

Vytvorenie lokálneho kanála pre spracovanie voľne dostupných dátových súd

Ročníkový projekt - Daniel Tadeáš Dobrík - šk. rok 2025/2026

Report zimný semester

Plán práce na zimný semester

- Voľba vhodných voľne dostupných dátových setov, identifikácia chýb, v prípade potreby prispôsobenie dát
- Špecifikácia metód na opravu dát
- Implementácia metód na opravu dát (ETL proces), ukladanie spracovaných dát do DB

Priebeh práce počas zimného semestra

Prvou úlohou bolo nájsť a zvoliť vhodný dataset pre dané zadanie. Boli zvolené dva datasety namerané organizáciou Chicago Park District pomocou vodných a meteorologických senzorov v oblasti Michiganského jazera.

Následne prebehla analýza dát, ktorá spočívala v detekcii typov chýb v nameraných dátach a v určení frekvencie ich výskytu vo vybraných datasetoch. Keďže súčasťou zadania bolo vytvorenie lokálnej pipeline použiteľnej aj pre iné datasety podobného charakteru, pipeline bola navrhnutá tak, aby dokázala detekovať a opravovať aj chyby, ktoré sa priamo vo vybraných datasetoch nevyskytovali, ale môžu sa objaviť v podobných dátach.

Výsledný zoznam chýb, ktoré má pipeline detekovať a opravovať, je nasledovný:

- dáta prišli v iných časoch, než je očakávané, pričom v očakávanom čase chýbali
- dáta prišli v inom čase, než je očakávané, ale zároveň prišli aj v správnom čase
- duplicitné záznamy s rovnakými hodnotami
- duplicitné záznamy s rôznymi hodnotami
- nekorektné hodnoty mimo bežného rozsahu
- chýbajúce hodnoty pre niektoré atribúty
- chýbajúce celé záznamy pre niektoré timestampy

Ďalším krokom bolo slovné zadefinovanie jednotlivých typov chýb a spôsobov ich opravy. Ako vstupný formát bol zvolený CSV súbor, a to aj s ohľadom na možné budúce zameranie projektu na spracovanie dát v reálnom čase.

Na základe zadania projektu bol pre návrh pipeline zvolený ETL model. V konkrétnej implementácii je celý proces riadený inštanciou triedy Transformer, ktorá vytvára inštancie tried Extractor a Loader. Trieda Extractor zabezpečuje čítanie vstupného súboru, zatiaľ čo trieda Loader rieši ukladanie výsledných hodnôt do databázy.

Implementácia pipeline prebehla v programovacom jazyku Python. Z implementačných riešení stojí za zmienku spôsob zhromažďovania dát. Keďže sa vo vstupných datasetoch nachádza viacero senzorov, bolo potrebné vytvoriť samostatný akumulátor pre každý senzor, do ktorého sa zhromažďujú príslušné merania, ktoré Transformator dostáva po riadkoch od inštancie triedy Loader.

Z pohľadu generickosti pipeline umožňuje konfiguračný súbor nastaviť veľkosť chunku, maximálny počet vynechaných hodnôt určených na aproximáciu, počet a názvy senzorov CSV súboru, počet a názvy stĺpcov CSV súboru, ako aj bežný rozsah hodnôt pre jednotlivé atribúty.

Výsledok práce

Výsledkom práce je lokálna pipeline, ktorá vo väčšine všeobecných prípadov správne detekuje a opravuje vopred definované typy chýb podľa stanovených postupov. Pipeline sa dokáže prispôbiť viacerým vlastnostiam vstupného datasetu a výsledok transformácie ukladá do databázy.