

# Jízdní řády

Lukáš Riedel

Specifikace

## Obsah

<b>1</b>	<b>Úvod do problému</b>	<b>2</b>
1.1	Dostupné zdroje . . . . .	2
<b>2</b>	<b>Architektura desktopové aplikace</b>	<b>3</b>
2.1	Timetables data preprocessor . . . . .	3
2.2	Timetables core library . . . . .	4
2.3	Timetables main application . . . . .	6
2.4	Timetables mobile data preprocessor . . . . .	6
<b>3</b>	<b>Architektura mobilní aplikace</b>	<b>6</b>
<b>4</b>	<b>Návrh datových struktur</b>	<b>7</b>
4.1	Datové struktury vytvořené preprocessingem dat . . . . .	7
4.2	Problém s časem . . . . .	11
	<b>Reference</b>	<b>12</b>

Poznámka: Text psaný *kurzívou* je velmi hrubý popis částí softwaru, které budou implementovány až v rámci bakalářské práce. V ročníkovém projektu (respektive zápočtových programech na C++ a C#) tedy tyto části zahrnuté nebudou.

# 1 Úvod do problému

Cílem projektu je navrhnout a implementovat sadu aplikací, přičemž uživatel bude koncový software vnímat pouze jako jednu aplikaci. Aplikace, která bude mít user-friendly grafické rozhraní a jejíž cílová platforma bude Win32, bude uživateli v offline režimu umožňovat vyhledávání spojení v rámci Pražské Integrované Dopravy a získávat informace o odjezdech z daných zastávek. Po přechodu do online režimu navíc bude moci sledovat aktuální situaci v provozu, tzn. mimořádnosti v dopravě a informace o výlukách. Online režim bude nutný pro občasnou aktualizaci jízdních řádů, kterou bude aplikaci provádět sama v pravidelných intervalech (jednou za několik dnů), popřípadě, pokud dojde k překročení tohoto intervalu, kdykoliv se cílový stroj připojí k internetu. Jízdní řády budou platné na několik dnů dopředu, jakmile se bude blížit jejich expirace, uživatel bude s touto skutečností obeznámen. *Do budoucna se počítá i s vytvořením mobilní aplikace pro cílovou platformu Android, která bude pracovat online. Bude spolupracovat s výše zmíněnou aplikací tím způsobem, že v mobilní aplikaci nebude téměř žádná logika, vše potřebné bude obstarávat desktopová aplikace, mobilní aplikace bude sloužit zejména pro zadávání a zobrazování výsledků z desktopové aplikace a bude s ní komunikovat po síti prostřednictvím vhodně navrženého protokolu. Nicméně bude možnost přejít do offline módu tím způsobem, že si uživatel bude moci stáhnout jízdní řády pro jeho oblíbenou trasu, které budou dostupné na několik dní dopředu, a bude mít tedy možnost vyhledávat spojení offline.* Součástí bude webová prezentace produktu napsaná v HTML5 a CSS3, kde budou všechny důležité informace.

## 1.1 Dostupné zdroje

Jelikož je aplikace cílená pro Pražskou Integrovanou Dopravu, jsme nuceni jako hlavní zdroj dat použít informace o jízdních řádech v GTFS[1] formátu. Tento formát, navržený společností Google, je velmi variabilní. Cílem je optimalizovat řešení pro Prahu, nicméně aplikace také poskytne uživateli možnost jednoduše změnit cílové město - jízdní řády v tomto formátu jsou totiž poskytovány i pro jiná velká (nejen) evropská města, například Londýn, Paříž nebo Řím. Algoritmů na hledání v jízdních řádech je celá řada, většina z nich používá nějakým způsobem modifikovaný Dijkstrův algoritmus. Toto řešení bude využívat RAPTOR [2] algoritmus, vyvinutý společností Microsoft, který nepotřebuje prioritní frontu a je dostatečně rychlý a efektivní. Algoritmus pracuje v kolech, kde v  $k$ -tém kole najde nejrychlejší spojení do všech zastávek, do nichž se můžeme dostat pomocí  $k - 1$  přestupů. Z toho plyne, že nehledáme spojení pouze do jedné cílové zastávky, nicméně do všech zastávek v grafu dopravní sítě. To ale nevadí, protože časová složitost zůstane stále velmi dobrá (lineární vůči trasám všech linek, každou procházíme pouze jednou) a spotřeba paměti nebude na dnešní poměry také nijak zásadní, neboť různá spojení rekonstruujeme pouze třemi pointery - odkaz na jízdu jedné konkrétní linky (tzv. trip), odkaz na nástupní stanici v daném tripu a odkaz na výstupní stanici v daném tripu. To znamená, že jedno spojení o  $k$  přestupech spotřebuje (bez

overheadů všech použitých kontejnerů a tříd) pouze 24k bytů paměti na x64 platformě, 12k bytů paměti na x86 platformě. Ze software-inženýrského hlediska bude možnost jednoduše změnit vstupní (GTFS) formát, a sice přepsáním pouze jedné části aplikace, která se stará o parsování do jakéhosi mezi-formátu. Vše ostatní nebude třeba modifikovat.

## 2 Architektura desktopové aplikace

Cílová platforma aplikace bude Win32. Hlavní jádro aplikace bude psané v C++, vše ostatní v C#. Aplikace nebude obsahovat žádné externí knihovny, řešení bude používat pouze funkce z STL (C++) a BCL (C#). Aplikace by se dala rozdělit na 4 části.

### 2.1 Timetables data preprocessor

Tato část aplikace bude napsaná v C#, ve finální aplikaci bude představovat jediné vlákno. Spustí se při prvním spuštění aplikace na daném stroji a následně v pravidelných intervalech, kdy bude k dispozici připojení k internetu a kdy aplikace nebude moc využívána (jednou za několik dní) kvůli aktualizacím dat. Bude kontrolovat validitu dat a její činnost by se dala rozdělit na tři části:

1. Stahování GTFS dat z předem navoleného zdroje dat a jejich následnou dekomprimaci. Z toho plyne, že podmínkou pro zdroj dat je ten, aby data byla v zip formátu. Předpokládá se, že URL ke zdroji dat nebude zadrátováno do kódu, ale bude možnost ho nastavit ručně. Předpoklad je takový, že URL ke zdroji dat bude v nějakém (pro to určeném) textovém souboru, tudíž nebude možnost měnit jízdní řády interaktivně přímo z aplikace. Tento princip dá pokročilejšímu uživateli svobodu zvolit si, v rámci jakého města bude aplikace pracovat, a neznalému uživateli nedá možnost aplikaci rozbít.
2. Jak již bylo řečeno, GTFS formát je velmi variabilní. Některá povinná data jsou pro náš účel zbytečná, stejně tak se najdou nepovinná data, která potřebujeme. Příkladem nepovinných dat jsou například časy přestupů mezi jednotlivými zastávkami (tzv. footpaths). Tato část aplikace bude obecný GTFS formát přizpůsobovat tak, aby se spojení mohla vyhledávat dostatečně rychle. To znamená, že pokud bude soubor obsahující footpaths chybět, dojde k jejich dopočítání na základě GPS souřadnic zastávek. Předpokládá se, že vzhledem k tomu, že pomocí GPS souřadnic můžeme určit pouze vzdušnou vzdálenost mezi dvěma souřadnicemi, tak bude zvolena dostatečně malá průměrná rychlost chůze s tím, že budeme uvažovat pouze footpaths s dobou chůze do 10-15 minut. Dalším příkladem činnosti této části může být například to, že v GTFS formátu jsou (povinně) uvedeny pouze údaje o jednotlivých nástupišťích. Uživatel typicky nechce vyhledávat odjezdy pouze z jednoho nástupišťe, ale ze všech nástupišť v rámci jedné zastávky (stanice). Aplikace tedy projde všechny nástupišťe a na základě přesné shody jmen vytvoří jednotlivé zastávky (stanice) sdružující všechny nástupišťe, které pod ně spadají. Do třetice uvedeme fakt, že aplikace vhodně předpočítá různé indexy a uvede data do stavu, abychom k požadovaným datům po načtení do paměti mohli efektivně přistupovat a aby nedocházelo k žádným memory leakům. Stručně

řečeno, tato část načte obecná GTFS data do paměti a výstupem aplikace budou data (opět soubory), která budou plně přizpůsobena pro činnost hlavní aplikace. Předpokládá se, že na průměrně výkonném stroji poběží tato aplikace (vlákno) nejvýše 20 sekund (+ čas potřebný na stažení zip archivu) jednou za tři dny.

3. Smazání zip archivu a GTFS databáze. Tyto soubory již nebudeme potřebovat, budeme mít vytvořené vlastní.

Důvod pro zvolení C# je takový, že vzhledem k povaze aplikace není třeba takový výkon, aby musela být psaná v C++. Navíc, jelikož bude hlavní část aplikace napsaná rovněž C#, dojde ke snazšímu provázání těchto dvou aplikací. V neposlední řadě můžeme zmínit, že dekomprimace souborů lze provádět přímo z .NET kódu, čímž zlepšíme přenositelnost kódu - totiž v C++ žádná funkce na dekomprimaci souborů není a práce se soubory je tam rovněž náročnější. Na code-readability a přenositelnosti zde záleží více, když není potřeba žádný extrémní výkon. Výhoda této aplikace je rovněž ta, že ušetříme čas načítání hlavní aplikace (předpoklad zrychlení je 300%), potažmo i něco málo z paměti. Popřípadě, kdybychom chtěli aplikaci uzpůsobit i jiným formátům, než je GTFS, stačilo by upravit pouze tuto aplikaci a do žádné jiné bychom vůbec nemuseli zasahovat.

Během sepisování této specifikace došlo k microbenchmarkům na datech pro Prahu u obou přístupů, a sice načítání dat pomocí knihovny z meziformátu a pak přímo z GTFS formátu. Samotný GTFS formát zabírá na disku 105 MB, zatímco meziformát pouze 56 MB. Po načtení GTFS formátu do paměti aplikace využívá 160 MB, zatímco po načtení meziformátu pouze 120 MB - je to dáno tím, že můžeme využívat efektivnější datové struktury (kvůli předpřipravené indexaci můžeme využít vector místo mapy, nějaké struktury můžeme úplně vypustit...). Ještě uvedeme, že čas inicializace dat z GTFS formátu je 12 sekund, zatímco čas inicializace z meziformátu pouze 3,3 sekundy - to je proto, že si spoustu věcí předpočítáme již do meziformátu, ušetříme počet alokací díky rezervaci místa ve vektorech atd. Zbývá podotknout, že čas přípravy meziformátu je společně se stahováním GTFS dat zhruba půl minuta.

## 2.2 Timetables core library

Tato část aplikace bude napsaná v C++ a bude plně optimalizovaná na výkon. Důvod pro zvolení C++ je tedy zřejmý. Aplikace bude zahrnovat dva základní druhy požadavků:

1. Zobrazení prvních  $n$  odjezdů z dané zastávky v daný datum a čas. K tomu asi není co víc dodávat, v databázi se najde zastávka odpovídající danému jménu, která bude obsahovat reference na všechny nástupiště, které pod ni spadají. Každé nástupiště v sobě obsahuje údaje o všech odjezdech všech linek za jeden den. Při procházení odjezdů je nutné dbát, jestli je odjezd v daný datum relevantní, neboť o víkendy, státní svátky atd. platí jiné jízdní řády než ve všední dny. I na toto se pamatuje, totiž každý trip obsahuje položku, kdy daný trip operuje (tzv. services). Najde se prvních  $n$  odjezdů splňujících kritéria a vrátí se ve formě kontejneru obsahující nějakou user-friendly datovou strukturu, jejíž obsah půjde snadno zobrazit z GUI.

2. Nalezení n nejrychlejších spojení mezi dvěma zastávkami v daný datum a čas, přičemž bude možnost navolit maximální počet přestupů a zvolit si průjezdné body. K tomu se bude využívat RAPTOR [2] algoritmus, o němž již byla řeč. Slušelo by se podotknout, že z testování společností Microsoft vyplynulo, že při vhodné implementaci a vhodných strukturách dokáže tento algoritmus nalézt libovolné spojení na grafu Londýna do 8 milisekund. Výstup bude podobný jako v podobném případě.

Předpoklad je takový, že tato část bude s hlavní částí aplikace (myšleno s tou částí s GUI) provázána pomocí interoperability C++/CLI se C#. Bude se jednat o jedno vlákno, které v sobě bude mít nekonečný cyklus, který bude zpracovávat požadavky od uživatele z GUI *a později i ze sítě, až bude hotová mobilní aplikace* a bude je umisťovat do dvou front. Fronty budou dvě, a sice pro zobrazení odjezdů ze zastávek (tzv. departure board) a vyhledání spojení (tzv. router). Každá fronta bude mít defaultně přiřazené jedno vlákno, které bude zpracovávat požadavky v ní. *V případě velkého množství požadavků se spustí (popřípadě i více) nové vlákno a požadavky se budou rovnoměrně rozdělovat. Toto se týká pouze požadavků ze sítě, předpokládá se, že z desktopové aplikace přijde pokaždé právě jeden požadavek.* Zpracovávání požadavků bude vypadat zhruba takto, bez újmy na obecnosti uveďme příklad pro departure board:

1. Přijde požadavek na zobrazení N odjezdů ze zastávky s názvem X v čas Y. Nekonečný cyklus to zaznamená, vytvoří objekt DEPARTUREBOARD s argumenty N, X, Y a odkazem na data. V konstruktoru dojde k inicializaci objektu, tzn. že se nastaví čas Y, počet odjezdů N a v datech se najde odkaz na zastávku s názvem X. *Pokud se jednalo o požadavek ze sítě, rovněž se uloží odchozí brána, IP adresa příjemce, cílový port a další údaje potřebné pro síť.* Pokud inicializace selže, například se nepodaří nalézt zastávku, požadavek se zruší, respektive se přesune do výstupní fronty, a bude si v sobě držet chybový stav, aby mohl informovat uživatele o příčině neúspěchu. V případě úspěšné inicializace se objekt zařadí do vstupní fronty.
2. Požadavek je ve vstupní frontě a přijde na něj řada. Zpracovávací vlákno na něm zavolá proceduru bez návratové hodnoty OBTAINDEPARTUREBOARD. Tato metoda je noexcept, požadavek se po provedení této metody za všech okolností zařadí do výstupní fronty. Zpracovávaný objekt se rozšířil o časy odjezdů, které jsou uloženy v paměti ve formě vhodných kontejnerů (viz. výše). Nebo nerozšířil, když nebyl nalezený žádný odjezd z dané zastávky.
3. Požadavek je ve výstupní frontě a přijde na něj řada. Data spočítaná v předchozím bodě se odešlou na patřičný výstup (tedy *síť nebo GUI*).

Deklarace třídy DEPARTUREBOARD tedy může vypadat například takto.

```

1 class DepartureBoard {
2 private:
3     std::vector<Timetables::Structures::Departure> foundDepartures;
4     const Timetables::Structures::Station& station;
5     const Timetables::Structures::DateTime& earliestDeparture;
6     const std::size_t count;
7 public:

```

```

8   DepartureBoard(const Timetables::Structures::DataFeed& feed,
9   const std::wstring& stationName, const Timetables::Structures::DateTime&
10  earliestDeparture, const size_t count) : earliestDeparture(earliestDeparture),
11  count(count), station(feed.Stations().Find(stationName)) {}
12
13  void ObtainDepartureBoard();
14
15  inline const std::vector<Timetables::Structures::Departure>& ShowDepartureBoard()
16  { return foundDepartures; }
17  };

```

Pro třídu ROUTER to bude dosti podobné.

## 2.3 Timetables main application

Aplikace bude zcela jistě naprogramována v C# a bude využívat buď technologie WinForms nebo WPF. Bude nějakým (již popsáním) způsobem obsahovat obě výše zmíněné aplikace. Aplikace bude uživateli poskytovat příjemné grafické rozhraní, kde jak pro zobrazování odjezdů, tak pro vyhledávání spojení, bude k dispozici "našeptávač" pro nalezení zastávky. Bude k dispozici hezky formátovaný RSS feed pro mimořádnosti v dopravě a informace o výlukách, který bude k dispozici pouze v online módu. Pokud bude aplikace v offline módu a bude vyžadována aktualizace dat, aplikace na tuto skutečnost upozorní. V nastavení bude možnost zvolit si barevný motiv (předpokládají se základní barvy) a do budoucna zde bude i možnost lokalizace aplikace, tzn. přepínání mezi anglickým a českým jazykem.

## 2.4 Timetables mobile data preprocessor

*Toto bude rozšíření desktopové aplikace. Aplikace bude v C# a v rámci hlavní aplikace se bude jednat o jediné vlákno. Vstupem bude požadavek na vytvoření dat pro jednu trasu. Aplikace spustí vyhledávání spojení, typicky by se měla vyhledat spojení ve všech dnech v týdnu, ve všech denních dobách a i nějaké alternativy. Na základě výsledků tato aplikace vytvoří data pro danou trasu - na základě vygrepování všeho potřebného z hlavních dat, aby se na těchto datech mohl routovací algoritmus spouštět a vždycky (v každý den, každou denní dobu) mezi dvěma zadanými zastávkami našel několik optimálních spojení. Tato data se následně po síti pošlou do mobilní aplikace. Opět se předběžně počítá nějakou frontou, aby bylo možné, v případě velkého množství požadavků, data nějak uspokojivě zpracovávat a na požadavky odpovídat.*

## 3 Architektura mobilní aplikace

Aplikace napsaná v C# pro cílovou platformu Android bude velmi podobná jako ta desktopová s tím rozdílem, že bude obsahovat pouze aplikaci popsanou v sekci 2.3. To znamená, že veškeré požadavky na zobrazení informací o zastávkách nebo spojeních se budou řešit online a to tím způsobem, že prostřednictvím vhodně navrženého protokolu, předběžně se počítá s aplikačním protokolem nad TCP, pracujícím na nějakém volném portu s komunikací (návratové kódy) podobné protokolu HTTP, bude posílat data do desktopové aplikace, kde se budou vyhledávat

*spojení a po síti se vracet zpět do aplikace mobilní. Mobilní aplikace bude v její základní verzi pouze takový zobrazovač, prohlížeč... Bude také obsahovat našeptávač, RSS feed i nastavení - stejně jako desktopová aplikace.*

*Bude zde i možnost uložit si oblíbené trasy, které se budou moci vyhledávat offline. To bude velmi pravděpodobně realizováno tak, že se po síti pošle požadavek do desktopové aplikace a aplikace ze sekce 2.4 ho zpracuje. Nalezne nejvhodnější spojení i nějaké alternativy, z dat vygrepuje nutné údaje o odjezdech z potřebných zastávek a ve formátu ze sekce 2.1 je pošle do mobilní aplikace. Výhodou tohoto přístupu je fakt, že kompletní data jsou poměrně hutná, pro Prahu mají zhruba 100 MB. Pokud se vygrepuje pouze to potřebné pro jednu trasu, budeme se pohybovat ve velikostech jednotek MB (spíše stovky kB), možná i méně. V mobilní aplikaci, pokaždé když budeme chtít vyhledat nějaké oblíbené spojení, se on-demand načtou tato data do paměti a spustí se odlehčená verze routovacího algoritmu, který bude zahrnut přímo v mobilní aplikaci a bude velmi pravděpodobně přepsán do C#. Jelikož budeme procházet velmi odlehčený graf, bude C# výkonnostně dostačující a vyhneme se problémům s interoperabilitou. I tato data budou mít nějaký datum expirace, po expiraci dat již nebude oblíbená trasa k dispozici a bude nutné data zaktualizovat. Na to bude připravené tlačítko, které zařídí vše potřebné.*

## 4 Návrh datových struktur

### 4.1 Datové struktury vytvořené preprocessingem dat

Data se z GTFS formátu načtou do paměti v přirozené formě. Musíme tak činit sekvenčně, neboť na sobě jednotlivá data různě závisí. GTFS formát je velmi obecný, pro nás mají význam pouze tyto položky:

1. Calendar - Obsahuje údaje o jednotlivých services. Díky tomu víme, kdy daný trip operuje. Každá service je indexovaná svým ID. Dále obsahuje binární hodnoty od pondělí do neděle, kde 1 signalizuje, že daná service v daný den operuje a 0 analogicky pro opačný případ. Můžeme reprezentovat buď polem boolovských proměnných, nebo pomocí bitů například v charové proměnné. Jelikož je počet všech services poměrně malý (většinou max 100-200), obětujeme 1 kB paměti pro reprezentaci v poli, abychom dosáhli lepší code-readability. Rovněž zde nalezneme dvě data, která nám udávají, v jakém rozmezí jsou jízdní řády platné.
2. Calendar Dates - Jeden záznam v tomto souboru nám říká něco o případných mimořádnostech v provozu dané service. Vždycky musí obsahovat ID dané service, datum mimořádnosti a typ mimořádnosti. Typ bude v našem případě pouze přidána/odebrána v daný datum, proto to můžeme reprezentovat boolovskou proměnnou. Při čtení tohoto souboru nevytváříme žádnou novou strukturu, pouze rozšiřujeme strukturu vytvořenou v předchozím bodě o mimořádnosti.
3. Routes - Obsahuje nejstručnější informace o provozu jednotlivých linek, jako je krátký název dané linky (například A, 22, 135...), dlouhý název linky (například Depo Hostivař - Nemocnice Motol), barvu linky v HEX formátu a typ linky, přičemž my si vystačíme

s osmi typy (Tram = 0, Subway = 1, Rail = 2, Bus = 3, Ship = 4, CableCar = 5, Gondola = 6, Funicular = 7). Každá linka je rovněž indexována nějakým stringovým identifikátorem, proto je pro linky nejvhodnější strukturou unordered map/Dictionary, zatímco pro services to může být vector/List.

4. Stops - Obsahuje údaje o jednotlivých nástupišťích, indexace pomocí stringů. Pro nás hrají roli pouze název zastávky a GPS souřadnice, konkrétně zeměpisná šířka délka. Opět by bylo vhodné uchovávat si informace v unordered mapě/Dictionary. Při čtení dat je nutné dbát položku o typu lokace, kde uvažujeme pouze lokace typu 0, což je právě nástupiště. Zbývající dvě možnosti jsou stanice (ty si vytvoříme dle svých kritérií sami) a vchody/východy do/ze stanic.
5. Trips - Trip je trasa linky s časy odjezdu. Typicky máme daleko více tripů než linek. Například odjezd metra A ze stanice Depo Hostivař v 7:00:00 je jiný trip než odjezd téže linky ze stejné stanice v čas 7:10:00. Tripy se indexují přirozenými čísly, tudíž nám jakožto struktura stačí vector/List. Dále zde máme odkazy na jednotlivé linky v podobě jejich identifikátorů, to stejné platí pro services. U tripů nám stačí si uchovávat jen headsign, což je název cílové stanice ("to, co je na příjezdějící tramvaji/metru/autobusu"). Trasu určíme z následujícího bodu.
6. Stop Times - Toto je nejrozsáhlejší položka všech dat a obsahuje údaje o odjezdech. Každá položka obsahuje identifikátor tripu, identifikátor zastávky, o kolikátou zastávku v řadě se jedná a časy příjezdu a odjezdu z dané zastávky. Ke každé zastávce přidáme ukazatel na každý stop time, který je asociován k dané zastávce (snadné vyhledávání departure board) a to stejné učiníme pro každý trip, jemuž daný stop time přísluší. Tak jsme schopni zrekonstruovat trasu tripu.

Struktura pro GPS souřadnice bude navíc obsahovat statickou metodu pro počítání doby chůze mezi dvěma body. Defaultní průměrná rychlost je nastavena na 0,5 m/s a metoda využívá Haversine formulí.

Pro správné fungování RAPTOR algoritmu ještě potřebujeme vytvořit trasy pro tripy, které sdílí stejné zastávky. Údaje o linkách přímo použít nemůžeme, protože například ačkoliv linka 22 v trase Vypich - Nádraží Strašnice sdílí stejné zastávky, tak tatáž linka v trase Bílá Hora - Nádraží Hostivař je v podstatě linka úplně jiná. Toto uděláme v rámci preprocessingu. Položku routes si označíme jako routes info a vytvoříme skutečnou položku routes, kde jedna route bude obsahovat posloupnost zastávek in-order, dále odkaz na route info a odkazy na všechny tripy, seřazené podle času odjezdu. To si přetřídíme již v rámci preprocessingu dat. Pro algoritmus také potřebujeme vědět, jaké routes prochází jednotlivými stops. Budeme procházet všechny route a u každé zastávky do ni přidáme odkaz na danou route. Struktura pro tento přístup může být klidně vector/List. Toto si ale nemusíme předpočítávat, protože by to bylo stejně efektivní, jako kdybychom to počítali přímo při inicializaci dat z C++.

Proto to budeme dělat pokaždé při spuštění aplikace.

Rovněž musíme dopočítat footpaths, tedy dobu přestupu mezi jednotlivými zastávkami. Původní plán byl takový, že GTFS data, která tuto volitelnou položku obsahují, ji budou také využívat. Nakonec vyšlo najevo, že bude lepší vše předpočítat ručně, aby to bylo konzistentní a měli jsme časy mezi všemi zastávkami, které potřebujeme. Jediná nevýhoda tohoto



přístupu je, že čas běhu se společně s rostoucími zastávkami kvadraticky zvyšuje. Zdroj dat pro Pražskou Integrovanou Dopravu to neobsahuje. Toto je nejpomalejší část parsovacího algoritmu, protože časová složitost vytváření těchto footpaths bude kvadratická, neboť musíme počítat vzdálenosti mezi každými dvěma zastávkami. Haversine formule potřebuje ke své činnosti goniometrické funkce, proto je tato část algoritmu skutečně velmi časově náročná. Údaje o footpaths budeme uchovávat u každého nástupiště (stop) ve formě mapy/Sorted-Dictionary a to tak, že klíčem bude intová hodnota, za jak dlouho jsme schopni při chůzi průměrnou rychlostí dojít do zastávky, která se schovává pod tímto klíčem. Pro ušetření paměti budeme do paměti ukládat pouze hodnoty nepřevyšující 600 sekund.

Po naparsování GTFS formátu do paměti aplikace dojde ke vhodnému výpisu dat do nových souborů, aby se daly použít k výpočtům. To bude probíhat v podstatě tak, že původní soubory okleštíme na nutné minimum (viz. předchozí text), popřípadě vytvoříme soubor s footpaths a rozhodně vytvoříme soubor se stanicemi (stations). Při výstupu nástupišť (stop) tedy bude navíc každý záznam obsahovat intový identifikátor, kterým budou indexovány stanice v nově vzniklém souboru. Dále soubor s routes přeměníme na routes info, soubor routes vytvoříme podle výše zmíněných pravidel a do každého tripu navíc přidáme identifikátor na danou route. Každá route navíc bude obsahovat ID na příslušné route info. Dále na první řádku každého nového souboru, pro nějž budeme později vytvářet v paměti kontejner, připišeme počet záznamů, abychom si na danou velikost mohli kontejner před načítáním dat přizpůsobit a ušetřit tak čas i počet alokací. Nově vytvoříme soubor s jediným záznamem, v němž bude na první řádce napsáno, kdy jízdní řády expirují (konec platnosti první service v datech). Variabilitu GTFS při parsování ošetříme tím způsobem, že využijeme první řádky souboru, kde je pokaždé napsáno, jaká data daný sloupec obsahuje. Již na prvním řádku tedy ověříme, jestli daný soubor obsahuje vše, co potřebujeme. Následně, protože nikde není psáno, v jakém pořadí mají sloupce vystupovat, vytvoříme Dictionary s klíčem string a hodnotou int, kde string je název daného sloupce a int je index, tedy o kolikátý sloupec se jedná. K datům budeme přistupovat přes tento slovník. Vytvoříme interface IDataFeed, který bude obsahovat seznam všech souborů, které musí výsledná data obsahovat (viz. níže) a metodu CreateDataFeed, která tyto soubory vytvoří. Pro každý soubor vytvoříme abstraktní třídu, která bude definovat datové položky v daném souboru a způsob, jakým se mají vypisovat. Parsování GTFS souborů vyřešíme tak, že pro každý nový soubor vytvoříme třídu GtfsXXX, kde XXX je název abstraktní třídy, od níž se bude dědit. Takže abstraktní třídy budou obsahovat vše potřebné pro výpis nového formátu, zatímco zděděné třídy budou obsahovat vše potřebné pro parsování GTFS formátu. Děláme to takto proto, abychom umožnili snadné rozšíření o parsování nějakého nového formátu do formátu potřebného pro práci knihovny pro jízdní řády. Celou práci aplikace bude obsluhovat jedna generická funkce, kde se jako typ funkce zadá formát, z kterého se mají data parsovat. Respektive, tento formát bude reprezentován třídou, v našem případě GtfsDataFeed, která bude implementovat rozhraní IDataFeed. Konkrétně by vše popsané mohlo vypadat třeba takto:

```

1 public interface IDataFeed {
2     Calendar Calendar { get; }
3     CalendarDates CalendarDates { get; }
4     RoutesInfo RoutesInfo { get; }
5     Stops Stops { get; }
6     Stations Stations { get; }
7     Footpaths Footpaths { get; }
8     Trips Trips { get; }
9     StopTimes StopTimes { get; }
10    Routes Routes { get; }
11    string ExpirationDate { get; }
12    void CreateDataFeed(string path);
13 }
14
15 public static class DataFeed {
16     public static void GetAndTransformDataFeed<T>() where T : IDataFeed {
17         Downloader.GetDataFeed("temp_data/");
18         IDataFeed data = (T)Activator.CreateInstance(typeof(T), (string)"temp_data/");
19         data.CreateDataFeed("data/");
20         Downloader.DeleteTrash("temp_data/");
21     }
22 }

```

Následuje seznam souborů, které bude nový formát obsahovat s názvy sloupců (ty ovšem nikde v souborech napsané nejsou, jedná se o "implementační tajemství"). Pro připomenutí platí, že na první řádce každého souboru je vždy počet záznamů, které následují. Záznamy mají vždy fixní velikost (fixní počet sloupců). Je to tak pro ušetření počtu alokací při načítání do vektorů a také proto, abychom mohli načíst všechna data pomocí for cyklu. Záznamy od sebe nejsou odděleny newliny, vše je napsané v jedné řádce. Jednotlivé sloupce jsou od sebe odděleny středníkem.

Calendar - ServiceID, Monday ... Sunday, ValidSince, ValidUntil  
CalendarDates - ServiceID, Date, TypeOfExtraordinaryEvent  
RoutesInfo - RouteInfoID, ShortName, LongName, MeanOfTransport, Color  
Stops - StopID, Name, Latitude, Longitude, ParentStationID  
Stations - StationID, Name  
Footpaths - Duration, FirstStop, SecondStop  
Trips - TripID, RouteInfoID, ServiceID, RouteID, Headsign  
StopTimes - TripID, StopID, ArrivalTime, DepartureTime  
Routes - RouteID, RouteInfoID, NumberOfStops  
ExpirationDate - ValidUntil

*Co se týče nutného offline fungování mobilní aplikace, budou k dispozici pouze názvy zastávek a linky, které jimi projíždějí. V mobilní aplikaci tedy bude, co se dat týče, soubor routes info. Budeme pro ni muset vytvořit speciální soubor stations, kde jen přidáme identifikátory všech routes info, které projíždí nějakým nástupišťem dané stanice. Toto vše vytvoří aplikace na preprocessing dat v rámci aktualizace jízdních řádů. Těmto souborům vytvoří nějaký timestamp, do kdy jsou data validní. Mobilní aplikace požádá o tato data při prvním spuštění a poté pokaždé, když dojde k expiraci dat na základě timestampu. Pokud nebude k*

*dispozici připojení k internetu, mobilní aplikace upozorní na možnou expiraci základních dat.*

Později, na základě profilace výkonnosti algoritmu, se ukáže, jestli bude nutné si do paměti předem připravit i různé indexy, například kolikátá je zastávka X v pořadí route, nebo jestli bude dostatečně efektivní využít lineární přístup hledání optima/minima/maxima, popř. binární vyhledávání (pro nějaké jiné položky, například hledání ideálního času odjezdu). Toto teď vůbec nejsem schopen odhadnout, bude to hrát roli až při existenci mobilní aplikace, kdy může na server chodit několik stovek požadavků za sekundu. Pak by samozřejmě bylo ideální mít algoritmus co nejrychlejší.

## 4.2 Problém s časem

Je zde problém s časem v GTFS formátu, a sice že můžou obsahovat i data převyšující čas 23:59:59. V GTFS formátu je čas 25:15:00 a 1:15:00 ekvivalentní, a oba to jsou validní časy. Je to tak proto, že v GTFS se o datech (téměř) vůbec nemluví, pouze o času a dnech v týdnu. Podle dne v týdnu určíme, kdy daný trip operuje. Nicméně může existovat trip, který vyjíždí v operační den ve 23:50:00 a do cílové stanice dojíždí další den v 0:10:00. Kdyby onen trip v ten další den neoperoval, byl by zde problém, protože by nikdy nedojel do cílové stanice. Proto je toto řešeno tak, že trip odjíždí v operační den ve 23:50:00 a dojíždí v "ten samý" operační den ve 24:10:00, což je v reálné situaci samozřejmě nedefinovaný čas. Při operacích s časem v C# používáme klasickou a přirozenou strukturu `Datetime` (*až později při psaní mobilní aplikace budou dořešeny detaily*), zatímco při práci v C++, kde dochází k vyhledávání odjezdů a spojení, si pro snadnou práci zavedeme vlastní strukturu, která bude představovat rozšíření struktury `tm` o naše vlastní požadované funkce. Tato struktura bude rozlišovat datum a čas, kde datum bude veden v typu `time_t` a bude se s ním pracovat za užití funkcí `mktime` a `localtime`. Bude se jednat vždy čistě jen o datum. Čas bude uchováván v `long long int`u a zde se bude rozlišovat čas z GTFS. Datový typ `long long int` byl zvolen proto, aby mohly být efektivně implementovány operace s časem, funkce typu `AddDays` a `AddMinutes` se zápornými hodnotami a ani nevadí, že v případě času, kde den má 86400 sekund, nevyužijeme typ `size_t`, nebo nějakou `unsigned` variantu `int`u. Při hledání spojení tedy budeme k časům ze `stop-timů` přidávat i data, abychom mohli spolehlivě určit, jestli takové spojení v daný den vůbec existuje (opět, hromadná doprava většinou jezdí o víkend jinak než ve všední dny). To uděláme tak, že si časy budeme uchovávat v přirozené formě a na základě časů u `stop-timů` zpětně určíme, jestli trip v daný den operuje. Vůbec by se to nemuselo dělat takhle složitě, v MHD většinou linky nejezdí déle než den. Nicméně chceme, aby knihovna byla v budoucnu použitelná i pro nějakou větší oblast, například celou republiku nebo Evropu. Toto je ale čistě programátorský problém, uživateli se vše bude zobrazovat v "normálních" časech.

## Reference

- [1] GTFS Format Reference  
<https://developers.google.com/transit/gtfs/reference/>
- [2] RAPTOR: Round-Based Public Transit Routing Algorithm  
[https://www.microsoft.com/en-us/research/wp-content/uploads/2012/01/raptor\\_alenex.pdf](https://www.microsoft.com/en-us/research/wp-content/uploads/2012/01/raptor_alenex.pdf)