social-network-data-migration Documentation

Release latest

May 20, 2022

CONTENTS

System for migrating social network data from heterogeneous sources to a graph database

CHAPTER

ONE

CONTENTS:

1.1 Specyfikacja funkcjonalna

System do migracji danych dotyczących sieci społecznych z heterogenicznych źródeł do grafowej bazy danych

Autorzy: Gabriel Kępka, Piotr Makarewicz

1.1.1 Migracja danych

Aplikacja konsolowa do migracji z bazy PostgreSQL

Kryteria akceptacji:

- część projektu realizująca migrację z bazy PostgreSQL będzie uruchamiana z linii poleceń
- jako parametry uruchomienia aplikacja przyjmie ścieżkę do pliku konfiguracyjnego z parametrami połączeń do baz danych oraz ścieżkę do pliku konfiguracyjnego ze zdefiniowanym sposobem mapowania

Przykład:

```
java Migrator db.properties mapping.json
```

Aplikacja konsolowa do migracji z pojedynczego pliku CSV/XML

Kryteria akceptacji:

- część projektu realizująca migrację z pojedynczego pliku CSV/XML będzie uruchamiana z linii poleceń
- jako parametry uruchomienia aplikacja przyjmie ścieżkę do pliku konfiguracyjnego z parametrami połączeń do baz danych oraz ścieżkę do pliku z danymi

Przykład:

```
java Migrator db.properties data.csv
```

Plik konfiguracyjny z parametrami połączeń do baz PostgreSQL i Neo4j

Kryteria akceptacji:

- ustalono jednolity format pliku konfiguracyjnego dla połączeń z bazami PostgreSQL i Neo4j
- użytkownik ma możliwość podłączenia się do obu baz, podając jedynie nazwę pliku konfiguracyjnego

Plik konfiguracyjny będzie w formacie Java Properties z kluczami:

- `postgresHost` - adres serwera PostgreSQL
- `postgresDB` - nazwa użytkownika PostgreSQL
- `postgresUser` - hasło do bazy PostgreSQL
- `postgresPassword` - nazwa bazy PostgreSQL
- `neo4jHost` - adres bazy Neo4j
- `neo4jUser` - nazwa użytkownika Neo4j
- `neo4jPassword` - hasło do bazy Neo4j

Przykład:

```
postgresHost=localhost
postgresDB=socialdata
postgresUser=sna_user
postgresPassword=password
neo4jHost=localhost
neo4jUser=neo4j
neo4jPassword=password
```

Plik konfiguracyjny z mapowaniem między schematem bazy relacyjnej a schematem docelowym

Kryteria akceptacji:

- ustalono jednolity format pliku konfiguracyjnego dla mapowania, gdy zbiorem źródłowym jest baza relacyjna SQL

Mapowania będą definiowane w pliku o formacie JSON:

```
{
  "nodes": [
      <node_mapping>,...
  ],
  "edges": [
      <foreign_key_edge_mapping>|<join_table_edge_mapping>,...
  ]
}
```

<node_mapping> jest obiektem JSON reprezentującym mapowanie rekordów tabeli SQL na węzły w bazie Neo4j. Pola obiektu JSON:

`sqlTableName` - nazwa tabeli w bazie SQL

`nodeLabel` - etykieta węzła w bazie Neo4j

`mappedColumns` - obiekt JSON, w którym klucze to nazwy kolumn tabeli `sqlTableName`, a wartości to nazwy odpowiadających im atrybutów węzła

<foreign_key_edge_mapping> jest obiektem JSON reprezentującym mapowanie powiązania kluczem obcym SQL na krawędź w bazie Neo4j. Pola obiektu JSON:

`edgeLabel` - etykieta krawędzi w bazie Neo4j

`foreignKey` - łańcuch znaków w formacie `table.column` oznaczający tabelę i kolumnę klucza obcego w bazie SQL

`from` - nazwa tabeli odpowiadającej węzłowi, z którego ma być poprowadzona krawędź

`to` - nazwa tabeli odpowiadającej węzłowi, do którego ma być poprowadzona krawędź

<join_table_edge_mapping> jest obiektem JSON reprezentującym mapowanie powiązania tabelą łącznikową SQL na krawędź w bazie Neo4j. Pola obiektu JSON:

`edgeLabel` - etykieta krawędzi w bazie Neo4j

`joinTable` - nazwa tabeli łącznikowej

`from` - nazwa tabeli odpowiadającej węzłowi, z którego ma być poprowadzona krawędź

`to` - nazwa tabeli odpowiadającej węzłowi, do którego ma być poprowadzona krawędź

`mappedColumns` - obiekt JSON, w którym klucze to nazwy kolumn tabeli `joinTable`, a wartości to nazwy odpowiadających im atrybutów krawędzi

Przykład (dla bazy Salon24):

```
{
 "nodes": [
   {
     "sqlTableName": "authors",
     "nodeLabel": "Person",
     "mappedColumns": {
       "id": "id",
       "bloglink": "blog_url",
       "name": "name"
```

(continues on next page)

(continued from previous page)

```
      }
    },
    {
      "sqlTableName": "posts",
      "nodeLabel": "Post",
      "mappedColumns": {
        "id": "id",
        "categoryno": "categoryno",
        "content": "content",
        "date": "timestamp",
        "link": "url",
        "title": "title"
      }
    },
    {
      "sqlTableName": "comments",
      "nodeLabel": "Comment",
      "mappedColumns": {
        "id": "id",
        "content": "content",
        "date": "timestamp",
        "salon_id": "salon_id",
        "title": "title"
      }
    },
    {
      "sqlTableName": "tags",
      "nodeLabel": "Tag",
      "mappedColumns": {
        "name": "tag_name",
        "id": "id"
      }
    }
  ],
  "edges": [
    {
      "edgeLabel": "IsAuthorOf",
      "foreignKey": "posts.author_id",
      "from": "authors",
      "to": "posts"
    },
    {
      "edgeLabel": "IsTaggedWith",
      "joinTable": "posts_tags",
      "from": "posts",
      "to": "tags",
      "mappedColumns": {}
    }
  ]
}
```

Interaktywne przejście przez tworzenie mapowania między schematem bazy relacyjnej a docelowym

Kryteria akceptacji:

- użytkownik może zdefiniować te same mapowania, co za pomocą plików konfiguracyjnych, przez interakcję z aplikacją konsolową
- aplikacja umożliwia użytkownikowi ustalenie, że określone tabele lub kolumny nie będą importowane
- **aplikacja podpowiada użytkownikowi i umożliwia wybór dostępnego:**
  - schematu docelowej bazy grafowej
  - typu wierzchołka dla danej tabeli
  - atrybutu wierzchołka dla kolumny tabeli
  - typu krawędzi dla klucza obcego
  - typu krawędzi dla tabeli łącznikowej
  - atrybutu krawędzi dla kolumny tabeli łącznikowej

Plik konfiguracyjny z mapowaniem między listą krawędzi w pliku XML a schematem docelowym

Kryteria akceptacji:

- ustalono jednolity format pliku konfiguracyjnego dla mapowania, gdy zbiorem źródłowym jest plik XML z grafem w postaci listy krawędzi
- użytkownik może wybrać w pliku jeden z dostępnych schematów bazy grafowej
- **użytkownik może ustalić w pliku mapowanie między tagiem XML a:**
  - typem wierzchołka
  - atrybutem wierzchołka
  - typem krawędzi
  - atrybutem krawędzi
- użytkownik może ustalić w pliku, że określone tagi XML nie będą importowane lub są tagami zewnętrznymi dla właściwych danych

Interaktywne przejście przez tworzenie mapowania między listą krawędzi w pliku XML a schematem docelowym

Kryteria akceptacji:

- użytkownik może zdefiniować te same mapowania, co za pomocą plików konfiguracyjnych, przez interakcję z aplikacją konsolową
- aplikacja umożliwia użytkownikowi ustalenie, że określone tagi XML nie będą importowane lub są tagami zewnętrznymi dla właściwych danych
- **aplikacja podpowiada użytkownikowi i umożliwia wybór dostępnego:**
  - schematu docelowej bazy grafowej
  - typu wierzchołka dla odpowiedniego tagu XML

  - typu krawędzi dla odpowiedniego tagu XML
  - typu atrybutu krawędzi dla odpowiedniego tagu XML wewnątrz tagu odpowiadającego krawędzi
  - typu atrybutu wierzchołka dla odpowiedniego tagu XML wewnątrz tagu odpowiadającego wierzchołkowi

Plik konfiguracyjny z mapowaniem między listą krawędzi w pliku CSV a schematem docelowym

Kryteria akceptacji:

- ustalono jednolity format pliku konfiguracyjnego dla mapowania, gdy zbiorem źródłowym jest plik CSV z grafem w postaci listy krawędzi
- użytkownik może wybrać w pliku jeden z dostępnych schematów bazy grafowej
- aplikacja pozwala na wczytywanie zarówno plików CSV z etykietami kolumn, jak i bez
- **użytkownik może ustalić w pliku mapowanie między kolumną a:**
  - typem wierzchołka
  - atrybutem wierzchołka
  - atrybutem krawędzi
- użytkownik może ustalić w pliku, że określone kolumny nie będą importowane

Interaktywne przejście przez tworzenie mapowania między listą krawędzi w pliku CSV a schematem docelowym

Kryteria akceptacji:

- użytkownik może zdefiniować te same mapowania, co za pomocą plików konfiguracyjnych, przez interakcję z aplikacją konsolową
- aplikacja umożliwia użytkownikowi ustalenie, że określone kolumny nie będą importowane
- **aplikacja podpowiada użytkownikowi i umożliwia wybór dostępnego:**
  - schematu docelowej bazy grafowej
  - typu wierzchołka dla odpowiedniej kolumny
  - typu atrybutu krawędzi dla odpowiedniej kolumny
  - typu atrybutu wierzchołka dla odpowiedniej kolumny

Zawężenie przedziału czasowego przy imporcie danych

Kryteria akceptacji:

- aplikacja umożliwia użytkownikowi filtrowanie importowanych danych po jednym lub więcej atrybutach reprezentujących datę i czas
- aplikacja umożliwia użytkownikowi ustalenie przedziału czasowego dla importowanych danych

Rozszerzenie istniejącego grafu

Kryteria akceptacji:

- aplikacja pozwala na import nowych danych do już istniejącego grafu

Miary podobieństwa węzłów

Kryteria akceptacji:

- aplikacja rozpoznaje, gdy dane importowane pochodzą z tego samego źródła, co dane w bazie grafowej. Wtedy aplikacja wyznacza miarę podobieństwa między odpowiednimi węzłami
- miara podobieństwa węzłów jest wyznaczana na podstawie wybranych przez użytkownika atrybutów węzłów

Scalanie grafu wejściowego i docelowego

Kryteria akceptacji:

- **gdy dane importowane pochodzą z tego samego źródła, co dane w bazie grafowej:**
  - aplikacja pozwala użytkownikowi zdecydować, powyżej jakiej wartości miary podobieństwa scalić odpowiednie węzły, a poniżej której uznawać je za osobne
  - w przypadku konfliktu wartości między atrybutami scalanych węzłów aplikacja pozwala użytkownikowi wybrać czy woli zachować wartości źródłowe czy docelowe

1.1.2 Analiza sieci

Zawężenie przedziału czasowego przy analizie sieci

Kryteria akceptacji:

- aplikacja umożliwia użytkownikowi filtrowanie danych wejściowych do danego algorytmu SNA po jednym lub więcej atrybutach reprezentujących datę i czas
- aplikacja umożliwia użytkownikowi ustalenie przedziału czasowego dla danych wejściowych do danego algorytmu SNA

Wybór i wykonanie algorytmu analizy sieci

Kryteria akceptacji:

- aplikacja umożliwia użytkownikowi wybór jednego z dostępnych algorytmów analizy sieci
- aplikacja wykonuje algorytm SNA i zapisuje wyniki w tej samej bazie, co dane wejściowe lub w nowej bazie, w zależności od tego, co ustali użytkownik

Dostępne algorytmy SNA

Kryteria akceptacji:

- **aplikacja pozwala na uruchomienie następujących algorytmów / obliczenie następujących parametrów:**
  - Density
  - Clustering coefficient
  - Degree centrality
  - Closeness centrality
  - Betweenness centrality
  - PageRank
  - Degree distribution

Eksport do formatu JSON lub CSV

Kryteria akceptacji:

- użytkownik ma możliwość eksportu grafu z wynikami analiz do pliku w formacie JSON lub CSV