISBN 978-1-5090-5510-4.

[30] Zikopoulos, P. a Eaton, C. *Understanding big data: Analytics for enterprise class hadoop and streaming data*. McGraw-Hill Osborne Media, 2011. ISBN 0071790535.

[31] Zito, P., Amato, G., Amoroso, S. a Berrittella, M. The effect of Advanced Traveller Information Systems on public transport demand and its uncertainty. *Transportmetrica*. Informa UK Limited. jan 2011, zv. 7, č. 1, s. 31–43. DOI: 10.1080/18128600903244727. Dostupné z: https://doi.org/10.1080%2F18128600903244727.