Západočeská univerzita v Plzni
Fakulta aplikovaných věd
Katedra informatiky a výpočetní techniky

Diplomová práce

ETL nástroj pro konverzi OpenStreetMap dat do datové struktury nástroje TrafficModeller

Plzeň 2021 Vaněk Jakub

Místo této strany bude zadání práce.

Prohlášení

Prohlašuji, že jsem diplomovou práci vypracoval samostatně a výhradně s použitím citovaných pramenů. V knize jsou použity názvy programových produktů firem apod, které mohou být ochrannými známkami nebo registrovanými ochrannými známkami příslušných vlastníků.

V Plzni dne 29. listopadu 2021

Vaněk Jakub

Poděkování

Placeholder - k nalezení v souboru thesiskiv.cls

Abstract

Configuration of Software Components Generator from Feature Models. Goal of this thesis is to generate feature model from grammar of tesa language written in Xtext and to generate final source code from chosen features of this model. This thesis describes and uses knowledge about feature modeling and generative programming. In the opening part of the thesis reader is introduced with feature modeling, tools used for feature modeling, generative programming and Xtext framework. In the later part implementation design and final solution is described.

Abstrakt

Cílem této práce bude implementovat nástroj vykonávající obousměrný konverzní algoritmus mezi daty s datovou strukturou OpenStreetMap a nástrojem TrafficModeller. V úvodní části bude představen proces ETL, popsány datové struktury OpenStreetMap a nástroje TraMod. Dále bude popsána implementace a výsledné řešení.

Obsah

1 Úvod

Diplomová práce se bude zabývat sběrem a přípravou dat pro nástroj TrafficModeller, který byl vyvinut na Západočeské Univerzitě v Plzni. Tento nástroj umožňuje pomocí svého API jednoduše a rychle testovat různé scénáře dopravy na specifikovaném území.

K naplnění datové struktury tohoto nástroje budou použita data z OpenStreetMap (dále jen OSM), které jsou volně dostupné široké veřejnosti. K naplnění datové struktury nástroje TrafficModeller bude potřeba získat data z OSM, poté je transformovat do příslušné datové struktury nástroje TrafficModeller a následně je do něj nahrát.

V úvodní části této práce bude popsán proces ETL. Dále budou popsány přístupy k harmonizaci dat se zaměřením na geodata. Následně bude popsána datová struktura dat z OSM a budou detekovány části relevantní pro nástroj TrafficModeller.

V další části bude navržen a implementován nástroj vykonávající konverzní algoritmus, který data převede z datové struktury OSM do datové struktury TraMod. Součástí nástroje bude také zpětná konverze dat do datové struktury OSM obohacených o výstupy nástroje TraMod. Nástroj bude otestován a bude detailně okomentován.

V závěru práce budou popsány a kriticky zhodnoceny dosažené výsledky.

2 ETL

V současném obchodním světě se různé společnosti potýkají s roustoucím množstvím sbíraných a uchovávaných dat a s potřebou těmto datům co nejlépe porozumět. Tato data mohou být ve společnostech používána k omptimalizaci firemních procesů, sledování účinnosti firemních strategií, objevování nových nevyužitých příležitostí a mnohému dalšímu. Využití dat k rozhodovacím procesům ve společnosti nazýváme *Business intelligence*. Business intelligence chápeme jako soubor technologií a procesů, které uživatelům umožňují přístup k datům a analýzu dat za účelem podpory rozhodování [Howson].

Jednou ze základních komponent BI je tzv. *datový sklad*. William Inmon datový sklad definuje takto: "*A data warehouse is a subject-oriented, integrated, nonvolatile and time-variant collection of data in support of management's decision*". Jedná se o zvláštní typ databáze, která je používána pro datové analýzy. Inmon ve své definici používá 4 důležité charakteristiky takové databáze:

- *subject-oriented* - orientovaný na subjekt - Datový sklad je orientovaný na subjekt, protože poskytuje informace o konkrétním subjektu namísto probíhajících operací organizace. Těmito subjekty mohou být zákanící, dodavatelé, prodej, výnosy apod. Datový sklad se nezaměřuje na probíhající operace, ale zaměřuje se na modelování a analýzu dat za účelem rozhodování.

- *integrated* - integrovaný - Ze všech charakteristik je právě tato tou nejdůležitější. Data jsou do datového skladu integrována z více různých zdrojů, které mohou mít rozdílnou strukturu, názvosloví, jednotky apod.. Taková data tedy musí být očištěna a transformována tak, aby se do datového skladu nahrála v jedné konzistentní podobě a byla tak umožněna jejich analýza.

- *non-volatile* - stálost - Tato charakteristika značí skutečnost, že data v datovém skladu jsou neměnná. Do datového skladu data pouze nahráváme a následně k nim přistupujeme, nikdy je však neaktualizujeme. Data jsou do datového skladu nahrávána v podobě statického záznamu, který reflektuje stav v daném čase. Pokud se tedy takový stav změní, není v databázi aktualizován, ale je nahrán nový záznam opět reflektující stav v novém čase a starý záznam je pro analytické účely v datovém skladu zachován.

- *time-variant* - časová variabilita - Data jsou do datového skladu nahrána tak, že reflektují stav v přesně daném čase. Tyto stavy jsou tedy v datovém skladu zachovány a tím je umožněno získat přesný stav systému v jakémkoliv okamžiku.

K analýze dat je tedy nejdříve třeba vytvořit datový sklad. Datový sklad je tvořen procesem zvaným *ETL*. Zkratka ETL reprezentuje populární třífázový proces, při kterém jsou data z jednoho či více heterogenních zdrojů nahrána do datového skladu. těmito třemi fázemi jsou fáze extrakce (z angl. *extraction*), transformace (z angl. *transform*) a nahrání dat (z angl. *load*). Běžným označením pro prostředky ETL je rovněž datová pumpa. V následujících částech budou popsány jednotlivé fáze procesu ETL.

2.1 Extrakce dat

Prvním procesem, který je používán při výstavbě datového skladbu je proces zvaný *extrakce*. Poté co určíme cíl datového skladu a stanovíme jeho strukturu, je třeba identifikovat zdrojové systémy, ze kterých budou data do datového skladu extrahována. Tyto zdrojové systémy se od sebe navzájem liší. Mohou se lišit například ve své struktuře, či formátu uložení. Běžnými formáty, ve kterých jsou data uložena mohou být například relační databáze, formát XML či JSON, ale může se jednat o jakékoliv jiné systémy pro uložení dat.

Klíčovým požadavkem v této fázi je, aby byla všechna data ze zdrojových systémů nahrána v požadovaném čase. S tím se pojí hned několik problémů. Zdrojové systémy mohou být například dočasně nedostupné. Může se také stát, že zdrojový systém není uzpůsoben k tako rozsáhlé extrakci dat a požadovaná zátěž pro něj může být z různých důvodů nepřípustná a je třeba hledat náhradní řešení (zálohy systému, čtení pouze části dat apod.). Dalším problémem, na který je možné při extrakci dat narazit může být například příliš velký objem dat, kdy u těchto procesů není výjimkou objem dat v řádu několika GB denně.

Extrakce probíhá ve dvou fázích [https://www.cs.colostate.edu/etl/papers/Thesis.pdf] . V první fázi probíhá tzv. úplná extrakce. Při této fázi jsou data extrahována poprvé a je tedy nutné je extrahovat kompletně celá. Druhá fáze je tzv. inkrementální. Tato fáze nastává ve chvíli, kdy se ve zdrojových systémech objeví nová nebo modifikovaná data. Nová nebo modifikovaná data je třeba identifikovat a odlišit od takových, které již procesem extrakce prošla dříve [Kimball]. Identifikace nových dat může být problematická. Můžeme k ní využít tři přístupy.

- Logy v databázi - V této technice mohou být použity logy DBMS. Tyto logy jsou použity pro nalezení přidání nebo změny dat ve zdrojové databázi.
- Triggery - Na každé tabulce ve zdrojové databázi jsou vytvořeny triggery, které jsou automaticky spuštěny při přidání či změně dat ve zdrojové databázi pomocí DML (Data Manipulation Language).
- Časová razítka - Některé databáze používají sloupce pro časová razítka, která specifikují čas ve kterém byl daný řádek naposledy modifikován. Pomocí těchto sloupců lze jednoduše identifikovat změnu ve zdrojovém systému.

Pokud však zdrojovým systémem není relační databáze, není možné takové přístupy použít. Je tedy třeba manuálně nalézt způsob, jak identifkovat přidaná či změněná data a ty následně extrahovat.

Cílem této fáze je tedy identifikovat relevantní informace ve zdrojových systémech a nahrát je do jediné struktury či formátu, která bude vhodná pro fázi transformace.

2.2 Transformace dat

2.3 Nahrání dat

2.4 Konkrétní případ ETL

3 Závěr

Literatura